Spatial models of linguistic data in Africa and beyond

D i s s e r t a t i o n

zur

Erlangung des akademischen Grades

Doktor der Philosophie

in der Philosophischen Fakultät

der Eberhard Karls Universität Tübingen

Vorgelegt von

Miri Mertner

aus

Kopenhagen, Dänemark

2025

**Gedruckt mit Genehmigung der Philosophischen Fakultät
der Eberhard Karls Universität Tübingen**

Dekanin:	**Prof. Dr. Angelika Zirker**
Hauptberichterstatter:	**Prof. Dr. Gerhard Jäger**
Mitberichterstatter:	**Prof. Dr. Christian Bentz**
Tag der mündlichen Prüfung:	**28.08.2025**

Universitätsbibliothek Tübingen, TOBIAS-lib

THE UNIVERSITY OF TÜBINGEN

DOCTORAL THESIS

Spatial models of linguistic data in Africa and beyond

Author:
Miri MERTNER

Supervisor:
Prof. Dr. Gerhard JÄGER

July 28, 2025

THE UNIVERSITY OF TÜBINGEN

Abstract

Department of General Linguistics/Seminar für Sprachwissenschaft

Doctor of Philosophy

Spatial models of linguistic data in Africa and beyond

by Miri MERTNER

The spatial distribution of linguistic diversity and structural elements of language, like morphosyntactic and phonological features, is a rich source of knowledge for those interested in historical linguistics and the evolution of language (Nichols, 1992). Bayesian spatial models are a promising way to uncover these patterns. However, modelling linguistic data in space comes with a unique set of complexities and considerations. These include questions of how to represent the geographic locations of languages, how to measure the distance between them, and how to account for variations in topography. In this thesis, I present a series of case studies on African languages which will illustrate different Bayesian spatial models, all of which simultaneously incorporate information about language history in the form of phylogenies or language families. Recognising that the application of spatial models in linguistic typology is a relatively recent development, I will provide a general overview of some common models used in spatial statistics and describe the approaches which have been used in this thesis. Environmental and social factors have long been thought to influence the distribution of linguistic diversity in time and space (Currie and Mace, 2009; Nettle, 1998; Nichols, 1992). I will use a novel combination of methods to show that the factors which impact recent diversification are distinct from those which impact the maintenance of diversity over time. Following that, I will examine areal patterns of structural convergence between African languages and uncover systematic variation in the diffusibility of structural elements of language. Lastly, I will examine geographic patterns of data sparsity and discuss their implications for future statistical studies in linguistics.

Acknowledgements

This work would not have been possible without the help of many others. I joined the University of Tübingen as part of a project funded by the ERC Grant 834050 under the Horizon 2020 research and innovation programme. My advisor, Gerhard Jäger, first helped me to become literate in the mostly foreign-to-me language of Bayesian statistics. Since then, he has provided me with sound advice, supported my interests in a diverse range of fields and topics, and allowed me a great deal of academic freedom to pursue what I was interested in while still giving me a good idea of which direction to go in. Thanks to him, I was also able to participate in an interdisciplinary research environment for a year, which showed me in very concrete ways how the study of language evolution benefits immensely from collaboration across disciplines and people with diverse interests. I thank him for his patience and continual support.

Someone who changed the trajectory of my PhD for the better is Matías Guzmán Naranjo, with whom I had the pleasure to work in Tübingen for a short while. I have benefitted incredibly from his expertise and generosity in sharing his knowledge with me. He is one of the people who truly works with the goal of improving the field, and who continually investigates new, unexplored territory in the field of spatial models. He has helped me with many parts of the work in this thesis, notably providing extremely helpful feedback on the work presented in Chapter 4 and helping to code the statistical model in Chapter 5. I thank him deeply for his time, knowledge, technical expertise, insightful conversations, and collaborations (of which there will hopefully be many more).

My colleagues, Johannes Dellert, Johannes Wahle, and Isabella Boga, have also been incredibly helpful, especially in those early stages of my PhD, where they took time out of their busy schedules to support my learning of many new technical skills. I owe much of my understanding of Bayesian statistics to Johannes Dellert. I would also like to sincerely thank Christian Bentz for our many interesting conversations over coffee, which were always inspiring, and for the feedback he gave me which sparked the idea that led me to change a major aspect of the work presented in Chapter 3.

I would also like to mention some of those whose work greatly inspired me during my studies and who I later had the pleasure to meet. I wish to thank Nicholas Evans, whose books and articles had inspired me for a long time before I met him in Tübingen. His work, along with the work of Johanna Nichols, was the primary reason I decided to study the diversity of language through the lens of linguistic typology, and when we met, the insights he shared in person helped me write many parts of

this thesis. (One of his comments even inspired the idea for a short story I wrote for a competition called *Language Evolves*; I went home to write it the same evening that we met, and it became one of the runners-up). Likewise, my work owes much to that of Dan Dediu, whose work on the many interesting intersections between genetics, anatomy, the environment and language evolution continues to inspire and inform me, and with whom I hope to collaborate in the future.

I have also had the pleasure of meeting a colleague who became a friend, Frederic Blum, who I would like to thank both for your feedback and your company during conferences, of which I hope there will be many in the future. From my time at the DFG Center 'Words, Bones, Genes, Tools', I would like to thank Alexandros Karakostis. Collaborating with you was a pleasure and an inspiration, and it taught me a lot.

During my time at Leiden University, I had the pleasure of being a student of Victoria Nyst, who I thank for introducing me to richness of the world of signed languages and for sparking a thought process on how to include them in typological studies (which I have not given up). I would also like to thank Maarten Mous for his insightful teachings on African linguistics; much of what he taught me still informs my work now.

Another important group of people whose contribution I'd like to acknowledge includes those who document and describe languages, as well as those who create, curate, and maintain databases. Without your valuable work, which forms the basis of so much research and so much of our knowledge, work such as mine would not be possible.

Of course, these acknowledgements wouldn't be complete if I didn't thank my friends, in particular Suri and Laurence, who have always showed up for me during my time in Tübingen, and my family, for their unwavering support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Languages are spoken in a wide variety of geographic locations by communities of highly variable sizes. Some are spoken across large areas in which few other linguistic communities reside; others are spoken by a small community of speakers in a single village or town. Many languages are not restricted to a single location but can be spoken alongside multiple other languages in several different areas. Any researcher interested in representing the locations of languages in space will also have noticed the high level of variation in the climate and terrain surrounding these languages and may have considered how to measure geographic distance in a way that accounts for the fact that the same distance is likely more difficult to traverse in a mountainous area than in temperate grasslands. These complexities (and there are many more) can make the task of coming up with an adequate abstract representation of languages in space for the purposes of statistical modelling seem intractable. This is further complicated by the recognition that language locations themselves are only a proxy for what we are really interested in detecting, which is social interaction between speakers of different languages. The kind of interaction that leads to changes in one or more of the speakers' languages, which we call 'contact effects', is usually regular and sustained over long periods of time, and often involves some level of multilingualism. Thus, locations in space act not only as a proxy for the social interaction between speakers of different languages, but also for the probability that these interactions have been occurring regularly enough over a long period of time to impact the languages involved.

Many interesting questions can be raised concerning the link between languages and the environment in which they are spoken. These range from questions about the role of the environment in shaping patterns of linguistic diversity (Hua et al., 2019; Nettle, 1998; Nichols, 1992; Urban, 2020) or areal convergence (Hammarström and Donohue, 2014) to hypotheses around the way in which the environment could directly or indirectly influence the sound systems of languages (Bentz et al., 2018; Everett, 2013; Urban and Moran, 2021; Wang et al., 2023). The relationship between geographic distance and linguistic diversification has also been probed quantitatively

using data on dialects as well as distinct languages (Holman et al., 2008; Nerbonne, 2013; Nerbonne and Heeringa, 2007; Wieling, Nerbonne, and Baayen, 2011). More generally, languages are influenced not just by their geographic surroundings but also by their cultural and social environment (Bentz et al., 2015; Currie and Mace, 2009; Epps, 2020; Lupyan and Dale, 2010; Trudgill, 2011). Some have proposed links between genetics and the sound systems of languages, as well as interactions between genes, cultural variables, the environment, and colour lexicons (Dediu and Ladd, 2007; Josserand et al., 2021).

One of the challenges associated with disentangling the impact of external factors on language or detecting linguistic areas using quantitative methods lies in the complex interplay between geographic space and social interaction. Most linguistic datasets include information about where languages are spoken in the form of latitude and longitude coordinates on a map, which fail to capture differences in the spatial extent of languages. However, it could be argued that although there are some languages, like English or Spanish, which are spoken across very large geographic areas in multiple disconnected locations, these kinds of language distributions are the exception rather than the norm, especially from a historical perspective. At the same time, groups of nomadic or seminomadic hunter-gatherer and pastoralist groups have existed throughout history and prehistory, and their spatial extent may be far more accurately reflected by a polygon than a point location. Because of their nomadic lifestyle, these peoples could come into contact with communities which appear to be very far away based on point locations only. This reflects a related point, which is that language locations are not static over time, as communities may move, migrate, or be displaced by other communities or environmental changes. Movement thus adds yet another layer of complexity.

Some of the common distance measures used in linguistic typology do not account for the presence of variable terrain elevation or other topographic features (Guzmán Naranjo and Mertner, 2022; Murawaki and Yamauchi, 2018; Ranacher et al., 2021). Some recent studies have taken steps to remedy this by calculating distances across a topographic surface, which means that two language communities living in an area characterised by even terrain will have a shorter distance between them than two language communities which are separated by a mountain (Guzmán Naranjo and Jäger, 2022; Guzmán Naranjo, Mertner, and Urban, 2024). However, it is worth examining the question of how much geographic detail is necessary when doing statistical studies with spatial data, and to adjust the level of detail according to the region under study and the topic of interest. Not enough work has been done comparing the performance and results of different types of spatial models for linguistics, although this is an emerging field, and there is a lack of information on when to use

one spatial model over another. Although there is a rich literature on Bayesian spatial models for use fields such as spatial econometrics, epidemiology, and ecology which has contributed to some linguistics studies like Wieling, Nerbonne, and Baayen (2011) and Winter and Wieling (2016), there is no comprehensive overview of spatial modelling techniques written for linguists. In recognition of this, the goal of Chapter 2 of this dissertation is to provide a general introduction to a diverse range of spatial models and their advantages and limitations for the purpose of modelling different kinds of linguistic data. Additionally, I will present a brief comparison of some of the methods introduced in the chapter, in recognition of the lack of studies comparing the performance of different spatial models using the same dataset.

The influence of terrain on the dynamics of social interactions between linguistic communities is a prominent field of interest for linguistic typologists. The idea that terrain could be an essential part of defining the linguistic profile of an area was made influential Nichols (1992), who observed that many of the most linguistically diverse areas of the world are found in mountainous regions, and called these accretion zones or *residual zones*. The question of whether these areas simply happen to be mountainous, or whether the terrain has a direct influence on the diversity of languages in these areas, continues to draw interest (Comrie, 2008; Hua et al., 2019). The impact of variable terrain on linguistic diversity has implications for how terrain impacts contact dynamics, too, as it predicts the extent to which terrain causes communities to be isolated from each other. If they are isolated for long enough in order for significant differentiation to happen, this suggests that the terrain effectively prevents regular contact (Huisman, Majid, and Hout, 2019).

While geographic barriers such as mountains, water, and rough terrain can undoubtedly impede contact between groups, this is not always the case in practice, as human groups have been known to traverse incredibly challenging terrain and to travel long distances across the ocean. Speakers of Austronesian languages managed to settle in a large number of distant islands in the Pacific, and speakers of Bantu languages migrated across a vast area including rainforests, mountains, and desert, all of which could be considered difficult to traverse in their own way (Bostoen, 2020). This renders it even more challenging to include prior assumptions about the extent to which terrain influences the probability of social interaction between language groups. For Africa, Hammarström and Donohue (2014) find that geographic features like mountain ranges do not seem to predict patterns of areal convergence, which suggests that variable topography may not pose as much of a barrier to contact as previously thought. Moreover, increased attention has been drawn recently to the role of social factors, like culture, political organisation, and language attitudes, in the maintenance of distinct languages in an area and, in some cases, diversification

(Comrie, 2008; Evans, 2010; Lüpke, 2016; Mansfield, Leslie-O'Neill, and Li, 2023; Pakendorf, Dobrushina, and Khanina, 2021). Because social and environmental factors are linked, their impact on linguistic diversity should ideally be tested simultaneously (McElreath, 2020). Chapter 3 of this dissertation uses a spatial model to disentangle the role of cultural and environmental factors in shaping the distribution of languages in Africa. An additional goal of the study is to shed light on whether the same underlying processes govern the maintenance of deep-time diversity in an area as opposed to more recent diversification. The results have implications for the importance of including variable terrain elevation in spatial models of contact focusing on Africa, and on our understanding of the relative importance of culture and geography for shaping patterns of diversity.

The majority of language contact throughout history likely occurred locally in multilingual settings (Evans, 2017). A recent quantitative study provided support for the idea that contact effects on structural and phonological features are predominantly local (Guzmán Naranjo, Mertner, and Urban, 2024). The range within which areal diffusion can be detected is likely to depend on the particular linguistic features or lexical items which are included in the model. The famous list of basic concepts by Swadesh (1955) is based on this idea, holding that some parts of the lexicon are less prone to replacement or modification by language contact than others. Although this idea predates the work by Swadesh (1955) (see Kaplan (2017) and List (2018) for an overview of the topic), his list of concepts, which are expected to be similarly stable throughout time across a diverse range of languages, has undoubtedly been the most influential. Although the list itself has been criticised, adapted, and updated continually, lists of universally or culturally specific stable concepts have continued to be used across a variety of disciplines, notably for the automatic inference of language trees, with promising results (Bowern and Atkinson, 2012; Brown et al., 2008; Jäger, 2013, 2018).

Similarly, some structural elements of language appear to be more borrowable than others in the sense that they require a longer period of contact between speakers, or a more extensive degree of multilingualism or bilingualism, in order to be impacted by contact (Thomason and Kaufman, 1988). Quantitative studies drawing on typological data have found support for this idea, as some features have been found to form geographic clusters, while others seem to be transmitted predominantly through common inheritance; some may also evolve more slowly over time than others (Dediu and Cysouw, 2013; Nichols, 1992; Wichmann, 2015). Other studies have called this into question, re-examining some of the features thought to be areal and still finding a greater phylogenetic signal for those features (Dunn et al., 2011; Skirgård et al., 2023). At the same time, interest in language contact and its effects

more generally has been rising, and some have argued that it may play a bigger role in language change than previously thought (Guzmán Naranjo and Becker, 2021). All this indicates that there is a need for more research on this topic. Research on quantitative methods which can effectively disentangle the effects of contact and inheritance on language structure using large-scale or regional datasets is ongoing, and the study in Chapter 4 of this dissertation aims to add to the growing body of work in the field (which includes, among others, work by Dediu (2011), Dediu and Cysouw (2013), Murawaki and Yamauchi (2018), Neureiter et al. (2022), Nichols (2003), Ranacher et al. (2021), and Wichmann (2015)).

Any comment on the borrowability of linguistic features needs to be made in opposition to genealogical stability. Likewise, the linguistic effects of language contact are most often detected in opposition to the role of shared ancestry. Language contact itself also has a temporal dimension, since groups of speakers do not stay in the same location over time. Additionally, languages may be influenced differently by interaction with speakers of related languages than unrelated languages (Matras, 2007). Family-internal contact is even more difficult to detect and disentangle from processes of inheritance, in part because it may be easier for contact to effect structural changes when the languages involved are more structurally similar to each other (Thomason, 2010), and in part because it becomes more challenging to rule out inheritance as an explanation for observed similarities. Although the main focus of this dissertation is on the role of space as a proxy for the social interactions that lead to language contact, all of the models presented in this dissertation also include controls for phylogenetic effects, which were adapted to the relevant case studies. Whenever possible, phylogenetic regression was used in order to account for the structure of entire language trees, rather than only modelling genus- or family-level effects (see Chapter 14 of Villemereuil and Nakagawa (2014) for an introduction to the method). One of the issues with using an existing phylogeny to control for shared ancestry is that, for many languages, there is some degree of uncertainty regarding their classification. These are issues which lie outside the scope of this dissertation, but in the conclusion in Chapter 6, I will provide a more comprehensive overview of some of the possibilities for future work to account for them.

Large-scale databases on the phonology, morphosyntax, and lexicon of diverse languages have facilitated a wealth of studies into cross-linguistic diversity (Dryer and Haspelmath, 2013; List et al., 2022; Moran and McCloy, 2019; Skirgård et al., 2023). Databases which include data specific to particular areas or language families have also made crucial contributions to this field (such as Dellert et al. (2019), Gerardi, Reichert, and Aragon (2021), and Segerer and Flavier (2011-2021), among others). Both kinds of databases are essential to the quantitative study of linguistic typology.

However, researchers in this field are almost guaranteed to be faced with the issue of missing data, which shows up in several different ways. It can manifest as sparse coverage of linguistic features, short or incomplete word lists, and a lack of detailed data on the geographic locations of languages. Understanding the underlying processes which predict the distribution of data sparsity can help us mitigate biases induced by missing data (McElreath, 2020). Previous studies have controlled for such biases using balanced sampling (Bakker, 2010; Miestamo, Bakker, and Arppe, 2016), and significant progress has been made in the development of geographic and phylogenetic controls which allow for the inclusion of all the available data (Becker, Guzmán Naranjo, and Ochs, 2023; Guzmán Naranjo and Becker, 2021). However, no study has investigated the geographic distribution of missing data in linguistic typology explicitly before. The use of statistical controls depends on the assumption that missing data is biased according to specific underlying processes, so studying its distribution can help inform future work on bias control.

In the following section, I will provide an introduction to language contact and linguistic diversity in Africa, which forms the basis for all but one of the case studies in this dissertation. Chapter 2 will introduce the literature on Bayesian spatial models with reference to their relevance for modelling linguistic data. A brief case study of the areal distribution of labial-velar consonants in Africa will be presented in order to compare some of the models discussed in the chapter, which will finish with an overview of the different methods' advantages and limitations. In Chapter 3, I will re-examine some of the hypothesised links between linguistic diversity and cultural and environmental factors in the context of Africa's history, especially in an attempt to disentangle the factors behind variations in language density as opposed to phylogenetic diversity. Chapter 4 will address the question of whether some structural features are more prone to areal diffusion than others using a novel approach, which can also detect areal patterns that could be indicative of structural convergence in Africa. Lastly, in Chapter 5, I will examine the geographic distribution of missing data in the World Atlas of Language Structures (Dryer and Haspelmath, 2013) and discuss its implications for missing data bias, as well as the methods which can be used to mitigate its effects.

1.1 An areal view of Africa

"One tantalizing question that has been raised a number of times asks whether Africa as a whole constitutes a linguistic area. This would of course imply a great deal of contact over a widespread area likely for an extended period of time. That such a question has been raised so many times points again to the lumping tradition in

African linguistic studies. It also suggests that contact might be just as important if not more important than genetic inheritance for understanding the structures and relatedness of African languages" (Childs, 2010, p. 695).

Africa is the continent with the highest genetic diversity in the world and around 2,000 distinct spoken languages, with some variations depending on how they are counted, and yet most of these can be divided into a few major families, or phyla (Hammarström, 2018). It should be noted, too, that Africa has a relatively large number of signed languages compared to other parts of the world, partly due to high incidences of genetic deafness concentrated in specific areas, although many of them are now highly endangered (Hammarström, 2018; Zeshan and Vos, 2012). These signed languages do not belong to any of the major phyla, and so their existence adds to the overall picture of linguistic diversity on the continent (Nyst, 2010).

Africa is home to one of the most linguistically diverse and dense regions of the world, an area in and around Cameroon (Nettle, 1996). Why this area in particular is home to such a large number of distinct languages remains unknown and has been attributed to both the environment and sociolinguistic factors (Di Carlo, Esene Agwara, and Ojong Diba, 2020; Nettle, 1996). Another puzzle is that Africa, the continent with the longest history of human habitation in the world, has fewer small language families than more recently settled areas like South America (Blench, 2013) (although the exact number of families has been called into question by researchers like Güldemann (2018a)). Language contact is one of the reasons why the historical and linguistic landscape of Africa is so challenging to uncover. Thus, the study of contact and its effects at the micro- and macro-level is increasingly being recognised as a necessary step towards achieving a better understanding of language histories, although that is by no means the only reason why contact is a valuable research avenue (Heine and Nurse, 2008) Thus, Africa has some unique characteristics, including its distinctive areal distribution of linguistic diversity and structural features as well as sociolinguistic dynamics. These are the key themes of the case studies in this dissertation. Additionally, Africa is a historically under-researched area, particularly in the domain of language contact (Childs, 2010).

The validity of the four main phyla defined by Greenberg (1963) – Niger-Congo, Nilo-Saharan, Afro-Asiatic, and particularly Khoisan – has been disputed, but there is no doubt that they remain highly influential (Güldemann and Fehn, 2014; Heine and Nurse, 2007). Greenberg (1959) also identified some of the first areal patterns in Africa, as well as later examining the areal distribution of specific phonological and morphosyntactic features (Greenberg, 1983). Much (if not all) of the subsequent work on both areal and genealogical linguistics in Africa builds on these sources. However, the classification of African languages faces a number of challenges, chief

among them a lack of documentation and language endangerment (Hammarström, 2018). The effects of language contact, which are considerable in Africa, can also obscure genealogical relationships (Di Carlo and Good, 2023). The classification of some language groups, like Mande, Songhay, and Ubangi, remains disputed, with some arguing for their belonging to larger stocks while others prefer to treat them as individual language families, and much of this debate is centred around what can be attributed to contact as opposed to inheritance or other factors (Blench, 2013; Güldemann, 2008, 2018a). However, I will not delve into the debates around the genealogical classification of African languages, as there is a wealth of literature on this which the reader can refer to (see e.g. Blench (2013), Dimmendaal (2008b), Güldemann (2018a), and Hammarström (2018), and references therein). Instead, I will give a brief overview of the study of language contact and areal linguistics in Africa (of which more comprehensive overviews can be found in Dimmendaal (2001, 2020), Güldemann (2018b), Heine and Nurse (2007), and Leyew (2008) and references therein), focusing on what the study of areal patterns in linguistic diversity and language structures can contribute to our understanding of African linguistics and history.

Although the effects of language contact can obscure deep-time relationships, it is also an important source of information on African history and prehistory. The volume by Heine and Nurse (2008) emphasised the importance of studying areal patterns in African linguistics, not just as a way to improve genealogical classification but as a goal in itself. They also argued that the importance of grammatical replication, or the transfer of functional categories between languages without the transfer of forms, has been overlooked to the detriment of the field (Heine and Kuteva, 2003); the same phenomenon was expanded upon and given a slightly different name in Matras (2007, 2010), who called it *pattern replication*. Pattern replication can be indicative of intensive, deep-time contact between communities, but it is notoriously difficult to identify unequivocally as the result of contact because patterns can be similar across languages for many other reasons, even when common inheritance can be ruled out (Bickel, 2017). However, the areal distribution of patterns can be used as a way to test the hypothesis of contact against other possibilities, like cognitive biases. If a pattern is found in a cluster of geographically contingent languages belonging to different families, and if that pattern is uncommon outside that area, it becomes easier to conclude that it probably spread through areal diffusion (Heine and Nurse, 2007, p. 7).

Areas of convergence can provide a window into history and prehistory (Bickel, 2020). As an example, Güldemann (1998) identified a possible convergence area in the Kalahari Basin, and subsequent work on the area suggested it could be the

remnants of a large and ancient linguistic area that stretched all the way to the Rift Valley in Tanzania. This was proposed to explain the striking distribution of click consonants in Africa. Notably, they are found in the languages of the Kalahari Basin typically referred to as Khoisan, which is widely understood not as a genealogical unit, but as a convenient term for those languages in Africa which make extensive use of click consonants as part of their phoneme inventories and which do not belong to either the Bantu or Cushitic language families (Vossen, 2013, p. 49). Click consonants are also found in a few Bantu languages in southern Africa, including Zulu and Xhosa, which have had extensive contact with certain Khoisan language groups, and two languages in the Rift Valley area, Hadza and Sandawe (Kießling, Mous, and Nurse, 2008; Sands and Gunnink, 2019). Language contact provides a compelling explanation for the presence of click consonants in this small, specific subset of African languages (Vossen, 2013).

The Rift itself is considered an accretion zone and is the only place in Africa where all four major phyla meet, thus possibly providing a window into deep-time diversity (Nichols, 1992). The area between the Kalahari Basin and the Rift may have been a zone of impressive linguistic diversity and possibly large-scale patterns of convergence in the past, and some of its characteristics could show up as substrate effects in the Bantu languages spoken there today, although such effects have not been detected with any degree of certainty (Bostoen, 2020). It does not help that the languages of Central Africa are in dire need of further documentation. For many of the large linguistic areas described in Africa (such as the Macro-Sudan belt (Güldemann, 2008) and Ethiopian linguistic area (Crass and Meyer, 2007; Ferguson, 1976)), the boundaries are typically not well-defined, and often the areas themselves are disputed (Tosco, 2000). This is a common property of linguistic areas, which are perhaps best viewed not as binary classifications but as diffuse entitites comprising core and a periphery, the boundaries of which may change over time (Haspelmath, 2001). Linguistic areas can thus overlap in both time and space: peripheral members of a linguistic area may have been core members of a different linguistic area in the past, which likely left some traces in the typological profile and lexicons of those languages. Moreover, these traces can be informative, and thus, it may be equally interesting for the historical linguist to examine the parts of linguistic areas which resist discretisation.

Although I will focus on methodology in many parts of this dissertation, I also aim to draw on the rich body of literature on African linguistics, sociolinguistic and cultural dynamics of contact, and micro-level variation in order to inform the methods and the interpretation of their results. One of the goals of this dissertation is to consider the concerns of language experts in the formulation of novel statistical

methods, and to make the technical literature on spatial models more accessible to those who have a background in linguistics.

Chapter 2

Bayesian spatial models for linguistic typology

The purpose of this chapter is to provide a general introduction to the field of spatial modelling for linguistics using a Bayesian statistical framework. I will discuss why modelling spatial relationships between languages is crucial to the study of linguistic typology and how geographic space can be represented for the purposes of statistical modelling. This will be followed by an overview of some common spatial modelling approaches, including ones which have been applied previously to the study of linguistic typology and language evolution, and some which have not. Recognising the lack of systematic comparisons between different kinds of spatial models and representations of spatial relationships between languages, I will present a case study on labial-velar consonants in Africa in order to compare some of these models in terms of their efficiency, results, and predictive performance. Some of the difficulties associated with comparing different types of models will also be discussed. Following the case study, I will explore some of the general limitations and challenges of the methods presented in this chapter. I will also mention some additional methods for spatial modelling which have not been included here but which may be relevant to researchers seeking to explore alternative modelling strategies.

2.1 Why should we model space?

Languages are shaped by a variety of interacting forces, including but not limited to cognitive biases, external factors, genealogical inheritance, and language contact (Bickel, 2017). Understanding the distinct effects of these forces in order to uncover cross-linguistic tendencies, as well as tendencies which appear to be specific to particular areas or lineages, is one of the long-standing goals of linguistic typology (Nichols, 1992). Historical linguists share an interest in disentangling the effects of inheritance and contact in particular, as the impacts of contact can obscure the signs

of shared ancestry, even leading to the misclassification of some languages. Another major goal of linguistic typology has been the detection of universal cognitive biases and the limits of variation in human language (Bickel, 2007; Song, 2013). However, the focus of the discipline has gradually shifted, as it has been argued that the most interesting aspect of human language is the breadth of its diversity, and that perhaps more attention should be devoted to studying how and why such diversity arises, in favour of a focus on its limits (Evans and Levinson, 2009). In recent years, typologists have been less concerned with what is universal and more with the study of variation itself, with Bickel (2007, p. 239) succinctly summarising modern typology's central question, "what's where why?"

Space has always been central to typology, made explicit in the *where*. Similarly, the *why* implies a temporal dimension, as this question cannot be answered without reference to the history of the languages in an area. The answer to *why* specific languages, sounds or structures are prevalent in a given location is often to be found in the past, related to the ancestry of languages, the dynamics of evolution, and the way in which past interactions between speakers across space have left their mark on the languages they speak. The central themes of linguistic typology thus go hand in hand with the historical linguist's goal of understanding how languages evolve over time, and space has been a crucial part of this since the inception of the field (Greenberg, 1963; Nichols, 1992).

Quantitative data on diverse languages provides the foundation for linguistic typology, and in recent years, it has become ever more accessible and broad in scope. Greenberg (1963) is largely credited with founding the field of linguistic typology, and the influential studies by Bell (1978), Dryer (1989), and Nichols (1992) developed its central methodology of stratified sampling. Sampling has been used to avoid drawing biased conclusions about cross-linguistic tendencies because some large language families, such as Indo-European and Niger-Congo, happen to be overrepresented in the available data for historical reasons which have nothing to do with universal biases. Building on methods by Dryer (1989), Nichols (1992) distinguished between *genera*, language families with a time depth of around 4,000 years, and *stocks*, language families with time depths beyond 5,000 years. Nichols (1992) sampled one language per stock and genus, although she sampled up to six languages for very large and diverse lineages, such as Indo-European. The method of *balanced sampling* to avoid bias has since been widely used and has continued to contribute valuable insights into the nature of cross-linguistic variation (Bakker, 2010; Miestamo, Bakker, and Arppe, 2016). However, this approach has been criticised because it necessitates the exclusion of valuable linguistic data, which is already a sparse representation of the actual scope of possible human language in the past and present (Cysouw, 2011).

It may lead to the representation of language families and areas as more homogenous than they really are, obscuring within-family variation which could be informative in itself. This has become especially relevant as the scale of the available data has increased over the years (Dryer and Haspelmath, 2013; Moran and McCloy, 2019; Skirgård et al., 2023).

The earliest language sample developed by Bell (1978) focused on avoiding over-sampling from the same families or genera, with the assumption that language families are likely to be internally similar due to common inheritance, an assumption which may not hold across families (Cysouw, 2005). Prior to the large-scale analysis of cross-linguistic contact situations by Thomason and Kaufman (1988), which showed that any aspect of language can be impacted by language contact, the field had received less attention than the study of shared ancestry as the underlying cause of structural similarity between languages (see Hickey (2010) and references therein). Several studies have since shown, with reference to specific areas as in Aikhenvald (2002) or using language samples as in Matras and Sakel (2007), that language contact can impact languages and their structures at any level over time, and it can lead to the diffusion of common features across large geographic areas (Bickel, 2017, 2020).

Dryer (1989) recognised the effects of large-scale areal diffusion when he proposed a division of the world into macroareas. These macroareas can be included in statistical models as predictors or in a Bayesian hierarchical regression model as a random effect, an approach which has been used by Jaeger et al. (2011) and others. This method has been employed by several studies since then and provided a novel way to retain more data than is possible when using balanced sampling, while controlling for some of the effects of language contact. However, this approach makes two important assumptions which, in the context of some typological studies, may be difficult to defend. The first assumption is that there are no instances of contact across macroareas. Although researchers such as Hammarström and Donohue (2014) have re-defined the boundaries of macroareas in order to maximise independence between them, there may still be some instances of contact between macroareas. For example, we know that languages spoken in Asia, as well as Arabic, have had extensive contact with Swahili, a Bantu language spoken on the island of Zanzibar as well as across eastern Africa (Mugane, 2015). Additionally, there is some inconsistency in different experts' judgments of what should constitute a macroarea. For instance, Güldemann (2018b) argues that Africa and the Arabian peninsula should be considered a single macroarea, called Afrabia, but this division is not commonly adopted by typologists, who tend to include the Arabian peninsula in the macroarea of Eurasia. The second assumption made when using macroarea alone as a control for contact is that the languages within a macroarea have possibly converged due to areal diffusion.

This assumption is generally well-founded. Languages within macroareas are likely to be more similar than languages across macroareas. For Africa, Leyew (2008, p. 35) succeeds in finding a set of morphosyntactic, lexical, and phonological properties which are 'not found at a comparable quantitative magnitude in languages outside the area', indicating that Africa could be considered a linguistic area. However, this strategy may not always be adequate, as it cannot detect areas of convergence within macroareas. Additionally, it cannot detect contact-induced divergence (Mansfield, Leslie-O'Neill, and Li, 2023). Thus, this method has limitations which are not shared by other kinds of spatial models.

As an interest in the effects of language contact and areal diffusion at the micro- and macro-level continues to rise, there has been an increased recognition of the need for better models of the geographic space in which languages are spoken and interact with each other (Guzmán Naranjo, Mertner, and Urban, 2024; Ranacher et al., 2021). Models including a representation of space to control for areal effects have been gaining popularity in quantitative studies of language change, linguistic typology, and historical linguistics (such as the studies by Cathcart et al. (2018), Guzmán Naranjo and Becker (2021), Guzmán Naranjo and Mertner (2022), Guzmán Naranjo, Mertner, and Urban (2024), Kauhanen et al. (2018), Murawaki and Yamauchi (2018), Neureiter et al. (2022), Nikolaev (2019), and Ranacher et al. (2021), and others). However, these models can still be difficult to fit in practice, and choosing an appropriate model is not trivial. Modelling geography is highly challenging, given the presence of variable terrain elevation, differences in climate, bodies of water which may either facilitate or hinder human travel, and other environmental features such as forests and deserts. In addition to the complexity of the terrain itself, extra-linguistic factors are known to influence how languages are impacted by contact. These include sociolinguistic hierarchies and prestige, the level of bi- or multilingualism between speakers, the duration of contact, the manner of social organisation of language communities, cognitive or functional biases in the brain, and structural similarities or differences between the languages involved which may prohibit or facilitate borrowing (Matras and Sakel, 2007). Additionally, we know that languages are spoken across highly variable geographic extents, with some languages spoken in a small area and others spoken across larger areas. In contrast, languages are often represented as point locations in databases (Dryer and Haspelmath, 2013; Hammarström et al., 2023; Skirgård et al., 2023). However, it would be effectively impossible to include all of these factors in a spatial model. When formulating statistical models, we necessarily rely on a degree of abstraction which allows us to detect underlying trends and patterns in large amounts of data which may be difficult to detect by human judgment alone. McElreath (2020, p. 13) reiterates a statistics

aphorism which summarises this reality: 'All models are wrong, but some are useful'.

This chapter aims to somewhat remedy the difficulty associated with choosing a spatial model within a Bayesian framework for a study in linguistic typology, providing an overview of some of the available methods, their advantages and suitability for particular data types, and their limitations. Note that all of these methods require some experience in R, **brms**, and (for some of the more advanced models) Stan (Bürkner, 2017; Carpenter et al., 2017).

The chapter will be structured as follows. First, in Section 2.1.1, I will describe some common tests used in a variety of fields such as spatial econometrics, ecology, and epidemiology, which can help us determine whether a spatial model is necessary for a given dataset. In Section 2.2, I will discuss some of the ways in which we can represent space in a model when we have either point locations (like latitude/longitude coordinates, which are generally available for all languages in (Hammarström et al., 2023)) or language areas in the form of polygons, such as those which can be accessed through *Ethnologue* (Eberhard and Fennig, 2023). When either data type is available, it is possible to represent the spatial relationships between languages as a *neighbourhood matrix* (also referred to in the literature as the neighbour graph or spatial weights matrix) (LeSage and Pace, 2009; Ver Hoef, Hanks, and Hooten, 2018; Wall, 2004). These structures can be used in *autoregressive models*, which are common in the fields of spatial econometrics, political science, and ecology, among others. These models are described in section 2.4.

An alternative approach to spatial modelling of point locations is to use one or more latent *Gaussian processes* (GPs), a method introduced for bias control in linguistic typology by Guzmán Naranjo and Becker (2021). This will be described in section 2.5. Using a GP can mitigate some of the disadvantages of using points as a representation of language locations in space, and they provide a highly flexible way of modelling language contact, although this comes at a computational cost. Chapters 4 and 5 of this thesis use GP models.

In section 2.6, a method for modelling the spatial signal of covariate data (such as extra-linguistic factors in a study examining the relationship between a linguistic variable, such as tone, and a non-linguistic variable, such as temperature) will be introduced. This method is used in Chapter 3 for cultural and environmental covariate data.

Section 2.9 will discuss some methods which have not been presented in detail here but which are worth mentioning in an overview of spatial models. Lastly, section 2.10 will conclude the chapter.

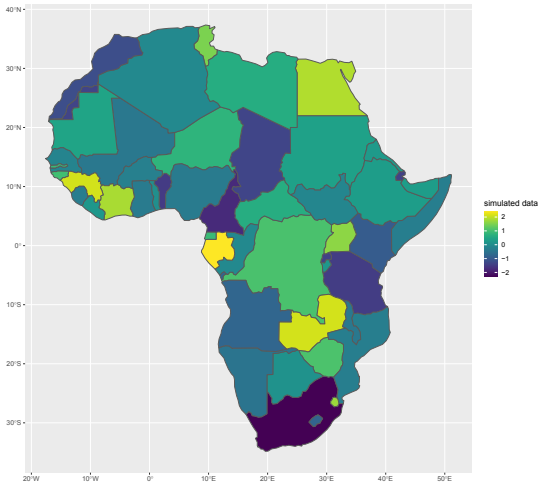
2.1.1 Spatial autocorrelation

"Spatial autocorrelation exists whenever a variable exhibits a regular pattern over space in which its values at a set of locations depend on values of the same variable at other locations. Spatial autocorrelation is present, for example, when similar values cluster together on a map." (Odland, 1988, p. 7)

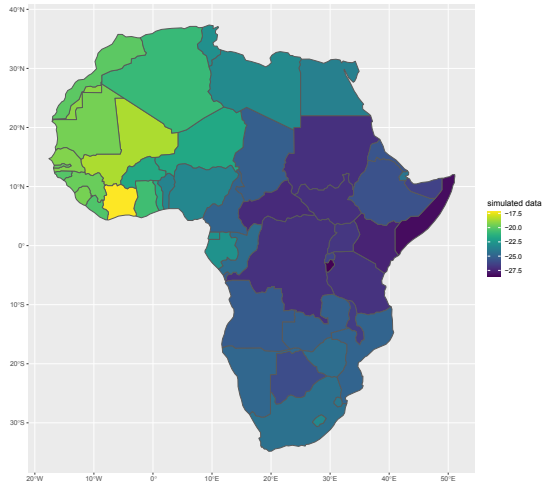
Because languages influence each other across time and space, they violate an important assumption of many statistical models, namely that data points are *independent and identically distributed*, or *i.i.d.* Related languages are likely to be more similar than unrelated languages, and so are languages which are spoken in close geographic proximity to one another. However, this may not be the case for all grammatical, lexical or phonological properties, nor may it hold equally across all these properties. Nichols (1992) was among the first to demonstrate that some properties of language exhibit a greater degree of similarity within areas and families than others. Because areal dependence between languages is thus a matter of degrees, it can be helpful to quantify this before beginning analysis. The following section will outline some of the ways in which *spatial autocorrelation* can be quantified. Spatial autocorrelation can also describe situations in which the values of data points on a map are more *dissimilar* than would be expected by chance. For example, contact between languages can, in some cases, lead to divergence rather than convergence (Mansfield, Leslie-O'Neill, and Li, 2023). In these cases, there would be *negative* spatial autocorrelation. The clustering behaviour of simulated data at different levels of spatial autocorrelation (set by the parameter ρ) in a range from -1 to 1, with 1 being the highest possible level of spatial autocorrelation, is illustrated in Figure 2.1.

Two of the most common indexes are Moran's I and Geary's C, both of which measure *global* spatial autocorrelation (Chun and Griffith, 2013). For a given dataset with locations on a map, Moran's I and Geary's C return a single value indicating how strong the degree of spatial autocorrelation is across the entire dataset. The difference between the two is that Geary's C is more sensitive to negative spatial autocorrelation (Chun and Griffith, 2013, p. 12). Because of this, local outliers (points whose values diverge from the values of neighbouring points, which are otherwise similar) have a greater effect on the value returned by Geary's C. Moran's I and Geary's C are mathematically complementary and, when local outliers for a variable are not highly influential, they typically indicate similar levels of spatial autocorrelation (Donegan, Chun, and Griffith, 2021). They can thus be used in combination with each other.

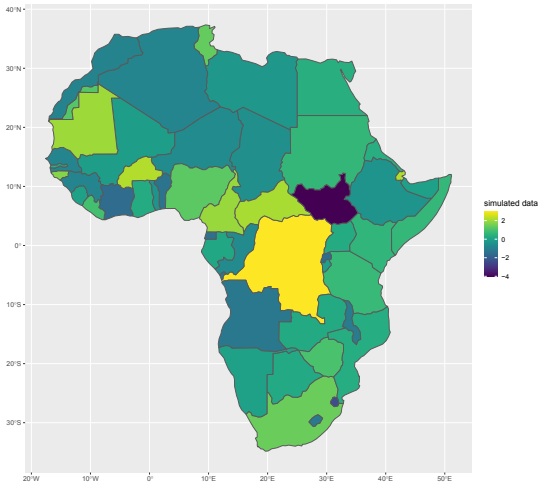
Some datasets may contain small clusters of high spatial autocorrelation as well as some which exhibit low or negative spatial autocorrelation. Global Moran's I and Geary's C cannot capture this kind of variation. Thus, when the degree of spatial



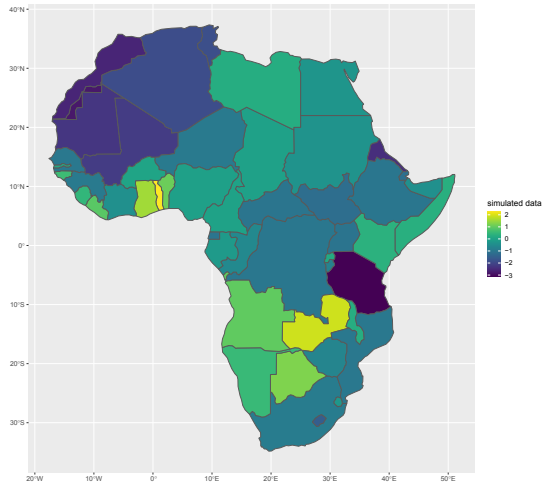
(A) $\rho = 0$ (no spatial autocorrelation; no spatial clustering)



(B) $\rho = 0.99$ (complete spatial autocorrelation; maximum clustering)



(C) $\rho = -0.99$ (negative spatial autocorrelation; diverging values)



(D) $\rho = 0.5$ (moderate spatial autocorrelation; some clustering)

FIGURE 2.1: Spatially autocorrelated data simulated on a map of African countries.

autocorrelation is highly variable, *local* metrics may be more useful. These LISAs (local indicators of spatial autocorrelation) were introduced and developed by by Anselin (1995), Getis and Ord (1992), and Ord and Getis (1995). Both Geary's C and Moran's I can be calculated locally as LISAs. As with the global versions of these indexes, Moran's I is suitable for detecting overall spatial clustering behaviour, while Geary's C is useful for detecting local outliers.

It is important to note that some of these statistics are not designed for use with categorical or binary data. When the available data is binary, a common approach is to use *join-count statistics*, which provide a measure of spatial autocorrelation by counting how many neighbours share the same value and how many have diverging values. The number of neighbours which share the same binary value is compared to the number of neighbours which would be expected to share the same value under random chance.

All of the indexes discussed here are implemented in the R package **geostan**, which also provides functions for visualising the results (Donegan, 2022; Donegan, Chun, and Griffith, 2021).

Testing for spatial autocorrelation, regardless of the index used, requires a representation of the spatial relationships between the languages. Some common ways of measuring geographic distance and representing the positions of languages in space are discussed in the next section.

2.2 How should we model space?

The question of how to model space is a challenging one which lacks a straightforward answer. In a statistical model, we cannot represent reality exactly as it is; in fact, this would be undesirable, as the point of modelling is to use abstractions in order to uncover underlying tendencies in a complex world. Thus, this section will describe some of the abstractions which are commonly used to represent spatial relationships between locations. To begin with, we can simply distinguish between data for which we have locations in the form of *points*, usually given as longitude/latitude coordinates, and *areas* in the form of polygons or grid cells.

2.2.1 Distances and neighbours

Representing spatial relationships between the locations of languages is not trivial. Point locations for languages are widely available and more easily accessible than data on language areas. Although these locations are an abstraction from reality, points may provide an adequate (and perhaps the only available) way of representing

the geographic relationships between languages in some cases. To measure the distance between them, some studies have relied on Euclidean distances, which do not take topography or the curvature of the Earth into account, but can provide an adequate approximation of the actual geographic distances between languages (Guzmán Naranjo and Mertner, 2022; Murawaki and Yamauchi, 2018; Ranacher et al., 2021). Topographic distances, which involve calculating the shortest distance between two points while taking terrain elevation into account, provide a more realistic way to measure the geographic proximity between languages (Guzmán Naranjo and Jäger, 2022). In lieu of topographic distances, which can be time-consuming to calculate for large datasets, Great Circle distances, which are calculated 'as the crow flies' and take into account the curvature of the Earth's surface, are a good alternative as they perform almost as well as topographic distances in the model tested by Guzmán Naranjo and Jäger (2022). Models with topographic distances are used in Chapters 5 and 6 of this thesis. As an example, let D be a matrix of topographic distances between a hypothetical set of languages L1, L2, and L3, given in kilometres.

$$\mathbf{D} = \begin{array}{c} \\ L1 \\ L2 \\ L3 \\ L4 \end{array} \begin{array}{c} \begin{array}{cccc} L1 & L2 & L3 & L4 \end{array} \\ \begin{pmatrix} 0 & 50 & 400 & 300 \\ 50 & 0 & 700 & 30 \\ 400 & 700 & 0 & 100 \\ 300 & 30 & 100 & 0 \end{pmatrix} \end{array}$$

A common way of representing the relationships between languages in space is to use a *neighbourhood matrix* (also called a neighbour graph, spatial weights matrix, adjacency matrix, or spatial connectivity matrix) which is often represented as $\mathbf{W}$ in mathematical notation. When $\mathbf{W}$ is binary, w_{ij} indicates whether location i and location j are neighbours, in which case $w_{ij} = 1$. Otherwise, $w_{ij} = 0$ (LeSage and Pace, 2009). $\mathbf{W}$ can also be weighted, such that w_{ij} includes information about the the distance between location i and location j. For example, one could devise a weighting scheme such that w_{ij} reflects the distance between pairs, with higher values for neighbours which are closer together. Defining which pairs of locations in the weights matrix $\mathbf{W}$ are neighbours can be based on a distance threshold beyond which languages are no longer expected to influence each other. All $w_{ii} = 0$, i.e., the same language is not considered a neighbour of itself. Therefore $\mathbf{W}$ must have zeros along the diagonal. Taking the same distances as the one used to specify $\mathbf{D}$, we can now specify a binary neighbourhood matrix, $\mathbf{W}_\mathrm{b}$, such that languages beyond a distance threshold of 500km are not considered neighbours:

$$\mathbf{W_b} = \begin{array}{c} \\ L1 \\ L2 \\ L3 \\ L4 \end{array} \begin{array}{c} \begin{array}{cccc} L1 & L2 & L3 & L4 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

This is a common way of deciding which languages are neighbours ($w_{ij} = 1$) and which are not ($w_{ij} = 0$). For example, Murawaki and Yamauchi (2018) use a distance threshold of 1000 km. Languages within this radius are considered neighbours; any languages which fall outside it are not. The chosen threshold has a significant impact on model results, and the choice is necessarily somewhat arbitrary, as there is likely little difference between the probability of a language pair which is 980 km apart being in contact compared to a pair of languages which is 1003 km apart. There is also a difference between travelling 1000 km in the Kalahari desert and travelling the same distance in the rainforest.

Specifying a fixed, uniform threshold within which languages are no longer expected to influence each other is challenging, as some languages may be in contact across long distances, while in other cases, contact only occurs between immediate neighbours over short distances. Thus, when using thresholds of this kind, we need to make sure they are based on a realistic understanding of contact. For instance, an inverse distance weighting scheme can be applied on top of a distance-based binary neighbourhood matrix or binary matrix of k nearest neighbours, such that languages which are further apart exert less influence on each other. Using the binary matrix $\mathbf{W_b}$ and the distances in D shown earlier, we can now apply an inverse distance weighting scheme such that neighbours which are further apart are given a lower value (and thus influence) in the weighted version of $\mathbf{W}$, $\mathbf{W_w}$, which could look like this:

$$\mathbf{W_w} = \begin{array}{c} \\ L1 \\ L2 \\ L3 \\ L4 \end{array} \begin{array}{c} \begin{array}{cccc} L1 & L2 & L3 & L4 \end{array} \\ \begin{pmatrix} 0 & 0.02 & 0.002 & 0.003 \\ 0.02 & 0 & 0 & 0.03 \\ 0.002 & 0 & 0 & 0.01 \\ 0.003 & 0.03 & 0.01 & 0 \end{pmatrix} \end{array}$$

For model inference, $\mathbf{W}$ is typically row-standardised such that all rows sum to 1, which means the values would vary from the ones shown above. In a row-standardised $\mathbf{W}$, the weights depend on the number of neighbours each language has.

An alternative approach to distance thresholds is to use a k nearest neighbours algorithm, whereby each language is assigned a fixed number of neighbours (k) regardless of the actual geographic distance between them. This approach was used by Kauhanen et al. (2018) and can also be combined with weighting schemes, like an inverse distance weighting scheme, in order to reflect the distance between language pairs. However, we do not expect all languages to have the same number of neighbours, as is the assumption when using k nearest neighbours. In linguistically dense areas, the 10 nearest neighbours of a given language would not even encompass all of its immediate neighbours, whereas 10 nearest neighbours in an area like North Africa could result in the model assuming contact between languages which are thousands of kilometres apart. Adaptive k nearest neighbours algorithms provide a promising way to adjust k according to language density, such that languages in denser areas are assigned more neighbours. These methods are better equipped to handle variation in the spatial density of data points.

An important consideration is that spatial weights can be set according to prior information or hypotheses about the sociolinguistic dynamics of contact. In this way, neighbourhood matrices or graphs are a flexible way for researchers to 'build in' their prior knowledge of how language contact works, or to test different hypotheses about contact against each other through comparing models with different neighbourhood structures.

2.2.2 Areal data

When a variable is distributed on a regular grid or arbitrary lattice (like polygons), the weights matrix $\mathbf{W}$ is typically defined based on *contiguity*. There are two main ways to define contiguity over a set of grid cells, known as rook and queen contiguity, which are illustrated in Figure 2.2. When working with regular lattices like grids, queen contiguity tends to be the default choice. However, for irregular polygons like the ones representing language areas, *polygon contiguity* is defined as any intersection or overlap between polygons. Thus, if any part of a language territory intersects or overlaps with the border of another, they will be considered neighbours.

Figure 2.3 shows what happens when polygon contiguity alone is used to determine which language territories are neighbours and which are not[1]. These polygons (shown in purple on the map) from SIL/Ethnologue represent the geographic extent of languages, and they allow language territories to overlap (Eberhard and Fennig, 2023). It is evident that large language territories have the most neighbours, and they can be neighbours of languages which are far away in terms of the geographic

[1]The R package **spdep** function, **poly2nb**, was used.

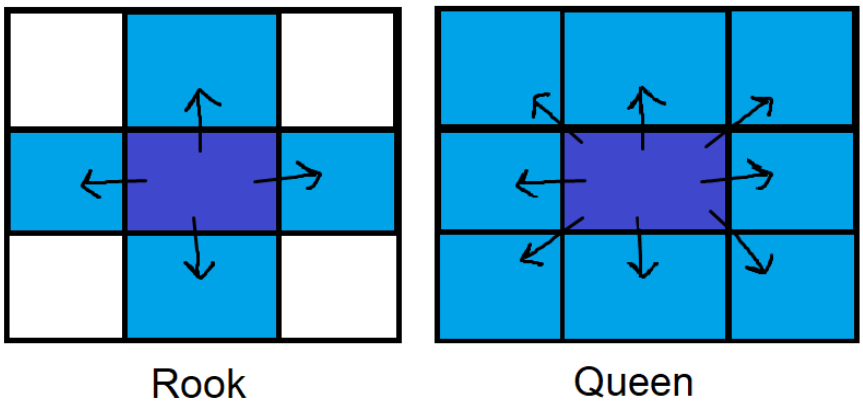


FIGURE 2.2: Rook vs. queen contiguity for regular grids.

distance measured between the centroids of the polygons. Because polygons are irregularly spaced, in most datasets some languages will lack neighbours entirely as their borders do not intersect with those of any other language, even though the distance between them may be relatively short. Some languages will have neighbours which create a disconnected cluster, like when languages are spoken on an island. Although such a representation of neighbourhood may be realistic in some ways, it can create mathematical issues when a graph with many isolated data points and disconnected clusters is used in a spatial model (Wall, 2004). While polygon contiguity is a useful tool when such data is available for languages, it may be a mistake to assume that languages are only in contact if their territories intersect or overlap. To solve these issues, in Chapter 3, I combine polygon contiguity with distances, creating a distance-weighted spatial graph which is fully connected and symmetric. Thus, different ways of specifying neighbourhood can be combined when defining $\mathbf{W}$.

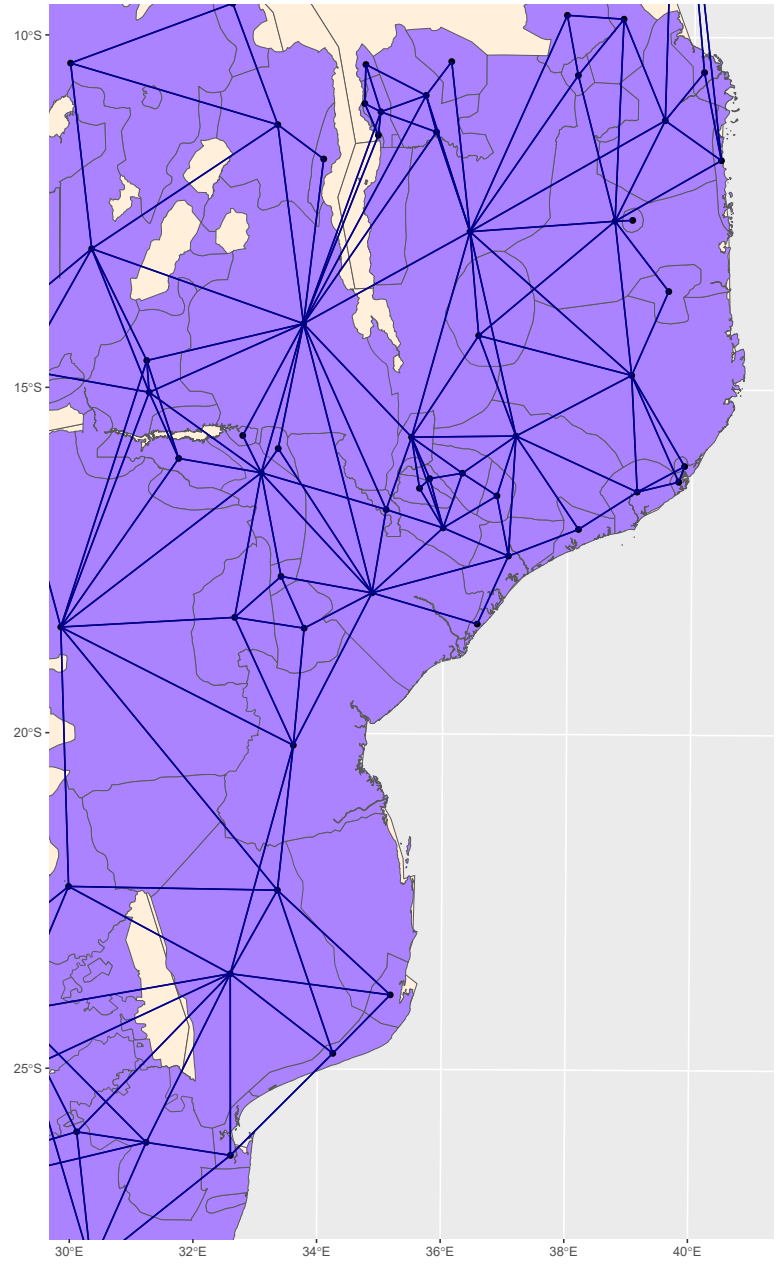


FIGURE 2.3: A neighbour graph based on polygon contiguity for all the languages in Mozambique. Language polygons are shown in purple; the land mass is shown in beige. The centroids of the polygons are depicted as points. The lines show which polygons are defined as neighbours. A line going from the centroid of one polygon to another means that the territories of these polygons intersect or overlap.

2.3 Types of spatial models: An overview

Once we have established the presence of spatial autocorrelation in a linguistic variable of interest, we can start thinking about what kind of spatial model to use. This section will describe some well-known methods for modelling spatially autocorrelated data, many of which will be used in the later chapters of this thesis. Where this is the case, a link to the relevant chapter will be included.

2.4 Autoregressive models

Spatial autoregressive models are a common class of statistical model for the analysis of spatially autocorrelated data in fields such as econometrics, ecology, sociology, epidemiology, and many more (LeSage and Pace, 2009). These models essentially work by estimating the degree of spatial dependence given a spatial graph or neighbourhood matrix which indicates which data points are neighbours and which are not. Depending on how the matrix is defined, additional information can be included, such as the distance between languages. The specification of the neighbourhood/weights matrix $\mathbf{W}$ influences the results of these models and thus requires careful consideration (Cressie, 1994; Wall, 2004).

In this section, I will cover two models which can be used to model spatial dependence given a weights matrix $\mathbf{W}$. This first is the simultaneous autoregressive (SAR) model (Whittle, 1954). The SAR model was introduced to model the *spatial lag* of Y, the dependent variable. The second, developed slightly later, is known as the conditional autoregressive (CAR) model (Besag, 1974). Like the SAR model, this model was introduced to model spatial dependence in a variable Y. Choosing between a SAR and CAR model is not trivial, particularly because they are often used for seemingly similar studies and they both provide a way to deal with the issue of spatially autocorrelated data. However, they operate under very different assumptions, which are important to bear in mind (Ver Hoef, Hanks, and Hooten, 2018).

The fundamental difference between SAR and CAR models lies in their specification. SAR models assume a *global* spatial autocorrelation structure, meaning that every cell is assumed to influence every other cell, though immediate neighbours are usually more influential. Thus, the estimated spatial influence of each cell on the other decays as spatial relationships become more distant. This should not necessarily be interpreted in terms of geographic distance, but as the effect that neighbours of neighbours of neighbours (and so on) have on each other in the model. Global spatial spillover effects "impact the neighbors, neighbors to the neighbors,

neighbors to the neighbors to the neighbors, and so on. Local spillovers represent a situation where the impacts fall only on nearby or immediate neighbors, dying out before they impact regions that are neighbors to the neighbors" (LeSage and Pace, 2019, p. 3). When global spatial dependence is high, this decay happens more slowly. The strength and extent of spatial spillovers is controlled by the parameter ρ. In a model with no covariates, the SAR model would be implemented as follows:

$$y = \rho \mathbf{W} y + \epsilon, \tag{2.1}$$

where ρ quantifies global spatial dependence, $\mathbf{W}$ is the spatial weights matrix, and ϵ is an error term.

CAR models, on the other hand, assume a *local* spatial autocorrelation structure, where the value at location i is conditional on the values of its immediate neighbours. This is known as the *Markov property*, and CAR models are also referred to as a specific type of *Markov random field* (MRF). According to Ver Hoef, Hanks, and Hooten (2018), the spatial CAR model is typically specified as follows:

$$Y_i | y_{-i} = \mathcal{N}\left(\sum_{j=1}^{n} w_{i,j} y_j, m_{i,i} \right), \tag{2.2}$$

where Y_i is a random variable at the ith location, y_j is its realised value, y_{-i} is the vector of values for all y_j where j is not equal to i, $\mathbf{W}$ is the spatial weights matrix with $w_{i,j}$ as its i, jth element, and n is the total number of points in the model. $\mathbf{M}$ is a diagonal matrix with positive diagonal elements $m_{i,i}$, which describes the variance. The values of $m_{i,i}$ may depend on the values of the ith row of $\mathbf{W}$. Overall, what this formula tells us is that the conditional mean of Y_i is a weighted linear combination of the values at neighbouring locations, without taking into account the values of every location in $\mathbf{W}$. Because the spatial dependence induced by a CAR model only includes the immediate neighbours of a given location as defined by $\mathbf{W}$, this means that we can use sparse representations of $\mathbf{W}$, which makes computations more efficient. Additionally, this property renders it more flexible than the SAR model when it comes to including disconnected clusters of languages, which can arise when using language polygons or a low distance threshold to create $\mathbf{W}$.

2.4.1 Weights and asymmetry

Both CAR and SAR models depend heavily on the specification of $\mathbf{W}$. Some have proposed methods for inferring $\mathbf{W}$ from data; I will not be describing these approaches as they have not been used in this thesis, but they will be discussed briefly in Chapter 6. An alternative strategy is to compare the model using cross-validation using different plausible specifications of $\mathbf{W}$. It is also crucial to draw on prior knowledge of plausible contact ranges, the terrain, and the social dynamics of the area under study. One way of incorporating such knowledge is to set spatial weights. For example, inverse distance weighting schemes provide a simple way to build in the assumption that languages which are close together are more likely to influence each other than languages which are far apart. This is commonly used for point locations, but it can also be applied on top of a neighbourhood matrix defined using polygon contiguity (for an example, see in Chapter 3 of this thesis). If there is reason to think that other factors facilitate or hinder contact between languages, these can also be operationalised as weights. The only drawbacks to doing so are 1) the possibility of incorporating incorrect assumptions directly into the model structure, and 2) the increased complexity of the model. Both of these can be mitigated through model evaluation and comparison.

Asymmetry is a more difficult concept to tackle. Many contact situations are inherently asymmetric, and the effects of such contact on the languages involved will reflect this (Matras and Sakel, 2007). Some of the factors which can cause asymmetric contact effects include prestige and hierarchical relationships between languages, differences in community size or political or military power, and cognitive biases which may favour one linguistic variant over another (Trudgill, 2010). Therefore, a spatial model which allows for asymmetric contact relations would be a welcome addition to the field. However, most of the existing methods cannot incorporate asymmetry in the spatial relationships between languages. A CAR model requires a symmetric weights matrix $\mathbf{W}$ (Besag, 1974). Thus, if the influence that languages have on each other across space is likely to be asymmetric, then a SAR model is a better choice, as it does not require a symmetric $\mathbf{W}$. However, the nature of the asymmetry must be specified prior to model fitting; it cannot be inferred by the model. For example, it would be possible to assign a larger weight (spatial influence) to languages with a larger community size, modelling the assumption that larger language communities are likely to exert a greater degree of influence on the smaller languages around them than vice versa.

2.5 Gaussian processes for point locations

Conceptually, including a a latent *Gaussian Process* (GP) in a model means making the assumption that languages which are closer together in space are likely to be more similar than languages which are further apart. It is a way of detecting the underlying or hidden spatial patterns which influence the observed variables (McElreath, 2020). These spatial patterns can be thought of as a 'hidden surface' which is not directly observed but is inferred by the model. Let's say the observed variable is the presence or absence of a phoneme. The presence of this particular phoneme is strongly predicted by its presence in languages within 100 km of each other, but there are some exceptions. A few languages, perhaps due to other factors like common inheritance, lack the phoneme despite being within a 100 km radius of languages which have it. However, the underlying spatial pattern inferred by the model should still reflect a high probability of languages being similar within a range of 100 km. This is how a latent GP can detect underlying spatial trends given noisy data.

One of the useful features of a GP is that the spatial range within which data points are expected to be similar is inferred from the data (Williams and Rasmussen, 2006). In the hypothetical example mentioned above, the model infers similarity between the values of languages which are located within 100 km of each other, but for a different feature, this radius could be 300 or 600 km (Guzmán Naranjo and Mertner, 2022). This concept is often referred to in the literature as the *smoothness* of the function (Williams and Rasmussen, 2006). In practice, smoothness refers to how gradually similarity between observations decays as the distance between them increases. Because of this property, it is not necessary to specify prior to model fitting which languages should be considered neighbours, as it is when using an autoregressive model. Thus, using a latent GP allows us to avoid the issues associated with defining the spatial weights matrix $\mathbf{W}$. This makes a GP uniquely suitable for modelling contact situations in which the expected range of language contact or areal convergence is unknown.

In more technical terms, a GP is an extension of the Gaussian (normal) distribution to high-dimensional spaces. It can handle *non-linear dependencies* between data points in Bayesian regression models (Williams and Rasmussen, 2006). Spatial autocorrelation is an example of a non-linear dependency, so GPs are well-suited to modelling spatial data. A GP can capture the covariance (expected similarity) between all the points in a dataset. In a spatial latent GP model, the covariance between two points represents how similar observations are expected to be based on their proximity in space. There are multiple formulas for calculating covariance in a GP, and these are referred to as *kernel functions*. The kernel function calculates the

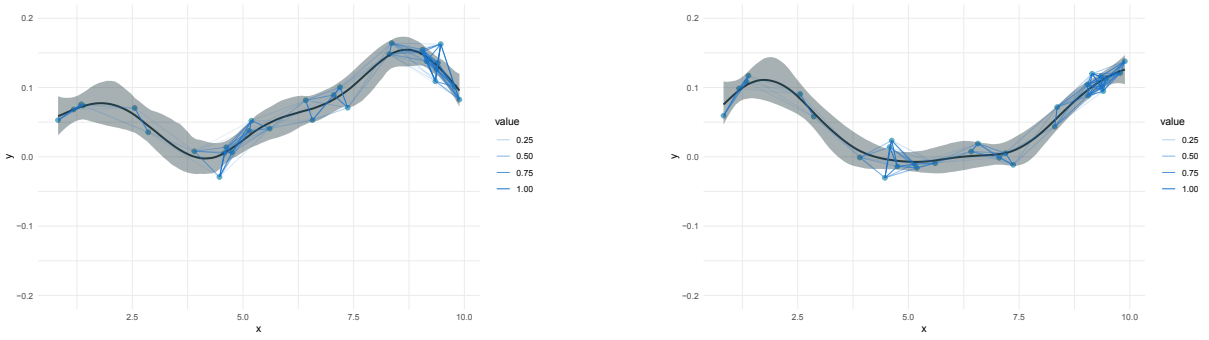
expected similarity between observed data points. Thus, it controls how much data points influence each other across space (Williams and Rasmussen, 2006).

A common default choice of kernel function is the *squared exponential* kernel, which assumes that the similarity between points decreases gradually as distance between them increases. In other words, it assumes a smooth function. However, this assumption is not always warranted, especially when modelling spatial data. The similarity between languages could decay sharply rather than gradually past a certain distance. To solve this issue, the Matérn class of kernel functions was developed for spatial forestry data, which is likely to show sharp discontinuities and thus violate the smoothness assumption of the squared exponential kernel (Matérn, 2013; Stein, 2012). The Matérn class of kernel functions is better suited to modelling these kinds of discontinuities in the data.

A GP has two hyperparameters. The first is commonly referred to as the *length-scale* or *horizontal scale*, and it describes how quickly the correlation between two points (languages) decreases as the distance between them increases. A large value of the horizontal scale indicates that data points may be correlated across long distances. For linguistic data in space, this indicates the geographic range within which languages show similarity. It can be interpreted as the areality or areal extent of a feature.

The second hyperparameter is called the *vertical scale* and is the marginal standard deviation of the GP. It is also frequently referred to in the literature as the *amplitude*. It can be thought of as the expected level of variation in the data across space. Thus, a higher amplitude indicates greater variability in the feature values of languages at different locations. This could manifest as spatial clusters of feature values which diverge heavily from the values at other locations. When amplitude is high, we would expect to see 'hotbeds' for particular features with sharp boundaries around them. In contrast, a low amplitude could indicate a greater level of randomness in the distribution of feature values. Spatial clusters may not be as obvious or clearly demarcated. The amplitude is unrelated to and should not be confused with the degree of spatial autocorrelation.

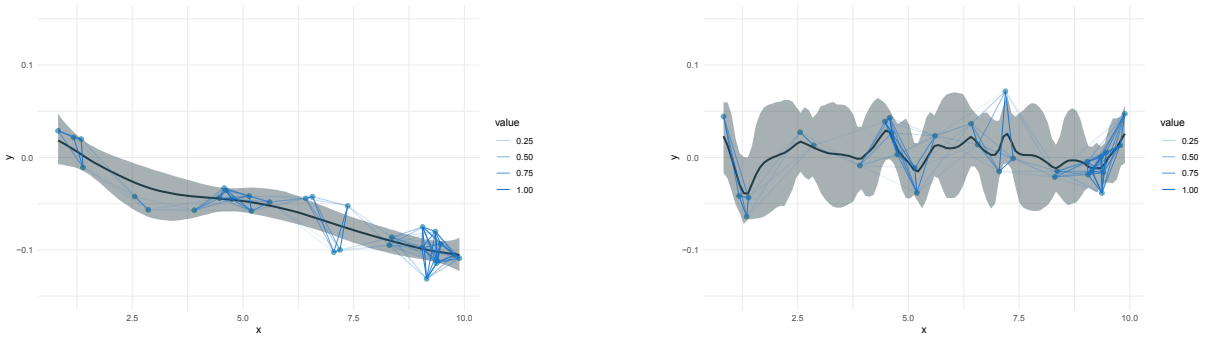
The effects of tweaking these hyperparameters for a simulated data example are shown in Figure 2.4 and 2.5. The data was drawn from a Gaussian process model given fixed values of the hyperparameters, after which a model was fitted to the data. The plots show the conditional effects and 95% confidence intervals of the fitted models. Additionally, the correlation matrix given the relevant values of the horizontal and vertical scale parameters is plotted in the form of lines between the points. The thickness of the line indicates the strength of potential correlation between two points. The correlation matrix is determined by the value of the horizontal scale,



(A) Horizontal scale = 1, vertical scale = 1.

(B) Horizontal scale = 0.5, vertical scale = 1.

FIGURE 2.4: Effects of varying the horizontal scale parameter on simulated data, with the correlation matrix shown as lines between points.



(A) Horizontal scale = 0.5, vertical scale = 1.5.

(B) Horizontal scale = 0.5, vertical scale = 0.1.

FIGURE 2.5: Effects of varying the vertical scale parameter on simulated data, with the correlation matrix shown as lines between points.

as a larger scale increases the distance along x at which points are expected to be correlated.

Because the parameters of a GP model interact with each other and can be difficult to interpret directly, it is crucial to visualise the results on a map. One way of doing so, which is used by Guzmán Naranjo and Mertner (2022) and Guzmán Naranjo, Mertner, and Urban (2024), is to draw predictions from the GP after model fitting using the inferred parameters. These predictions can then be plotted on a map. This is helpful in visualising, evaluating, and interpreting the underlying spatial patterns inferred by the GP.

2.6 Spatially lagged covariates

All of the methods so far have been designed to handle spatial lags or spatial autocorrelation in Y, the dependent variable. However, it can be equally interesting and relevant to model the spatial lag of X (SLX), the explanatory variable(s). This is referred to as a *spatial spillover effect*, which has been defined as "the marginal

impact of a change to one explanatory variable in a particular cross-sectional unit on the dependent variable values in another unit" (Elhorst and Halleck Vega, 2017, p. 2). Thus, in addition to modelling the direct effect of X at a particular location on Y at the same location, the effect of X at neighbouring locations can also impact Y at the original location. This can be an especially valuable addition to a model when X is spatially autocorrelated.

Intuitively, this holds for some of the variables which have been hypothesised to impact linguistic diversity and structural or phonological features, like temperature. In Chapter 3 of this thesis, I examine the effect of variables such as political complexity and climate on linguistic diversity. The climate of the area in which a language is spoken is not independent of the climate at neighbouring locations. A language spoken in a hot area with a high level of rainfall is likely to be surrounded by languages with a similar climate, which means this variable will be spatially autocorrelated.

Additionally, it makes sense to consider the values of neighbouring locations when we think these could have an impact on the variable of interest. To give a slightly different example, linguistic diversity could easily be impacted just as much by the terrain elevation of its neighbours as the terrain elevation of the area within which it is spoken. This might be the case if the terrain of neighbouring languages provided a barrier to travel, causing isolation and thereby differentiation or the maintenance of distinct languages.

SLX terms can be incorporated into a regression model using the following formula:

$$Y = \mathbf{W}X\gamma + X\beta, \tag{2.3}$$

where $\mathbf{W}$ is a row-standardised spatial weights matrix, $\mathbf{W}X$ is the mean value of covariate X, and γ is a coefficient vector. A function for the inclusion of SLX terms in spatial Bayesian regression models is implemented in the R package **geostan**, the documentation for which also provided the formula above (Donegan, 2022, p. 74).

2.7 Case study: Labial-velar consonants in Africa

In this section, I will briefly compare two of the methods presented here, not just in terms of their predictive ability but also in terms of the results obtained from each model type. The dataset which forms the basis of this case study was compiled by Idiatov and Van de Velde (2021), who examine the areal distribution of of labial-velar stops (henceforth LV stops) in Africa. The data itself came from RefLex, which

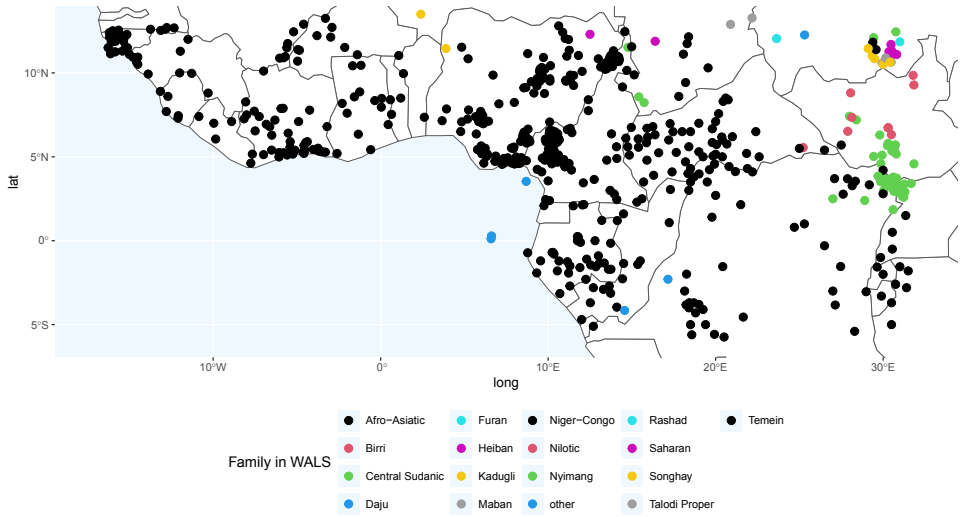
combines multiple sources of lexical data across Africa (Segerer and Flavier, 2011-2021). LV stops, like /kp/ and /gb/, are a class of co-articulated phonemes "that are produced with almost simultaneous gestures of velar and labial closure" (Idiatov and Van de Velde (2021, p. 72), citing Ladefoged and Maddieson (1996, pp. 332–43)).

Languages with LV stops form an interesting case study of contact effects, as they are found across a diverse set of language families in a geographically contingent area in Africa. Meanwhile, LV stops are exceedingly rare outside this area (Clements and Rialland, 2008). Their presence was a key factor in the proposal of a large linguistic area called the Macro-Sudan belt, which stretches from Senegal in western Africa to the Central African Republic (Güldemann, 2008). Another study by Clements and Rialland (2008) identifies a very similar geographic area of phonological similarity, notably including LV stops as one of its defining features, which they call the Sudanic belt. LV stops are mainly concentrated in West Africa and parts of central Africa, although a few of the languages which have them are spoken further east. The geographic extent of languages with these consonants matches the borders of the proposed Macro-Sudan belt quite well (a map of the area is shown in Güldemann (2018b, p. 473)). Because of their conspicuous areal distribution, their spread has frequently been attributed to language contact. Their origins are still disputed, with some arguing for an innovation- or inheritance-based explanation coupled with diffusion through language contact, while others hold that they may have been a feature of an earlier substrate language and thus should not be reconstructed to any of the proto-languages of the region. Regardless of their origins, which Idiatov and Van de Velde (2021) describe in more detail in light of their findings, contact undoubtedly played a major role in the spread of LV stops.

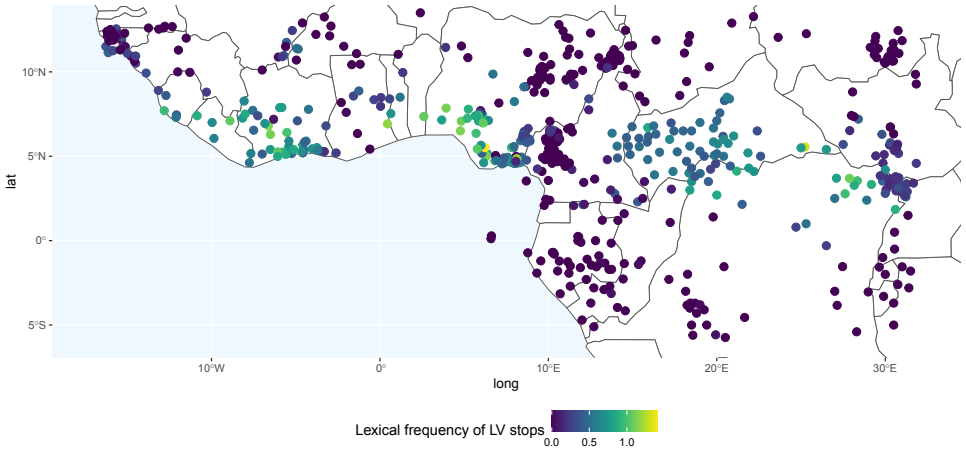
In their original study, Idiatov and Van de Velde (2021) calculate the lexical frequency of LV stops in the lexicon of the languages which have them. Because detailed lexical data is only available for some of the languages in the dataset, this results in a somewhat smaller but still rich sample of genealogically diverse languages in Africa. In this study, I will use the lexical frequency of LV stops in each language as the outcome variable, rather than a binary indicator of whether a language has LV stops or not. As argued by Idiatov and Van de Velde (2021), lexical frequency provides a more detailed measure of how entrenched LV stops are in the languages that have them, thus allowing for a more detailed analysis of their areal distribution.

I excluded languages which were outside the maximum geographic extent of languages with LV stops, resulting in a sample of 581 languages in a geographic area stretching from the coast of Guinea-Bissau to the Central African Republic. I did not exclude languages without LV stops, but restricted the area under study in order to avoid a highly skewed distribution resulting from the inclusion of languages across

southern and eastern Africa with no LV stops. The language sample is shown on a map in Figure 2.6a, and the normalised data on the lexical frequency of LV stops in the language sample is shown in Figure 2.6b. In order to stay as close to the original study as possible, I followed Idiatov and Van de Velde (2021) in using a varying intercept for language family, which is defined as the family affiliation given in WALS (Dryer and Haspelmath, 2022), to control for shared ancestry. This remains a common approach in quantitative typology, although some studies have argued for the use of phylogenetic regression instead (Guzmán Naranjo and Becker, 2021).



(A) The point locations of the languages in the study, coloured by their family affiliation as given in WALS.



(B) The point locations of the languages in the study, coloured according to the lexical frequency of LV stops. High values are in yellow; low values are in dark purple.

2.7.1 Spatial weights

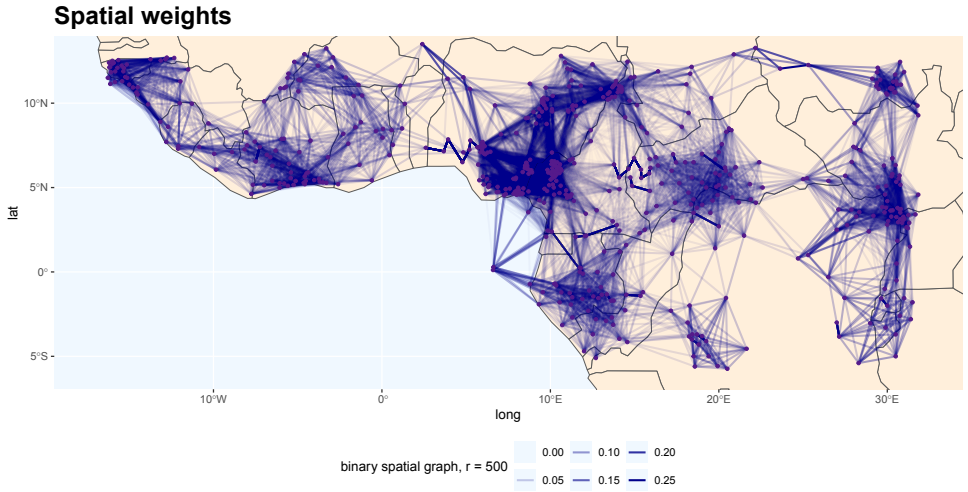
The main goal of this case study is to compare a simultaneous autoregressive (SAR) model using different weights matrices, $\mathbf{W}$, with a Gaussian process (GP) model. For the SAR models, three ways of defining $\mathbf{W}$ were compared. The first was a binary graph with a distance radius of 1000 km, such that any languages within that radius are considered neighbours ($w_{ij} = 1$) and any languages outside that radius are not neighbours ($w_{ij} = 0$). The resulting graph is depicted on a map in Figure 2.7a. The second graph likewise uses a radius of 1000 km to define with languages are neighbours and which are not, but instead of assigning w_{ij} a binary value based on neighbourhood, exponential weights were calculated based on Great Circle distances, indicating how far away the languages are from each other. Closer languages have a higher weight and more distant languages have a lower weight, as depicted in Figure 2.7c. The third graph was derived in the same way as the second, using a radius of 500 km instead of 1000 km, as shown in Figure 2.7c. This graph looks very similar to the previous one because the exponential weights are very low when languages are more than 500 km away from each other, resulting in faint connections between the languages. I also tested these against an exponentially weighted $\mathbf{W}$ with a smaller radius of 300 km, which is not depicted on the map but is shown in the model comparison section.

For the GP model, I used Euclidean distances, which are less realistic as a representation of space than other distance measures but are still widely used, as in e.g. Guzmán Naranjo and Mertner (2022) and Ranacher et al. (2021). Because a GP estimates the expected range of spatial effects, there was no need to define $\mathbf{W}$ or to use a distance threshold. However, a GP takes significantly longer to run[2].

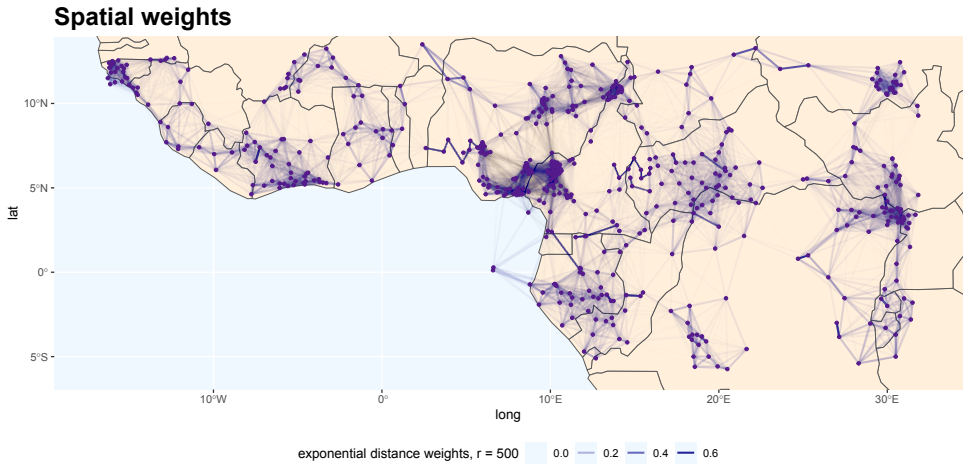
2.7.2 Model evaluation

For this case study, the GP model, even with Euclidean rather than circle distances, by far outperformed the SAR models with variable distance thresholds. The SAR models with exponential distance weights also by far outperformed the model with a binary neighbour graph, indicating the importance of including weights. In contrast, the differences between the distance thresholds r when using a weighting scheme were minimal. All the models also included a varying intercept for language family. The model with only family-level intercepts, without any spatial controls, had the worst performance by far when compared to the models which included both. The results of the cross-validation using PSIS-LOO (Pareto-smoothed importance sampling) are

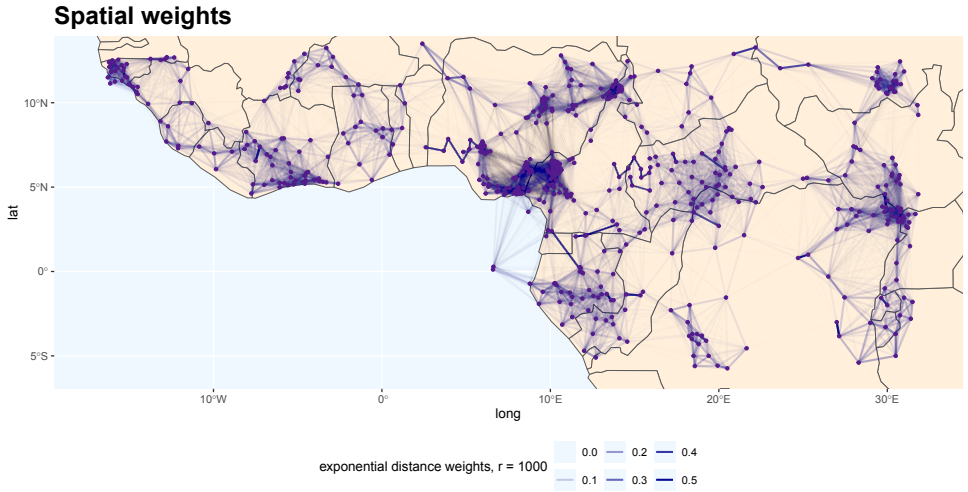
[2]The SAR models finished sampling in approximately 4 hours, while the GP model finished sampling in around 3 days.



(A) A binary spatial graph in which neighbours are defined as language points located within a radius of 500 km of each other.



(B) A spatial graph with exponential distance weights and neighbours defined within a radius of 500 km. The transparency of the line represents the weights. Neighbours are defined as language points located within a radius of 500 km of each other.



(C) A spatial graph with exponential distance weights and neighbours defined within a radius of 1000 km. The transparency of the line represents the weights.

Model	**ELPD difference**	**SE difference**
GP + family	0.0	0.0
SAR (r = 500, exp. weights) + family	-195.0	34.9
SAR (r = 300, exp. weights) + family	-195.3	34.6
SAR (r = 1000, exp. weights) + family	-196.7	35.0
SAR (r = 1000, binary) + family	-307.2	39.0
Family only	-508.0	43.1

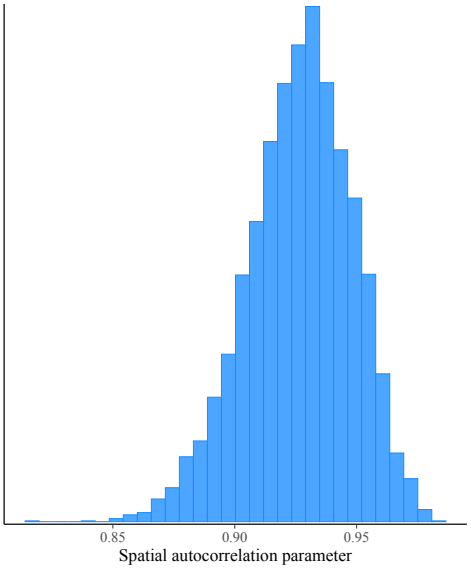
TABLE 2.1: Comparison between GP and SAR models for LV stops.

shown below (Vehtari, Gelman, and Gabry, 2017; Vehtari et al., 2024). PSIS-LOO is an approximation of leave-one-out cross-validation, a common way of evaluating the out-of-sample predictive accuracy of Bayesian models, which is largely used as a metric of how well the model fits the data (McElreath, 2020). Using an approximation makes it more tractable to compare complex models. When using PSIS-LOO, the models are compared in terms of ELPD (expected log predictive density), a metric defined by Vehtari, Gelman, and Gabry (2017, p. 2) which is an approximation of out-of-sample predictive accuracy. It cannot be interpreted in absolute terms, only in relation to other models. A higher ELPD is considered better. When doing comparisons, the best model has an ELPD of 0.0, and the models which follow it have negative values that indicate the magnitude of the difference in expected predictive accuracy between them.

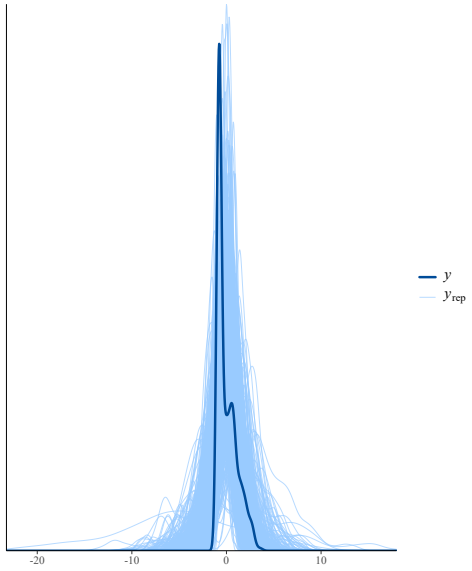
2.7.3 Comparison of results

Here, I will present the results of the best SAR model alongside the results of the GP model. A SAR model estimates a single parameter that controls the degree of spatial dependency in the data. In this case, a very high degree of spatial autocorrelation was estimated ($\rho = 0.93$), confirming the areality of LV stops. Draws from the posterior distribution of the SAR model for ρ and the outcome variable are shown in Figure 2.8.

This tells us that LV stops are very likely to have spread through areal diffusion. However, a drawback of the SAR model is that it fails to detect local areal clusters. This is likely because of the global spatial autocorrelation structure, which is much less flexible than a GP. In the predictions drawn from the GP model in Figure 2.9, we can see that the strongest hotbed for the lexical frequency of LV stops in the lexicon occurs in a region bordering Nigeria and Benin, which could be the region where these stops first originated (Idiatov and Van de Velde, 2021). A slightly weaker, disconnected



(A) Draws from the SAR ρ parameter.



(B) Draws from the posterior predictive distribution; y_{rep} shows the posterior draws and y is the original variable.

FIGURE 2.8: Posterior draws from the SAR model.

hotbed occurs further west around Côte d'Ivoire, and there is a still slightly weaker one in further east which includes part of the Central African Republic.

For this dataset, a GP far outperformed the SAR model, but AR models still can be useful to control for spatial autocorrelation in a computationally tractable way. They can also give a relatively straightforward estimate of the degree to which a feature is spatially autocorrelated after controlling for other variables, like family membership. If the goal is to produce detailed visualisations of where exactly a linguistic feature is expected to cluster in space, then a GP is the best suited option. In this case study, the GP also had a much higher expected predictive accuracy, which might be due to the high level of spatial heterogeneity in this dataset. Nonetheless, I would recommend their use whenever it is feasible for the size of the dataset. For SAR models and AR models more generally, I would recommend using distance weights rather than a binary graph, and comparing different specifications of the graph, as these can significantly impact the results. Using a more detailed phylogenetic control, including a covariance matrix indicating how closely related the languages in the study are, could also improve these models.

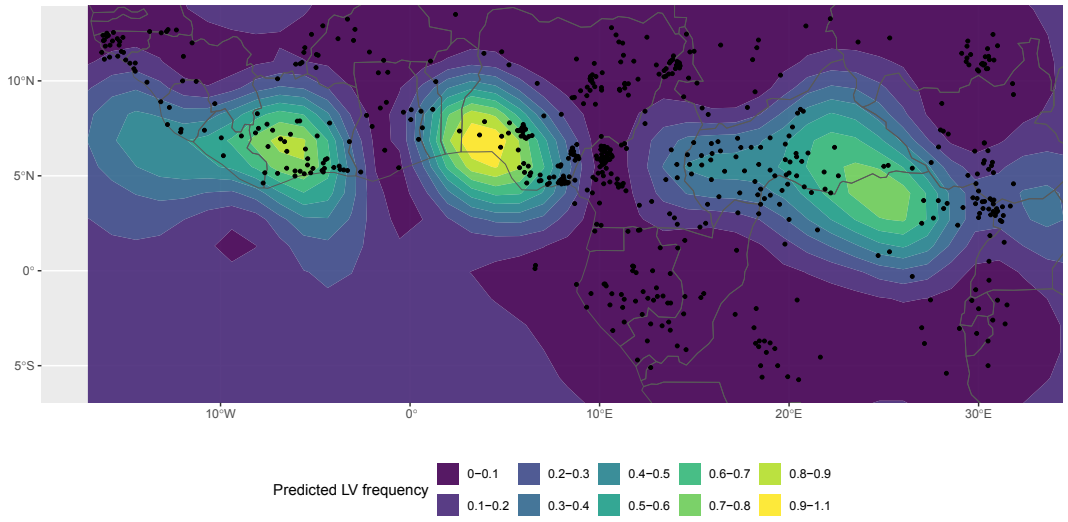


FIGURE 2.9: The predicted distribution of the lexical frequency of LV stops drawn from the latent GP model.

2.8 Limitations and challenges

Gaussian processes and autoregressive (AR) models are both common choices for modelling spatial data, and they each have advantages and limitations. The main advantage of a GP over AR models is that they do not require the prior specification of the spatial weights matrix $\mathbf{W}$. Moreover, a latent GP can identify areal clusters in a more flexible way than an AR model and, as shown, they are significantly better at predicting unseen data, which has positive implications for their use in missing data imputation (Guzmán Naranjo, Mertner, and Urban, 2024). However, all of this comes at a computational cost. Exact GPs are computationally intensive and therefore, they are prohibitively expensive to apply to large datasets. This limits their applicability to large-scale studies. However, approximation methods for GPs exist and can provide a way around the long computation time. The basis function approximation for GPs presented in Riutort-Mayol et al. (2022) is implemented in the R package **brms**, and it was coded in Stan for Chapter 6 of this dissertation[3]. The main disadvantage of this approximation method is that it requires the use of Euclidean distances, which means sacrificing some of the geographic realism of the model. It also requires the number of basis functions to be specified prior to model fitting, which is non-trivial (for practical recommendations on this, see Riutort-Mayol et al. (2022)).

The flexibility of a GP can make the interpretation of results challenging. Unlike in CAR/SAR models, none of the estimated parameters in a GP directly respond to the degree of spatial autocorrelation. Because of this, it is not enough to examine

[3]This was done with help from Matías Guzmán Naranjo.

the posterior estimates and uncertainty intervals of the parameters; it is also crucial to visualise the results of the GP using spatial interpolation methods, which can be challenging to code. These visualisations are crucial to understanding the model results and even then, they are not always straightforward to interpret.

Another issue is that GPs require locations to be represented as points, not areas, precluding the use of polygon data with GP models. However, this is a field experiencing fast developments and this limitation of may be overcome in the future. Using a realistic, well-founded measure of geographic distance can mitigate this issue (Guzmán Naranjo and Jäger, 2022). Additionally, asymmetry cannot be directly incorporated into the covariance matrix used for a GP, as this must be symmetric, which is a potential limitation when modelling contact effects which may be asymmetric (Williams and Rasmussen, 2006). However, see Guzmán Naranjo et al. (2025) for an alternative way to include asymmetry in a GP model while keeping the covariance matrix symmetric.

In AR models, areal data can be used directly to define spatial contiguity, which means that these models may be more suitable when polygon data is available. When polygons are not available, some researchers use triangulation methods, such as Voronoi tessellation, in order to derive approximate polygons from the point locations at which languages are spoken (Aurenhammer and Klein, 2000). These can then be used to define neighbouring languages based on which of their borders intersect. However, there is little consensus on whether the resulting polygons are a realistic representation of the geographic extent of languages (Kälin, 2017). In real life, language territories can and do overlap with each other, which is not the case when using Voronoi tessellation. In particular, it is uncertain whether these approximated polygons are better than using a realistic measure of distances between point locations, such as topographic distance.

Arguably the main limitation of AR models lies in the need to define the spatial relationships between languages through the specification of a neighbour graph or matrix, $\mathbf{W}$. Often, we do not know prior to model fitting which languages have been in contact and which have not. Thus, although these models are less computationally intensive to fit than a GP, if there is any uncertainty around $\mathbf{W}$ (which is usually the case), it may be necessary to conduct more extensive model comparison and evaluation in order to make sure that the specification of $\mathbf{W}$ will not cause biased results. As done in the case study, one way to do this is to compare a set of models with different specifications of $\mathbf{W}$, which are based on plausible distance thresholds for contact. This can be used as an indirect way to infer the most likely spatial range within which language communities are in sufficiently regular contact to effect linguistic change. Thus, the issues with specifying $\mathbf{W}$ can be mitigated through

model evaluation as well as using the existing literature on the dynamics of language contact and areal diffusion in order to create the most realistic $\mathbf{W}$ possible (see e.g. Bickel (2017), Evans (2020), Lüpke (2016), Mansfield, Leslie-O'Neill, and Li (2023), Nichols (2020), Nikolaev (2019), and Pakendorf, Dobrushina, and Khanina (2021), among many others).

2.9 Discussion

Because this chapter cannot describe every possible method for modelling data in space, the following section is intended to give a brief outline of some alternative approaches to the ones which have been discussed so far. These approaches have not been implemented as part of this thesis, but may be useful for the reader interested in exploring further spatial modelling possibilities.

Generalised additive models (GAMs), as implemented by Idiatov and Van de Velde (2021) and Winter and Wieling (2016) and others, are outside the scope of this thesis. Similar to GPs, a GAM allows for the use of splines, which provide a less computationally intensive way to estimate non-linear dependencies between data points (McElreath, 2020). However, they suffer from the same issues with interpretability as GPs, so high-quality visualisations are crucial for understanding their results (Idiatov and Van de Velde, 2021).

Geographic distances can also be used to specify a covariance matrix prior to model fitting, given a kernel function and a set of parameters, as done by Skirgård et al. (2023) using Great Circle distances. Skirgård et al. (2023) also make use of a recent method for approximate Bayesian model inference, implemented for R users in **R-INLA**, which allows the user to fit fast GPs and other spatial models. Unlike Stan, these models do not require sampling via MCMC chains as they result in an approximate posterior rather than an exact one, and are thus extremely fast. However, this interface lacks the diagnostics provided by Stan which can alert us to model misfit, misspecification, and mathematical errors, which may cause issues to go unnoticed. Nonetheless, it is a promising framework for spatial modelling of very large datasets which should be noted here as an alternative to more traditional MCMC sampling methods.

The focus of this chapter, and this dissertation, is on the spatial modelling of synchronic language data. The temporal dimension of linguistic variation and change is included as a control in all the models presented in this thesis, but these models do not explicitly model evolution like phylogenetic models do, which is a potential limitation (Jäger, 2013).

2.10 Conclusion

This chapter has been intended as a mostly conceptual overview of some of the Bayesian spatial modelling techniques which are applicable, and in some cases have already been used, in studies on linguistic typology. Details of their implementation in R have been provided where possible. Spatial models, like any statistical method, should be selected based on the type of geographic data and the amount of data available, as well as the goal of the study. The main goal of this chapter has been to illustrate the breadth of possibilities in the field of Bayesian spatial models and how these methods can be used in quantitative studies of language contact and related fields in which controlling for the effects of contact might be necessary. It should be clear that there is no 'one size fits all' solution to modelling the effects of language contact and areal diffusion. Having a diverse set of methods at our disposal is necessary if we are to tackle a diverse set of questions.

Chapter 3

Cultural and environmental correlates of linguistic diversity in Africa

In this chapter, I will use a Bayesian spatial model to investigate some of the cultural and environmental correlates of linguistic diversity in Africa. One of the goals of the study is to examine whether the variables which influence linguistic diversification are distinct from those which influence the emergence and maintenance of phylogenetic diversity over time. A methodological innovation is that the model contains spatially lagged predictor variables, which can capture covariate effects of spatial neighbours on linguistic diversity (Donegan, Chun, and Griffith, 2021). This method has not been applied to the study of diversity before. The results suggest that some of the effects of cultural and environmental variables on language density and phylogenetic diversity are indeed different, suggesting that it may be helpful to distinguish between quantitative measures of diversity which focus on how different languages are as opposed to those which focus on how many languages are spoken in a given area. An additional goal of this study is to evaluate previous hypotheses and findings on the variables which impact linguistic diversity in the context of Africa using a new methodology (Currie and Mace, 2009; Hua et al., 2019; Huisman, Majid, and Hout, 2019; Nettle, 1996, 1998). The results provide support for many of the previous findings, with the caveat that they depend on how we define linguistic diversity in quantitative studies.

3.1 Introduction

Around 7,000 distinct languages are currently spoken across the globe, although the exact number remains elusive (Eberhard and Fennig, 2023; Hammarström et al., 2023; Nettle and Romaine, 2000). In large part, this has to do with the ongoing extinction of the world's languages, many of which have never been described and about which very little is therefore known (Evans, 2010). Another issue with assigning

a number to the world's languages is that the line between languages and dialects is famously blurry. Should two lects which are mutually intelligible but which have different names and are spoken by distinct communities be counted as two distinct languages? The answer to this question impacts how we measure and study linguistic diversity, including how it arises and how it is lost, questions which have occupied researchers in linguistics for several decades and which continue to be relevant today (Bromham et al., 2022; Currie and Mace, 2009; Evans, 2010; Hua et al., 2019; Nettle, 1998; Nichols, 1992).

Linguistic diversity is unevenly distributed across the world, as illustrated in Figure 3.1. However, linguistic diversity is not just unevenly distributed in the sense that, if we divide the land mass of the world into grid cells of equal area, some grid cells may have hundreds of languages while others have only one or two, or none at all. The uneven spread of languages is clear in the distribution of speakers, too: the fifteen largest languages in the world are spoken by almost half of the world's population (Nettle and Romaine, 2000). In reality, languages are not discrete entities; the majority of the world's population speaks at least two languages, and thus, most of the world's languages exist in simultaneous relation to other languages (Evans, 2017). In many cases, this relationship is hierarchical or diglossic, with one language being used in formal settings like politics, bureaucracy, and education, while other languages spoken by the same population are restricted to informal social settings (Nettle and Romaine, 2000). In this way, the distribution of languages in space is inextricably linked to the sociolinguistic settings in which they are spoken. Languages are also spoken in diverse geographic environments; some are spoken in tropical areas, where linguistic diversity tends to accumulate alongside biodiversity (Nettle, 1998), while others are spoken in desert environments or mountainous areas. The environment has been hypothesised to impact linguistic diversity in direct and indirect ways (Hua et al., 2019; Huisman, Majid, and Hout, 2019; Nettle, 1998). The purpose of this study is to estimate and disentangle the impact of cultural and environmental predictors of linguistic diversity, using two different ways of quantifying linguistic diversity in order to shed light on distinct diachronic processes of diversification.

The map in Figure 3.1, in depicting only the number of languages within each grid cell, obscures an important facet of linguistic diversity, namely phylogenetic diversity. This is especially relevant for Africa. Africa is the continent with the longest history of human habitation, with over 2,000 languages spoken across the continent. It may seem surprising in light of this that Africa has been characterised by some as lacking in phylogenetic diversity (Blench, 2013). Compared to North and South America, Africa has almost double the number of extant languages, but while the Americas combined have over seventy isolates and over forty small phyla, the

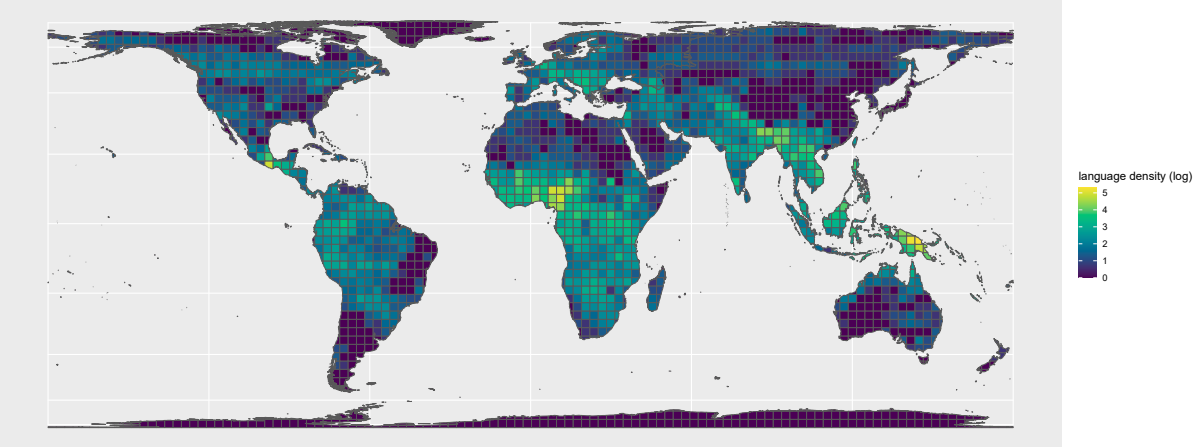


FIGURE 3.1: Global language density on an equal-area projection divided into 300 by 300 km grid cells.

majority of Africa's languages have been grouped into the four major phyla proposed by Greenberg (1963), and there are few undisputed isolates, according to Blench (2013, p. 44).

The established genetic groupings of Africa are not uncontested. Dimmendaal (2008b) and Güldemann (2018a) favour a more diverse view of African languages than Blench (2013). Some of the internal divisions within Niger-Congo, such as the Atlantic group, could be areal rather than genealogical groupings, and other subgroups which Greenberg (1963) included in the phylum have arguably weak evidence for this affiliation. Dimmendaal (2008b) argues for the treatment of these groups, Mande and Ubangi, as independent language families, at least until stronger evidence for a Niger-Congo affiliation can be presented. This view is reflected in the uncertain classification of groups such as Mande, Ubangi and Songhay in Hammarström et al. (2023). Uncertainty around genealogical classifications can make it more difficult to measure phylogenetic diversity, which may explain why language density has been the focus of several prior studies like Currie and Mace (2009) and Hua et al. (2019), while Bromham et al. (2022) take a different approach by measuring language endangerment rather than diversity. In contrast to previous approaches, the present study will treat language density and phylogenetic diversity as potentially driven by distinct processes.

3.1.1 Previous findings and hypotheses

A striking pattern emerges when one visualises the distribution of global language density: languages tend to be clustered most densely along the Equator, where the climate is tropical. The most linguistically dense area in the world is found in modern-day Cameroon in Africa, which is located near the Equator. Tropical areas also have the largest number of distinct flora and fauna in the world, and indeed, language diversity has been found to correlate with biodiversity (Nettle, 1998). Tropical climates are characterised by an abundance of natural resources, the availability of which varies little from season to season. On the other hand, seasonal climates are characterised by a high level of ecological variability, such that resource abundance is followed by resource scarcity. Borrowing terminology from economics literature, Nettle (1996, p. 413) referred to the level of ecological variability and availability of subsistence resources in the natural environment as *ecological risk*. Based on this observation, Nettle (1996) hypothesised that low ecological risk is central to the maintenance of small, distinct linguistic communities. The hypothesis holds that the minimum viable community size differs between areas based on their ecological risk; if a community is too small for its environment, it will be more likely to increase contact and exchange with its neighbours, or to increase its spatial extent. This mutual dependence between communities may be facilitated by the presence of a shared language and linguistic identity, which means that they could lead to the gradual loss of linguistic diversity through convergence or language shift.

Globally, small language communities are not rare. In Arnhem land, some languages have been spoken in communities counting around 70 speakers for as long as anyone can remember (Evans, 2010, p. 8). Humans have likely existed in small groups surrounded by communities speaking distinct languages for the vast majority of our evolutionary history (Evans, 2010, pp. 10–11). Small communities of speakers can and do have intensive contact with each other in the form of widespread multilingualism and tight-knit social networks, which may also be accompanied by material exchanges and marriage patterns such as exogamy (Pakendorf, Dobrushina, and Khanina, 2021). This contradicts the idea that intensive contact, exchange and collaboration between linguistic groups necessarily leads to a loss of linguistic diversity. On the other hand, it could be argued that small-scale multilingualism is uniquely viable in areas of low ecological risk. A well-studied area of this kind of multilingualism, described by Di Carlo, Esene Agwara, and Ojong Diba (2020), is the Lower Fungom area of Cameroon, which has a tropical climate. However, small-scale multilingualism is also found in areas which are not tropical, including highland Daghestan in the Caucasus, and it has been argued that in some cases, strong connections between small societies have been forged as a means of protection from resource scarcity

and other external risks (Pakendorf, Dobrushina, and Khanina, 2021). Thus, low ecological risk may contribute to the viability of small linguistic communities, but cultural norms which resist the subsumption or merging of linguistic identities could also be a powerful force acting against language loss and shift (Comrie, 2008; Evans, 2010).

Hua et al. (2019) investigated the impact of environmental variables on global language density and found that climate is the strongest environmental driver of language diversity. This provides some support for the ecological risk hypothesis. The authors applied their model to a large global dataset without focusing on potential variation between the different macroareas of the world. Since the drivers of language diversity may not be the same everywhere, it is worthwhile to re-investigate these hypotheses with reference to specific areas. Another reason to re-test some of the same hypotheses is that Hua et al. (2019) only investigated the impact of environmental variables. If cultural factors simultaneously impact linguistic diversity, some environmental effects may appear stronger or weaker once these other variables are controlled for (McElreath, 2020).

Another hypothesis about how the environment drives variation in patterns of language diversity, which was also investigated by Hua et al. (2019), is the *isolation hypothesis*. Geographic features such as mountains, valleys, and bodies of water can isolate speech communities from each other, leading to diversification over time; this is analogous to observed processes of differentiation by geographic isolation in genetics (Huisman, Majid, and Hout, 2019; Orsini et al., 2013; Wright, 1943). As an example, the astounding linguistic diversity of the Caucasus has been linked to its highly variable terrain, although it has been argued that its cultural norms have been just as important, if not more so, in shaping the linguistic landscape of the area (Comrie, 2008).

It should be clear that it is not straightforward to disentangle the impact of geography from the impact of social factors on linguistic diversity. This is particularly true for cultural practices which depend directly on the environment, like subsistence strategies. Other cultural features may be impacted indirectly by the environment, although the nature and strength of this impact is not well understood. One aspect of culture which may be indirectly impacted by the environment is *political complexity*, i.e., whether societies organise themselves as autonomous bands or villages or larger chiefdoms or states (Murdock and Divale, 1999). Political complexity could be linked to subsistence strategy, which in turn depends on the environment. It has been argued that the practice of agriculture facilitates the long-term storage of food, which could make it easier for essential resources and thus for political power to accumulate in the hands of a few individuals (McIntosh, 1999). However, it is unclear whether

there is indeed a causal link between political complexity and agriculture, especially since 'agriculture' encompasses a broad spectrum of cultivation and food production practices of varying levels of intensity and utilising diverse crops, not all of which necessarily lead to the creation of surplus (Bostoen, 2020; Haudricourt and Dibie, 1987).

Africa is home to some of the most linguistically dense areas of the world, and according to the Ethnographic Atlas, most societies in Africa rely at least partly on agriculture for food production (Gray, 1999; Murdock, 1967). This includes communities in some of the most linguistically dense and diverse areas. Historically, the most well-known and impressive spread of a single language family in Africa is the Bantu Expansion, and the Bantu-speaking populations who migrated cannot be characterised as agriculturalists (Bostoen, 2020). In general, the link between language spread and food production has been contested (see Diamond and Bellwood (2003) and Renfrew (1992) in support for this idea, and Bostoen (2020) for a rebuttal of it in relation to the Bantu Expansion). Focusing on both African and Eurasian languages, Currie and Mace (2009) examined the influence of political complexity and subsistence strategy on language area, concluding that both political complexity and, to a lesser extent, agriculture, correlate with a larger language area. This suggests that the effect of political complexity on language spread is not merely a result of its hypothesised (and contested) relationship with agriculture. In other words, political complexity appears to facilitate the spread of linguistic groups regardless of their dominant subsistence strategy (Currie and Mace, 2009).

3.1.2 Diachronic perspectives on linguistic diversity

It is not clear whether the maintenance of linguistic variation and phylogenetic diversity over time is driven by the same processes as the coexistence of several closely related languages in an area, as the latter can be indicative of recent language diversification. Following Nichols, 1992, p. 13, some *spread zones* (areas of low genetic and structural diversity, often dominated by languages belonging to a single family) could be linguistically dense without being linguistically diverse in the same way as *accretion zones* (areas in which structurally and genealogically diverse languages have likely coexisted for a very long time).

In this study, two separate but related processes will be modelled: first, the maintenance of a high level of phylogenetic diversity over time, and second, the co-existence of many distinct languages in close proximity to each other. Although these are certainly related, they could be shaped by different diachronic processes. For example, in areas where many closely related languages are spoken, this state

may indicate relatively recent diversification from a common ancestor. Differentiation between these languages could have taken place due to factors like social or geographic isolation, environmental changes, neutral drift, or processes of contact-induced divergence (Bickel, 2017; Mansfield, Leslie-O'Neill, and Li, 2023).

The diachronic processes which give rise to residual zones includes the gradual addition of languages over time which contributes to overall genetic and structural diversity in an area (Nichols, 1992). The earliest habitants of this eastern African residual zone were likely the ancestors of the speakers of Hadza and Sandawe, although the exact time depth is not known (Kießling, Mous, and Nurse, 2008). These languages were, and in some cases still are, referred to as Khoisan languages in the literature, following Greenberg (1963). Now, Khoisan is a term that encompasses a group of languages which share certain features, most notably click consonants, and which are distinct from neighbouring larger language families (Vossen, 2013). There is still significant uncertainty regarding their internal classification, which may also explain the continued use of the term. Most scholars treat Hadza as an isolate (Güldemann and Vossen, 2000), while the status of Sandawe is uncertain, with some positing a genealogical relationship between Sandawe and Khoe-Kwadi on the basis of the similarities between them (Güldemann and Elderkin, 2010), while others maintain that it is an isolate (Sands, 1998).

The area was likely populated in waves by speakers of Cushitic languages followed by Niger-Congo languages and Nilotic languages, all of which contributed to the diversity of the area (see Dimmendaal (2020) and Dimmendaal, Crevels, and Muysken (2020) for a fuller discussion of African migration pattern). Although these groups likely displaced speakers of languages that might have been related to the earlier populations, overall, the area appears to have been characterised by a gradual increase in diversity rather than its loss. This diversity has then subsequently been maintained, and this could be linked to the absence of stable hierarchical relations or to isolation as a result of rugged terrain (Kießling, Mous, and Nurse, 2008). Thus, the conditions which lead to the emergence and maintenance of residual zones over time may be different to those which lead to linguistically dense zones comprising related languages, although there is undoubtedly some overlap between the two (Nichols, 1992).

3.1.3 Overview of the chapter

This study aims to investigate the ideas presented above using a new Bayesian spatial modelling approach, incorporating a larger number of cultural predictor variables than previous studies and using new ways of measuring linguistic diversity. The

following section will first present the data sources used for the cultural and environmental predictor variables. Some different possible ways of measuring linguistic diversity will be discussed, along with their advantages and disadvantages, with the conclusion that language density and phylogenetic diversity can be treated as distinct outcome variables. Following this, the results of the best models will be described, taking model uncertainty into account, and the different models will be compared. The study will conclude with a discussion of the results in the context of the literature, as well as some remarks on the methodology and its potential, and possible directions for future research.

3.2 Materials and methods

3.2.1 Language territories

The data on the locations of the languages in Africa was licensed from the *World Language Mapping System*[1] as part of the Summer Institute of Linguistics (SIL), which gathers data on language locations through an extensive network of field researchers (Eberhard and Fennig, 2023). This dataset includes the spatial extent of the world's languages in the form of digital polygons. Polygons provide a better representation of the geography of a language than points, as languages are spoken across areas rather than at single point locations. Language areas often overlap or intersect, which is also the case in the polygon dataset. These polygons can also be used to calculate the size of language territories and retrieve detailed information about their climate and topographic features, while taking the entire language area into account.

3.2.2 Environmental data

The environmental variables were retrieved from the *WorldClim* database through the R package **raster**. This database is open source and includes global information on climatic variables like temperature and rainfall. The variables which are included in this study are listed in Table 3.1.

The idea that geographic isolation leads to greater diversification or maintenance of diversity predicts that regions with a greater variability in elevation should have a higher level of linguistic diversity. This is tested by adding terrain ruggedness index (TRI) to the model (Riley, Degloria, and Elliot, 1999). Rather than measuring elevation alone, TRI is a measure of terrain *variability*. Thus, it provides a good measure of how difficult it is to travel within an area, given its terrain. For example, some highlands

[1] More details on the dataset can be found at https://www.worldgeodatasets.com/language/.

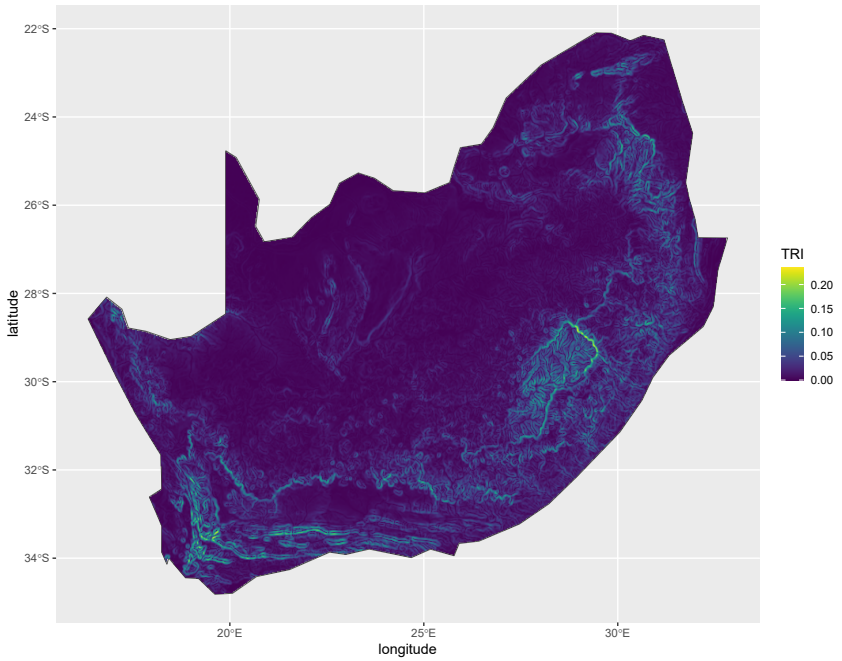


FIGURE 3.2: Terrain Ruggedness Index calculated for South Africa, Lesotho and eSwatini. Lighter colours indicate areas of more rugged (variable) terrain while dark colours represent areas of more even terrain.

may have high terrain elevation but little variability, meaning that travelling within that area would not require people living there to ascend or descend mountains in order to access a nearby location. This kind of area would have a lower TRI than an area comprising both valleys and mountains, which would require a person to travel across or around a mountain in order to get from one valley to another. Calculating TRI requires a raster Digital Elevation Model (DEM), which is a digital representation of the Earth's topographic surface. A raster is made up of grid cells representing specific locations. TRI quantifies the level of topographic heterogeneity in an area by measuring elevation differences between a grid cell and its eight neighbouring grid cells (Riley, Degloria, and Elliot, 1999). An example of what TRI looks like for the surface of South Africa is shown in Figure 3.2. This is then averaged across each of the polygons in the dataset in order to obtain a measure of mean TRI for each language territory. The mean TRI for each language (depicted as the centroids of their territories[2]) is shown in Figure 3.3.

[2]This is due to restrictions on visualising and sharing some of the polygons in the dataset.

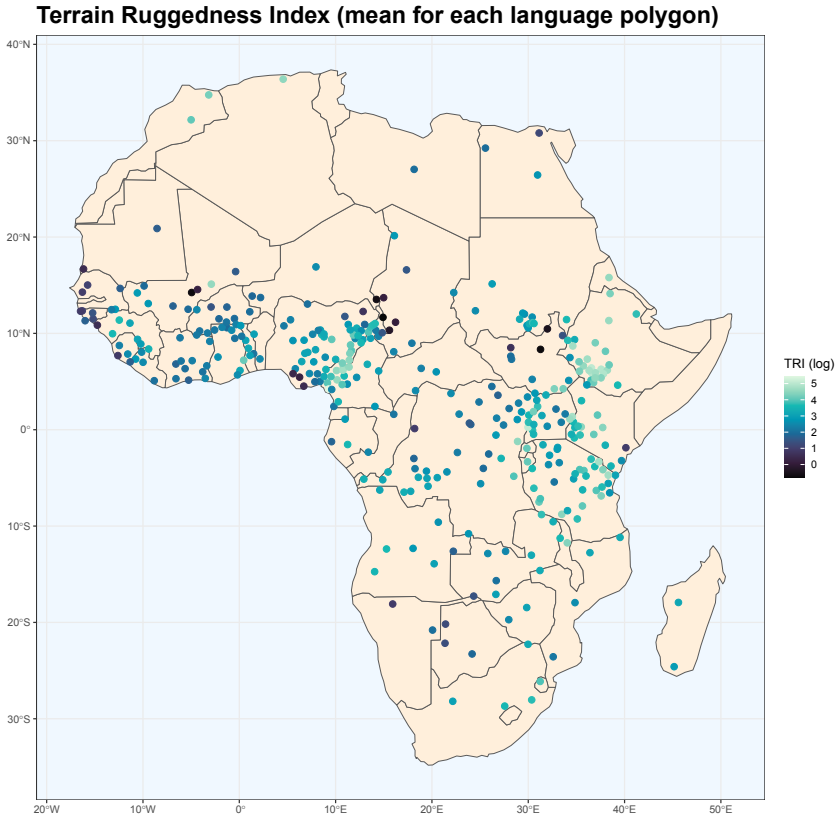


FIGURE 3.3: Mean Terrain Ruggedness Index (TRI) for each language polygon. Lighter shades indicate a higher TRI.

Variable name	**Description**
Temperature	Annual mean temperature
Diurnal range	Mean difference between monthly maximum and minimum temperatures
Annual range	The difference between annual maximum and minimum temperatures
Temperature seasonality	The standard deviation of monthly mean temperatures
Annual precipitation	The amount of rainfall per year
Precipitation seasonality	The difference between the maximum and minimum annual rainfall
Terrain ruggedness	Variation in terrain elevation within a polygon area

TABLE 3.1: All the environmental variables included in the study.

3.2.3 Cultural data

The cultural data was downloaded from D-PLACE, an online database which includes two large-scale cultural datasets (Kirby et al., 2016). For the purposes of the present study, only data from the Ethnographic Atlas was used, originally compiled by Murdock (1967) and later corrected by Gray (1999). The data source for Kirby et al. (2016) is the codebook published in Murdock and Divale (1999). Importantly, imposed regimes such as colonial states are excluded from the Ethnographic Atlas, as Murdock (1967) attempted to capture historical, pre-colonial ways of life in Africa. Whether this endeavour was successful is an ongoing debate, and it is important to note that it is necessarily an approximation of past states (Cogneau and Dupraz, 2014; Herbst, 2000; White and Brudner-White, 1988). Nonetheless, in the absence of any other database with similar data and coverage, the Ethnographic Atlas has been used as a reflection of historical ways of life in Africa in several studies (Michalopoulos and Papaioannou, 2013; Nunn, 2008; Nunn and Puga, 2012).

The list of selected cultural variables, all of which are ordinal, is shown in Table 3.2. Variable selection was done on the basis of previous literature, cross-validation, and data coverage, excluding variables in Kirby et al. (2016) which had a large number of missing values for African societies. Preliminary model comparison using *Pareto smoothed importance sampling* (PSIS-LOO) was done in order to select the best performing set of predictor variables which were also relevant to the literature (Vehtari, Gelman, and Gabry, 2017; Vehtari et al., 2024). All of the variables in the final set are ordinal, meaning they have discrete values representing states going from lower to higher values along a meaningful scale. For example, political complexity is ordered from 1 (autonomous local communities) to 5 (complex states). Community hierarchy is ordered from 2 (independent families) to 4 (nuclear families nested within extended families or clan-barrios) (Murdock and Divale, 1999). Settlement strategy was transformed by merging some of its levels in the original dataset into a measure of how sedentary societies are, ranging from 1 (nomadic societies) to 4 (societies living in large permanent settlements).

The inclusion of cultural data significantly limits the sample size of this study. There are 2,067 language polygons in Africa according to the SIL dataset, but only 366 of these are in D-PLACE (Kirby et al., 2016). These languages, their locations, and family affiliations according to Glottolog are depicted in Figure 3.4 (Hammarström et al., 2023).

Variable name	**Description**
Political complexity	Levels of jurisdictional hierarchy beyond the local community
Community hierarchy	Levels of jurisdictional hierarchy of the local community
Settlement strategy	The level of settlement permanence (nomadic to permanent settlements)
Agriculture	The level of dependence on agriculture for subsistence

TABLE 3.2: All the D-PLACE variables included in the study.

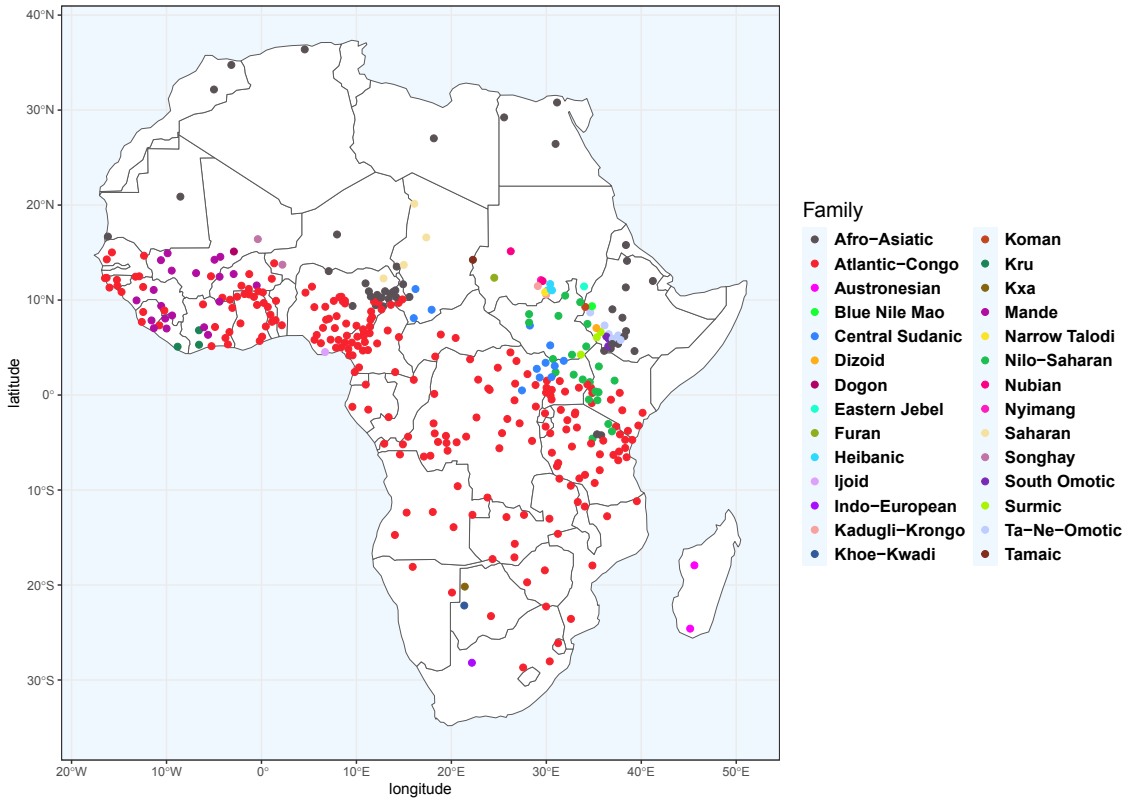


FIGURE 3.4: Languages in the sample with their point (centroid) locations. The colours represent language families.

3.2.4 Quantifying linguistic diversity

Previous quantitative studies of the factors which facilitate or reduce linguistic diversity have used different ways of measuring diversity. Thus, it can be difficult to compare results across studies. For example, Hua et al. (2019) measured linguistic diversity by counting the number of language polygons which overlap with each grid cell, while Currie and Mace (2009) measured the density of languages in a given area. Both of these strategies result in a measure of linguistic diversity which does not include any information about how different or similar the languages are. Huisman, Majid, and Hout (2019) use an approach which quantifies the lexical difference between languages, which is similar to the method which will be used here.

Based on language density alone, Africa appears to be more linguistically diverse than South America, as seen in the world map in Figure 3.1. However, as discussed previously, Africa has fewer isolates and phyla than South America, so it is typically considered less genealogically diverse (Blench, 2013). Thus, genealogical diversity (or lack thereof) might be obscured when we measure linguistic diversity by counting languages.

When devising a measure of linguistic diversity, the first thing which should be defined is the spatial extent of that measurement. Using grid cells is one way of making sure that the chosen measure of linguistic diversity is applied over a consistent geographic range, as in Hua et al. (2019). However, using grid cells as data points means needing to average the predictor variables within each cell. This makes sense for environmental variables, as the mean temperature within a given area is informative. For cultural variables, however, this would be more problematic, partly due to data sparsity. Dividing Africa into 300 by 300 km grid cells would result in many cells which contain only one or two language polygons which are in D-PLACE (Kirby et al., 2016). Those same grid cells might contain a large number of languages for which no information is available. This means that the aggregate level of a cultural variable, like political complexity, for a grid cell with several languages could be determined by only one or two languages in that grid cell. Because of this, averaging over the values of sociocultural variables within grid cells may lead to arbitrary results or obscure real-life diversity. Thus, in order to keep as much language-specific cultural data as possible, I chose to devise a measure of linguistic diversity that could be calculated per language rather than across an arbitrary spatial range.

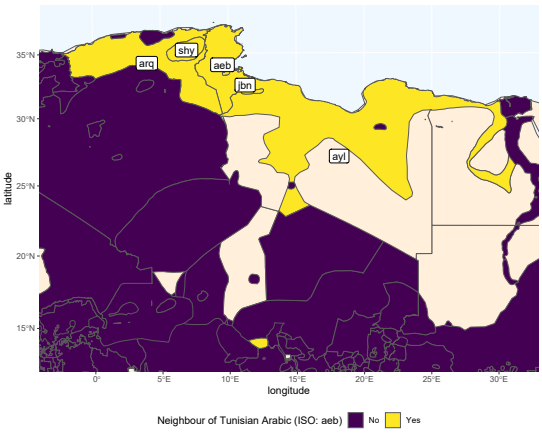
Drawing on the idea of language ecology developed in Mufwene (2001) and used in a study by Bromham et al. (2022), a measure can be devised which represents the diversity in the immediate vicinity of a given language community. This idea involves seeing languages and their speakers within the context of their surroundings (or language ecology), which encompasses social as well as environmental factors.

The social context includes the number of distinct languages in a society's immediate vicinity, and how different those languages are from each other.

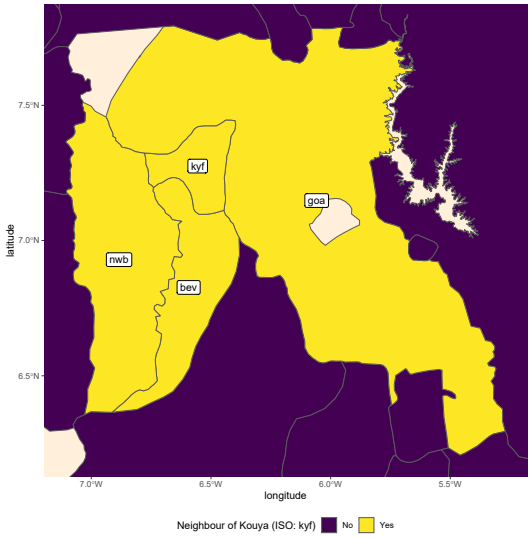
For the purposes of this study, all the language territories which intersect or overlap with another language polygon were included in the ecology of that language. Two territories intersect if their borders touch, and they overlap if they share part of the same territory. Using this definition of language ecology, two different measures of linguistic diversity can be derived. First, for each language, the language territories which intersect it are retrieved. These are counted to obtain a measure of the language density surrounding a given language territory. After that, I retrieve information about how different the languages within the given language ecology are from each other. *Levenshtein distances*, also called *edit distances*, are a well-known way of measuring how different languages are from each other given a word list of stable concepts, like a Swadesh list (Holman et al., 2008). However, Jäger (2013) outlines some potential issues with using edit distances as a proxy for phylogenetic diversity and introduces *PMI distances* as an alternative measure of aggregate linguistic differences which is better suited to handling data sparsity and detecting related languages. Thus, PMI distances are used here as a proxy for phylogenetic diversity. Crucially, automated measures of linguistic difference do not rely on expert judgments of language relatedness, which can be an advantage when these judgments differ (as they do between e.g. Blench (2013) and Dimmendaal (2008b)).

Figure 3.5a illustrates how language territories are counted as part of the ecology of a language, in this case Tunisian Arabic. We can see that Algerian Arabic, Libyan Arabic, and two Berber languages are counted as part of this language ecology. When counting the language density surrounding Tunisian Arabic, the number of languages would thus be four. Some polygons are discontinuous, like that of Libyan Arabic, which is why there is a yellow polygon which appears to be disconnected from the rest of the yellow polygons in the plot. In some cases, a single language is spoken in more than one area, and that language will be represented as a discontinuous polygon.

In Figure 3.5b, the same concept is applied to a smaller language territory, Kouya, a Kru language spoken in West Africa. Kouya's territory intersects with that of three other languages, two of which also belong to the Kru family, and one of which is a Mande language and thus unrelated to Kouya. Because of this, although Kouya is surrounded by fewer languages than Tunisian Arabic, it could be quantified as an area with a more phylogenetically diverse ecology. These maps highlight another potential issue: it would be problematic to directly compare the ecology of a language with a large spatial extent, like Tunisian Arabic, to that of a smaller language without controlling for the effect of territory size. Large language territories are more likely



(A) Capturing which language polygons are neighbours of Tunisian Arabic (Afro-Asiatic, Semitic).



(B) Capturing which language polygons are neighbours of Kouya (Kru, Beteic).

FIGURE 3.5: In these plots, the neighbouring polygons counted as part of the language ecology are yellow. The polygon centroids are labelled with their language ISO codes; polygons which are not neighbours of the target language are dark purple. The beige areas represent the land mass of Africa where there are no language polygons.

to overlap with or intersect a larger number of languages. Because of this, language area is included as a control in the model.

The measures of phylogenetic diversity and language density are shown for the entire language sample used in the model in Figure 3.6. The sample includes 366 languages, which are depicted as point locations representing the centroid of each polygon on the map.

3.2.5 Transformation of predictor variables

One of the issues with including a set of interdependent variables in a single model is that the correlations between them can bias the results and lead to high uncertainty estimates. This is referred to as *multicollinearity* (McElreath, 2020, p. 162). A common method for reducing multicollinearity is *Principal Components Analysis* (PCA). This reduces complex data to a few key components which can explain most of the patterns in the data. Crucially, these components are maximally different from each other, which means that they are less correlated than the raw variables. For the climate variables, the PCA resulted in two main principal components, similar to the result obtained in Bentz et al. (2018). The first component mainly relates to temperature, with higher values indicating higher temperatures and a tropical climate. The second component relates to temperature and rainfall seasonality. The

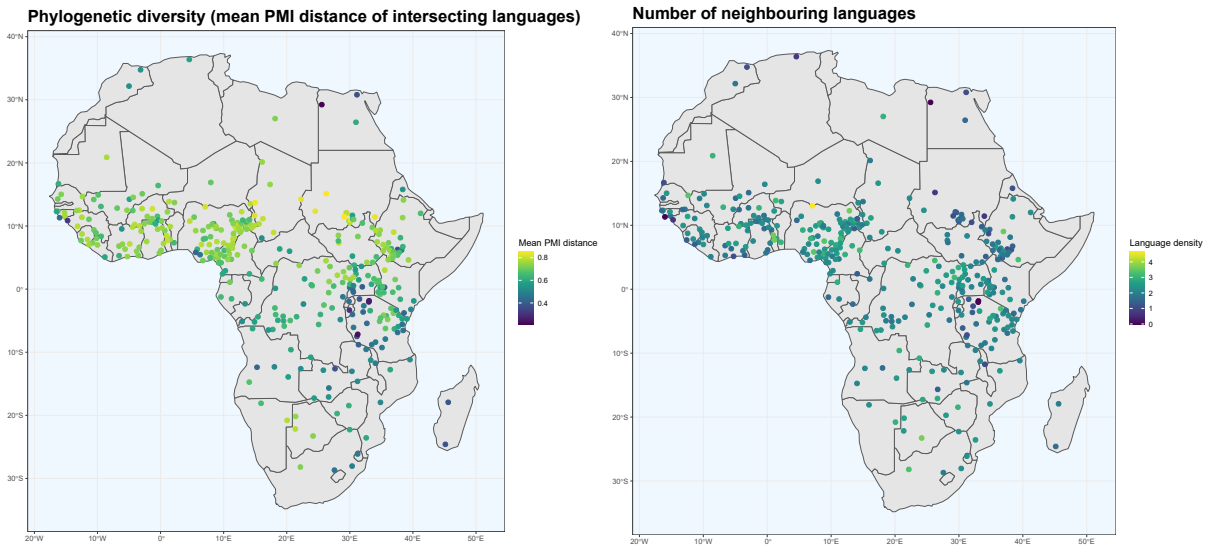


FIGURE 3.6: Neighbouring language density (left) and linguistic diversity (right) of each language polygon, depicted here as points. High values are shown in yellow, low values are in dark purple, and the land mass of Africa is light grey.

PCA is presented in more detail in Appendix B and is discussed in the results. Terrain ruggedness was not included in the climate PCA as this would not be expected to correlate meaningfully with climatic variables.

Since the cultural variables are discrete and ordinal, a different approach is necessary for these. I ran an *ordinal PCA* as implemented in the R package **Gifi**, which first converts the ordinal data into continuous values which preserve the original order before applying a PCA to them (Gifi, 1990; Mair, De Leeuw, and Groenen, 2022). The PCA can reveal some of the ways in which the variables covary. Lastly, the PCs were transformed using *varimax rotation* for better interpretability (Kaiser, 1958). This process revealed some interesting patterns in the data[3]. For example, a society's settlement strategy covaries with its level of dependence on agriculture. Thus, the first principal component indicates how much a society depends on agriculture and how sedentary it is. Low values mean that a society depends mostly on agriculture for subsistence and lives in permanent settlements. The second principal component indicates how many levels of hierarchy exist within the local community, which is also an ordinal variable ranging from independent, nuclear families to extended families and clans. The third component represents the levels of jurisdictional hierarchy beyond a local community, or political complexity.

The resulting values for the principal components are depicted in Figure 3.8. These principal components (PCs) will be used in the models in place of the original

[3]More detail and more plots of the PCA results are shown in Appendix B.

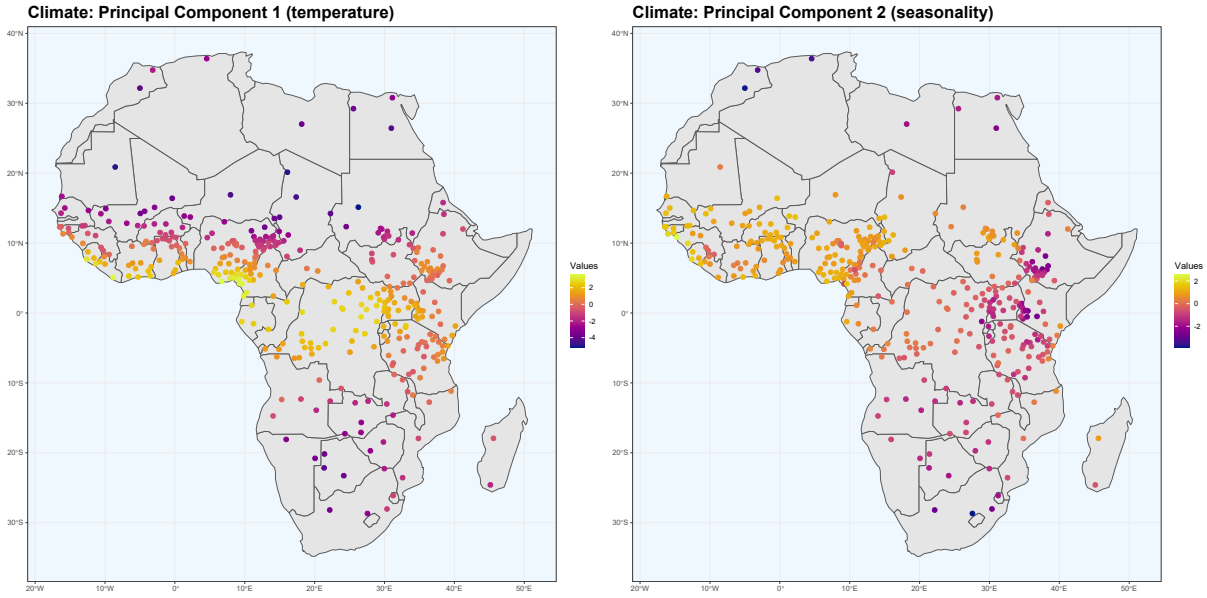


FIGURE 3.7: Scores derived from the principal components analysis of the climatic data. Lighter values indicate a higher score along the component. There is clear spatial clustering near the Equator for the first PC, while the second PC reveals a gradient between east and west.

variables. Model comparison was performed to ensure that the PCs performed as well as or better than the original ordinal variables.

3.2.6 Phylogenetic decorrelation

A potential issue with the sociocultural variables from D-PLACE (Kirby et al., 2016) is that they are likely to be phylogenetically non-independent. Just like words and grammatical features, cultural traits can be inherited when populations diversify and

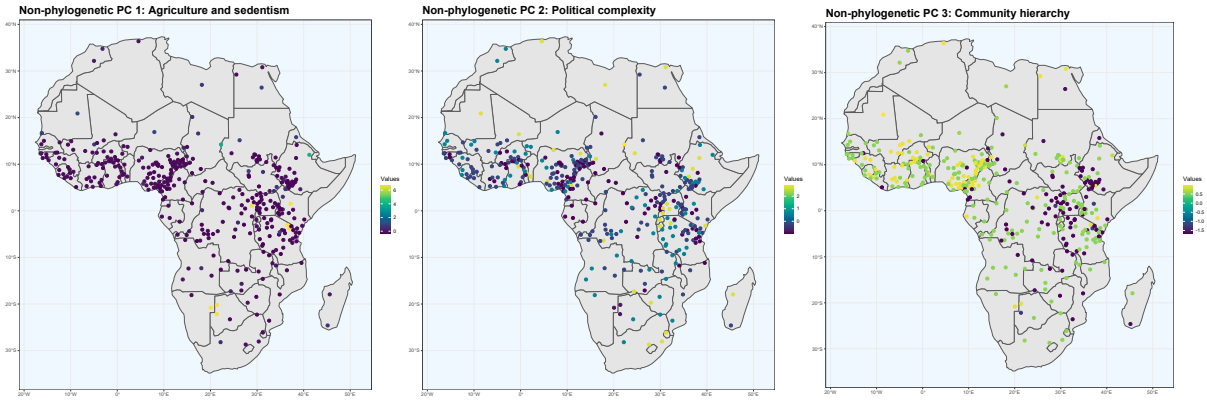


FIGURE 3.8: Scores associated with the three rotated principal components derived from the Ethnographic Atlas data. Dark purple represents low values while yellow represents high values along each component.

thus stay stable over time in related populations. This was demonstrated quantitatively for North American cultures by Towner et al. (2012). Controlling for language family in the model would not solve this issue, since in this case, phylogenetic non-independence affects the predictor variables, not the outcome. This means that, in order to properly control for the effect of common inheritance, these variables should ideally be phylogenetically decorrelated before model inference. The method for decorrelating the ordinal variables was implemented in Stan by Gerhard Jäger[4] and involves using a phylogenetic model to model the tree-like evolution of each ordinal variable as a continuous latent variable. Thus, the original ordinal variables are converted into continuous latent variables, which can be extracted. The residuals can also be extracted, and these should represent the parts of the latent variable which cannot be predicted based on the phylogenetic model, thus indicating where the phylogenetic signal fails to predict the distribution of variables.

A model will be run which includes both of these extracted variables, which should essentially filter out the phylogenetic signal. The results of this model will be compared to the results of a model using the predictor variables without phylogenetic decorrelation. Additionally, the same PCA was applied to the phylogenetically decorrelated variables, so that the predictors would be comparable across models.

3.2.7 The spatial model

Spatial autocorrelation exists when data points which are close together in space are more similar than data points which are far apart; it can be observed as clusters of similar values on a map (Odland, 1988, p. 7). The presence of spatial autocorrelation in a variable of interest can lead to biased inference if it is not properly controlled for in a statistical model. Spatial autocorrelation can be detected prior to model fitting using Moran's I, also called the Moran coefficient (Anselin, 1995; Chun and Griffith, 2013; Cliff and Ord, 1981). Positive values indicate that spatial autocorrelation is present, values close to zero mean that there is no spatial autocorrelation, and negative values indicate that data points which are close to each other tend to be less similar than expected given random chance. Figure 3.9 confirms the presence of spatial autocorrelation for both of the outcome variables used in this study: language density and phylogenetic diversity. Language density has a high Moran coefficient of 0.89, while phylogenetic diversity has a somewhat lower Moran coefficient of 0.59. Both of these confirm the presence of strong spatial autocorrelation, which is to be expected since these variables were derived from neighbouring languages.

[4]The code and a detailed description of the method can be found at https://profgerhard.de/ordinal_decorrelation/.

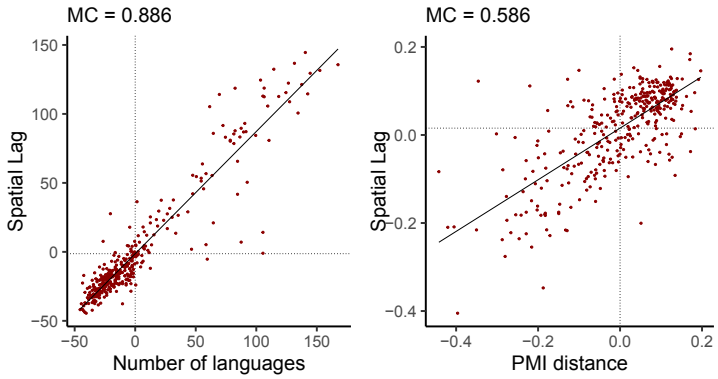


FIGURE 3.9: Moran's I and scatterplot for language density measured as the number of languages (left) and phylogenetic diversity measured as the mean PMI distance between neighbouring languages (right).

As covered in Chapter 2 of this thesis, there are two main ways to control for spatial autocorrelation in statistical studies. One of these involves adding a latent Gaussian process (GP) to the model, which uses point locations for the languages in the study (Guzmán Naranjo and Becker, 2021; Guzmán Naranjo and Mertner, 2022; Guzmán Naranjo, Mertner, and Urban, 2024). A matrix of the geographic distances between those languages is then used to calculate the expected similarity between them. Any distance metric can be used as long as it satisfies certain mathematical constraints; for example, distances which take terrain variability into account can be used (Guzmán Naranjo and Jäger, 2022). However, a GP model cannot be applied directly to language areas (Williams and Rasmussen, 2006).

Since the locations of languages in this study are polygons rather than point locations, a different modelling strategy is more suitable, namely autoregressive (AR) models (Besag, 1974; LeSage and Pace, 2009). There are two main types of AR models, and the one used here is a conditional autoregressive (CAR) model, which is a popular choice in spatial econometrics, ecology, and related fields[5] (LeSage and Pace, 2009; Ver Hoef, Hanks, and Hooten, 2018).

The CAR model was defined by Besag (1974) to handle spatial autocorrelation in Y, the outcome variable. It can estimate the strength of spatial autocorrelation and thus 'filter out' the spatial signal in order to avoid biased inference for spatial data. However, it can be equally interesting and relevant to model the spatial lag of X (SLX), the explanatory variable(s). This is referred to as a *spatial spillover effect*, which has been defined as "the marginal impact of a change to one explanatory variable in a particular cross-sectional unit on the dependent variable values in another unit" (Elhorst and Halleck Vega, 2017, p. 2). Thus, in addition to modelling the direct

[5]For further information about the different types of autoregressive models and when to use them, see Chapter 2 of this thesis, or refer to Cressie (1994), LeSage and Pace (2009), and Ver Hoef, Hanks, and Hooten (2018), and references therein.

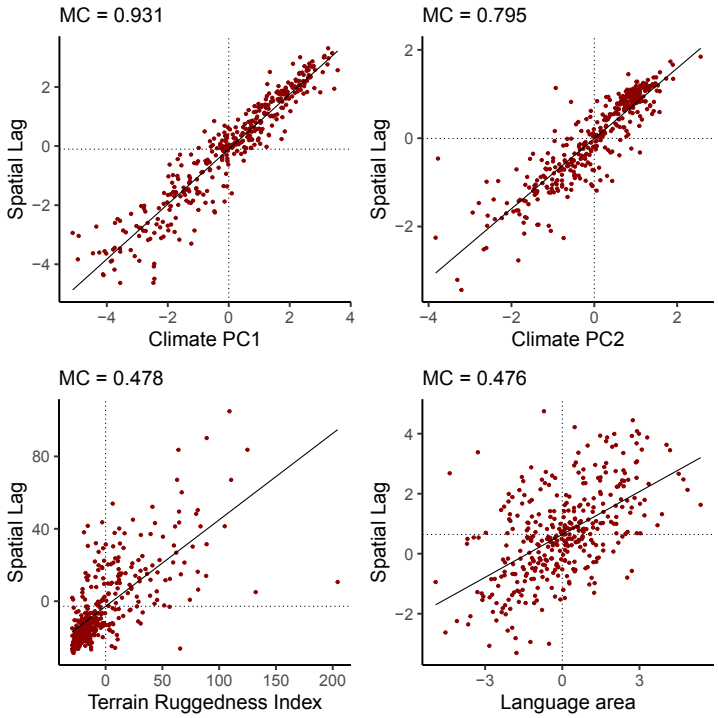


FIGURE 3.10: Moran plots for environmental variables and language area. The x axis represents the values of the variables; the y axis is the spatial lag.

effect of X at a particular location on Y at the same location, the effect of X at neighbouring locations can also impact Y at the original location. This can also help account for spatial autocorrelation in X. To find out whether the predictor variables in this case are spatially autocorrelated, Moran's I was calculated for each of them. As can be seen below, the environmental variables are highly spatially autocorrelated, while the principal components of the sociocultural variables show moderate spatial autocorrelation.

As can be seen in Figure 3.11, the phylogenetic decorrelation method appears to have removed most of the spatial autocorrelation, too. This aligns with what we know about language contact and relatedness, which is that related languages also tend to be geographically closer together.

Following the notation conventions in McElreath (2020), the model and priors were specified as follows:

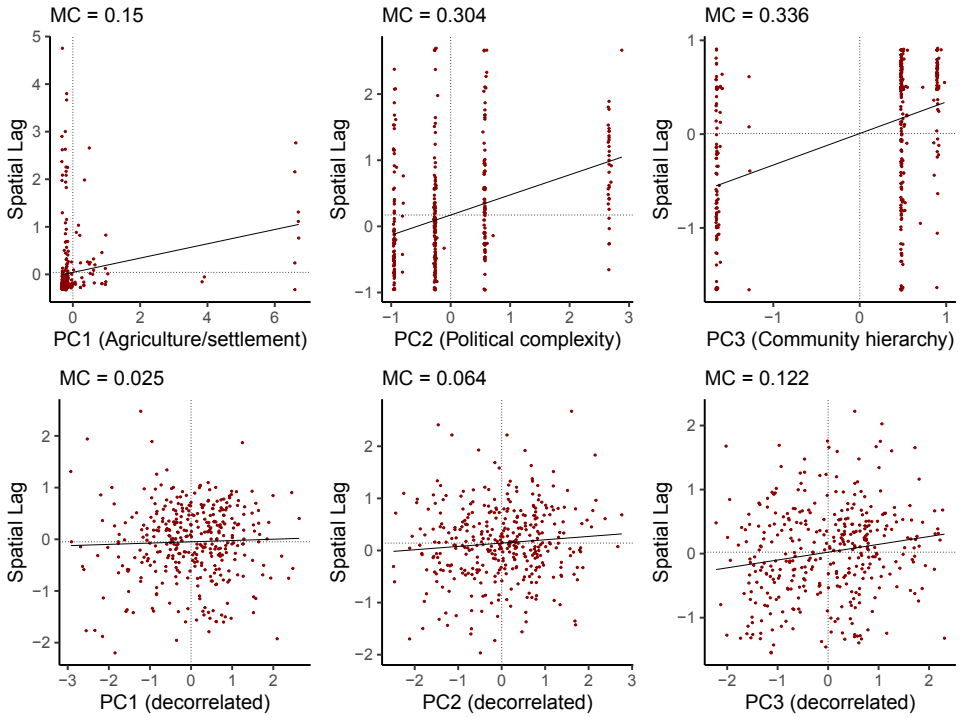


FIGURE 3.11: Moran plots for the principal components of the cultural variables (top row) and the phylogenetically decorrelated cultural variables (bottom row). The x axis represents the values of the variables; the y axis is the spatial lag.

$$Y = \rho \mathbf{W} Y + \mathbf{W} X \gamma + X \beta + \mu + \epsilon \tag{3.1}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.2}$$

$$Intercept \sim \mathcal{N}(0, 1) \tag{3.3}$$

$$\beta \sim \mathcal{N}(0, 3) \tag{3.4}$$

$$\rho \sim \mathcal{N}(0, 0.5) \tag{3.5}$$

$$\sigma \sim Exponential(1) \tag{3.6}$$

where y is a vector of data $y = (y_1, ..., y_n)$, β is a vector of coefficients which all share the same prior, and ρ is a spatial autocorrelation parameter which indicates the strength of spatial autocorrelation in the outcome variable. Phylogenetic diversity was measured as a continuous variable, and thus a normal distribution was used. Language density was measured as language counts, so a Poisson distribution was used, as recommended by McElreath (2020). The model was implemented using the R package **geostan**, which was also used to calculate Moran's I (Donegan, 2022; Donegan, Chun, and Griffith, 2021).

3.2.8 The weights matrix

In order to control for spatial autocorrelation, a CAR model requires the prior specification of a spatial weights matrix **W** (Wall, 2004), which determines which languages are considered neighbours. Spatial autocorrelation is calculated based on how similar neighbours are, and languages which are not neighbours are expected to exert no spatial influence on each other.

The weights matrix for this study was defined using polygon contiguity, such that languages which have intersecting or overlapping territories are considered neighbours. Some of the languages in the sample are isolated in the sense that there are no languages which directly border their territory. Those languages were connected to their nearest neighbour via a minimum spanning tree using Great Circle distances between the centroid of each language polygon. Great Circle distances are 'as the crow flies' distances which take the Earth's curvature into account (Guzmán Naranjo and Jäger, 2022). Following this step, the distances between all the polygon centroids were calculated and an inverse distance weighting scheme was applied, such that neighbours which are closer to each other are expected to influence each other more than neighbours which are further apart. This matrix and the inverse distance weights were calculated using the R packages **spdep** (Roger Bivand, 2022) and **geostan** (Donegan, Chun, and Griffith, 2021). The resulting spatial weights matrix is shown in Figure 3.12.

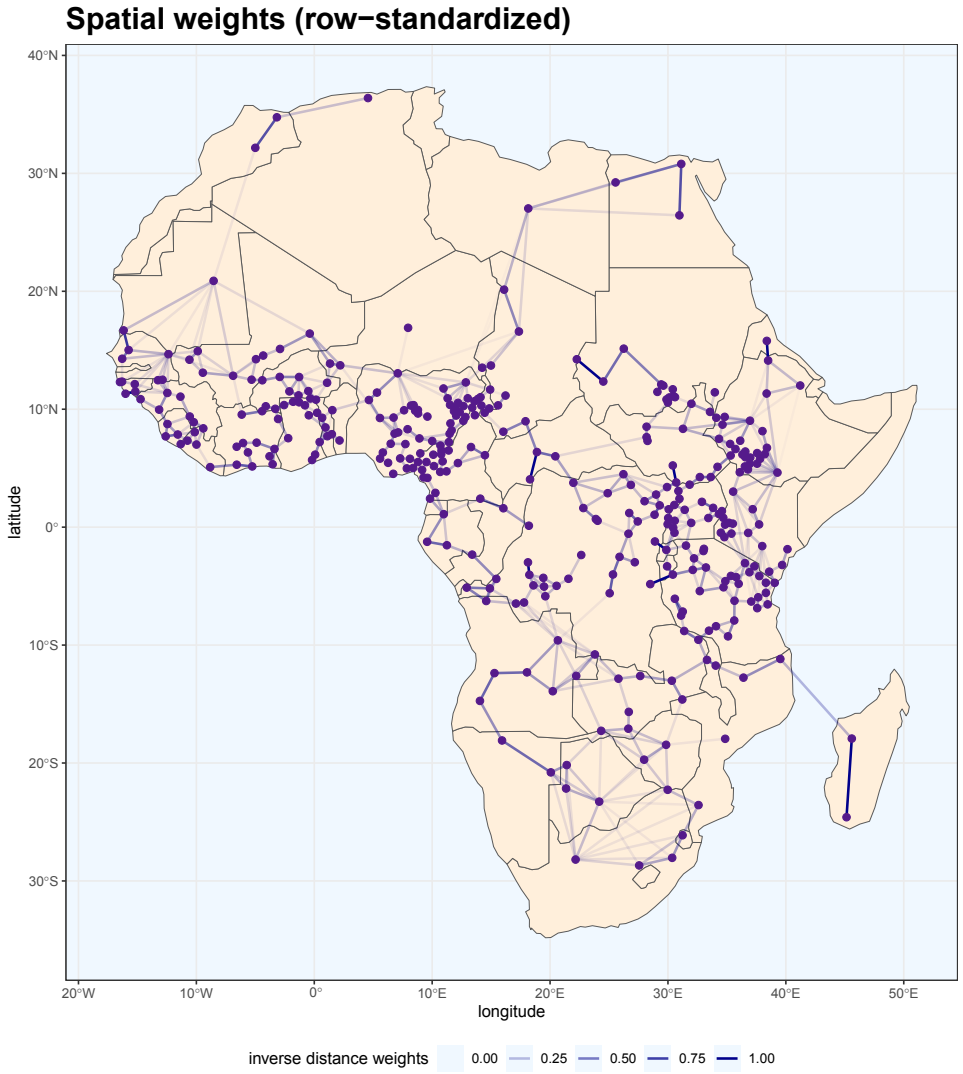


FIGURE 3.12: Distance and contiguity based spatial connectivity matrix for the language sample. The line thickness indicates how geographically close the languages are. Because the matrix is row-standardised, languages with only one or two neighbours have thicker lines despite being geographically further away than some other pairs.

3.3 Results

In this section, the results of the best models will be shown. This will be followed by a section explaining how the models were evaluated and selected. The differences between the model results will also be explored. In particular, I will be examining how the impact of the predictor variables on language density differs from their impact on phylogenetic diversity. I will also consider how the phylogenetic decorrelation of the predictor variables influences the results. The final list of predictor variables is given in Table 3.3. All of the variables in the final model are continuous and have been standardised such that they are on the same scale. According to Chapter 4 of

McElreath (2020), this is recommended when predictor variables have very different scales; for example, language area is given in km^2 while the principal components have arbitrary scales, e.g. from -1 to 6. These differences in the scale of variables could lead to skewed results. Standardisation is a way to avoid this.

Variable name	**Description**
EA PC1: Agriculture and settlement strategy	The degree of dependence on agriculture and permanence of settlements
EA PC2: Political complexity	Levels of jurisdictional hierarchy beyond the local community
EA PC3: Community hierarchy	Levels of jurisdictional hierarchy of the local community
Climate PC1	High mean annual temperature and high precipitation
Climate PC2	High temperature and precipitation seasonality
Language area	The area of language polygons in km^2
Terrain Ruggedness Index	Variability of terrain elevation within an area

TABLE 3.3: All the variables included in the model.

This section will be structured as follows. First, the posterior estimates of the models of phylogenetic diversity will be compared. A model was run with the phylogenetically decorrelated cultural predictor variables and one was run without the phylogenetic decorrelation method. Second, the models of language density will be compared in the same way, and the effects of decorrelating the cultural variables will be examined. In all the models, the rest of the predictor variables will remain the same. Following these comparisons, the coefficients of the spatially lagged (SLX) terms will be visualised and described. Lastly, the models will be evaluated and compared.

3.3.1 Posterior distributions of coefficients

As can be seen in the results in Figure 3.13, which depicts the estimated coefficients of all the predictor variables on phylogenetic diversity, the phylogenetic decorrelation of the covariates impacted some of the results of the model while others stayed the same. Political complexity has a negative effect on phylogenetic diversity in both

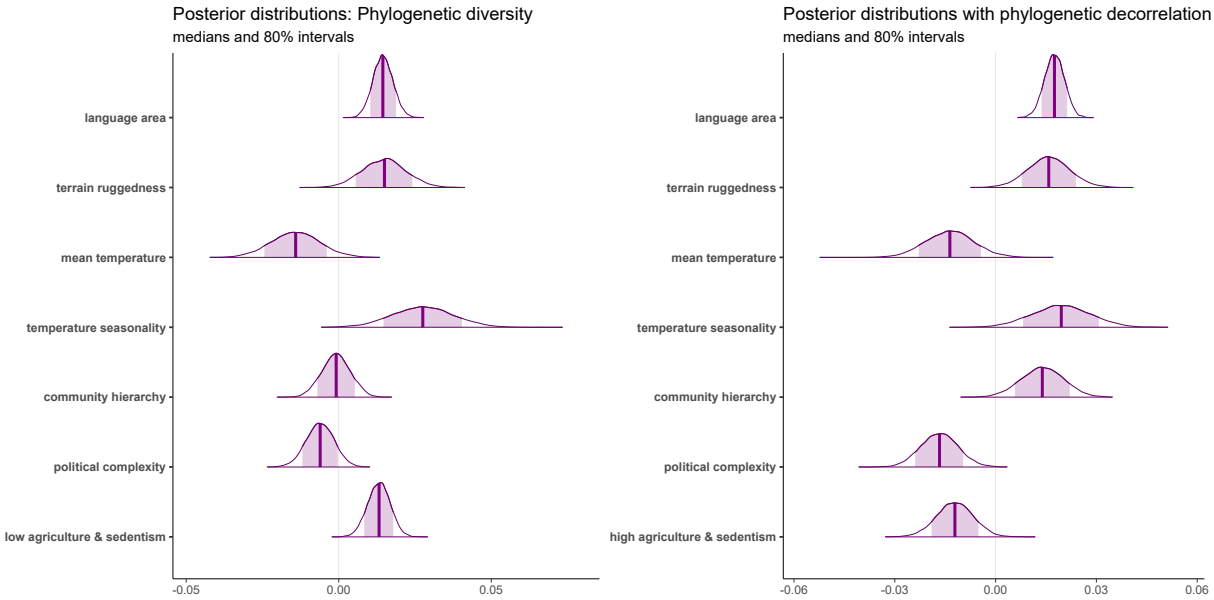


FIGURE 3.13: Posterior distributions of estimated covariate effects for phylogenetic diversity with the original variables (left) and with the phylogenetically decorrelated residuals (right); note that high values of 'temperature seasonality' indicate low seasonality and high precipitation.

models. This means that the negative effect of political complexity on phylogenetic diversity holds even when we control for the level of dependence on agriculture, settlement strategy, language area, and the phylogenetic signal. In fact, the estimated effect of political complexity is stronger in the model with the phylogenetically decorrelated residuals. The effect of agriculture and a more sedentary settlement strategy is consistently slightly negative across the two models[6]. Community hierarchy has no effect in the model without decorrelated residuals, but in the model which includes them, it has a positive effect. These results suggest that phylogenetic diversity is highest in ecologies with societies that have a low or moderate level of reliance on agriculture and a low level of political complexity. It is interesting that, in general, the estimated effects of the cultural variates are stronger in the model with the phylogenetically decorrelated residuals.

The variables which influence the phylogenetic diversity surrounding a language differ somewhat from those which influence the number of neighbours of that language, as Figures 3.13 and 3.14 show. The first major difference is that adding the

[6]For this principal component (agriculture and settlement), the estimated coefficients cannot be compared directly because the direction of the scores is different. For the decorrelated variable, a high score on the first principal component indicates a high level of dependence on agriculture and settlement permanence, while for the PCA-transformed original variables, a low score indicates the same. This matters only for interpreting the coefficient of this variable in the two models as the results are depicted side by side in Figure 3.13.

phylogenetically decorrelated variables has a strong impact on the results of the language density model. Once these are added, political complexity is estimated to have a negative effect on language density, which is in line with previous literature and with the results of the phylogenetic diversity model. In the model using the principal components without phylogenetic decorrelation, however, none of the sociocultural variables appear to affect language density.

In order to compare the factors which impact diversity and density, the model results which include the phylogenetically decorrelated residuals will be used. The main similarities and differences will be summarised first for the environmental variables, then for the cultural ones. After that, the effects of the spatially lagged covariates will be examined.

Terrain ruggedness was found to have a weak positive effect on both language density and phylogenetic diversity across models. Mean annual temperature was found to have a positive effect on language density that is slightly higher than that of terrain ruggedness. However, this effect is not present for phylogenetic diversity, where mean temperature even appears to have a weak negative effect. For both diversity and density, there is a moderate to high level of uncertainty around the estimated effects of temperature and rainfall seasonality, which is higher for phylogenetic diversity, meaning it is harder to draw conclusions about the effects of seasonality. Despite this uncertainty, the effect of low seasonality on phylogenetic diversity is reliably positive, while for language density, there is too much uncertainty around the coefficient to draw any conclusions about the effect.

For the cultural variables, political complexity has a negative impact on both density and diversity. The effect of community hierarchy is also consistently positive across the models. Meanwhile, agriculture and settlement strategy has no detectable effect on language density. It has a negative impact on phylogenetic diversity once the decorrelated residuals are added, although its effect is weaker than that of political complexity, as predicted by the findings in Currie and Mace (2009).

As mentioned in Section 2, these models include spatially lagged covariates (SLX terms), which should be considered alongside the non-spatially lagged covariates presented previously. SLX terms allow the covariate values of neighbouring languages can influence the surrounding language density or phylogenetic diversity of a given language. In some cases, these variables can act as a control for confounding factors, and in others, we might be interested in examining and interpreting their effects. For example, an important control variable is the size of the territory of neighbouring languages. Some language ecologies are small and yet contain a large number of languages for the size of the area. A language ecology with four languages in northern Africa likely spans a much larger area than a language ecology with

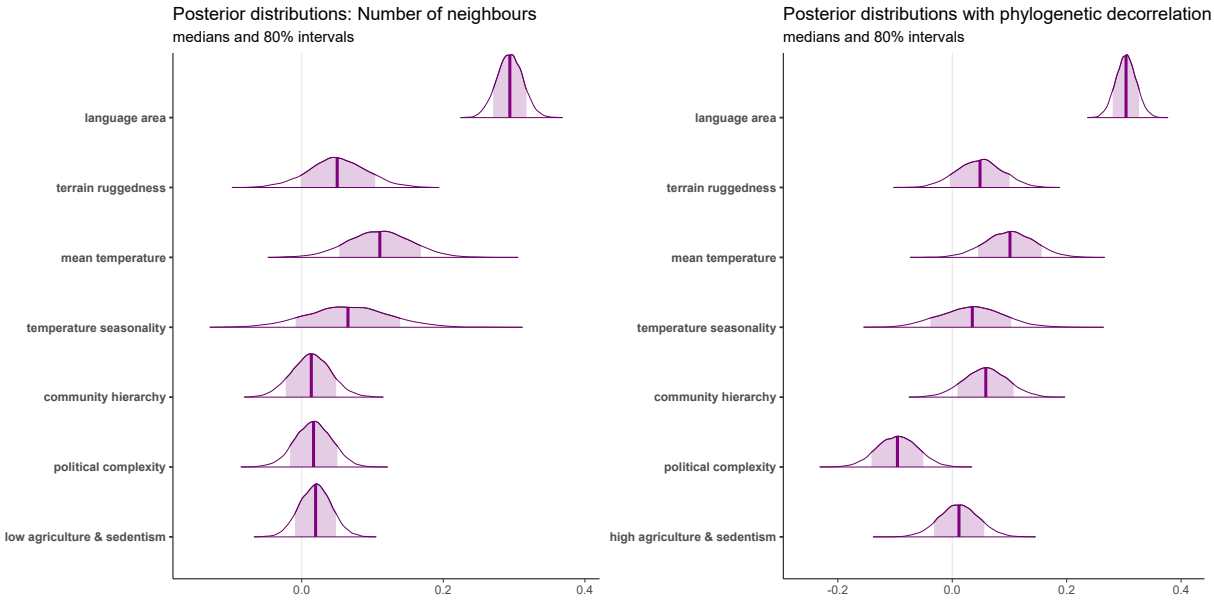


FIGURE 3.14: Posterior distributions of estimated covariate effects for language density with the original variables (left) and with the phylogenetically decorrelated residuals (right); note that high values of 'temperature seasonality' correspond to low seasonality and high precipitation.

four languages in Cameroon. These two ecologies should not be considered equally linguistically dense, and adding the area of the neighbouring polygons to the model is a way of ensuring that these variations in size will be controlled for.

Figure 3.15 shows the posterior distributions of the estimated coefficients of the SLX terms in the phylogenetic diversity model and the language density model. As is to be expected, the spatial lag of language area has a negative impact on both linguistic diversity and language density. Areas with many smaller language areas are likely to be more diverse and dense. Interestingly, the political complexity of neighbouring languages has a larger negative impact on phylogenetic diversity than the political complexity of the language itself, and this effect is stronger than that of language area. Likewise, the SLX terms for the other cultural variables have a greater impact on diversity than the variables which represent the cultural state of the language in question. This is likely in part due to the way of measuring language diversity used here, which relies on data from neighbouring languages. However, the overlap between the 2000 or so polygons which were used to quantify diversity and the 366 for which predictor variables were available is small, so this effect is unlikely to be caused only by that. A caveat is that some of these terms have a moderate to high degree of uncertainty, which suggests that some of the effects cannot be estimated reliably, perhaps due to collinearity. Temperature and seasonality are likely to be correlated with the level of reliance on agriculture, for example, and these variables also have some of the highest levels of uncertainty in the results. These

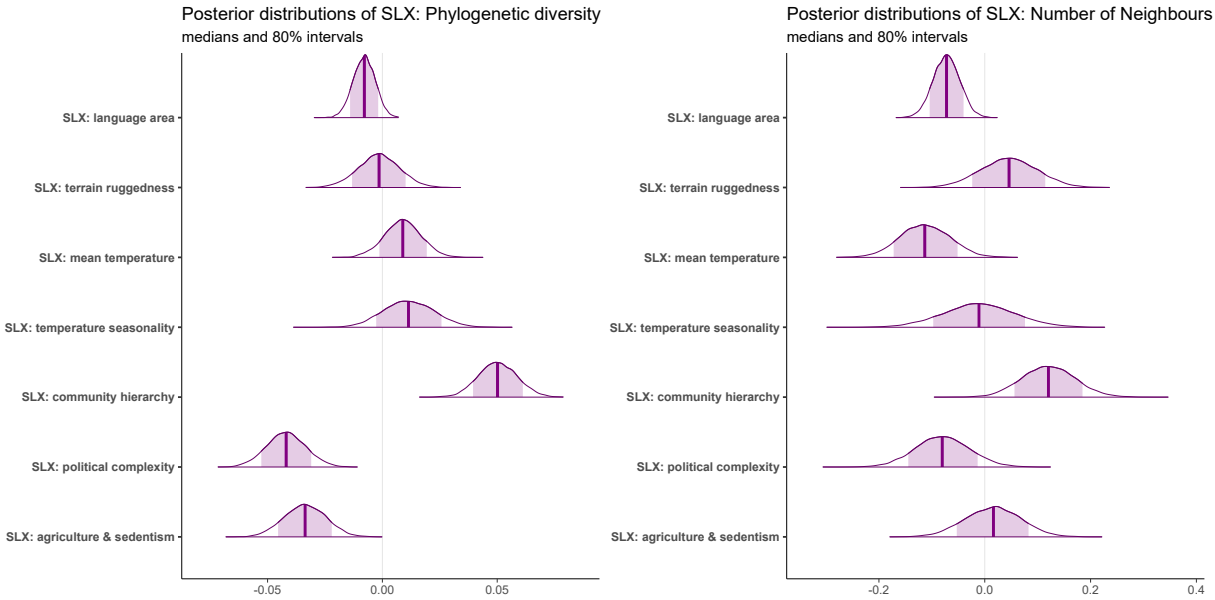


FIGURE 3.15: Posterior distributions of estimated SLX effects on phylogenetic diversity (left) and number of neighbours (right). High values of 'temperature seasonality' indicate low seasonality and high precipitation.

findings will be discussed further in Section 4.

3.3.2 Model evaluation

Posterior predictive checks and model evaluation using WAIC, as implemented in **geostan** and defined in Watanabe (2010), were used to evaluate the models. Similar to cross-validation, WAIC provides an estimate of how well the model is expected to predict new data. In **geostan**, the WAIC function also returns a penalty term which measures the number of effective parameters (**Eff. pars**) estimated by the model, which favours models with a smaller number of parameters, and log predictive density (**Lpd**). WAIC can only be used to compare models rather than providing an objective measure of predictive accuracy. The lower the WAIC value, the better the model. A model without a CAR term, a model without spatially lagged covariates (SLX terms), and a model with neither CAR nor SLX terms were evaluated against the model with both of these. The results are presented in Table 3.4 for the phylogenetic diversity model and in Table 3.5 for the Poisson model for the number of neighbouring languages.

As we see in Table 3.4, which shows the evaluation results for the models of phylogenetic diversity, the best model is the CAR model with SLX terms and phylogenetically decorrelated residuals. The next best model includes the decorrelated residuals without SLX terms, although this one performs significantly worse than the

Model	WAIC	Eff. pars	Lpd
Decorrelated PCs, SLX + CAR	-808.36	12.52	416.70
No SLX terms (decorrelated)	-734.49	6.83	374.08
Original PCs, SLX + CAR	-722.89	8.77	370.21
No CAR term (decorrelated)	-701.44	23.55	374.27
No CAR or SLX (decorrelated)	-647.16	14.04	337.62

TABLE 3.4: WAIC estimates for the models with phylogenetic diversity as the outcome variable.

best model. This suggests that for this model, the phylogenetic signal of the predictor variables is highly informative. It is also straightforward to choose the best model, as the difference between the best and second-best model is large. This contrasts with the results of the model evaluations performed for the language density model, shown in Table 3.5.

Model	WAIC	Eff. pars	Lpd
No SLX terms (decorrelated)	1979.22	128.41	-861.20
Original PCs, SLX + CAR	1981.63	130.45	-860.37
Decorrelated PCs, SLX + CAR	1981.79	128.49	-862.41
No CAR term (decorrelated)	2326.89	67.08	-1096.37
No CAR or SLX (decorrelated)	2372.13	48.83	-1137.24

TABLE 3.5: WAIC estimates for the models with neighbouring languages as the outcome variable.

For the model of language density, the best model was one without SLX terms or decorrelated variables. However, as the difference in WAIC, number of effective parameters and log predictive density between the three 'best models' is negligible, the model with decorrelated variables and both CAR and SLX terms was used to create the visualisations of the posterior distributions of the coefficients as well as the residuals. This was done in order to ensure that the results of the two models were as comparable as possible. When the differences in WAIC are very small, it is sometimes recommended to select the model based on a combination of the evaluation results and prior knowledge of the phenomenon under study. This result suggests that the SLX terms and phylogenetically decorrelated residuals do not add any information which increases the accuracy of the model's predictions of distribution of language density in Africa. This could indicate that the environmental variables are more important for predicting the distribution of language counts.

In Figure 3.16, the residuals for the selected models are plotted on a map. Note that these residuals are not directly comparable, as mean PMI distance is a continuous variable between 0 and 1, and the number of languages takes the form of counts. Residuals close to 0 indicates that the model can predict the mean PMI distance or

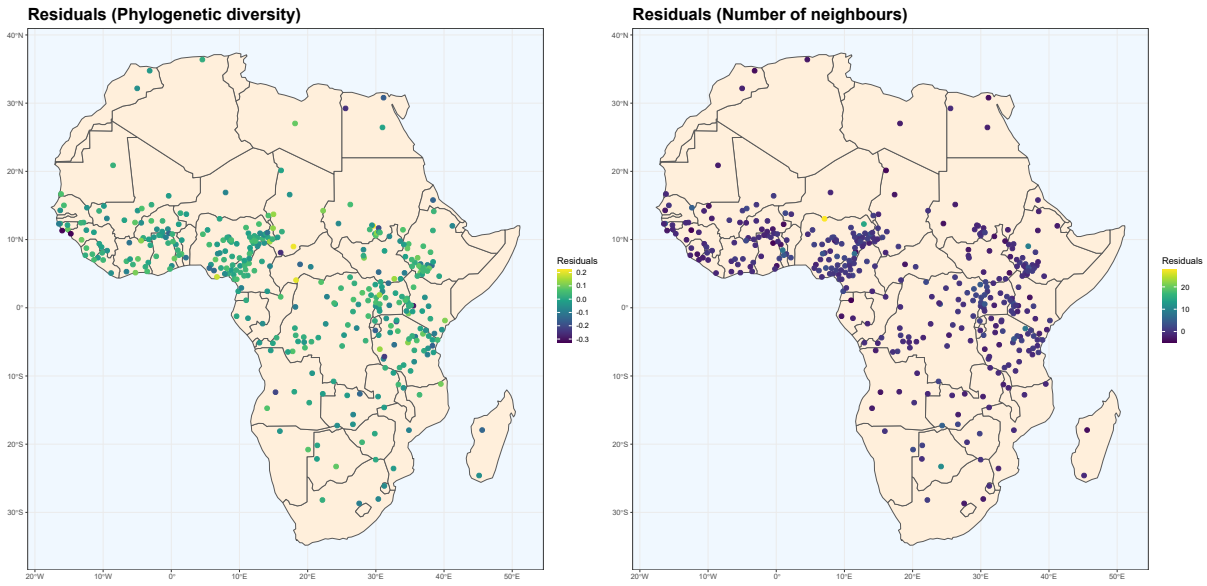


FIGURE 3.16: Model residuals plotted on a map from the phylogenetic model (left) and the language density model (right).

number of languages accurately. Negative residuals mean that the model predicted a value for that language that was too high, and positive residuals indicate that the model predicted a value that was too low. Thus, these maps can provide an insight into where the model's predictive abilities fall short.

In the plot of the language density residuals, we can see that most of the residuals are close to zero, but there are some outliers with high values. One data point has a residual value of 27, which means the model severely under-predicted the number of neighbouring languages for that data point. This data point is the Hausa language, which belongs to the Chadic family and is spoken in Nigeria. The model failed to predict the actual language density surrounding Hausa. Perhaps this is because Hausa does not conform to the general pattern of the languages around it, making it a spatial outlier. It could also be because it has a high value for political complexity and a low TRI, as these are estimated to negatively affect language density. Hausa in reality has a rich language ecology which is perhaps not adequately explained by the variables included in the model. In the residuals for the phylogenetic diversity model, the languages for which the model over- or under-predicted the degree of diversity are spatial outliers. For example, around Guinea-Bissau, there are two Atlantic languages with a low level of phylogenetic diversity in an area otherwise characterised by more diverse language ecologies.

The goal of modelling is not to perfectly predict every observed value, as this could be a sign of *overfitting*, which is also problematic (McElreath, 2020). However, model residuals can alert us to misspecification and highlight ways in which it could

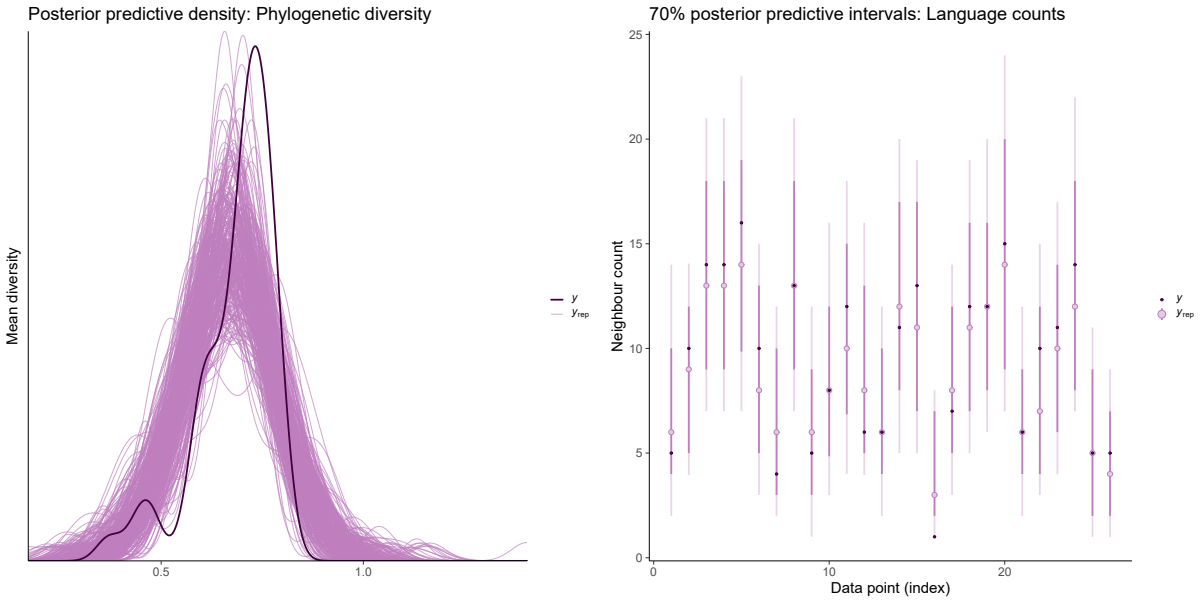


FIGURE 3.17: Posterior predictive draws from the phylogenetic model (left) and the language density model (right).

be improved. In this case, the effect of spatial outliers could be reduced by using a buffer around the centroid of each language polygon in order to capture more of the surrounding diversity of languages which have lower-than-expected levels of linguistic diversity as a result of being unusually small or being located in a coastal area. Choosing the spatial extent of such a buffer would be non-trivial, however, and so this is left to future work.

3.4 Discussion

In this section, the results of the phylogenetic diversity model will be discussed, followed by the results of the language density model. Then the differences between the two outcome variables and the extent to which they can be compared will be explored. Lastly, I will focus on the limitations of the study and how future studies could build on these findings and the methodology.

The sociocultural variables with the biggest impact on the phylogenetic diversity within the ecology of a given language are the spatially lagged (SLX) terms. This suggests that permanently settled, agricultural and politically complex societies can exert pressure on the language communities around them which, over time, may lead to a reduction in linguistic diversity. Out of the cultural variables, political complexity is the one with the strongest impact on the distribution of both language density and phylogenetic diversity, although its effect on diversity is stronger. This is in line with some previous hypotheses and findings (Currie and Mace, 2009; Skirgård, 2021).

This study included a variable describing the levels of hierarchy *within* the local community, which has not been included in previous studies. Community hierarchy can be described as the size of the family unit and the extent to which the extended family plays a role in social organisation. A high level of within-community hierarchy means that nuclear families are organised into larger units, such as extended families or clan-barrios (Murdock and Divale, 1999). The positive effect of this variable could be direct or indirect. Indirectly, it could be the case that in societies with a higher level of political complexity, the family plays a smaller role in a society's jurisdiction, as this role will be taken over by the chiefdom or state. This could lead community hierarchy to be in complementary distribution with political complexity. A more direct explanation could be that societies which rely on extended family units as a form of social organisation, perhaps comprising networks of family units, are more likely to rely on interactions and exchanges outside the local community, which could facilitate some of the kinds of dynamics observed in small-scale multilingual groups (Pakendorf, Dobrushina, and Khanina, 2021). In these situations, exchange and collaboration between groups is essential, but the distinct languages of these groups still play an important role in defining the social and cultural identities of their speakers (Di Carlo, Esene Agwara, and Ojong Diba, 2020; Lüpke, 2010). It is also possible that this form of social organisation is related to marital practices like exogamy, which are also frequently a feature of highly multilingual societies (see e.g. (Aikhenvald, 2002; Pakendorf, Dobrushina, and Khanina, 2021)), although this is speculative. Some information about marriage patterns is included in Murdock and Divale (1999), but this variable had more missing values than the others in this study and it was therefore excluded here.

Agriculture and settlement strategy appear to play a relatively minor role in shaping the phylogenetic diversity surrounding a language. As expected based on previous findings by Currie and Mace (2009), a higher level of reliance on agriculture by the language in question and its neighbours has a slightly negative impact on diversity. According to the model of language density, agriculture and settlement strategy have no impact on how many languages tend to coexist in a given area. This result may be specific to the African context, where much of the agriculture which is practised relies on crops such as tubers and tree fruits which are not amenable to long-term storage, as well as small-scale grain-based agriculture (Bostoen, 2020; Murdock and Divale, 1999). It is important to bear in mind that historically, the processes which shaped the distribution of languages across the world might be very different between macroareas.

The climate variables generally appear to have a stronger impact on language density than diversity. Low temperature seasonality and high precipitation has a

consistently positive effect on language density when controlling for dependence on agriculture, although the model estimates this effect with a considerable level of uncertainty. Low seasonality and high precipitation are both associated with tropical areas, which includes some of the most linguistically dense and diverse areas of the world. This provides some tentative support for the ecological risk hypothesis proposed by Nettle (1996). However, when considering the SLX terms, climate variables do not seem to have a reliable correlation with phylogenetic diversity in either direction. Mean temperature correlates negatively with diversity, as shown in Figure 3.13, but the spatial lag of the variable has a weak positive effect, as shown in Figure 3.15. Additionally, there is considerable uncertainty around the median, which suggests that the model cannot disentangle the effect of temperature from the other effects in the model, as high uncertainty can be a sign of multicollinearity (McElreath, 2020). More data would be needed to see if temperature has a consistent impact on diversity across areas, or if the effect is actually negligible once other variables are controlled for. Terrain ruggedness has a very weak positive effect on phylogenetic diversity, and its spatial lag has none, suggesting that isolation also has a relatively weak effect on diversity.

A similar picture emerges in the language density model: the estimated effect of temperature differs from the estimated effect of its spatial lag (the values of neighbouring language locations), which makes it more difficult to draw conclusions about its overall effect. In contrast, the impact of terrain ruggedness and its spatial lag on language density is reliably positive, albeit somewhat weak. It is not surprising that the effect is positive, but some might be surprised that it is not stronger, given that some of the most linguistically dense and diverse areas of the world are regions of high terrain variability, such as the Rift Valley and its surroundings in East Africa, as geographic isolation can contribute to linguistic divergence (Kießling, Mous, and Nurse, 2008). Childs (2010) also lists isolation by geography as one of the reasons why some Atlantic languages in West Africa have survived despite the encroachment of the larger, more powerful Mande languages. However, isolation occurs not just due to the presence of topographic barriers; it is also a result of communities living in a terrain which would be considered inhospitable for other reasons, making it less desirable for other groups to occupy it (Childs, 2010). Perhaps measuring how hospitable a given terrain or climate is for human habitation would be an alternative way of capturing this effect.

For the cultural variables, the results of the language density model have a higher level of uncertainty than those of the phylogenetic diversity model. It should also be noted that the model evaluation suggests that these variables are less influential than the environmental ones. Overall, the distribution of language density thus appears

to be influenced far more by environmental factors than by cultural ones, with the exception of political complexity.

The results of this study mostly support the results of previous studies while adding important nuance. Notably, Hua et al. (2019) found only a weak effect of landscape on language density, which aligns with the relatively weak but reliably positive effects found in this study. Thus, this study finds some support for geographic isolation as a driver of both phylogenetic diversity and language density. Additionally, Hua et al. (2019) found that a tropical climate had the strongest effect on language density, which was interpreted as providing support for the ecological risk hypothesis (Nettle, 1998). Similarly, in this study, climate variables like mean temperature and low seasonality had a positive effect on language density and diversity, although these effects were weaker than those found in the previous study.

This study can provide an insight into the effect of some environmental factors on linguistic diversity when controlling for cultural variables. For example, the positive effect of climate on linguistic diversity could in part be due to its effect on the ability of societies to practice certain forms of agriculture, and therefore, once agriculture is added to the model, it usurps some of the positive effect of climate[7]. However, this could also occur in the other direction, with the climatic effects usurping the effect of agricultural practices. This may have happened in the language density model, where agriculture was found to have no effect. This suggests that the impact of climate on language density is not only a result of the link between climate and certain types of agriculture.

It is promising that this study, despite its smaller sample size and scope, found effects which align with previous research. Broadening the scope of future studies to include as much data as possible would make it more comparable to previous literature. Another limitation of this study is the high level of uncertainty estimated for some of the effects. This is an issue which could be mitigated by adding more data, as the relatively small sample size could interfere with the model's ability to estimate effects reliably. Perhaps multicollinearity also plays a role in the uncertainty, and further exploration of dependencies between predictor variables could mitigate this.

3.5 Conclusion

This study has examined the effect of both cultural and environmental variables on phylogenetic diversity and language density in Africa using a spatial conditional autoregressive (CAR) model with spatially lagged covariates. The level of phylogenetic

[7]This is a phenomenon described in more detail in Chapter 6 of *Statistical Rethinking* by McElreath (2020).

diversity (or lexical differentiation) in the ecology of a language community appears to be most influenced by the societies around it, particularly their level of political complexity. Researchers wishing to further study the cultural or environmental variables which impact the dynamics of linguistic diversification might benefit from using spatially lagged covariates if they wish to capture this effect. Despite the difficulties associated with comparing the results of the models when using different types of outcome variables, it is clear that the factors which impact phylogenetic diversity and language density are related but not identical. Therefore, the choice of how to quantify linguistic diversity can be expected to impact model results, and language density may not always be an adequate proxy for phylogenetic diversity. One of the effects which remained consistent for both variables was that of political complexity, which was found to be the cultural variable with the strongest and most consistent impact on the distribution of linguistic diversity. However, as a whole, language density appears to be more heavily influenced by the climate variables included in this study, while the distribution of phylogenetic diversity seems to be driven more consistently by cultural factors as well as, to a lesser extent, terrain ruggedness. Future work could focus on testing these results with a larger dataset. It could be interesting to build on the idea that phylogenetic diversity might be driven by factors which are distinct from those driving diversification by testing whether they are shaped by different sets of variables entirely, since this study only compared models with the same set of variables.

Chapter 4

Variation in the areal diffusion of morphosyntactic features in Africa

In this chapter, I will build on previous work by Guzmán Naranjo and Mertner (2022) and Guzmán Naranjo, Mertner, and Urban (2024), presenting a modified version of the *multivAreate* models used in those studies. The primary goal of this chapter is to use this method to detect variation in the patterns of areal diffusion exhibited by different types of structural features in Africa. A lot has been said about the relative stability of structural elements of language (for example in Dediu (2011), Dediu and Cysouw (2013), Nichols (2003), and Skirgård et al. (2023)). Far less attention has been dedicated specifically to studying and comparing the areal patterns of linguistic features, although work on this has increased in recent years (Cathcart et al., 2018; Neureiter et al., 2022; Nikolaev, 2019; Nikolaev and Grossman, 2018; Ranacher et al., 2021; Wieling, Nerbonne, and Baayen, 2011). The central question of this chapter is whether the method presented here can detect variation in the spatial extent of different categories of linguistic features, which I will refer to as their *diffusibility* in order to distinguish this from the more common notion of borrowability in the literature (Matras, 2007; Thomason and Kaufman, 1988). Features which have diffused across large areas are likely to be easier to borrow, so diffusibility and borrowability are undoubtedly related, hence the relevance of this study for researchers interested in borrowability. A related but secondary goal of the chapter is to examine the areal patterns which arise for these different feature categories, for example comparing how gender/noun class systems and verbal categories cluster in space. These kinds of patterns can provide evidence for structural convergence due to language contact in Africa, and detect where certain sets of features are likely to have converged.

4.1 Introduction

Language contact is one of the main drivers of language change. When speakers of different language communities come into contact with each other, this can cause gradual changes to the lexicon, phonology, and morphosyntax of the languages involved through processes like adaptation to non-native speakers and bilingual transfer (Thomason, 2010). A set of related questions that has been raised repeatedly concerns the extent to which structural elements of language can be transferred between languages as a result of social interactions between their speakers, and under what conditions such transfer takes place (Muysken, 2010; Trudgill, 2011). This idea was famously formalised by Thomason and Kaufman (1988) in the form of a *borrowability scale*, which notably included structural features, like inflectional morphology, which were thought unlikely to be influenced by contact. The idea of a borrowability scale could account for the common observation that a lot of contact effects tend to occur in the same linguistic domains across languages, while also acknowledging that deeper structural changes, including ones which entirely change the typological profile of a language, can occur under the right circumstances. This furthered our understanding of language contact as a process that is subject to various constraints which can make certain types of transfer more or less likely, including cognitive biases, markedness, learnability, and external factors, while noting that these constraints are not absolute (Thomason, 2010).

Influential work by Nichols (1992) expanded on existing notions of borrowability using quantitative data on the typological profiles of a sample of diverse languages. She compared the areal distribution of features, particularly their geographic contingency, to their distribution within families and, based on her findings, concluded that some structural elements of language are more areal than others.

Conclusions about which kinds of grammatical features are inherently more stable or borrowable remain elusive, as previous scales are continually called into question by new research (see e.g. Skirgård et al. (2023) for a re-examination of the areal hypotheses proposed by Nichols (1992)). One of the challenges of the field lies in matching the research on specific instances of contact-induced change, which can provide insights into the possible but not necessary the probable, to the kind of data that is available for large-scale cross-linguistic analysis. An edited volume by Matras and Sakel (2007), which collected detailed standardised surveys on a diverse sample of languages in order to compare them systematically, made a valuable contribution to the field and added more detail to borrowability scales. However, this approach may not be feasible for much larger samples of languages, as it requires the availability of experts who are able to say with some certainty what elements of a language

have been changed as a result of contact, and which have not. In many cases, this kind of information is not available, which is why using statistical methods could be a promising way to detect patterns in larger samples of languages, including ones which are less well-documented.

The question of stability and borrowability has important implications for those who are interested in reconstructing deep historical relationships between languages. Language families are not identifiable past a time span of around 10-12,000 years, the time depth of the oldest currently known language family, Afro-Asiatic. However, some genealogical relationships between languages are undoubtedly far more ancient than that. This attests to the fact that all features in a language can eventually change and/or undergo replacement (Nichols, 2003). Following Nichols (2003), I will use the term *stability* to refer to the rate at which a feature changes or is replaced, whether through contact or other mechanisms. This can also be framed as a propensity (or lack thereof) for replacement or change.

The concept of 'borrowability' encompasses an interacting set of dynamics which take place in language contact situations. The 'ease' with which a feature is borrowed refers in Thomason and Kaufman (1988) to the intensity of contact which is required for a particular feature to transfer from one language to another. Some features, like content words, are argued to require only casual or sporadic contact, while others require intensive contact in order to be borrowed. 'Casual' and 'intensive' contact could be defined as the extent to which proficiency in the other language (or languages) is prevalent in their speaker communities, the frequency with which speakers from these communities interact with each other, the time depth of contact between the languages, or some combination of all of these variables (Thomason, 2010). In this way, 'intensive contact' can describe a series of very different sociolinguistic situations, and 'contact intensity' can be social or temporal, or (more often) both (Muysken, 2010). Additionally, the borrowability of a feature may refer either to the frequency with which that feature appears to be transferred between languages, or to an implicational relationship between features, such that the borrowing of one feature becomes a necessary condition for the borrowing of another (Matras, 2007).

It is often difficult to say with certainty whether a particular structural pattern or linguistic feature has been borrowed. Ascribing shared features to contact necessitates ruling out other possible explanations, including shared inheritance, neutral drift, and cognitive, social or environmental biases (Bickel, 2017). Additionally, the structure of the languages involved in contact situations can influence which features are borrowed and how likely they are to be borrowed (Matras, 2007). Sakel (2007) makes a useful distinction between *matter borrowing* and *pattern replication*, which I will draw on in this chapter. Matter borrowing includes the phonological form

of the morphological material which is replicated from one language to another, while pattern replication involves the transfer of conceptual structures or categories without any transfer of forms. Over time, pattern replication can lead to structural convergence between languages. While matter and pattern borrowing can and often do co-occur, this is not always the case (see e.g. Aikhenvald (2002) for examples of grammatical convergence without lexical transfer).

As linguistic forms are more salient to the speakers and therefore more likely to be changed through conscious manipulation, they are also more prone to processes of contact-induced differentiation, which can obscure historical relationships (Di Carlo and Good, 2023). In contrast, contact-induced divergence does not seem to impact grammatical structures (Mansfield, Leslie-O'Neill, and Li, 2023). Because of this, structural convergence can be a particularly valuable source of information on the deep past of under-researched languages, like those of Africa (Heine and Nurse, 2007). It has also been argued that the kind of intensive contact that leads to structural convergence is the same as that which leads to the formation of linguistic areas (Trudgill, 2010).

Linguistic areas (or *Sprachbünde*) are typically defined (with varying degrees of specificity) as geographic areas within which languages show a level of structural similarity for which the most likely explanation is convergence due to language contact (Matras, 2011). There is little agreement as to what exactly constitutes a 'linguistic area', such as how many features need to be shared between the languages involved, whether a particular number of language families is required, or the extent to which all of those languages should share the same set of features (Campbell, 2017). Additionally, it is far from trivial to decide which languages are members of an area or not, and this is a topic rife with disagreement for specific areas (see e.g. Joseph (2010) for an overview of this for the Balkans). Some languages may be considered 'partial members' while others are 'full members' of a particular linguistic area (Haspelmath, 2001, p. 1504), although this arguably only adds to the general sense of vagueness around what and where a linguistic area is. It is equally difficult to clearly define their geographic boundaries, and some have called for the abandonment of this pursuit in favour of a focus on the history of the convergent features themselves and how they arose (Campbell, 2017), although the two are not mutually exclusive. The boundaries of proposed linguistic areas thus tend to be diffuse and can also overlap with other linguistic areas in space and time (Thomason, 2001).

Past contact between languages or past processes of language shift can sometimes be detected in synchronic distributions of structural patterns, conceptual categories, or linguistic forms across space (Sands, 2022; Sands and Gunnink, 2019). Perhaps it does not matter whether linguistic areas are merely a convenient way to describe the

existence of certain linguistic phenomena in a certain place and time, or reflections of historical realities which can lead to important insights about the processes of convergence which took place between the languages spoken there. As Dahl (2001, p. 1458) wrote, these areas can be seen as "the sum of many such binary relationships", referring to instances of language contact. The existence of linguistic areas can provide a hint as to the sum of contact events between the languages spoken there; a larger sum indicates more contact events, and more contact events provide evidence for a longer history of co-territoriality of the communities in these areas.

Quantitative methods are well-suited to investigating the related questions of feature stability and diffusibility. The phylogenetic stability of different kinds of linguistic features has been studied using Bayesian phylogenetic models, or models incorporating information about language geneaology and geography simultaneously (Dediu, 2011; Greenhill et al., 2010; Kauhanen et al., 2018; Murawaki and Yamauchi, 2018). These studies have found some support for the idea that some features are inherently more stable than others. Some of the studies which have examined the areality of linguistic features have relied on the specification of a neighbour graph that indicates which languages are in contact (neighbours) and which are not, as in Kauhanen et al. (2018), Murawaki and Yamauchi (2018), and Nikolaev (2019). This can be an advantage, as it allows for the flexible incorporation of prior knowledge about language relationships into the model, as discussed in Chapter 2. However, a limitation of such methods is that they cannot infer the geographic range of contact between languages. This means that if there is some uncertainty about the range within which languages are likely to be in contact or show effects of areal diffusion, a different method (or more flexible specification of spatial relationships between languages) may be more suitable. No quantitative study so far has used a spatial model to examine variation in the areal spread of different kinds of linguistic features, which could provide a new perspective on borrowability and help detect past and present convergence between languages. Studies on language contact and the areality of linguistic features have already provided interesting insights into language change, history, and past migrations, which can complement findings on language geneaology (Allassonnière-Tang et al., 2021; Idiatov and Van de Velde, 2021; Nikolaev, 2019, among others).

This study presents a Bayesian spatial model for quantifying variation in the areal spread of categories of linguistic features. The model is tested on a selection of morphosyntactic features in Africa, including nominal number, gender/noun class systems, bound verbal categories, verbal tense and aspect markers, and word order, all of which are retrieved from Grambank (Skirgård et al., 2023). The model builds on previous work by Guzmán Naranjo and Mertner (2022) and Guzmán Naranjo,

Mertner, and Urban (2024) using a latent Gaussian process (GP) within a multivariate probit model (*multivAreate*) to infer the strength and range of potential language contact. This model is extended to include multiple latent GPs which group features together based on category membership. For example, verb finality and verb mediality in canonical clauses both belong to the category of 'word order'. The purpose of this is to allow for the inference of variation in the overall diffusibility of different categories of morphosyntactic features, rather than for individual features, which may have a noisy signal. This means that the inferred connections between languages in space are more likely to reflect past or present contact between those speaker communities, as random variation in the areal distribution of individual features will be filtered out.

Phylogenetic relationships between languages will be controlled for using a global tree with branch lengths calibrated to archaeological data, indicating how closely or distantly related language pairs are Bouckaert et al. (2022). Inter-feature correlations are also controlled for and inferred simultaneously. The explanatory power of the spatial and phylogenetic components of the model will be compared using k-fold cross-validation, and differences between feature categories will be examined to determine whether some appear to be more or less areal than others.

4.1.1 Linguistic areas in Africa

The question of whether Africa itself is a linguistic area is one that has fascinated and puzzled researchers working on African languages (Childs, 2010; Leyew, 2008). While there are some features which only exist in Africa, such as click consonants, these features are also geographically restricted *within* Africa, such that they cannot be considered typical of the continent as a whole (Clements and Rialland, 2008; Creissels et al., 2008). Leyew (2008) therefore defines African features not by their uniqueness to Africa, but by their prevalence in languages spoken within Africa relative to those spoken outside it. Using this metric, Leyew (2008) compiled a list of morphosyntactic, lexical and phonological features which could define Africa as a linguistic area. Even more areal patterns emerge when considering Africa as a collection of geographic zones, which can be characterised in terms of their linguistic diversity as either *accretion zones* or *spread zones* (Nichols, 1992), and in terms of sets of convergent features. Clements and Rialland (2008) divided Africa into a set of 'phonological zones' which bear significant similarity to the later map by Güldemann (2018b) depicting Africa's linguistic areas, accretion and spread zones, and areas which fall neatly into none of these categories.

I will provide a brief summary of the main geographic zones into which Africa has been divided, while noting that even those who have participated in the goal of delineating these areas state that their boundaries are fuzzy (Güldemann, 2018b; Heine and Nurse, 2007). Güldemann (2018b, p. 473) identifies four accretion zones in Africa: the Nuba Mountains, the Ethiopian Escarpment, the Rift Valley, and the Dogon Plateau. The spread zones include the central sub-Saharan African Bantu spread zone caused by the series of migrations known as the Bantu Expansion (Bostoen, 2020), and the Afro-Asiatic spread zone of northern Africa. Much of what remains of the continent is classified as an area that Güldemann (2018b) calls the 'central transition sphere', which stretches all the way across the Sahel and into the East-Sudan Gregory Rift. Notably, the central transition sphere encompasses all of the four accretion zones that Güldemann (2018b) defines. The main feature of this area is that it cannot be characterised as a whole by a set of features the way linguistic areas (ideally) can, even though it does show some local patterns of contact-induced convergence. The Sahel, which is part of the central transition sphere, has been characterised by Güldemann (2018b) as an area comprising many smaller contact zones rather than as a linguistic area. However, others have described it as a possible large linguistic area (Caron and Zima, 2006a,b).

Since Greenberg (1959, 1983) proposed the existence of a large linguistic area in Africa which spans the entire area of West Africa south of the Sahel and stretches, in some definitions of it, all the way to East Africa, interest in this area has been marked (Clements and Rialland, 2008; Güldemann, 2008). It is possibly the most well-researched linguistic area in Africa and has been referred to as *the Macro-Sudan belt* by Güldemann (2018b) and as *the Sudanic belt* by Clements and Rialland (2008). I will use the term Macro-Sudan belt here, as the Sudanic belt refers specifically to a zone of phonological convergence rather than including more kinds of linguistic features.

The Macro-Sudan belt encompasses several areas of intensive language contact and multilingualism in Africa, including the Cameroonian Grassfields where Bantu and Bantoid languages are spoken, and which possibly includes the traces of a smaller past accretion zone (Di Carlo, Esene Agwara, and Ojong Diba, 2020; Good, 2013). Despite their high levels of linguistic diversity, which has been proposed as a necessary or common characteristic of linguistic areas (Tosco, 2000), accretion zones do not necessarily show high levels of contact-induced convergence. In fact, spread zones show some of the most interesting patterns of contact on the continent, such as the Jos Plateau in Nigeria and an area of small-scale multilingualism in Senegal (Lüpke, 2010, 2016). In contrast, the Nuba Mountains have been described as an area without much evidence of grammatical convergence, although some lexical

convergence has been described between the languages in the area, as well as more recent influence from Arabic (Manfredi, 2022). Likewise, the Rift Valley in Tanzania can be characterised in terms of some convergent features, like ejective consonants, but it has not been defined as a linguistic area (Kießling, Mous, and Nurse, 2008).

Chad-Ethiopia was first proposed as a linguistic area by Heine (1975, 1976). A set of features defining the Ethiopian linguistic area, which had a more limited geographic extent, was proposed by Ferguson (1976). This list became highly influential and has been variously contested (Güldemann, 2018b; Tosco, 2000) and supported (Crass and Meyer, 2007). This may have been a larger linguistic area in the past, but environmental changes and the influx of Arabic-speaking groups caused the Chadian and Ethiopian parts of the area to be separated from each other. Now, the area (if it is to be analysed as a single area) can be seen as comprising two separate convergent zones: one centred on the Horn of Africa and Ethiopia, and one with its centre in Chad.

Another large and ancient linguistic area that likely had a very different distribution in the deep past is the Kalahari Basin (Güldemann, 1998). There is some uncertainty as to whether it should be considered an accretion zone due to its high levels of genealogical diversity, particularly compared to the area surrounding it, which is part of the Bantu spread zone (Güldemann and Fehn, 2014). The Kalahari linguistic area has been characterised as a zone of contact-induced convergence based on a set of morphosyntactic and phonological features, including the famous click consonants (Clements and Rialland, 2008; Sands and Gunnink, 2019). The structural convergence between these languages that led them to be classified initially as a lineage named Khoisan (Greenberg, 1963), and the long history of habitation in the area, suggests that contact between these groups has been intensive and regular over a very long period of time. This area of intensive contact may have been much larger in the past. However, the migration of Bantu-speaking groups throughout central and southern Africa means that this is difficult to demonstrate. A conspicuous similarity, namely the presence of click consonants in two difficult-to-classify hunter-gatherer languages in Tanzania, has led to the hypothesis that this was a very large linguistic area in the past which was effectively divided in two by the Bantu Expansion (Güldemann, 1998). Such ideas show how areal distributions of structural features can be a window into the past.

4.1.2 An areal view of African morphosyntax

The effects of language contact on the morphosyntax of the languages of Africa is historically underexplored, but interest in it is increasing (Creissels et al., 2008). This

section will outline some examples of convergence at the morphosyntactic level in Africa, and briefly explain how researchers arrived at contact as the most likely explanation for specific instances of convergence. Where possible, the focus will be on pattern borrowing. First, word order will be discussed, followed by the nominal domain, including nominal number and classification systems, and lastly, the verbal domain, including tense/aspect marking and other verbal categories.

While the features which will be discussed and studied here are general enough to be included in a global typological database, many of the features which are used to diagnose convergent areas in the literature are ones which are highly specific to the local context and which may therefore be very rare outside Africa. As such, many of these are not included in large-scale typological databases like Grambank. Additionally, since typological data about grammatical systems tends to be binary, categorical, or ordinal, no information about specific forms is included. Although similarities in form, for example between noun class prefixes in the languages that have them, can provide convincing evidence of either inheritance or contact, we are confined to focusing on shared features or patterns when using cross-linguistic databases like Grambank. This kind of data lends itself well to uncovering areas of structural convergence, which is a worthwhile pursuit in its own right, as discussed previously.

Word order

The unmarked order of constituents (Subject, Object, and Verb) in canonical clauses within a language is often referred to simply as *word order*, which is a convention I will continue here. In the global sample of languages surveyed by Nichols (1992), word order was found to be an areally distributed feature, which indicates that it is prone to the impact of contact and diffusion. Indeed, the distribution of word order types in Africa provides a window into the past distribution of languages on the continent and hints at possible contact-induced convergence (Dimmendaal, 2020). Although the majority of African languages have SVO word order due to the extensive geographic spread of Bantu and other Niger-Congo speaking populations, there are some notable exceptions. Afro-Asiatic languages, spoken in the north and east of Africa, tend to have verb-final word order. In the area between the Horn of Africa and Lake Chad, verb-final word order is found in a number of distinct branches of Nilo-Saharan, which points to contact influence from nearby Afro-Asiatic languages and even hints at the existence of an ancient convergence area (Dimmendaal, 2020, p. 217). The area around Ethiopia, Eritrea, Sudan, South Sudan, and the Horn of Africa, encompassing the Nuba Mountains, is also posited as a convergence area elsewhere in the literature and is defined as a residual zone (following Nichols (1992))

due to the high number of languages from distinct lineages spoken in the area (Crass and Meyer, 2007; Güldemann, 2018b; Tosco, 2000).

Nominal categories

There are two main systems of nominal classification in Africa. The first is a type in which nouns are classified into two genders (masculine and feminine). This type is found in almost all branches of the Afro-Asiatic language family, as well as in some branches of Nilo-Saharan, including Nilotic, and a group of languages called Kadu which may either be an isolated family or a member of Nilo-Saharan. They are also found in several isolates, including Sandawe, Kwadi, and Hadza. It is also found in languages belonging to the Khoe-Kwadi family (Creissels et al., 2008, p. 115).

The second type of nominal classification system found in Africa is characterised by the absence of a masculine/feminine distinction, obligatory affixes on nouns which mark the noun class (which may also agree with other constituents in the clause), a relatively high number of noun classes (often including between ten and twenty or even more), and a semantically salient distinction between human and non-human and animate and inanimate nouns. In languages with this type of noun class system, nominal marking of number interacts with the marking of noun classes in intricate ways, as number cannot often be disassociated from noun classes (Creissels et al., 2008, p. 116).

This type of system is widespread across the continent due to the spread of Niger-Congo languages, especially those belonging to the Atlantic and Bantu families. These two branches are geographically far away from each other, suggesting a genealogical rather than areal origin (Creissels et al., 2008). Most linguists agree that noun classes were likely a feature of the earliest Niger-Congo languages, and that the languages within the family that lack them have lost them over time. While noun classes are thus unlikely to have developed as a result of language contact, there are observable contact effects in how they are used and in the form that they take. Although noun class systems are most likely to be inherited from a parent language, they are not unchangeable. Expansion and reduction of noun class systems are both attested processes of change in Bantu languages (Good, 2025). It is rare that Bantu languages lose their noun classes entirely; only one Bantu language, Komo, has been reported to have no noun classes at all (Good, 2025; Maho, 1999). However, it is not uncommon for Bantu and Bantoid languages spoken on the periphery of regions with a high concentration of languages from other families have reduced systems of nominal classification, for example retaining only two classes. This reduction is observed in areas where Bantu and Bantoid languages have contact with non-Bantu languages which lack noun class systems, suggesting this is a contact-induced

change (Van de Velde, 2019). In general, however, contact-induced changes to noun class systems appears to be relatively rare.

Verbal categories

There are few commonalities between African languages as a whole in the verbal domain, except that languages which lack verbal inflection systems entirely are exceedingly rare (Creissels et al., 2008). A high degree of specification in the domain of verbal tense, aspect and modality marking, and rich verbal inflection systems for marking other categories like applicatives, passives and antipassives, is widespread in Africa. Verbal inflection systems for the expression of negation are also especially common in Africa (Creissels et al., 2008, p. 104).

Rich tense marking systems are typical of Bantu languages, which tend to distinguish between tenses using verbal prefixes. These systems are not immune to contact effects. Contact with Bantu languages is the most likely explanation for the presence of a highly specified verbal inflection system for tense in a Nilotic language, Luo, as Nilotic languages typically express information about tense using adverbs of time rather than verbal inflection (Heine and Nurse, 2007, pp. 4–5). Interestingly, there is no similarity between the form of these verbal tense markers (which, unlike in Bantu languages, are positioned before the subject prefix) and tense markers used in the Bantu languages known to be in contact with Luo. Thus, it appears that Luo only borrowed the concepts from neighbouring Bantu languages without borrowing any of the specific forms. Instead, contact seems to have triggered a process of grammaticalisation of time adverbs already present in the language (Dimmendaal, 2001). This is thus a good example of structural convergence, which is the kind of borrowing which the method presented here should be able to detect, as it relates only to the presence or absence of categories in a language without any information about their position or phonological shape in the respective languages.

4.2 Method

The following section will first provide an overview of the data, showing all the features which I selected from Grambank and how they were grouped. Second, I will describe how the range of areal diffusion was inferred using a single latent Gaussian process (GP) for the features which were grouped together. Afterwards, the phylogenetic component of the model will be explained, followed by an overview of how the model combines all these different components.

4.2.1 Feature selection and categorisation

All the morphosyntactic features were retrieved from Grambank (Skirgård et al., 2023). Most features in Grambank are binary; those which were not were binarised prior to model fitting, as done in Guzmán Naranjo, Mertner, and Urban (2024). Features with more than 30% missing values were excluded from the sample. I followed the categorisation of features into the domains of word order, verbal categories, and nominal categories given in Skirgård et al. (2023). I further sorted the nominal categories into *number* and *gender/noun classes* and the verbal categories into *tense/aspect/mood markers* (TAM) and *bound verbal categories*. I chose to distinguish between TAM and verbal categories because the TAM features, as included in Grambank, do not need to be bound to the verb but can be marked by an auxiliary or particle, unlike the features included in the second category (Skirgård et al., 2023). If boundedness is something that makes borrowing more difficult, this could be a relevant distinction to make; if not, the estimated model results for the two categories should be very similar. The motivation behind sorting features into narrower categories was to see if variation between the overall diffusion patterns of features in the nominal and verbal domain would emerge naturally, or if they would be internally heterogenous.

Although it would be ideal to be able to include all Grambank features in the model for a comprehensive comparison, this is not possible due to computational constraints. Using an approximation method for the GP as done in Chapter 6 could make this possible for future versions of the model, but the approximation relies on the use of Euclidean rather than the more realistic geographic distances used here, which account for topography.

4.2.2 Grouped Gaussian processes

As described in more detail in Chapter 2, a latent Gaussian process (GP) is a flexible method which can capture dependencies between data points given a covariance matrix describing the distance between them. In this case, the covariance matrix contains pairwise topographic distances between all the languages in the sample. Thus, the model assumes languages which are geographically close together are likely to be more similar to each other than languages which are far apart (Guzmán Naranjo, Mertner, and Urban, 2024). Topographic distances, the shortest path between points while taking terrain elevation into account, were calculated using the R packages **gdistance** and **topoDistance**. If two languages are separated by an area with a highly variable topography, such as a mountain range, the distance between those languages will include the terrain elevation. Topographic distances were found to

Feature code	Description	Category
GB042	Singular marking on nouns	Nominal number
GB043	Dual marking on nouns	Nominal number
GB044	Plural marking on nouns	Nominal number
GB166	Paucal marking on nouns	Nominal number
GB051	Class assignment: sex	Gender/noun classes
GB052	Class assignment: shape	Gender/noun classes
GB053	Class assignment: animacy	Gender/noun classes
GB054	Class assignment: plant status	Gender/noun classes
GB192	Class assignment: phonology	Gender/noun classes
GB130a	SV order in intransitive clauses	Word order
GB130b	VS order in intransitive clauses	Word order
GB131	V-initial order in intransitive clauses	Word order
GB132	V-medial order in intransitive clauses	Word order
GB133	V-final order in intransitive clauses	Word order
GB136	Fixed constituent order	Word order
GB137	Clause-final negation	Word order
GB138	Clause-initial negation	Word order
GB082	Present tense marked on verb	Tense/aspect
GB083	Past tense marked on verb	Tense/aspect
GB084	Future tense marked on verb	Tense/aspect
GB086	Perfective/imperfective on verb	Tense/aspect
GB095	Core marking includes TAM distinctions	Tense/aspect
GB103	Applicative: benefactive	Bound verbal categories
GB107	Negation	Bound verbal categories
GB118	Locative/directional	Bound verbal categories
GB114	Reflexive	Bound verbal categories
GB115	Reciprocal	Bound verbal categories
GB147	Passive	Bound verbal categories
GB155	Causative	Bound verbal categories

TABLE 4.1: Grambank features included in the model

outperform walking distances and Great Circle distances by Guzmán Naranjo and Jäger (2022).

For this study, I built on previous work by Guzmán Naranjo and Mertner (2022), where we adopted a hyperparameter sharing approach across all the features in the model. The purpose of this approach was to reduce overfitting, as fitting a single latent GP to each feature in the model could lead the model to fit the noise (random variation) rather than the underlying tendencies in the data. In this study, instead of sharing the GP hyperparameters across all the features, I allow them to vary between feature categories (shown in Table 4.1). One of the GP hyperparameters, commonly referred to as the *horizontal scale*, is of particular interest in this context. As explained in more detail in Chapter 2, the horizontal scale can be interpreted as the rate at which the spatial correlation between two points (languages) decreases

	Makonde	**Wandala**	**Yasa**	**Goemai**
Makonde	1	0	0.7	0
Wandala	0	1	0	0.9
Yasa	0.7	0	1	0
Goemai	0	0.9	0	1

TABLE 4.2: Example covariance matrix between four of the sampled languages.

as the distance between them increases. For example, a high value indicates that data points are correlated across long distances, and a low value means that the model does not find evidence of long-distance correlations. As such, the horizontal scale represents the geographic range within which the model allows for potential structural convergence between languages. An additional hyperparameter is also inferred, called the *vertical scale* or the *standard deviation of the GP*. This parameter is described in more detail in Chapter 2.

While the horizontal and vertical scales are drawn from the same prior distribution for all the feature categories, they can otherwise vary freely between feature categories, thus allowing the model to uncover systematic variation in their spatial signal.

4.2.3 Phylogenetic regression with branch lengths

Following Guzmán Naranjo and Becker (2021) and Guzmán Naranjo and Mertner (2022), I use phylogenetic regression to control for non-independence between related languages (Villemereuil and Nakagawa, 2014). Instead of using the phylogeny from Glottolog (Hammarström et al., 2023), which does not have branch lengths which indicate how closely or distantly related language pairs are in time, I use the tree by Bouckaert et al. (2022), as this includes branch lengths, some of which have been calibrated based on archaeological data. The resulting covariance matrix between languages thus reflects not only the position of languages in the tree, but the point in time at which they diverged. Branch lengths were scaled according to the method by Grafen (1989), which is implemented in the R package **ape** (Paradis and Schliep, 2019). This builds in the reasonable assumption that languages which separated recently are likely to be more similar than languages which separated a long time ago. Completely unrelated languages have a correlation of 0. For example, in Table 4.2, we see that Wandala (Afro-Asiatic, Chadic) is completely unrelated to Makonde (East Bantu) and Yasa (Bantu, spoken in Cameroon), but it is closely related to Goemai (Chadic). The two Bantu languages here are slightly more distantly related than the two Chadic languages in this sample.

The idea behind phylogenetic regression is to allow each language to have its own intercept, which indicates how likely a given grammatical feature is to be present in the language, while constraining this intercept according to the covariance structure, which is provided by the tree. Thus, if a feature is present in almost all the languages in a given family (such as certain kinds of noun classes in Bantu), the intercept of Bantu languages will be positive for that feature. Thus, if a few Bantu languages do not follow the same pattern as the other Bantu languages, then the model will assign a larger portion of the variance to the spatial component of the model or the inter-feature correlations. This means that the model should pick up on contact-induced convergence where members of a family diverge structurally from the other members of that family. Isolates, like Hadza, will have an unconstrained intercept. So will languages which are the only representative of their family in the dataset.

4.2.4 Model overview

The model used here is a multivariate probit model, which was introduced in Guzmán Naranjo and Mertner (2022) and extended to allow for the inclusion of missing data in Guzmán Naranjo, Mertner, and Urban (2024). This model, called *multivAreate*, can infer correlations between an arbitrary number of binary features, which makes it very well-suited for quantitative studies in linguistic typology. Unlike many other methods, it does not assume that the features included in a multivariate model are independent. This means it can account for implicational universals of the kind identified by Greenberg (1966). Features which are known to be highly correlated or even mutually exclusive, like verb mediality and verb finality in canonical intransitive clauses, can be included in the same model without leading to issues such as multicollinearity[1].

For the purpose of illustrating how the model works, assume the model includes only two binary features (bearing in mind that the model can be extended to any number of binary outcomes). We will call these two binary variables Y_1 and Y_2. These could represent the presence or absence of two morphosyntactic features, like verb-initial and verb-medial constituent order. The model estimates two latent variables, $Y*_1$ and $Y*_2$, as coming from a multivariate normal distribution:

$$\begin{bmatrix} Y*_1 \\ Y*_2 \end{bmatrix} \sim MultiNormal\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho\sigma_1^2\sigma_2^2 \\ \rho\sigma_1^2\sigma_2^2 & \sigma_2^2 \end{bmatrix}\right) \tag{4.1}$$

[1]When multiple correlated features are included in multivariate models, this can lead to high uncertainty and misleading results, as described in McElreath (2020).

with the constraints:

$$Y_1 = \begin{cases} 1, & \text{if } Y*_1 >= 0 \\ 0, & \text{if } Y*_1 < 0 \end{cases} \tag{4.2}$$

$$Y_2 = \begin{cases} 1, & \text{if } Y*_2 >= 0 \\ 0, & \text{if } Y*_2 < 0 \end{cases} \tag{4.3}$$

where μ_1 and μ_2, and σ_1 σ_2 are the means and standard deviations of $Y*_1$ and $Y*_2$, respectively, and ρ is the correlation between them.

The full model contains a phylogenetic term for each outcome and a grouped Gaussian process. With these two terms the rest of model specification is as follows:

$$\mu_1 = \alpha_1 + \xi_1 + \eta_1 \tag{4.4}$$

$$\mu_2 = \alpha_2 + \xi_2 + \eta_2 \tag{4.5}$$

$$\alpha_1 \sim \mathcal{N}(u_1, 0.5) \tag{4.6}$$

$$\alpha_2 \sim \mathcal{N}(u_2, 0.5) \tag{4.7}$$

$$\xi_1 \sim \mathcal{N}(0, \sigma^2_{p1}\mathbf{A}) \tag{4.8}$$

$$\xi_2 \sim \mathcal{N}(0, \sigma^2_{p2}\mathbf{A}) \tag{4.9}$$

$$\eta_1 \sim MultiNormal(0, \Sigma_{GP}) \tag{4.10}$$

$$\eta_2 \sim MultiNormal(0, \Sigma_{GP}) \tag{4.11}$$

$$\Sigma_{GP} = K(\lambda, \sigma_{GP}, \mathbf{D}) \tag{4.12}$$

$$K_{i,j}(\lambda, \sigma_{GP}, \mathbf{D}_{i,j}) = \sigma^2_f \left(1 + \frac{\sqrt{5}|\mathbf{D}_{i,j}|}{\lambda} + \frac{5(\mathbf{D}_{i,j})^2}{3\lambda^2}\right) exp\left(-\frac{\sqrt{5}|\mathbf{D}_{i,j}|}{\lambda}\right) \tag{4.13}$$

$$\lambda \sim InvGamma(5, 5) \tag{4.14}$$

$$\sigma_{GP} \sim HalfNormal(0, 2) \tag{4.15}$$

$$\sigma^2_{p1} \sim HalfNormal(0, 0.5) \tag{4.16}$$

$$\sigma^2_{p2} \sim HalfNormal(0, 0.5) \tag{4.17}$$

where α_1 and α_2 are the intercepts of the features, the priors of which are calibrated to reflect global tendencies in feature values, represented here as being centred around a given informed prior u. σ^2_{p1} and σ^2_{p2} are the standard deviations of the phylogenetic intercepts ξ_1 and ξ_2, and $\mathbf{A}$ is the phylogenetic correlation matrix constructed from the tree.

The GP is captured by η_1 and η_2, which are latent variables sampled from a multivariate normal distribution. The spatial correlation is represented by Σ_{GP}. Although η_1 and η_2 are estimated for each feature separately, if they belong to the same group (e.g. verbal categories), then Σ_{GP} is shared between them. For each group of features, a different Σ_{GP} is estimated. Σ_{GP} is as a function of two hyperparameters, the horizontal scale λ and the standard deviation (vertical scale) σ_{GP}. These hyperparameters are shared between the two features. $\mathbf{D}$ is a distance matrix between all points, in this case with topographic distances, and $\mathbf{D}_{i,j}$ indicates the distance between points i and j. The kernel function K used here is a Matérn 5/2 kernel, which is a common choice in spatial statistics (Williams and Rasmussen, 2006). For a brief overview of kernel functions in Gaussian processes, see Chapter 2.

Like the model presented in Guzmán Naranjo, Mertner, and Urban (2024), this version of *multivAreate* can also use the estimated spatial correlations, phylogenetic effects, and correlations between features to impute missing values during model inference. This approach is recommended in cases of data sparsity by McElreath (2020) in favour of excluding all languages for which only incomplete data is available. This helps avoid biased inference induced by the non-random distribution of missing values. For a more detailed discussion of bias in the distribution of missing typological data, see Chapter 6.

4.3 Results

The results of the best model will be presented here. The first section will evaluate all the models which were tested and compare their predictive accuracy across all features as well as for individual features and categories. This will be followed by the results on variation in the estimated parameters for the different feature categories. Spatial predictions for the individual features will then be visualised, and some of these will be aggregated to show patterns of overall convergence within specific domains. Following that, the phylogenetic effects will be presented and briefly discussed. All of the generated plots which are not shown here, including the inter-feature correlations, can be found in Appendix B. The results will be contextualised in light of the literature on African linguistics and areal typology in the discussion (Section 4.4).

4.3.1 Model comparison

In this section, the results of model evaluation and comparison will be presented, as they provide important context for understanding the subsequent results. All the

Category	**GP**	**Phylo**	**GP + phylo**
Nominal number	0.53	**0.57**	0.55
Gender/noun classes	0.57	0.53	**0.62**
Word order	0.6	0.48	**0.61**
Tense/aspect	0.56	0.47	**0.59**
Verbal categories	**0.75**	0.5	**0.75**
Overall	0.6	0.5	**0.63**

TABLE 4.3: Mean balanced accuracy per feature category and model.

models were evaluated using k-fold cross-validation. The balanced accuracy with which the model predicted the presence or absence of linguistic features was then calculated. In order to shed light on possible differences in the areality and stability of different features and feature categories, these accuracies will be compared using a GP model and a phylogenetic model. Each of these will be tested against the model with both components. Interestingly, in the comparison between the model with only a GP against the model with only a phylogenetic component in Table 4.3, it is clear that the GP overall performs better. Only one feature category, nominal number, is predicted with a higher degree of accuracy by the phylogenetic model. This suggests two possibilities: firstly, that the distribution of these structural features within languages in Africa could more heavily influenced by areal diffusion than previously thought; and secondly, that the GP is likely able to pick up some of the phylogenetic signal, given that related languages are also often geographically close. Overall, the model with both a GP and phylogenetic regression performs best, although it matches the GP in predictive accuracy for verbal categories, and the phylogenetic model alone outperforms it for nominal number features. The mean balanced accuracy for the individual features for each of the models can be seen in Figure 4.1.

4.3.2 Variation in diffusibility

This section will discuss variation in the estimated horizontal and vertical scale parameters between groups of morphosyntactic features. The horizontal scale parameter λ is best thought of as the *potential* geographic range of a feature domain. A high value means that, once all the other variables in the model are controlled for, the model cannot *rule out* that areal diffusion occurred within the estimated range. Thus, this should not be interpreted as evidence *for* long-distance contact, necessarily, but a lack of correlations between language pairs beyond a certain distance can be

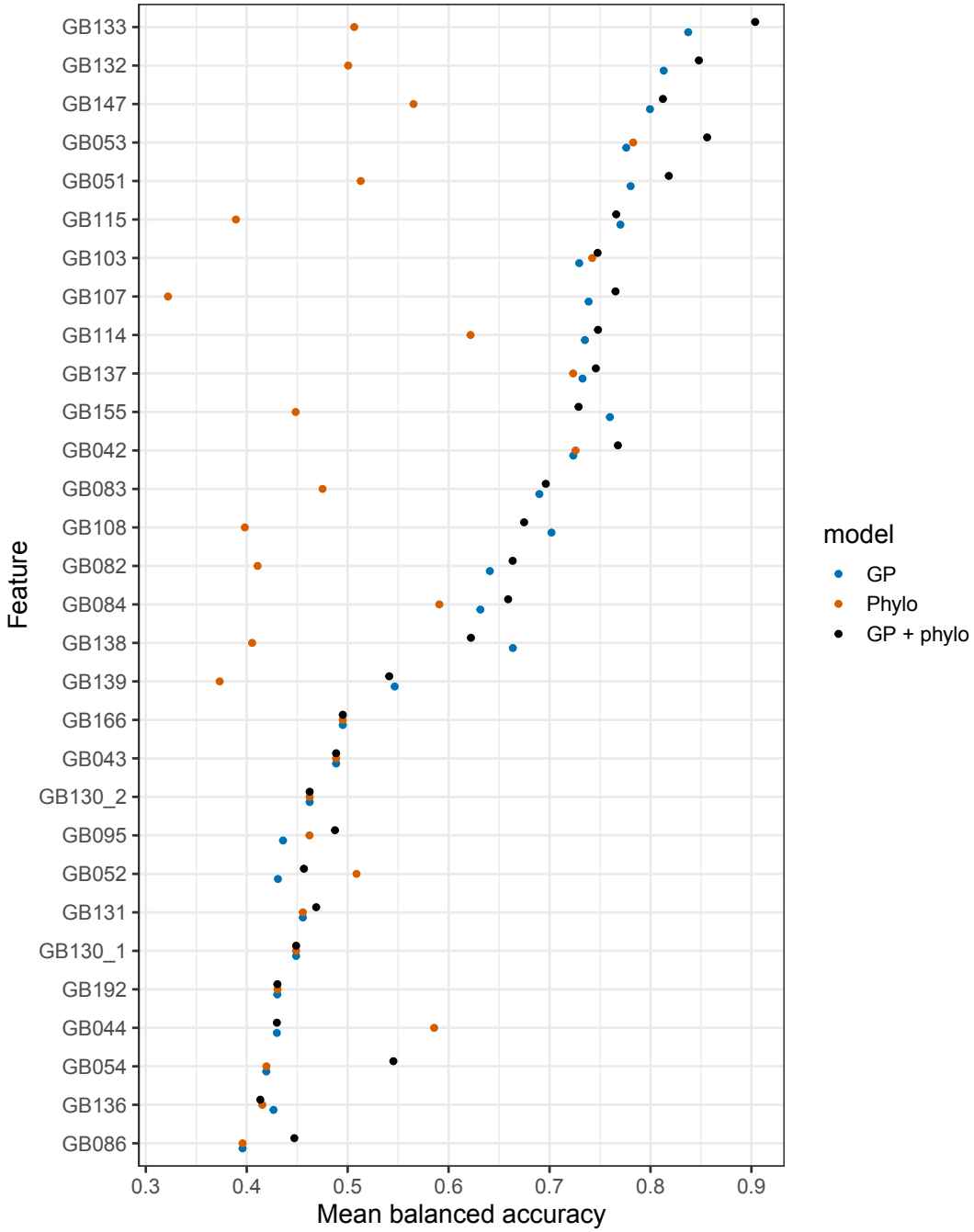


FIGURE 4.1: Mean balanced accuracy for the GP model, phylogenetic model, and combined model, shown for each feature individually.

interpreted as evidence *against* it. Thus, feature categories with a low λ have a smaller spatial extent, which can be interpreted as low diffusibility.

Figure 4.2 shows the confidence intervals of the estimated parameters for the horizontal and vertical scales for each feature group. It shows that, for the horizontal scale parameter, the estimates vary a great deal between the groups. There is also non-trivial variation in the degree of uncertainty for each group. TAM marking

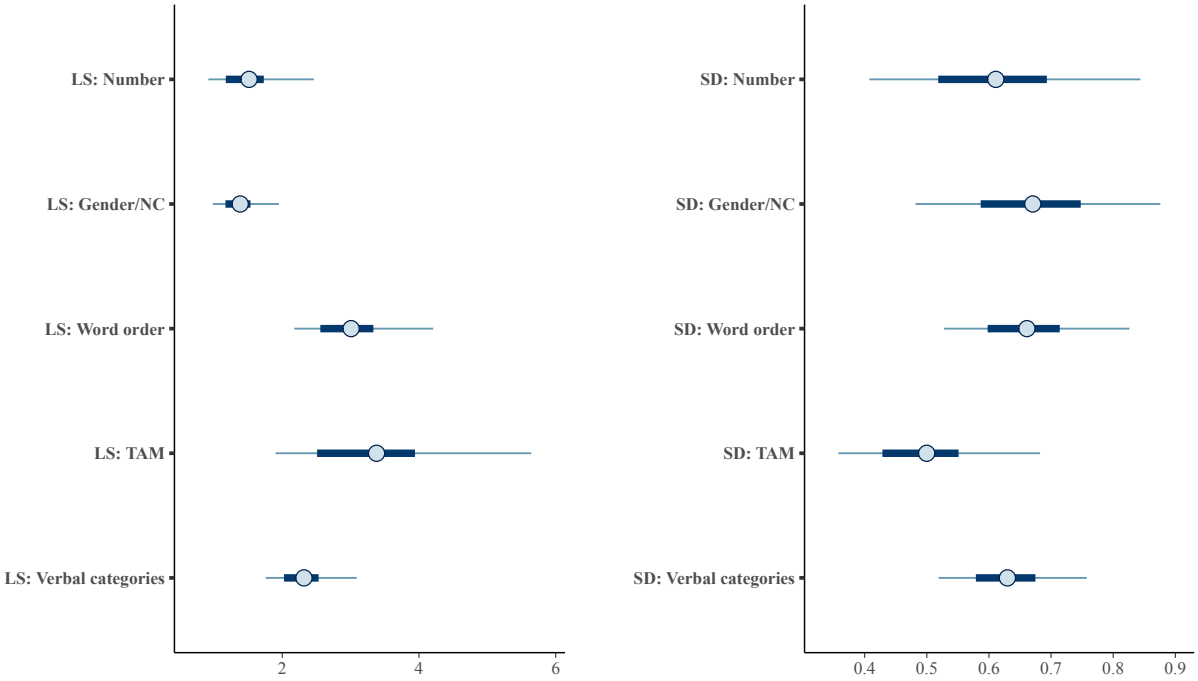
has the largest estimated horizontal scale parameter, but it also has the highest degree of uncertainty, which suggests that the features included in this category are spatially heterogenous. The median value of 3.3 indicates that the strongest convergence effects are likely to be found in a range of 330 km, and that languages outside the range of approximately double this value (660 km) are unlikely to show spatial convergence. Perhaps some of the features in this category have diffused over large areas while others have not. In this way, model uncertainty can provide a window into the spatial behaviour of feature domains. Overall, the rest of the feature categories can be ranked as follows: Word order has a somewhat larger range than bound verbal categories, followed by nominal number and gender/noun class systems, which have the smallest ranges. The groups can be tentatively ranked as follows in order of their potential range of diffusion in Africa (from high to low):

tense/aspect markers > word order > verbal categories > nominal number > gender/noun class systems.

These results will be discussed in more detail in Section 4.4.

The vertical scale parameter (estimates for which are also shown in Figure 4.2) is the marginal standard deviation of the GP, which is often called the amplitude. This is explained in more detail in Chapter 2, but a brief overview of the relevant information will also be given here. The amplitude describes the variability of the feature values of languages within space. A high amplitude indicates that there are spatial clusters of feature values which diverge sharply from the values at other locations. Thus, when the amplitude is high, we would expect to see 'hotbeds' for particular features with sharp boundaries around them. A lower amplitude could manifest as more diffuse boundaries around spatial clusters. It should be noted that amplitude of the GP is not necessarily related to the level of uncertainty around the horizontal scale. However, they can and should be interpreted together.

Gender/noun class systems have a high amplitude compared to TAM marking, although there is considerable uncertainty around the estimates. This indicates that gender/noun class systems exhibit a high level of spatial heterogeneity and clustering. Based on the estimates for the horizontal scale, areal clusters are likely to be small and distinct. Word order has a similar estimated amplitude, whereas nominal number and bound verbal categories both have a slightly lower amplitude. This suggests that the boundaries around areal clusters for these categories will be more diffuse. TAM marking has, by far, the lowest amplitude, and so the spatial clusters for this feature are likely to be large and very diffuse.



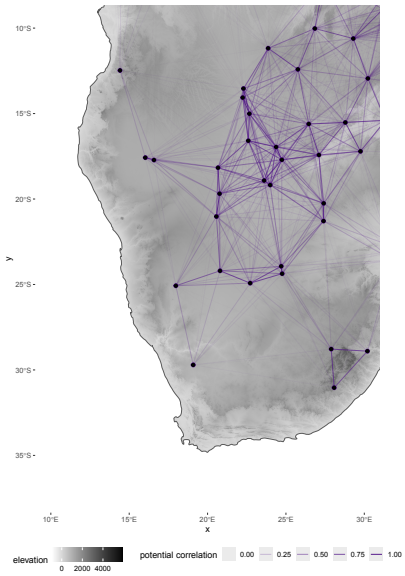
(A) Confidence intervals for horizontal scales per feature group.

(B) Confidence intervals for vertical scales per feature group.

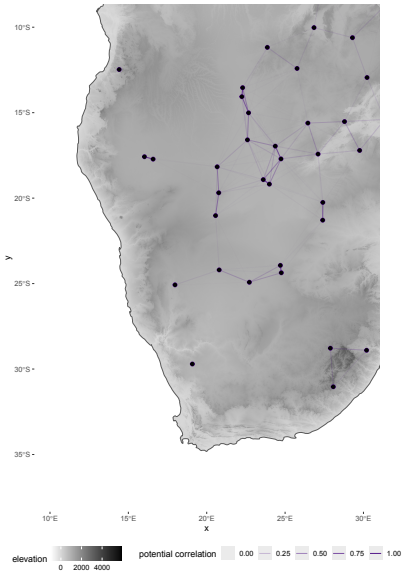
FIGURE 4.2: Confidence intervals for the horizontal (λ) and vertical (σ) scale parameters for each group of features.

Figures 4.3 and 4.4 visualise the inferred differences in the horizontal scale parameter for the feature category with the largest one (TAM marking) and the category with the lowest (gender/noun class systems). The lines represent the potential correlations between languages based on the topographic distances between them. These plots should not be interpreted as evidence for contact; rather, they represent the *potential* for contact to have occurred between the languages which are linked. Essentially, the model finds no evidence for convergence beyond the inferred geographic range. Because of this, these plots can provide evidence against long-distance areal effects, in this case for gender and noun class systems. This does not entirely preclude the possibility of diffusion across larger areas, however. Local contact relations can cause long-distance areal diffusion, for example if a linguistic category is transferred from one language to its immediate neighbours and then the neighbours of those languages also adopt the feature, and so on. For TAM features, on the other hand, the model cannot exclude the possibility of long-distance contact effects.

Visualising the spatial predictions drawn based on these parameters will help with their interpretability. These visualisations will be presented in the following section.

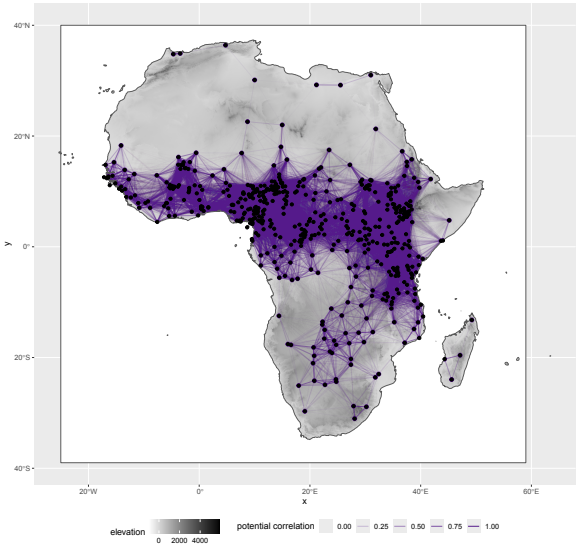


(A) Potential correlation between languages in the Kalahari for TAM markers ($\lambda = 3.4$).

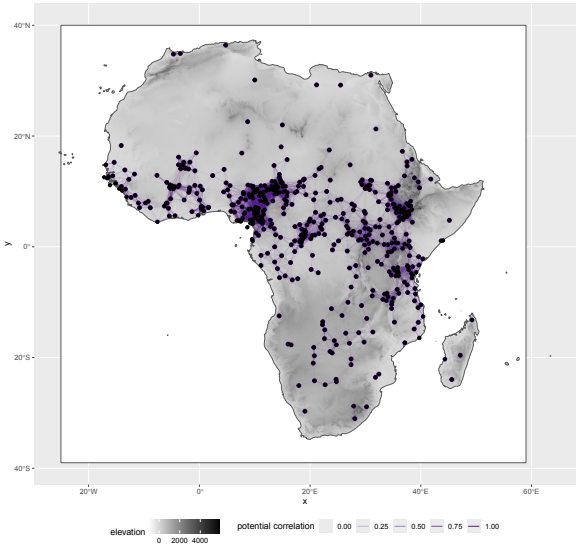


(B) Potential correlation between languages in the Kalahari for gender/noun class systems ($\lambda = 1.4$).

FIGURE 4.3: Potential spatial correlation between a sample of languages in southern Africa for the category with the largest extent (TAM) and the one with the lowest (gender/noun classes).



(A) Potential spatial correlations for TAM markers ($\lambda = 3.4$).



(B) Potential spatial correlations for gender/noun class systems ($\lambda = 1.4$).

FIGURE 4.4: Potential spatial correlation between all the languages in the dataset for the category with the largest extent (TAM) and the one with the lowest (gender/noun classes).

4.3.3 Spatial effects

Drawing predictions from the model allows for a more detailed visualisation of differences in the areal effects estimated for individual features and feature categories. Given the estimated parameters and covariance, predictions can be drawn for a grid of new locations across the surface of Africa. This allows us to visualise a surface of high and low probability for the presence of features in space. These predictions are drawn only from the spatial component of the model. Phylogenetic relationships between languages cannot be used by the model to improve the accuracy of predictions, as arbitrary locations in space have no phylogenetic affiliation, but they should be filtered out implicitly during model fitting.

Predictions are first drawn for each feature individually. These can then be aggregated in order to produce visualisations of the overall expected structural similarity between languages within a geographic area. This section will start by presenting some of the visualisations for individual features, then discuss the aggregated spatial effects. The effect of the informed priors becomes obvious when examining these plots, as they predict the baseline probability of the presence or absence of a feature. Because some features have a high overall probability of being present, like causatives (shown in Figure 4.7), areal clusters are defined by their absence rather than presence.

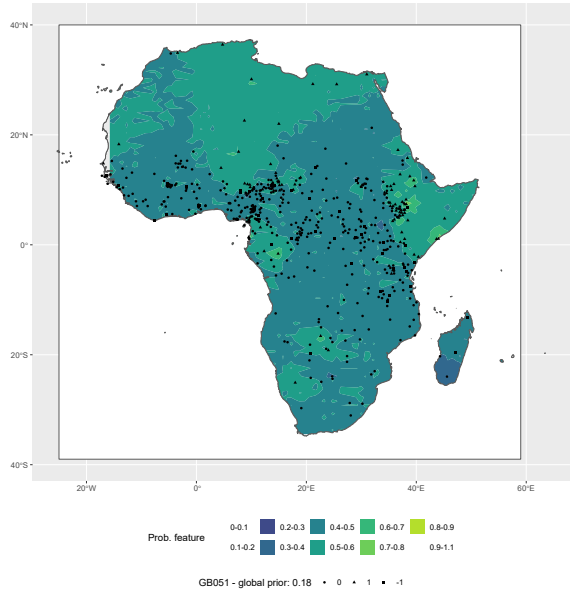
Figure 4.5 visualises the predictions drawn for the presence or absence of two different types of gender/noun class systems. As we can see, there are few areal clusters, as the model results suggest that noun classes are mainly clustered in language families. Where clusters are visible, these are very small. The informed prior sets the baseline for how probable the presence of a feature is; for animacy-based noun class systems, the prior is very low as this type of system is particularly rare outside Africa, and the map reflects this. Sex-based nominal classification systems are mainly found in a large area spanning the Afro-Asiatic spread zone in northern Africa and in the Horn of Africa, where Afro-Asiatic influence is prominent. There are three smaller clusters, one in Gabon which connects two Bantu languages Bubi and Lategе, which belong to different subgroups, and two in the Kalahari, although all the languages which share this feature belong to the Khoe-Kwadi lineage. Noun class systems in which sex is a factor in class assignment are uncommon in Bantu. Their presence in a few Bantu languages in Gabon could therefore be due to contact.

Figure 4.6 shows the predictions for verb-final constituent order in intransitive clauses and clause-final negation. Verb-finality has a relatively high overall probability of being present because of the informed prior, which indicates that this is a very common feature. The hotbeds of verb-finality are in the Horn of Africa, which is connected to Chad, and along the Western African coast around Côte d'Ivoire,

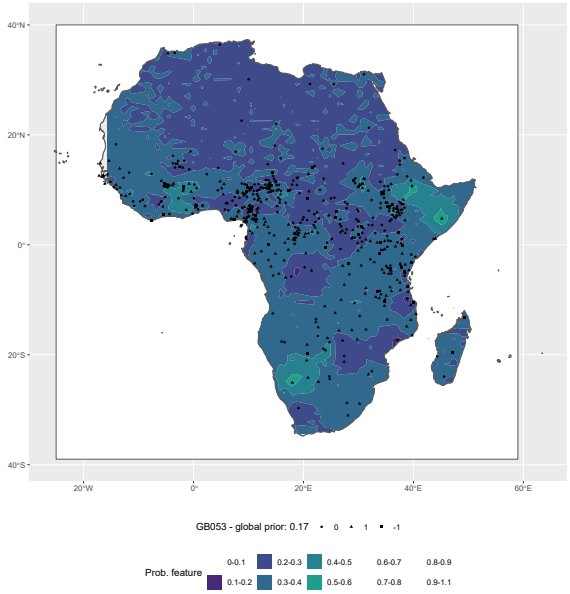
which is connected to Mali. Grambank restricts its feature on clause-final negation to standard negation, and makes no distinction based on optionality, as both obligatory and optional markers are included (Miestamo, 2008; Skirgård et al., 2023). Clause-final negation, as defined in Grambank, shows a spatially heterogenous pattern, with the most prominent hotbed between the border of the Central African Republic and South Sudan. Other, weaker hotbeds are found in Ethiopia and Kenya and in western Africa. A larger cluster connects sub-Saharan Central African Bantu languages in the DR Congo and Gabon. Cameroon has weak areal patterning for this feature, as southern Bantoid languages do not tend to have it. From this map, it looks as though clause-final negation is a highly areal feature that may have diffused across a large area from northern Central Africa. Either this process of diffusion over time resulted multiple discontinuous clusters, or the feature arose independently in several areas. The way the feature is clustered along central sub-Saharan Africa suggests that the feature may have spread along with some speakers of Bantu languages as they migrated southwards, but not in others. The fact that the likely origin point of the Bantu languages does not show a strong areal pattern for the feature could suggest that it was not a part of the earliest Bantu varieties (Idiatov, 2018; Koile et al., 2022). These areal patterns, although more diffuse, are not dissimilar to those found in Idiatov (2018), who provides a far more detailed discussion of this feature.

It can be useful to aggregate predictions across multiple features in order to get a better idea of the expected spatial similarity between languages given sets of features rather than individual features on their own (Guzmán Naranjo, Mertner, and Urban, 2024). In order to do this, I performed a principal components analysis (PCA) of the spatial predictions for different feature categories and depicted the values of the first principal component on a map, as this captures most of the variance. This should result in a measure of the overall predicted spatial similarity between languages. As we will see, some feature categories form small areal clusters, while others form larger clusters.

The aggregated predictions plot for all features, regardless of category, is shown in Figure 4.10, and Figures 4.8 and 4.9 show the aggregated predictions for gender/noun classes and verbal categories respectively. Two general patterns that emerge in these plots are, firstly, that the boundaries of areal clusters are highly diffuse, as predicted by Heine and Nurse (2007). The second general observation is that patterns of convergence appear to be predominantly local, but that these local patterns can combine into larger and more diffuse areas. Another recurring characteristic of these plots is that single languages are sometimes highlighted as having strong areal effects. This may happen when the values of those languages diverge from the values of the surrounding languages. In some cases, single-language spatial effects like this could

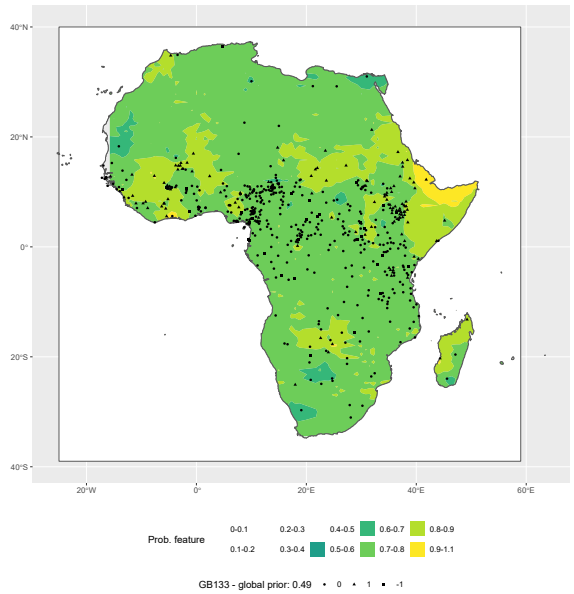


(A) Spatial effects for the presence of a gender/noun class system in which sex is a factor in class assignment.

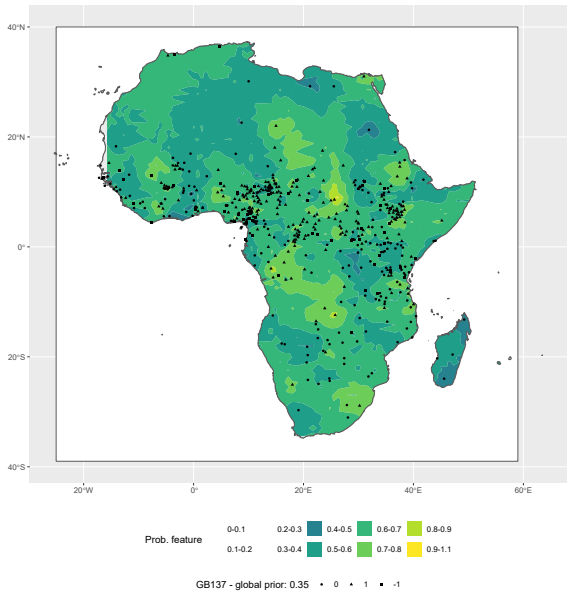


(B) Spatial effects for the presence of a gender/noun class system in which animacy is a factor in class assignment.

FIGURE 4.5: Spatial effects for the presence of sex-based and animacy-based gender/noun class systems. Yellow indicates a high probability of the feature being present in an area.

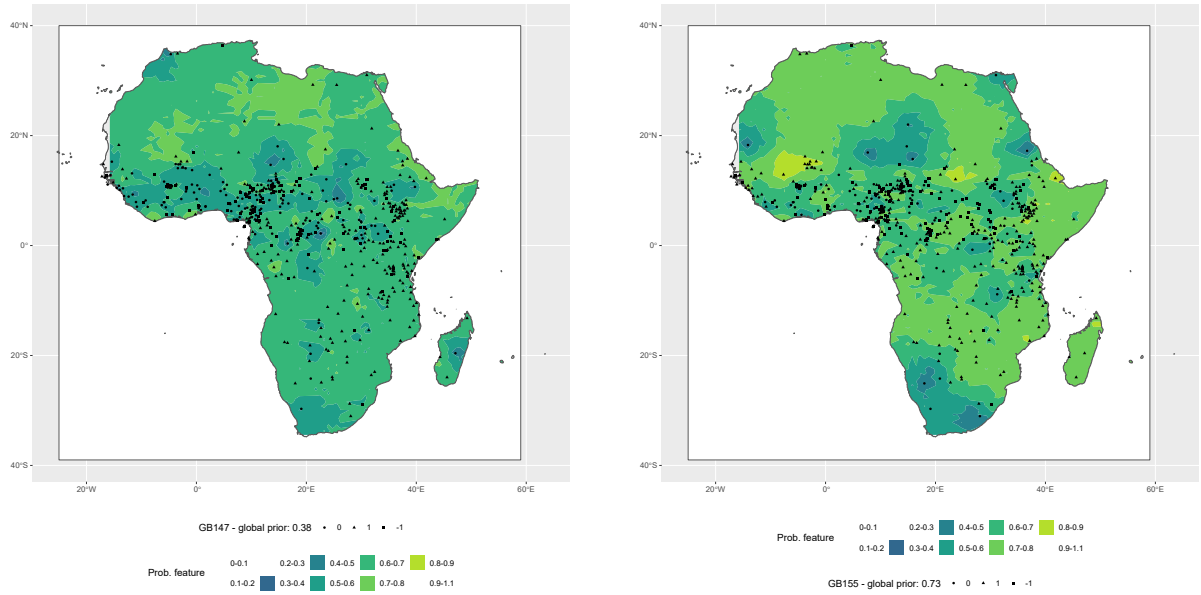


(A) Spatial effects for the presence of verb-final constituent order.



(B) Spatial effects for the presence of clause-final negation.

FIGURE 4.6: Spatial effects for the presence of verb-initial and verb-final order. Yellow indicates a high probability of the feature being present in an area.



(A) Spatial effects for the presence of a morphological passive marked on the lexical verb.

(B) Spatial effects for the presence of causatives formed by affixes or clitics on verbs.

FIGURE 4.7: Spatial effects for the presence of verb-initial and verb-final order. Yellow indicates a high probability of the feature being present in an area.

be due to random variation. The GP should filter out most of the variation which happens due to random chance, but it might be imperfect in some cases. These spatial effects could also represent areas of linguistic differentiation, for example a language which has a typological profile that is very dissimilar to its neighbouring languages.

For gender/noun class systems, the spatial clusters which emerge are generally fairly small (Figure 4.8). The largest is found in the Kalahari Basin and encompasses Nama (Khoe-Kwadi), Taa (Tuu), and, more peripherally, Herero (Bantu). The second largest cluster, which shows slightly weaker spatial effects than the Kalahari Basin, includes two languages in Somalia belonging to the Afro-Asiatic and Niger-Congo phyla, and it is diffusely connected to the languages of Ethiopia. For the smaller clusters, I will highlight those with the strongest spatial effects and which encompass more than one language. One of these encompasses a set of Mande and Atlantic languages in West Africa, groups which are known to be in sustained contact (Childs, 2010). The languages of the Nuba Mountains, which belong to at least two distinct phyla and many smaller families, also show a high level of spatial clustering within a relatively small geographic range. A third cluster is visible along the coast of Tanzania and stretches inland, and this one is unusual in that it only comprises Niger-Congo (mostly Bantu) languages. While family-internal contact in Bantu is well-attested,

it is unclear to what extent the model can disentangle convergence between closely related languages from common inheritance. Eastern Bantu languages do have some distinctive morphosyntactic features that are atypical for Bantu, including variations in their noun class systems (Edelsten et al., 2022). The model could attribute this family-internal variation to areal effects. The issue of family-internal convergence and its detection will be discussed further in Section 4.4.

The aggregated effects for verbal categories (Figure 4.9) show a different picture. Some of the same clusters are visible here as in the previous plot, so I will focus on the main differences. Firstly, spatial extent of the clusters is more variable than for gender/noun classes, and many are larger. There is a strong spatial effect around a language in the Horn of Africa, which could be a result of contact with the Arabian peninsula. This area is diffusely linked with a larger areal cluster of languages in Ethiopia, which stretches into Sudan. Small spatial clusters, sometimes linked, are found throughout western and central Africa, with stronger spatial clusters in the west where Atlantic and Mande languages meet. Additionally, a new, larger cluster is visible in this plot which encompasses a set of languages spoken in Niger and Chad, which include Saharan languages and varieties of Arabic. There is also a nearby cluster in South Sudan, the boundaries of which are difficult to separate from the Ethiopian zone. In the Kalahari, the cluster for this feature category includes Khoe-Kwadi, Tuu and southern Bantu languages.

When combining the predictions for all features, the picture that emerges has some key differences from the two previous plots. The main areas of convergence depicted include a northern African area comprising Arabic and Berber languages which seems to be somewhat connected to a Sudanic/Saharan area in Chad, which is part of the Sahel belt. The northern Arabic-Berber area is diffusely connected to the cluster of Saharan and Afro-Asiatic languages in Chad identified in Figure 4.9. A small areal cluster appears just south of that, on the border between the Sahel and the Cameroonian Bantu- and Bantoid-speaking area. The Horn of Africa shows spatial effects which diffuse into Ethiopia and Somalia. In southern Africa, the Kalahari Basin appears as a zone of convergence with its core around Khoe-Kwadi and Tuu languages. In West Africa and the central transition sphere of the Sahel, small clusters connect Atlantic and Mande languages. Although these clusters are small, their boundaries are so diffuse that it is difficult to delineate any particular area across western and eastern central Africa. These small, local convergent clusters are indicative of the kind of contact which Güldemann (2018b) characterised as typical of the central transition sphere. As a whole, they seem to possibly connect a larger area. Lastly, Madagascan languages also form an areal cluster, and some spatial effects from these languages can be observed along the coast of Mozambique or

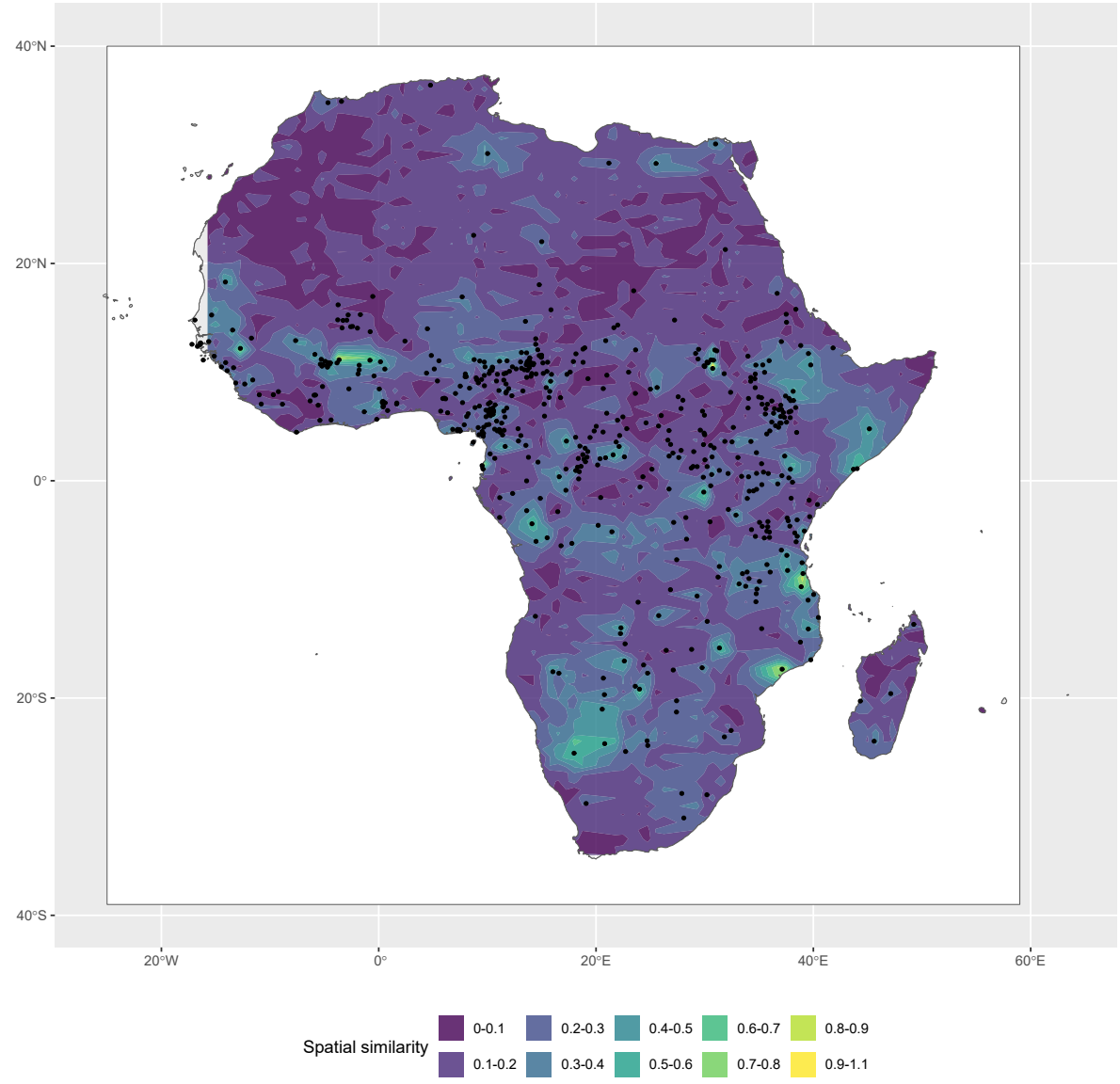


FIGURE 4.8: Aggregated spatial predictions for gender/noun classes.

Tanzania. This could be due to the fact that topographic distances do not account for the presence of water. Morphosyntactic similarities between Malagasy and Bantu languages have been attested, which could either be due to (early or more recent) contact with coastal Bantu languages or Bantu substrate effects (see Dahl (1988) in support of a substratum explanation, and Adelaar (2010) for a discussion of Bantu contact effects on Malagasy morphosyntax). There are also some weak spatial effects across central Africa which culminate in a smaller area around South Sudan, where Nilo-Saharan Sudanic languages meet Bantu languages.

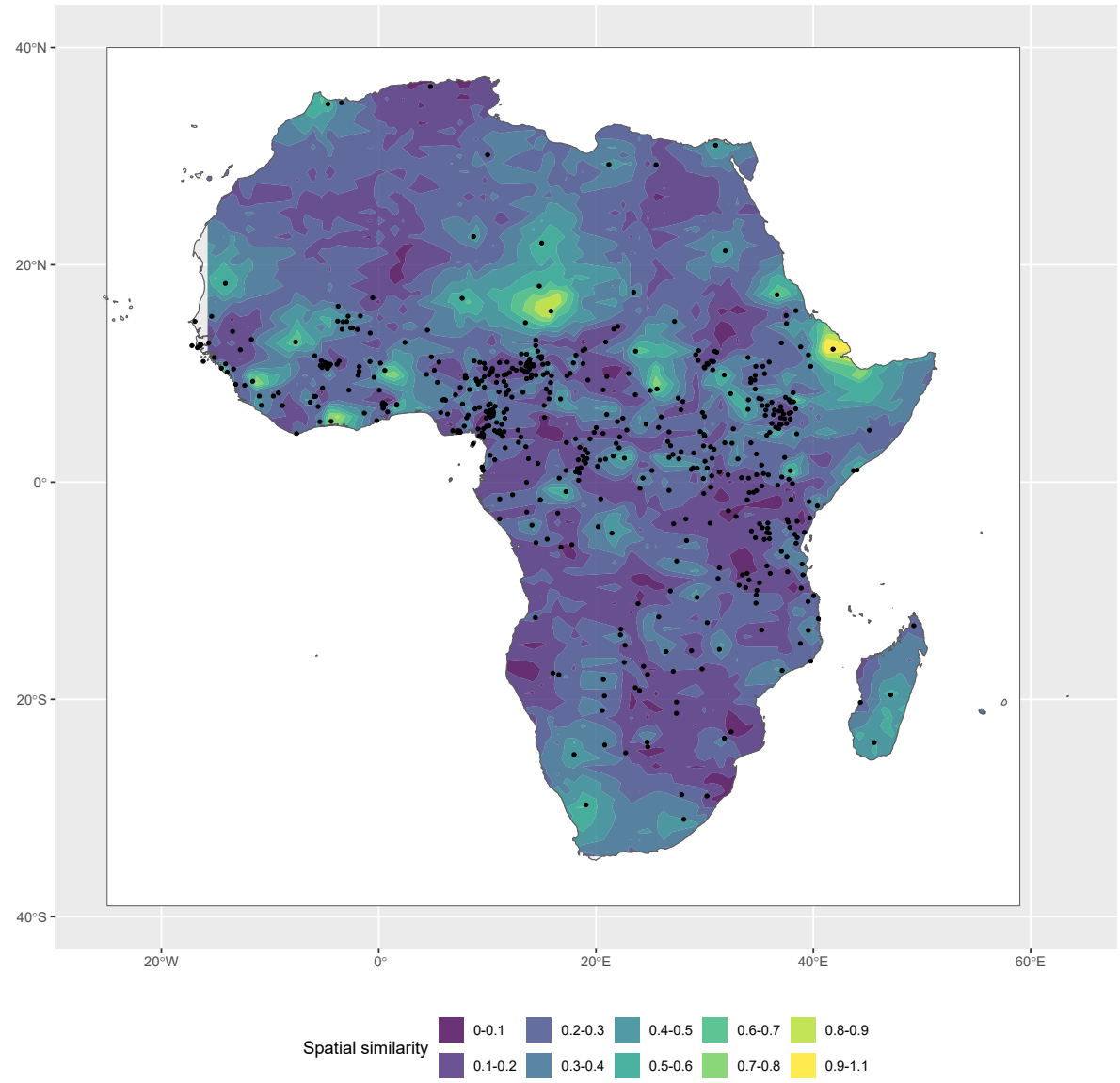


FIGURE 4.9: Aggregated spatial predictions for bound verbal categories.

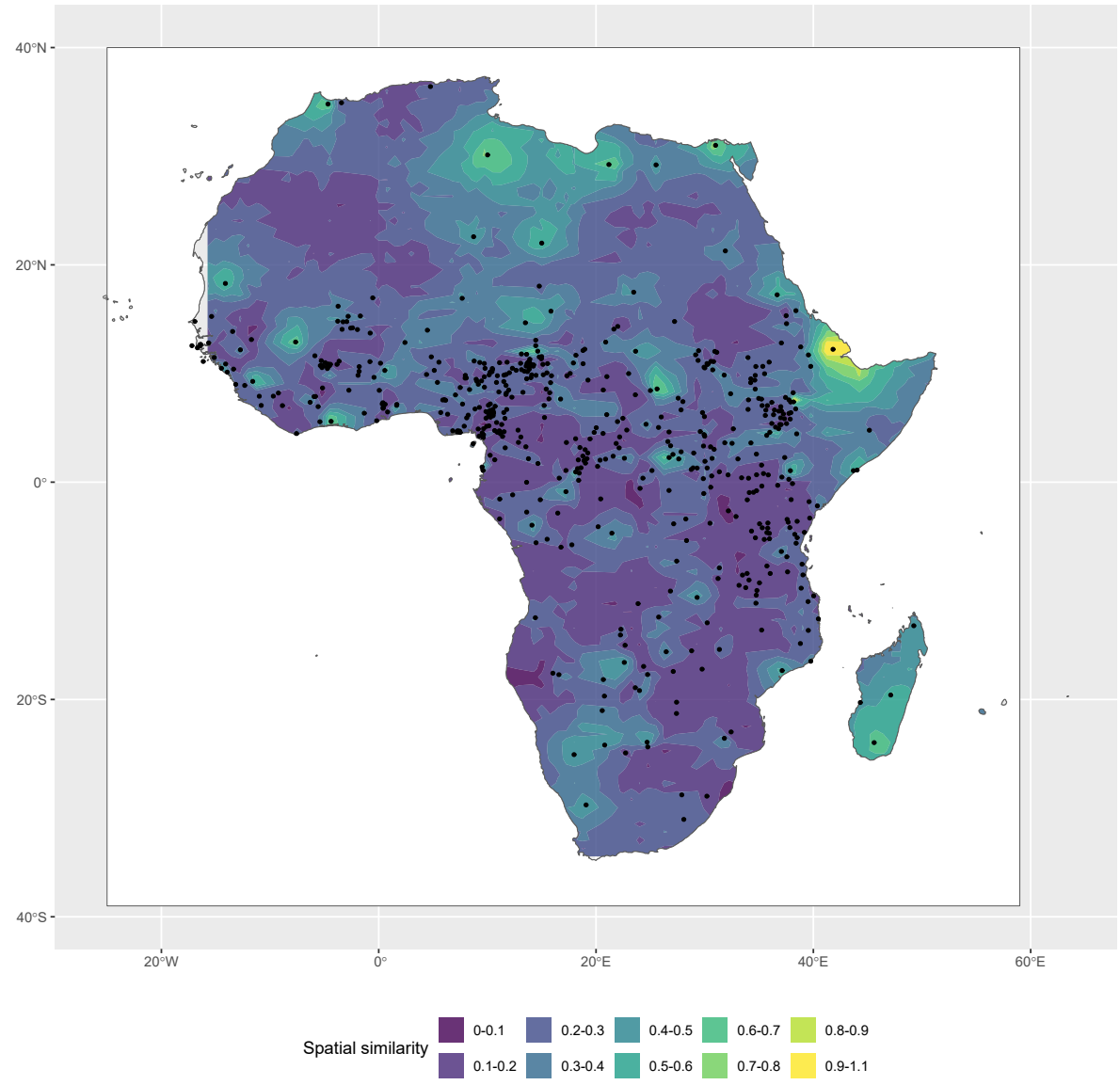
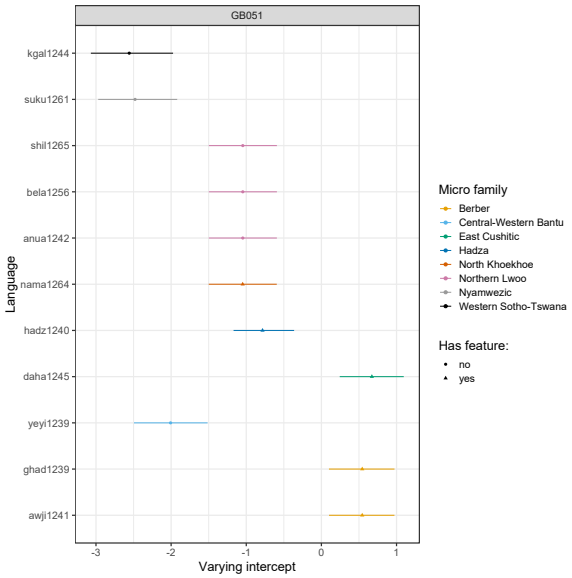
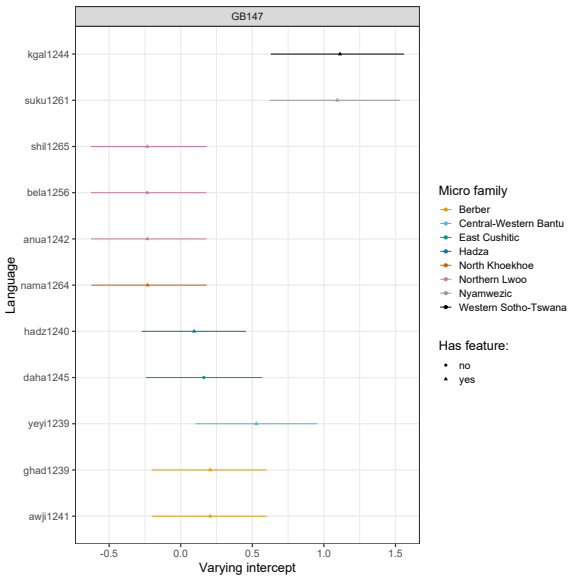


FIGURE 4.10: Aggregated spatial predictions for all features.



(A) Phylogenetic intercepts for the presence of a gender/noun class system in which sex is a factor in class assignment.



(B) Spatial effects for the presence of variations in marking strategies caused by TAM distinctions.

FIGURE 4.11: Phylogenetic intercepts for a gender/noun class feature and a TAM feature, coloured by Glottolog family.

4.3.4 Phylogenetic effects

There is a considerable degree of variation in the uncertainty of the phylogenetic effects estimated by the model for each feature. Gender/noun class systems have a lower level of phylogenetic uncertainty for the families shown here than variations in marking caused by TAM distinctions, as shown in 4.11. Higher uncertainty around the phylogenetic intercepts points to a greater degree of heterogeneity within families, which could be caused by contact-induced interference (although of course that is not the only possible source of phylogenetic uncertainty). Overall, this aligns with the finding that the phylogenetic model was less able to reliably predict TAM features than gender/noun class features. More visualisations of the phylogenetic effects for different features will be shown in Appendix B.

4.4 Discussion

In this section, I will discuss and contextualise the results in light of the literature. First, I will discuss the estimated overall variation between the diffusion of structural features belonging to different morphosyntactic categories. Then I will discuss the areal patterns in more detail before finishing the section with some general considerations related to the data and the methodology.

4.4.1 Diffusibility and borrowability

As I discussed in Chapter 2, one of the potential issues with using latent GPs in a model is that they can be challenging to interpret. This study is no different, but I have tried to present and visualise the results in as many ways as possible and to interpret them in relation to each other. In this section, I will discuss the results on diffusibility with reference to the spatial and phylogenetic effects wherever they become relevant. As mentioned earlier, I will not be conflating diffusibility (which is itself a shorthand for potential spatial extent) with borrowability. However, I will discuss the relationship between them, as well as what diffusibility and patterns of diffusion can tell us in their own right.

Unsurprisingly, the least diffusible set of features was gender/noun class systems; they are typically considered highly stable within lineages in Africa (Creissels et al., 2008). Although Grambank does not include information about the number of noun classes, but about their semantic underpinnings, like sex-based or animacy-based distinctions, these semantic features are also related to the number of noun classes, as sex-based systems often have a binary distinction between the two genders, as is common in Afro-Asiatic languages. When Bantu languages experience a loss or reduction of noun classes, this is often (but not always) attributed to contact, usually with languages that lack Bantu-like noun class systems (Van de Velde, 2019). The similarity in the estimated areal range of nominal number distinctions and gender/noun classes is likely due to the prevalence of noun class systems which are intertwined with number in Africa. The model should pick up on any correlations between them, such that if the presence of a number category is explained by the presence of a noun class category (or vice versa), less of the variance will be attributed to the spatial effects. Overall, these two categories were found to be the least diffusible ones. This can be generalised, in the context of the features included here, as a scale in which nominal features are less diffusible than the ones relating to the verbal domain.

The results of this study place bound verbal categories, including applicatives, causatives, and reciprocals, on the low end of the middle of the diffusibility hierarchy for the set of structural features tested here. Bound markers are traditionally considered some of the least borrowable features (Thomason and Kaufman, 1988), but it is unclear whether the same holds for the conceptual categories which can be encoded on the verb. Contact-based changes to systems of verbal marking can involve either the addition or loss of distinctions, but does not necessarily involve the transfer of forms. Categories might transfer more easily than forms between diverse families with different typological profiles due to structural incompatibilities. To build on an example mentioned earlier in this chapter, bilingualism in neighbouring Bantu languages is very common among Luo speakers, and one can imagine how

adverbs of time expressing concepts like 'the day before yesterday' could become grammaticalised through more frequent usage based on the distinctions found in the Bantu system (Heine and Nurse, 2007, pp. 4–5).

TAM markers (which do not have to be bound) were found to be more diffusible than bound verbal categories, which suggests that boundedness does matter for diffusibility. However, the uncertainty around the estimated diffusibility of tense/aspect suggests that is a high level of heterogeneity in this domain; perhaps some features are highly diffusible while others are not. Word order was found to have the largest range along with tense/aspect, and should perhaps be considered the 'most diffusible' category alongside it, given the high level of uncertainty. This is rather unsurprising, too, as previous researchers have argued that word order is highly areal in Africa (Dimmendaal, 2008a; Heine, 1976). Word order and TAM were also the two features with the lowest predictability based on the phylogeny alone, which aligns with the result that these types of features are more areal. However, this result is more robust for word order due to the lower estimated uncertainty.

4.4.2 Areal patterns

The identification of areal patterns is a secondary goal of this chapter, as I recognise that the delineation of linguistic areas typically relies on a mixture of morphosyntactic, phonological, and lexical data (see e.g. Güldemann (2018b, p. 481) for a list of proposed features diagnostic of the Macro-Sudan belt). Nonetheless, structural convergence could provide a window into past contact relations or past linguistic areas. Most linguists agree that convergence tends to happen in contact situations which are sustained, intensive, and in which interaction occurs regularly over a long period of time (Thomason and Kaufman, 1988). Structural features appear to change more slowly than the lexicon or the forms used to express morphosyntactic categories (Nichols, 2003), although this might not hold for all parts of the lexicon. Because of this, it may be worth examining the areal patterns found here in light of proposed linguistic areas like Chad-Ethiopia (Heine, 1975, 1976) or the Kalahari Basin (Güldemann, 1998).

The results of this study suggest that accretion zones are not necessarily convergence zones, and may in fact show fewer convergence effects than other areas. Güldemann (2018b) lists the Rift Valley in Tanzania as an accretion zone that forms a part of a larger area which he calls the Southern Gregory Rift, but it is not typically considered a linguistic area, even though its languages have some convergent features which cross genealogical boundaries (Kießling, Mous, and Nurse, 2008). It could be that these are simply different features to the ones included here, like ejective stops,

and that structural convergence cannot be detected at the level of cross-linguistically generalisable morphosyntactic features, like the ones which are in Grambank. The lack of detected convergence at this level could also be related to the same social or geographic mechanisms which maintained the diversity in these areas over time, which may include social attitudes to contact or the presence of topographic barriers (Dimmendaal, 2021). It could also be related to the structural dissimilarity between the languages themselves, which might make the incorporation of foreign structures more challenging, although plenty of counter-examples suggest that borrowing can happen even when the structures of the languages involved are incompatible, as in the case of Luo (Nilotic) borrowing tense and aspect categories from nearby Bantu languages (Niger-Congo), but incorporating them into the language as grammaticalised native adverbs of time which can be used as prefixes or clitics. This is in contrast to how tense/aspect markers function in Bantu, where they are clearly affixal and follow the verbal subject prefix rather than preceding it, as in Luo (Heine and Nurse, 2007, p. 5). The most likely explanation is that the data used here did not allow for the detection of the specific instances of convergence in this area, and it is still worth noting that areal convergence could not be detected for these specific features. This stands in contrast to the consonant inventories tested in Guzmán Naranjo and Mertner (2022), some of which showed clear convergence in eastern Africa, including the Rift.

It is interesting to examine the differences in the inferred areal patterns for the different sets of structural features included in this study. The languages of the Nuba Mountains, for example, show areal convergence exclusively in the domain of gender and noun class systems. The Nuba Mountains are an accretion zone and thus characterised by a high level of diversity, including Niger-Congo languages belonging to the Heiban and Talodi families, as well as Nilo-Saharan (Sudanic) and Kadu languages (Manfredi, 2022; Nichols, 1992). The role of diffusion in the area has been contested, with some observing similarities in the lexical and nominal domain that could be attributed to diffusion (Manfredi, 2022) and others arguing that the evidence for grammatical diffusion in the area is very limited (Dimmendaal, 2021). Based on the results shown in this chapter, convergence in this area may have occurred while being restricted to a very specific domain. The diffusion of noun class categories could have happened alongside the diffusion of lexical material.

There are some weak spatial effects across the central Bantu spread zone. They are likely weak because these languages are closely related, which will be reflected in the phylogenetic tree. Where the model infers family-internal areal patterns, this suggests that the languages are typologically unusual compared to other members of the family in ways that cannot be explained by the informed priors. This unusualness

could, as a very speculative hypothesis, result from past contact with speakers of languages which are no longer spoken in the area and which may have formed part of the prehistoric version of the Kalahari Basin, although prior attempts to detect such substrate effects have had limited success (Bostoen, 2020). Another explanation could be that the well-attested patterns of intensive language contact between speakers of different Bantu languages in central and southern Africa, as well as processes of diversification, have acted as triggers for structural changes which eventually spread.

4.4.3 Considerations and limitations

In general, it is not straightforward to map the inferred spatial clusters of categories of linguistic features to established or hypothesised linguistic macro-areas in Africa. Many of the local clusters of convergence align with patterns which have been discussed in the literature. On the other hand, some zones of contact between diverse families are conspicuously absent from the maps, including the Macro-Sudan belt. In analysing these results, I acknowledge that the selection of data sources and features is crucial, and that the absence of a convergent zone in these results only means that such a zone could not be detected based on this specific dataset using this specific model. Because Grambank is a large cross-linguistic database, it does not contain many of the cross-linguistically rare morphosyntactic features which are diagnostic of some of the proposed linguistic areas in Africa. Additionally, phonological features have been key to identifying contact areas in Africa. For example, labial-velar stops and advanced tongue root harmony helped first characterise the Macro-Sudan belt, clicks were essential in identifying the Kalahari Basin and contact between Khoisan and Bantu languages more generally, and the distribution of ejective stops, lateral fricatives, and affricates have provided evidence for contact between languages in the Rift (Clements and Rialland, 2008). As this study excludes phonological variables, it may not be surprising that the areas found do not map neatly onto linguistic areas in Africa in a general sense. Thus, when discussing the implications of these results, I note that they reflect possible patterns of morphosyntactic convergence only for the features included in this study.

A related issue is that convergence does not always result in a one-to-one mapping between typological features. Sometimes, contact triggers the reorganisation of some structural aspect of the recipient language in a way that does not match the structure of the source language. Because this method only detects congruences between binary variables, it cannot detect this kind of contact-induced change.

It should be noted that languages for which some feature values were missing were included in this model, and the missing values were imputed during model

fitting based on all the model components. Because of this, asymmetries in the distribution of missing values could influence model results and cause uncertainty or even weak spurious effects. For example, if a particular feature is missing with a higher frequency in a specific area or language family, the model could lack the necessary information to impute the right values. This should not be the case (as demonstrated by Guzmán Naranjo, Mertner, and Urban (2024), this model works even when data is very sparse, and in this study I restricted the selected features to ones which have relatively high coverage in Grambank). However, it is always a possibility for data sparsity to influence results; on a broader scale, linguists are working with a very small amount of the world's possible linguistic variation, based on how many extant languages exist now, how many have already died, and how many more have limited or absent documentation (Evans, 2010).

A methodological consideration is that it is difficult to extract information about how much variance (or importance) the model assigns to its different components. It would be useful to be able to quantify the proportion of variance assigned to the phylogenetic part, the spatial part, and the inter-feature correlations, as this would give us more insight into the areality and stability of these features, as well as highlighting potential issues with the model. For example, for the combination of all features, the model detects fewer large areas of convergence than would be expected based on the literature that characterises some of the included features, like word order, as a highly areal phenomenon in Africa (Heine and Nurse, 2007). This could be related to the fact that the word order features included here are highly correlated, and so those correlations might be sufficient, in the abstracted world of the model, to explain the distribution of word order in Africa. A large portion of the variance for those features might thus be assigned to the inter-feature correlations.

In some cases, the model might improperly assign more or less variance to phylogenetic effects, as the model has no information about how common a given feature is within families. When there is enough data on a family, the model can estimate this from the data, but when only a few languages from a family are included, the model may assign more explanatory power to common inheritance than an expert on that family would. Of course, all of this also depends on having a good language tree, and expert opinions differ on what this looks like for African languages (Güldemann, 2018a). Essentially, the different components of the model, including inter-feature correlations, informative priors, phylogenetic and spatial effects, may 'soak up' the variance in unpredictable ways. Ensuring that the model has as much prior information as possible is one way to mitigate this issue, as well as developing ways to extract the variance proportions from the model.

4.5 Conclusion

Different types of morphosyntactic features show distinct patterns of areal diffusion, which can be detected using the latest version of the *multivAreate* model presented here (Guzmán Naranjo and Mertner, 2022; Guzmán Naranjo, Mertner, and Urban, 2024). Some features diffuse over both large and small areas, while others appear to have a more restricted geographic range. Gender/noun class systems are the category of features with the smallest estimated range, followed by nominal number (which is often heavily intertwined with noun class systems in Africa), verbal categories, and word order and tense/aspect marking, the latter two of which have a relatively similar range. These findings align with previous literature on the areality of African morphosyntactic features, which hold that gender/noun class systems pattern according to shared ancestry, while many word order features pattern areally and have been used to identify potential linguistic areas in Africa (Greenberg, 1983; Güldemann, 2018b; Heine, 1976; Heine and Nurse, 2007; Idiatov, 2018).

The results presented in this chapter suggest that structural convergence of bundles of features predominantly happens locally. Individual features, like clause-final negation, can have a far greater areal spread. Indeed, some linguistic areas have been proposed on the basis of a small number of features, and at other times, linguistic areas are made up of several smaller areas with their own set of convergent features (Campbell, 2017). The convergence of sets of structural features across larger areas appears to be a result of those structures being transferred from one local contact situation to another, for example through a language being involved in social interactions with multiple language groups at the same time, which may or may not also be in contact with each other. In this way, groups form links between other groups which could result in widespread convergence over time, a situation which mirrors the wave model of language change and the diffusion of innovations (Heggarty, Maguire, and McMahon, 2010).

Future work could compare the diffusibility of morphosyntactic and phonological features. Particularly in Africa, phonological features have been crucial in the identification of areas of historic and prehistoric language contact, such as click consonants in southern Africa (Güldemann, 1998) or labial-velar consonants in western and central Africa (Clements and Rialland, 2008; Güldemann, 2008). It may also be valuable to compare the diffusibility of concept classes using cognate datasets such as that compiled for Bantu languages by Grollemund et al. (2015), or lexical datasets such as RefLex Segerer and Flavier (2011-2021). The model could be used to test the diffusibility of specific words, lexical domains, or concepts, and to detect potential loanwords in cognate datasets, as these can cause issues for language classification

using quantitative methods (Jäger, 2013; Neureiter et al., 2022). Additionally, areal patterns in concept classes and shared colexifications could reveal cultural connections between linguistic groups (Segerer and Vanhove, 2022).

Chapter 5

Geographic bias in the distribution of missing data

The issue of data sparsity is unavoidable for any researcher working in quantitative linguistics. It is especially relevant for linguistic typology, where researchers rely on cross-linguistic data on structural aspects of language to allow comparison between a large number of diverse languages. The interpretation of the results of these studies relies on the assumption that the sample of languages used is globally representative and unbiased. However, we know that the documentation of languages and their annotation in databases are driven by a number of non-random factors, including academic interest in specific languages, linguistic features, or families. Historical events have also played a major role in determining which languages have survived and which ones have been documented. However, so far, no study has demonstrated the presence of geographic biases in global patterns of data sparsity using a spatial model. Understanding the processes which shape the distribution of missing data can help inform statistical methods for reducing geographic bias in future studies. Using the World Atlas of Language Structures (Dryer and Haspelmath, 2022; Dryer and Haspelmath, 2013) as a case study, in this chapter I will demonstrate that the distribution of missing data in linguistic typology is shaped by both geographic and phylogenetic biases, and that the geographic bias appears to be stronger. The implications of this for future work includes emphasising the importance of controlling for geography in statistical studies.

5.1 Introduction

The study of linguistic typology, particularly when the goal is to uncover the universal tendencies or cognitive biases which have shaped the evolution of language, relies on cross-linguistically diverse data (Greenberg, 1963). Typological data on diverse languages has been used to shed light on linguistic prehistory, language contact, the

relationship between language and the environment, the stability of linguistic features over time, and interactions between human cognition, genetics, and language (including but not limited to the studies by Bentz et al. (2018), Dediu and Cysouw (2013), Dediu et al. (2021), Josserand et al. (2021), Nichols (1992), and Urban and Moran (2021)). The World Atlas of Language Structures, usually referred to as WALS, is an incredible resource for conducting this kind of work. In its original form as a published book, it contained information on structural aspects of 142 languages based on expert judgments (Haspelmath et al., 2005). In its more recent form as an online resource (Dryer and Haspelmath, 2013), it has since provided the data for a rich body of work in quantitative linguistic typology (like Atkinson (2011), Bentz et al. (2018), Bickel (2011, 2013), Dunn et al. (2011), and Jäger and Wahle (2021) and many more). As such, this database has been highly influential.

One of the main challenges associated with quantitative work in typology is data sparsity. Although there are between 7,000 and 8,000 extant languages in the world, only 2,000 of these are represented in WALS, and many of the included structural features are only documented for 1-15% of the languages included in the database[1] (Dryer and Haspelmath, 2022; Murawaki and Yamauchi, 2018). The only way to remedy this permanently is to document more languages. Additionally, it is important to ensure that the available information about documented languages is included in existing typological databases (which is something the creators of the Grambank database are currently working on (Skirgård et al., 2023)). Although documenting and describing more languages is the best way to solve the issue, statistical methods can be used as a first step to reduce the issues, like misleading results, which can be caused by data sparsity. Such methods depend upon an understanding of the generative process behind the distribution of missing data (McElreath, 2020).

Perhaps surprisingly, sparse data does not always pose an issue for statistical inference. Data can be missing from a dataset for many reasons, some of which may not cause any problems. For example, environmental data can be missing at certain locations or times due to random equipment failures or weather conditions. In this example, equipment failure would create a random pattern of missingness in the data. If we run a statistical model with this kind of data, where some values are missing at random, we would not expect to see a change in the results if we ran a separate model in which those missing values were included. On the other hand, data can also be missing in systematic ways, i.e., non-randomly. To build on the example above, environmental data could be missing because certain temperatures cause the equipment to fail. Because of this, there would be a disproportionate amount of missing values from locations with more extreme weather conditions, which means

[1] See the Supplementary Materials for a plot illustrating this.

that any statistical methods applied to such data would be ill-informed about what happens at locations with those temperatures. If information about the temperatures at these various locations is available, this can be included in the model as a control in order to prevent biased results. However, in many real-world cases, patterns of data sparsity are governed by non-random processes which are unknown to us or which have not been measured, and these cannot be included in a study. This kind of data is a common source of biased inference, as explained in Chapter 14 of McElreath (2020).

Language documentation is governed by several non-random processes which are difficult to measure. These include (but are not limited to) historical events, racial and cultural prejudices, geography, academic interest, and funding. This is by no means an exhaustive list, and all of these may be interrelated. Interest in producing a description of a language may increase if that language has a feature which is considered interesting or unusual from a linguistic standpoint, or if the language is an isolate or one of few surviving members of a language family. Accessibility might also play a role, with certain types of terrain and climate posing a potential barrier to travel and therefore research. On the other hand, certain kinds of climate and terrain also posed a barrier to colonisers, and this could be one of the reasons why some areas characterised by 'difficult terrain', like rainforests or mountains, are characterised by high levels of linguistic diversity (Nichols, 1992; Nunn and Puga, 2012). Thus, linguistic data may be missing for many reasons, but we can safely say that the process behind the current coverage of typological databases is not random. Additionally, large-scale typological data of the kind found in WALS (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023) requires not only a high-quality grammatical description of a language (as opposed to a word list), it also requires the author of the grammar to pay attention to cross-linguistic comparability. For example, some grammatical features are coded as missing in databases because the author does not explicitly mention the feature in the grammar, perhaps because the feature is not considered interesting or relevant for that language or family. This could cause another kind of non-random pattern in missing data which is driven by interest in certain features; a missing value could indicate that the feature is absent in a language, but it could also indicate that the feature is not considered interesting enough to mention for that particular language. All of this can cause issues for statistical inference. Using all the data we have without accounting for the missing data is very likely to induce bias (McElreath, 2020). If there is a geographic or lineage-specific pattern of missingness in the data, this will induce geographic or phylogenetic bias during statistical inference. Using a Bayesian approach is generally recommended as a way to mitigate this, as Bayesian models quantify model uncertainty, which can

indicate when results are not robust; however, the presence of biases in the data (or missing data) can obscure that uncertainty (Guzmán Naranjo and Becker, 2021).

It is unclear whether data sparsity is shaped primarily by geography, language family membership, or both. This study aims to uncover the nature of the bias that we are dealing with when doing large-scale typological studies in linguistics, particularly using WALS (Dryer and Haspelmath, 2022; Dryer and Haspelmath, 2013). Understanding the processes which generated the pattern of missing data can help inform the techniques used to reduce the effect of this bias in statistical inference (McElreath, 2020). While stratified sampling attempts to resolve this problem (Bakker, 2010; Dryer, 1989; Miestamo, Bakker, and Arppe, 2016), this unfortunately relies on discarding large amounts of data, which should be avoided whenever possible since the data we have already represents only a small sample of the world's possible languages (Nettle and Romaine, 2000). Alternative approaches include controlling for geographic proximity and ancestral relationships between languages, and these allow us to keep more of the available data (Becker, Guzmán Naranjo, and Ochs, 2023; Guzmán Naranjo and Becker, 2021; Guzmán Naranjo and Mertner, 2022). Another way to mitigate the effects of bias induced by missing data is imputation during model fitting, which is usually considered the best way to handle missing data in a statistical model (Guzmán Naranjo, Mertner, and Urban, 2024; McElreath, 2020).

The goal of this chapter is to investigate the distribution of missing data in linguistic typology, using WALS as a case study, in order to understand some of the underlying data-generating processes which have influenced the database that has formed the basis of so many previous studies. Not including Grambank, a newer database, is a limitation of this study. Some preliminary work on the spatial distribution of missing data in Grambank will be presented in Appendix B, but a full analysis of this will be left to future work.

5.2 Data

WALS consists of the following components: *languages* and *features*, or different structural aspects of language (Dryer and Haspelmath, 2022). It includes data from a variety of sources, such as grammars, and a large number of features from different linguistic domains. Examples of WALS features include the presence/absence of definite articles, the size of a language's consonant inventory, and the order of subject, object and verb. Some of these features have better coverage than others, and the discrepancies between them can be large. Languages can either have a value for a given feature, or they can be coded as missing. A language may have a value for the size of its consonant inventory, but lack a value indicating whether or not it allows

reduplication. In this way, different languages are most often annotated for different sets of features.

In this study, I use missingness as the response variable. That is, for all the languages in WALS, I count how many of the 192 features in the database have a value for that language. Some languages may have very low coverage, while others have almost complete coverage. In order to get a more complete picture of missingness, I also included all the languages in Glottolog (Hammarström et al., 2023). All the languages which are in Glottolog but not in WALS are coded as 0 (missing). This means that the data includes a very large number of zeros (for Glottolog languages not in WALS) and some count data (for the number of features annotated per language in WALS).

For the sake of completeness, all the languages in Glottolog were used; this means I did not exclude contact languages, signed languages, isolates, or unclassified languages (Hammarström et al., 2023). There is no need to exclude these, since the method can handle languages which resist treelike classification by assigning them an independent phylogenetic intercept. This allows them to vary freely without being constrained by the structure of the tree. Figure 5.1 shows all the languages included in the study, coloured by whether they are present in WALS or missing (Dryer and Haspelmath, 2022).

As we can see in Figure 5.2, the count data is massively overdispersed (left plot) and, overall, far more languages are missing from WALS than present (right plot). For languages which are in WALS, the most common number of annotated features is between 1 and 5. Relatively few languages are annotated for more than 50 or so features.

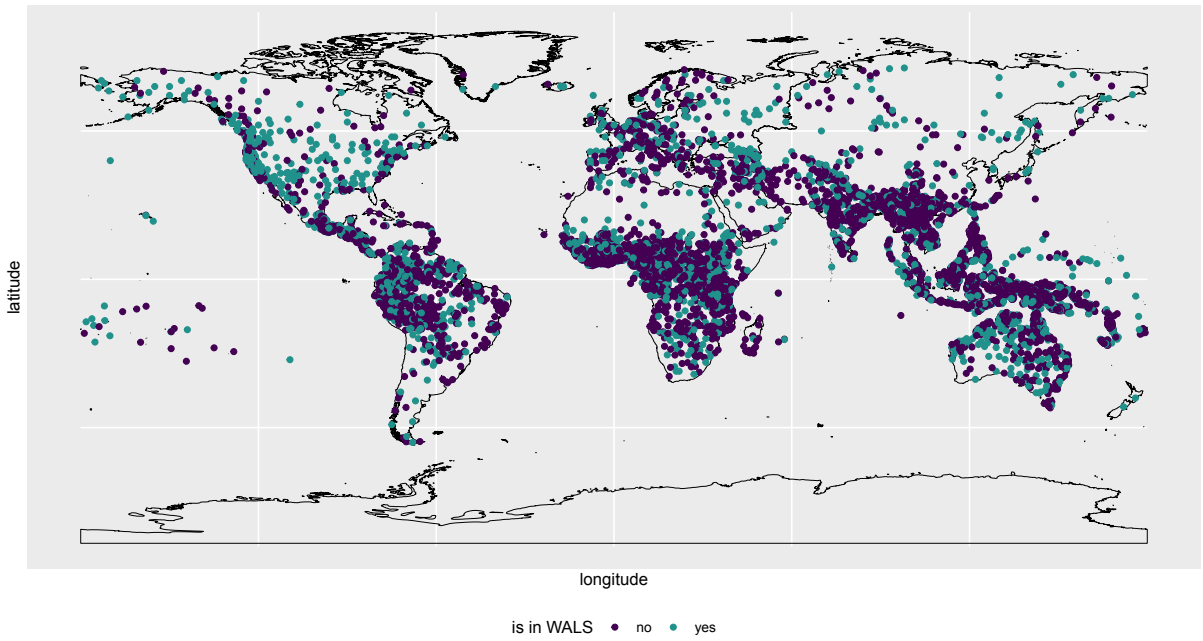


FIGURE 5.1: The languages in the study.

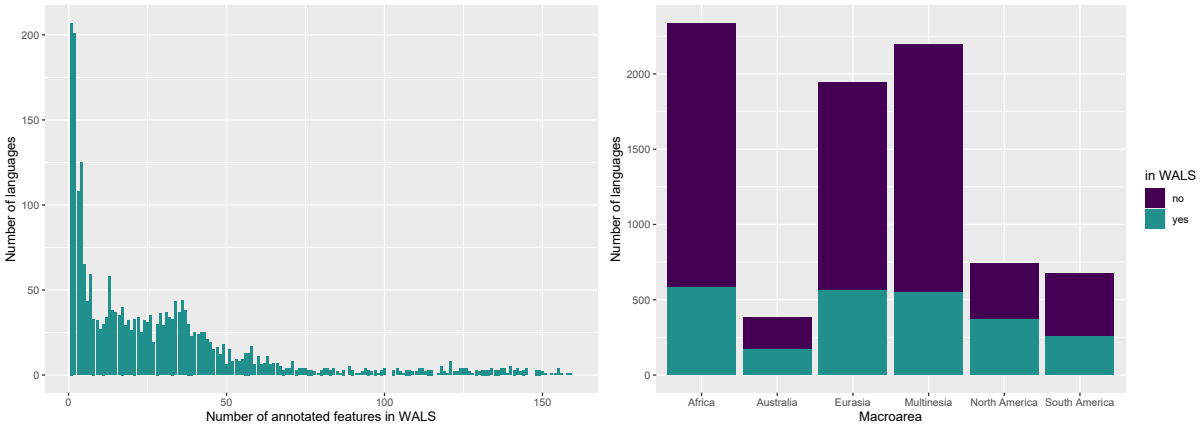


FIGURE 5.2: Count data for all the languages in WALS (left) and the number of languages which are present/absent in WALS for each macroarea (right).

5.3 Method

The method presented here builds on techniques proposed by Guzmán Naranjo and Becker, 2021 and Guzmán Naranjo and Mertner, 2022: using a Gaussian process to estimate areal effects, and phylogenetic regression to control for the effect of shared

ancestry. However, because there are too many data points to fit an exact GP as in the papers above, I use an approximate GP (Riutort-Mayol et al., 2022). This study is the first to develop a modelling approach for linguistic typology using a latent approximate GP. All models were fitted using R (R Core Team, 2020) and Stan (Carpenter et al., 2017)[2].

5.3.1 The hierarchical hurdle model: An overview

To model the number of features coded for each language in WALS, I use a negative binomial model. Negative binomial models are characterised by their ability to model highly skewed count data, which fits this case very well. As shown in Figure 5.2, WALS has a large number of languages which are coded for only a few features (between 1 and 20), and very few languages which are annotated for 50 features or more, which means that the data is highly skewed towards lower counts.

The factors which influence the probability of a language being in WALS are modelled as a separate process from the probability of a language which is already in WALS being annotated for a certain number of features. The model therefore has two separate components: 1) the negative binominal component, which models the number of features coded in WALS for those languages which are in it, and 2) a hurdle component, which models the probability of a language being missing from WALS.

Additionally, the model is hierarchical and has a varying intercept for each macroarea. This will allow us to find out whether, for example, languages in Africa generally have a higher probability of being in WALS than languages in Australia.

5.3.2 Approximate Gaussian processes

For the spatial component of the model, I use approximate GPs following the approach outlined by Riutort-Mayol et al., 2022 (their approach is based on mathematical background developed by Solin and Särkkä, 2020). In practical terms, this makes it possible to apply a GP, which is computationally intractable for large datasets[3]. Since Glottolog contains around 8,000 languages, there are as many data points in this sample, which means that Bayesian inference using an exact GP would be impossible (Riutort-Mayol et al., 2022).

The approximate GP approach uses the longitude and latitude locations of languages to calculate the Euclidean distance between them. As is the case generally for GP models, languages which are close together are expected to be more similar than languages which are far apart. The point at which the similarity between languages

[2]A link to the code can be found in Appendix A.

[3]An exact Gaussian process for 8,000 data points will never finish running on consumer hardware.

based on geographic distance begins to decay is estimated as the horizontal scale parameter of the GP. For example, in a previous study using exact GPs with Euclidean distances, languages in Africa were found to be more likely to share phonemes within a range of around 320km (Guzmán Naranjo and Mertner, 2022). Beyond this range, the expected similarity between the languages' phoneme inventories would be expected to decay rapidly until it becomes effectively zero. This is described in more detail in Chapter 2 and Chapter 4 of this thesis.

The type of approximation used here is called a *Hilbert Space approximate Gaussian process* (Riutort-Mayol et al., 2022), which I will refer to more simply as an approximate GP. The approach to approximating a GP developed by Solin and Särkkä, 2020 and Riutort-Mayol et al., 2022 relies on defining a number of basis functions, M, and a boundary condition, c. The number of basis functions M required depends on the degree of non-linearity of the function[4]. M also depends on c, as c needs to increase along with M. Riutort-Mayol et al. (2022, pp. 6–10) outline a useful set of possible starting values for M and c, explaining in more detail the relationship between them and the way in which they depend on each other. The accuracy of the approximation relies on having defined these values correctly for the data and for the true value of λ, also known as the *length-scale* parameter of a Gaussian process. Therefore, it is helpful to have an idea of the plausible range of λ based on prior knowledge of the data. This means there is a basis for making informed guesses about the reasonable range of λ estimates. Using this approach as a starting point, I set M and c according to the recommendations in Riutort-Mayol et al., 2022 and fine-tuned the values based on model testing. The resulting values were $M = 22$ and $c = 4$.

For computational reasons, each macroarea is modelled with a separate latent GP. There is also a conceptual reason behind this choice, as it means that the GP hyperparameters will be estimated separately for each macroarea, thus uncovering potential variation in the pattern of missing data between them.

5.3.3 Phylogenetic regression

To control for and estimate the effect of shared ancestry between languages, I use phylogenetic regression as explained by Villemereuil and Nakagawa (2014, Ch. 11); the same approach was used in Chapter 4 of this thesis. This approach is distinct from phylogenetic inference, which aims to infer phylogenetic relationships between languages, usually based on word lists (Dediu, 2011; Jäger, 2013). Phylogenetic regression assumes that we already have a phylogeny representing known relationships between languages. The phylogeny used here was extracted from Glottolog, which

[4]For spatial modelling, in general, we can expect a high degree of non-linearity, which is also discussed in Chapter 2.

draws on expert classifications of languages in order to build a global language tree (Hammarström et al., 2023).

The idea behind phylogenetic regression is similar to using varying effects for families or genera (Jaeger et al., 2011). However, rather than assuming each family has a different, uniform probability of having or not having a certain feature in a language, phylogenetic regression takes into account the entire structure of the tree. This means that closely related languages will be expected to show a greater degree of similarity compared to distantly related or unrelated languages. Additionally, this approach allows for the inclusion of languages without any known or extant relatives, as these are unconstrained by the structure of the tree.

5.4 Results

5.4.1 Model comparison

All the models were tested using *Pareto-smoothed importance sampling* as implemented in the **R** package **loo** (Vehtari, Gelman, and Gabry, 2017; Vehtari et al., 2024). This is an efficient method for estimating predictive errors, which can be compared between models to find the best fit. The model with both areal and phylogenetic effects outperformed the model with only areal effects, as well as the model with only phylogenetic controls and the model with no controls. This suggests that missing data in linguistic typology is generated by both phylogenetic and geographic processes, confirming that typological data is not missing at random.

The comparative metric used is ELPD (expected log predictive density) as defined by Vehtari, Gelman, and Gabry (2017, p. 2). Thus, the models are compared based on the difference in their expected predictive performance on new data.

Model	**ELPD difference**	**SE difference**
Approx. GP + phylo	0.0	0.0
Approx. GP	-94.7	12.0
Phylo	-126.7	15.1
Base	-247.7	21.1

TABLE 5.1: Model comparison using PSIS-LOO.

These results show that some macroareas have been disproportionately neglected in terms of documentation strength and/or annotation in WALS (Dryer and Haspelmath, 2022). As shown by the fact that the phylogenetic model alone can be used to predict the distribution of missing data, some language families have lower documentation or coverage independently of their areal distribution. Although the results

show phylogenetic as well as areal bias, the areal component alone is a better predictor of the distribution of missing data than the phylogenetic component, suggesting that missingness is more heavily influenced by geography. Therefore, I will focus mainly on the areal results in this paper, but a subset of the phylogenetic intercepts will also be shown and discussed.

5.4.2 Model evaluation

For the best model, I calculated the balanced accuracy of the hurdle predictions, which is a way of measuring the accuracy with which the model predicted both 0s (missing languages) and 1s (languages which are present in WALS). The estimated accuracy is shown below for each macroarea, along with the overall accuracy.

These predictive accuracies indicate that the model is better at predicting missingness in North America and Australia than in Africa and Multinesia. Because the model only includes information about language locations and families, we can infer that, for Africa and Multinesia, the distribution of missing data may be less heavily influenced by these factors than in North America. Perhaps other variables, like environmental configuration, could be particularly relevant for these areas; in Africa, for example, rugged terrain may have posed a barrier to colonisers that led to the current preservation of more languages (Nunn and Puga, 2012).

Macroarea	**Estimate**	**SE**
Africa	0.52	0.002
Eurasia	0.6	0.005
North America	0.7	0.012
South America	0.63	0.009
Australia	0.7	0.016
Multinesia	0.51	0.002
Overall mean	**0.61**	

TABLE 5.2: Balanced accuracy of posterior predictions drawn from the hurdle component of the model.

The root mean squared error (RMSE) is the difference between the model's predictions and the observed values for the count component of the model. For example, an RMSE of 18 for Africa means that the model systematically over- or underestimates the number of annotated features in WALS by 18 on average. Looking at the posterior predictive plots in Figure 5.3, it is clear that the model regularly overestimates the number of annotated features for languages in WALS. The model performs better when it comes to predicting the proportion of zeros (missing languages), as we can see in Figure 5.4. This suggests that the geographic and phylogenetic components,

as modelled here, are better predictors of whether or not a language is missing from WALS in the first place than they are of the number of annotated features once a language is included. Perhaps other factors, such as academic interest, quality of the available grammars, or variation in research traditions play a bigger role in causing some languages to be disproportionately well-annotated compared to others.

Macroarea	**Estimate**	**SE**
Africa	18.82	0.9
Eurasia	23.26	1.1
North America	27.33	1.3
South America	26.56	1.6
Australia	25.48	2
Multinesia	18.57	0.9
Overall mean	**23.33**	

TABLE 5.3: Root mean squared error (RMSE) of posterior predictions drawn from the negative binomial (count data) component

In addition to testing the model using PSIS-LOO (Vehtari, Gelman, and Gabry, 2017) and RMSE, I tested predictive performance by drawing samples from the posterior distribution using the parameters of the fitted model. These predictions can show whether the model is able to predict the distribution of missing data for each macroarea equally well or if there is significant variation between them. Although posterior predictions are likely to be more uncertain than predictions drawn using exact leave-one-out cross-validation, this approach was used because the size of the dataset would make exact approaches prohibitively computationally intensive to calculate.

In Figure 5.3, as expected based on the RMSE, we can see that the mean predicted counts are overall higher than the actual observed counts. These differences appear to be largest for Eurasia, South America, and Multinesia. This suggests that there are factors that negatively impact how many features are annotated in WALS which are not accounted for in this model, which is to be expected given how many variables can influence data sparsity. In contrast, for the proportion of zeros or languages missing entirely from WALS, shown in Figure 5.4, the model predictions are closer to the observed values, and the differences in the predictive accuracy for the different macroareas is relatively small. Eurasia is the only area for which the model predictions for the proportion of missing languages are lower than the observed proportion, which suggests that some additional variables not included in this model, like academic interest, have caused Eurasian languages to be more likely to be in WALS.

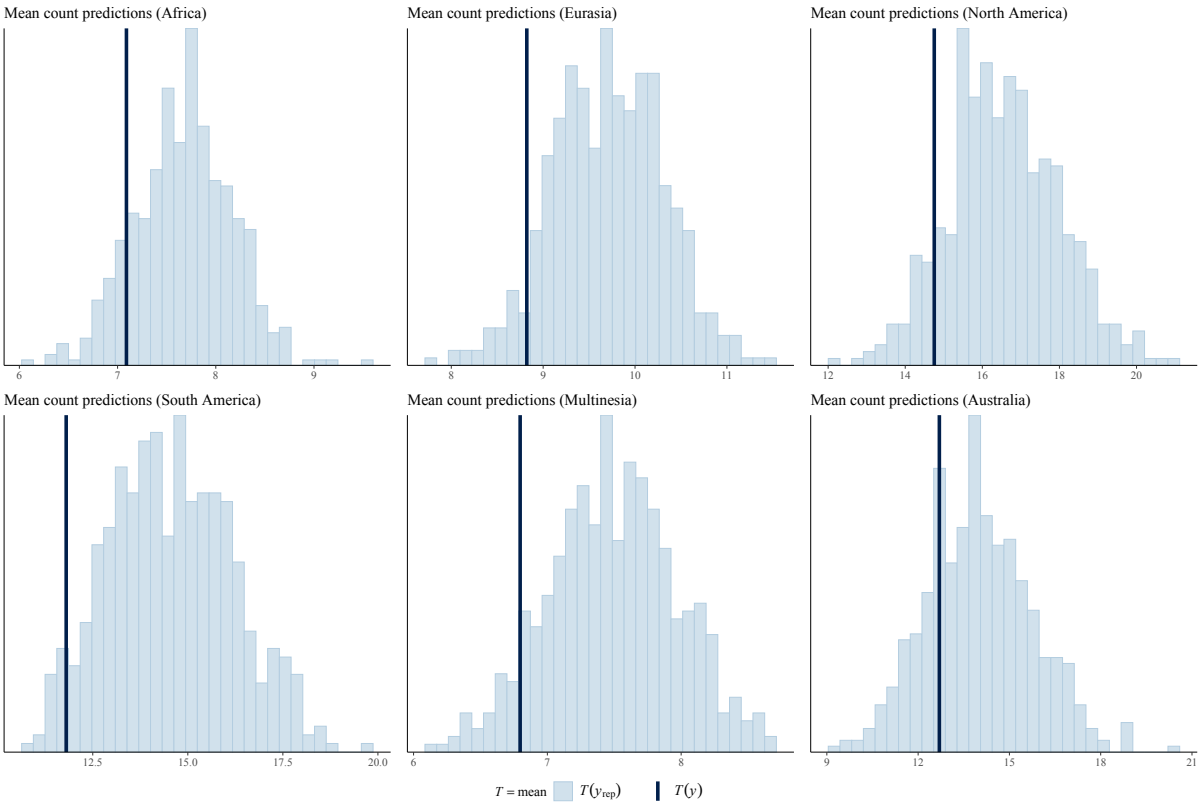


FIGURE 5.3: Posterior predicted counts (mean). The dark blue line indicates the observed value while the light blue bars represent samples from the posterior.

5.4.3 Areal predictions

Using the parameters of the fitted model, I drew predictions for new locations along a grid for each macroarea using only the spatial component of the model. In the hurdle plots, the colour scale goes from low (darker colours) to high (lighter colours) probability of a language being in WALS. In the count plots, the same colours represent how many features are predicted to be annotated for the languages in a given area.

As we can see in Figure 5.5, languages spoken in Africa have an overall low probability of being present in WALS, with the most common predicted probability lying between 0.2 and 0.3. The best coverage is predicted to be found north of the Sahara and on the southern tip of Africa. The model predicts the lowest coverage, with a probability of presence in WALS between 0.1 and 0.2, around Cameroon and Gabon. This is an area rich in languages, particularly those belonging to the Bantoid and Bantu families. Many of these languages have not been sufficiently documented. Bantoid languages in particular are in dire need of further study (Blench, 2015). The predictions from the negative binominal component of the model follow a very similar pattern to the hurdle predictions, meaning that areas where languages with a low expected level of annotation are spoken correlate with the areas in which many

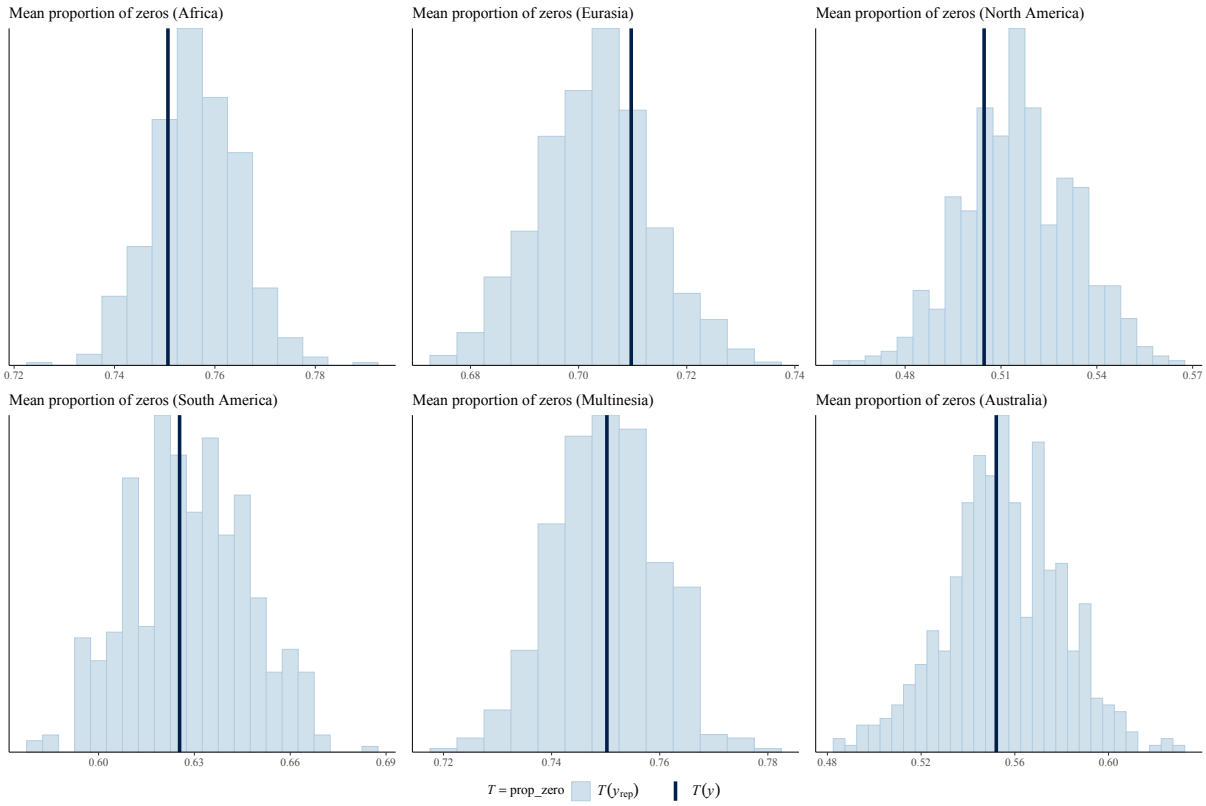


FIGURE 5.4: Posterior predicted proportion of zeros. The dark blue line indicates the observed value while the light blue bars represent samples from the posterior.

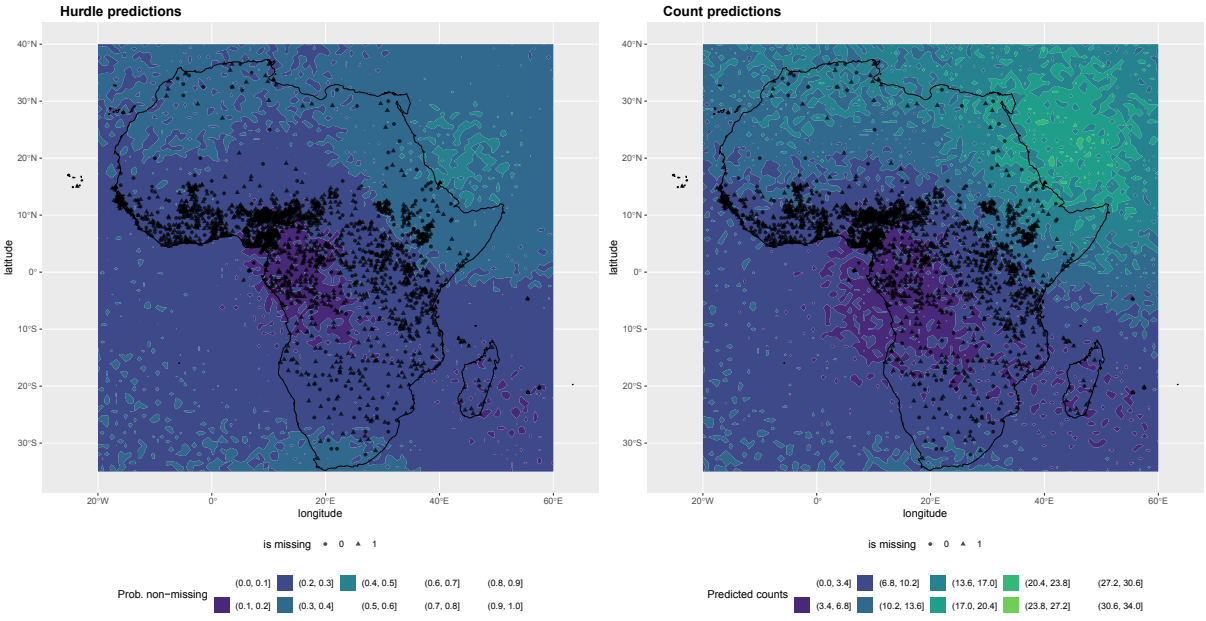


FIGURE 5.5: Spatial predictions for Africa

languages predicted to be absent from WALS.

For South America (Figure 5.6), we see a slightly higher overall probability of a

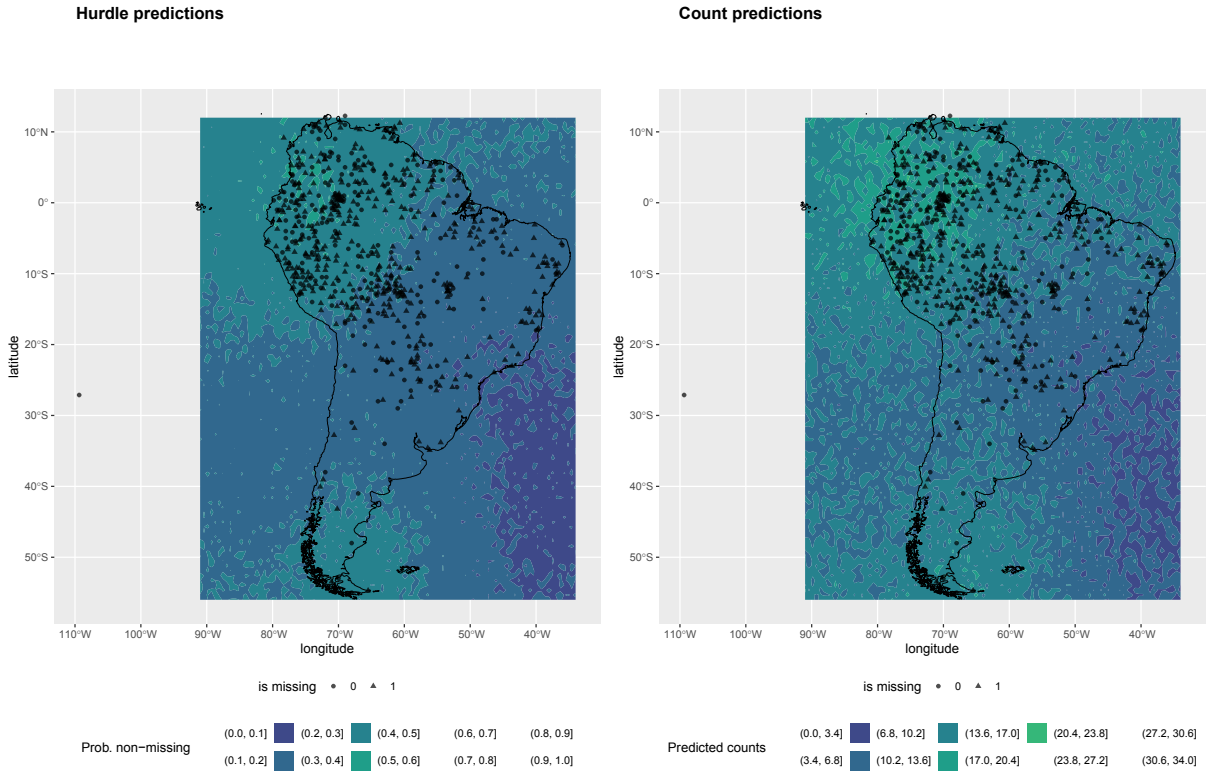


FIGURE 5.6: Spatial predictions for South America

language being present in WALS across the continent, with the most common values lying between 0.4 and 0.5, although this could also be due to the continent having a lower overall number of extant languages. Similar to the pattern in Africa, the areas with the highest probability of a language being in WALS correlate with the areas in which the best-covered languages are predicted to be. However, the highest probability predicted by the model is only 0.6. This locus of high probability is concentrated around the north-west coast of South America, encompassing parts of Ecuador, Venezuela and Peru. A lower probability of high data coverage and presence in WALS is predicted for Brazil.

North America (Figure 5.7) shows a higher overall probability of languages being in WALS than South America or Africa, with the highest predicted probability reaching 0.8. Again, the number of surviving languages on this continent could be low overall compared to how many there were in precolonial times. In both plots, we see a north-south and west-east cline which suggests that the languages of Central America and the Caribbean are less likely to be present in WALS than the languages in Canada or the United States. Interestingly, the count predictions drawn from the negative binomial component are slightly different from those from the hurdle component in this case. In the plot with count predictions, we see the highest

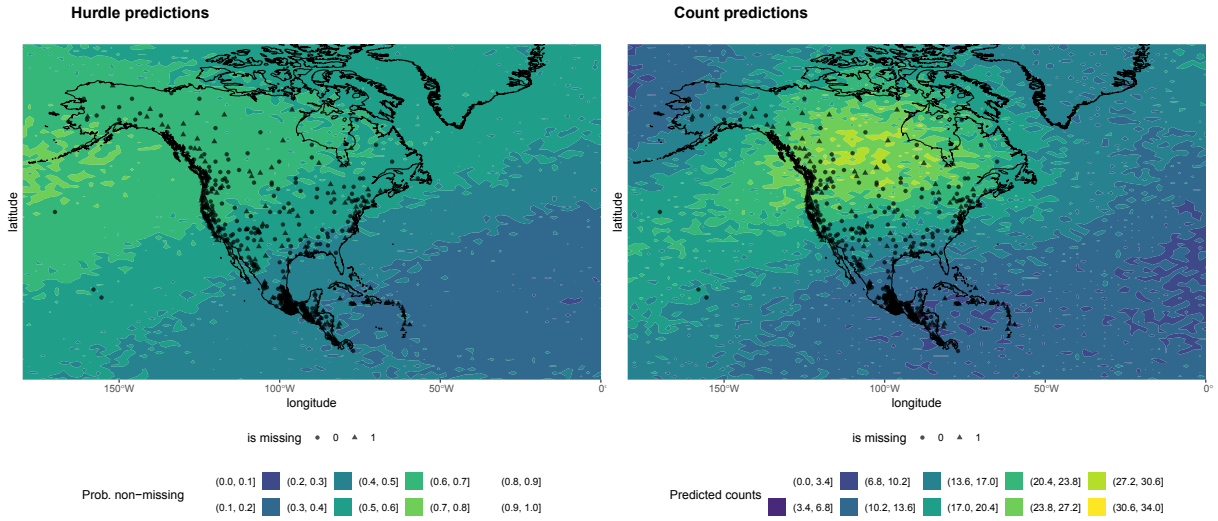


FIGURE 5.7: Spatial predictions for North America

predicted data coverage in Canada, a slight dip in predicted coverage for Alaska and the southern part of the United States, and the lowest predicted counts in Central America and the Caribbean.

In Eurasia (Figure 5.8), there is a clear north-south gradient, with the highest probability of a language being present in WALS in the north. In Europe, the predicted data coverage is lower for the Mediterranean and higher for Scandinavia, Central Europe, and the UK. In Asia, the gradient moves from the highest predicted data coverage in Russia, Japan and China, to a lower predicted coverage in the Middle East, India and South-east Asia.

In Australia, shown in Figure 5.9, the overall probability of a language being in WALS lies between 0.3 and 0.7 for the majority of the continent. This suggests that Australia's languages are better documented as a whole than those of South America and Africa. This may be due to Australia's current status as a relatively wealthy Anglophone country with a number of prestigious universities which have projects dedicated to the documentation of Australian languages; however, it could also be because there are fewer extant languages in Australia than in Africa or Eurasia, as so many of them have gone extinct. In addition, the majority of the indigenous languages in Australia today are endangered (Evans, 2010).

There is an overall higher probability of languages in the western and central part of Australia being in WALS. The area predicted to have the highest feature coverage in WALS is in the north of the continent close to Darwin. One of the reasons for this could be academic. Because this is where Australia was connected to Papua New Guinea, with which it formed a continent named Sahul in the distant past, this area is of special interest to scholars who are interested in understanding the distant

relationships between the people and languages of Australia and those of Papua New Guinea (Evans, 2020).

The last macroarea included in this study, Multinesia, spans the archipelago of Indonesia, the island of Borneo, some Oceanian islands including the Solomon Islands, and Papua New Guinea. In Figure 5.10, it is clear that Borneo has the lowest predicted probability of being in WALS, while the Solomon islands have the highest. In Papua New Guinea, we see a gradient from west to east with a slightly higher predicted coverage in the east.

5.4.4 Phylogenetic effects

The phylogenetic part of the model returns an intercept for each language, which is expected to be similar to the intercepts for related languages. An intercept close to zero indicates that phylogenetic effects are weak. A positive intercept indicates that languages in that family are likely to be relatively well-annotated in WALS, and a negative intercept indicates the opposite. As can be seen from the plots below, there is considerable uncertainty around many of the phylogenetic intercepts. This could suggest that the placement of a language in the phylogenetic tree is not a very reliable indicator of its being in WALS or having a high level of coverage in the database.

Figure 5.11 shows a sample of varying phylogenetic intercepts for Bantu languages on the left; these appear to have relatively high levels of uncertainty. For the Austronesian language sample, on the right, the uncertainty is still considerable, but appears to be lower. The level of variation in expected documentation is higher, too: for some Austronesian languages, the intercepts are positive (suggesting a higher probability for families in that subgroup to be well-annotated in WALS), whereas for Bantu languages, all the means fall into the negative, suggesting a lower probability that these languages will be well-annotated in WALS.

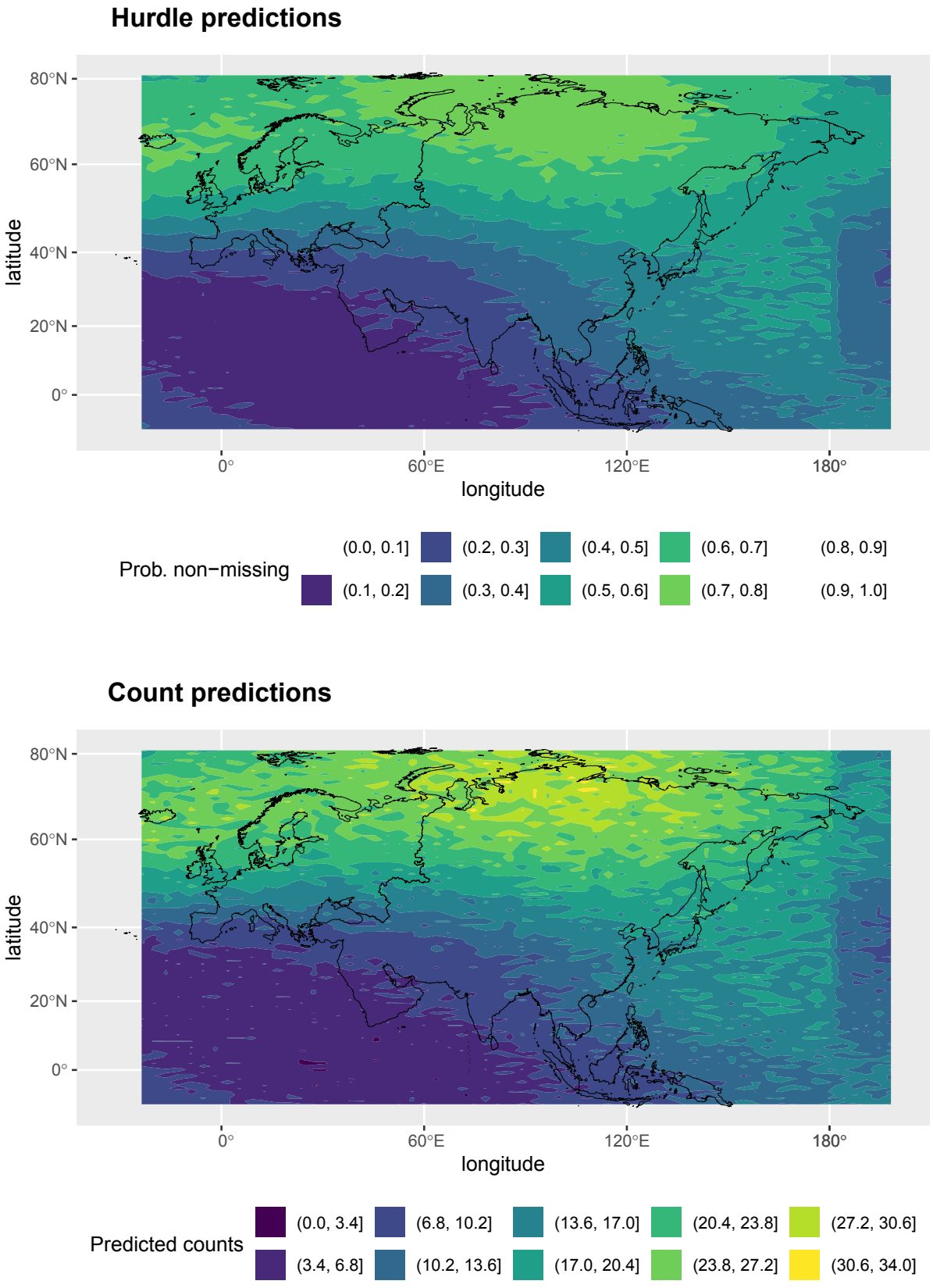


FIGURE 5.8: Spatial predictions for Eurasia (hurdle model)

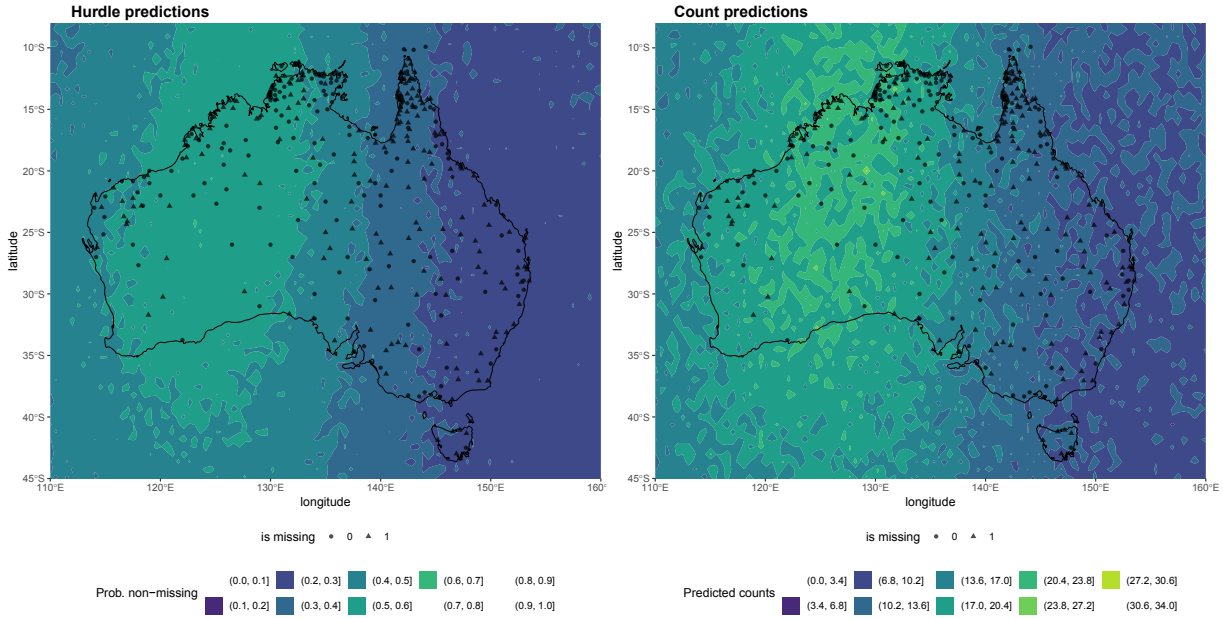


FIGURE 5.9: Spatial predictions for Australia

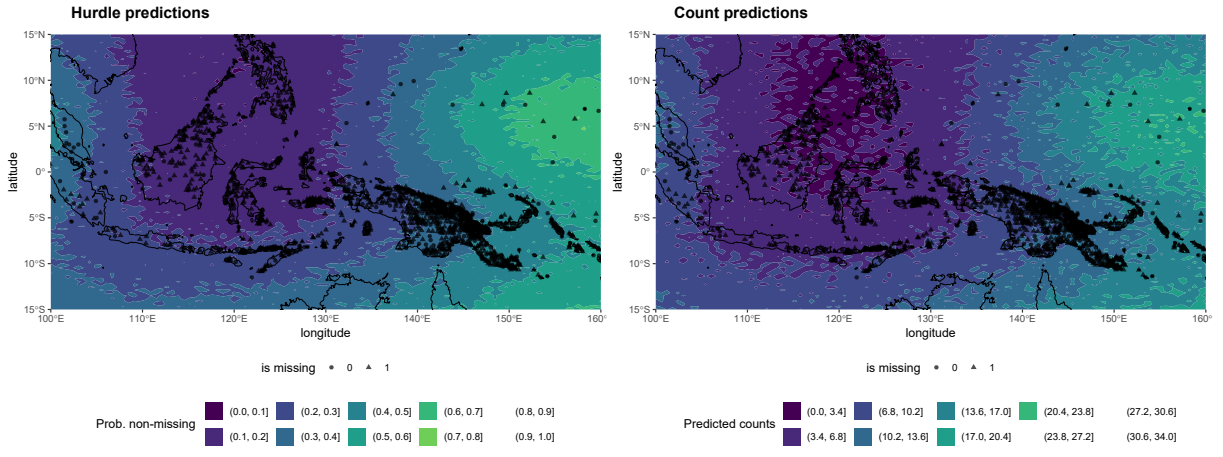
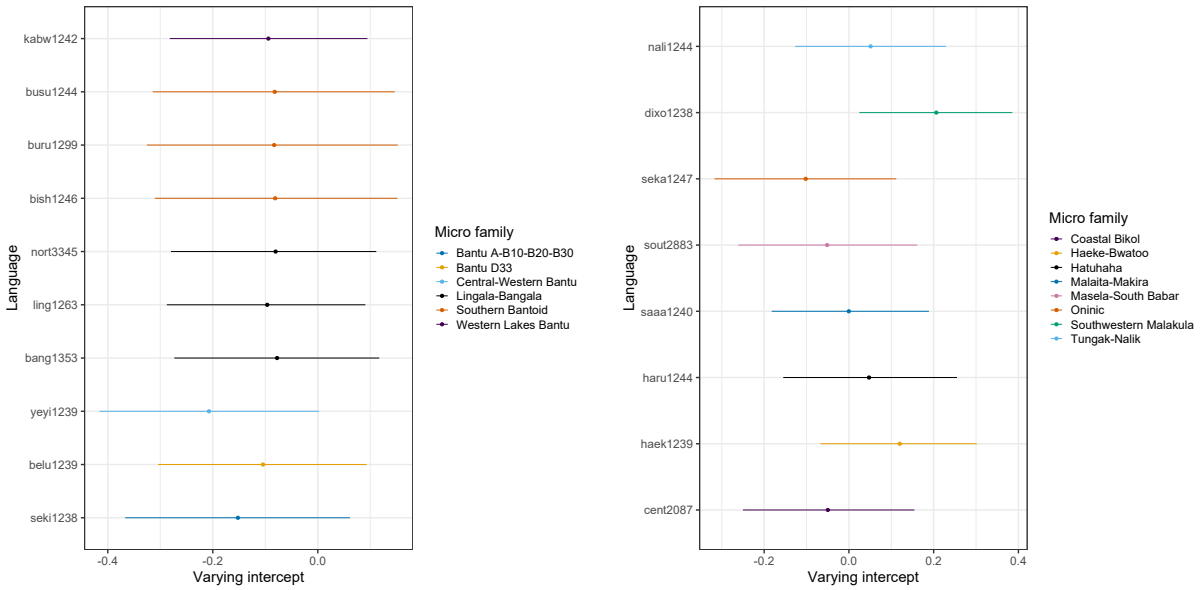


FIGURE 5.10: Spatial predictions for Multinesia



(A) Phylogenetic intercepts for a sample of Bantu languages.

(B) Phylogenetic intercepts for a sample of Austronesian languages.

FIGURE 5.11: Phylogenetic intercepts for a sample of Bantu and Austronesian languages.

5.5 Discussion

Languages spoken in some areas of the world are clearly more likely to be in WALS than others, and there is geographic variation in the level of data coverage within the database, too. North American languages have the best overall representation in WALS, followed by the languages of Eurasia, particularly the north, while languages spoken in Africa and South America are the least likely to be in WALS. Thus, these results show that data in WALS is not missing at random, and that missingness is both geographically and genealogically biased. As this model assigns a higher level of importance to geography, which indicates that the geographic bias shaping the distribution of missing data is stronger than the phylogenetic bias, the inclusion of spatial controls is particularly important for bias mitigation.

The results show an overall gradient from north to south, in that languages north of the Equator are generally more likely to be in WALS and to have a higher level of data coverage. Not surprisingly, the probability of a language being in WALS correlates with the number of features the model predicts to be coded for that language.

Bias occurs within macroareas as well as between them. When doing statistical studies on African data, we should be mindful that the northern and southern parts of the continent are more well-documented than the centre, east and west. This could be linked to environmental factors, as central Africa has dense rainforest coverage, which can make it difficult to access this area. Additionally, central Africa is known for being populated predominantly by Bantu-speaking communities. The lack of deep-time phylogenetic diversity may have caused some scholars to be less interested in documenting the under-researched languages of this region.

It should be noted that these results are necessarily based on the languages which are listed in Glottolog and their locations. The database cannot include all the languages which have ever existed, nor many of those which were made extinct due to colonisation and other historical processes. Additionally, if the locations of some languages are not accurate in Glottolog, this will cause problems for the predictive model. As discussed at length in the introduction of this thesis, we also know that languages are spoken across territories rather than point locations. Points can provide an approximation of where a language is spoken, but this is more complicated for cultures with nomadic lifestyles. For Australia in particular, this could cause unexpected results, since Australia is populated by many communities whose lifestyle is traditionally nomadic or seminomadic (Evans, 2020). Moreover, the spatial extent of a language could correlate with its probability of being well-documented, as they also likely have more speakers. Including the number of speakers in the model could be a way to control for and estimate the strength of this effect. In this study, no

additional explanatory variables were included, but there are many which might have influenced the distribution of missing data globally, some of which could plausibly be included. For example, for Eurasia, the highest predicted counts are centred on an area in the north of Russia. This may be a result of WALS having been compiled by a number of experts on Russian and Slavic languages. The model lacks the explanatory power of additional factors like academic interest, geography, or cultural variables, all of which could influence which languages are annotated for certain features in large-scale databases.

5.6 Conclusion

The contribution of this chapter is twofold. First, it has clearly demonstrated that data in linguistic typology is not missing at random, using WALS as a case study. Additionally, it has made a methodological contribution to the field of spatial modelling in linguistic typology by developing an approximate GP model. This makes it possible to apply a GP, which would otherwise be computationally intractable, to large-scale datasets as an areal control and a way to detect areal clustering of features.

The results show that the distribution of missing data is influenced by both areal and phylogenetic biases. This means that languages spoken in some macroareas have a higher probability of being in WALS than others. Within macroareas, there is geographic bias, too. For Eurasia and North America, this manifests as a gradient from north to south, where languages in the north are more likely to be in WALS than languages in the south. When doing statistical inference with this data, we are therefore likely to induce geographic and phylogenetic bias. Thus, it is important, as a first step, to use statistical techniques such as controlling for areal effects and missing data imputation to mitigate this (Guzmán Naranjo and Becker, 2021; McElreath, 2020). It is particularly crucial to include spatial controls, as missingness appears to be more biased along geographic lines than between language families.

The nature and strength of the bias induced by excluding missing data in linguistic typology is not well understood. We do not know whether statistical techniques like the ones mentioned above can successfully counteract the effects of heavily biased missing data. For example, if we only have three languages available in a family of a hundred languages, missing data imputation may not solve the problem and could itself be biased. There is likely a limit to how much statistics can help when we have heavy bibliographic bias like this, but we do not know where that limit lies. Similarly, no study has yet examined the strength of bias induced by excluding missing data from typological studies, or the extent to which methodological refinement can mitigate this. Controlling for bibliographical bias by including the quality of the available

grammatical description of languages is another possible approach to reducing bias induced by data sparsity (Becker, Guzmán Naranjo, and Ochs, 2023).

The creators of some of the newer databases, such as Grambank, are trying to remedy the issue of data sparsity (Skirgård et al., 2023). Other databases focus on more specific subsets of linguistic features or languages in order to achieve more fine-grained coverage using language-specific knowledge (e.g. Idiatov and Van de Velde (2021)). A combination of approaches, including funding further language documentation efforts and the curation and creation of cross-linguistic databases, as well as developing statistical techniques, is needed in order to ensure that our findings are as unbiased as possible. Additionally, it is essential that we replicate large-scale cross-linguistic studies in order to test whether their findings hold as new data and methods become available.

Chapter 6

Conclusion and outlook

Modelling space requires making a series of trade-offs between complexity and realism on the one hand, and tractability and interpretability on the other. Whenever we decide to include additional parameters in a model, the margin for misspecification and error increases. Because of this, it is typically considered good practice to use the simplest possible model that will capture the underlying trends in the data. However, it is far from trivial to decide what the simplest possible model should look like, and what is crucial information to include and what is not. In some cases, adding even a little bit of complexity improves model performance drastically, as was evident in the results of the case study presented in Chapter 2. In other cases, the inclusion of more information adds very little to the model's predictive accuracy or even makes it worse. This is an important point to bear in mind when considering potential directions for future work in this field.

In this chapter, I will start by summarise the findings of this dissertation, its overall contribution to the field, and the implications of its results. I will then provide some recommendations for possible directions that future work could take, of which there are many.

6.1 Summary

In this thesis, I have contributed a series of case studies to the field of spatial linguistic typology which make use of a diverse range of methods, some of which have not been applied to linguistic data before. I have tried to give as many details about their implementation as possible, and in Chapter 2, I gave some practical recommendations for when to use what kind of spatial model. Some of these were based on a brief case study of the areal distribution of labial-velar consonants in Africa based on datasets by Idiatov and Van de Velde (2021) and Segerer and Flavier (2011-2021). These were also accompanied by concrete recommendations for implementation in R, including some packages that make Bayesian statistical inference of spatial models

more accessible (Bürkner, 2017; Donegan, 2022). The code used in this thesis will also be freely available.

In Chapter 3, I used a novel combination of statistical techniques in order to disentangle the cultural and environmental factors which have shaped linguistic diversification in Africa. I devised two ways of measuring linguistic diversity which I found necessary to distinguish from each other: *density* was calculated by counting the number of neighbours of a given language territory, and *diversity* was calculated as the sum of the lexical differences between a language and its neighbours based on word lists, which acts as a proxy for phylogenetic diversity (Jäger, 2013). Language density and linguistic diversity were found to be correlated with some of the same variables, which indicates that their distribution has been shaped by some of the same diachronic processes. Political complexity, the degree to which communities organise themselves into larger jurisdictional units such as states, was found to be the variable with the most consistent negative impact on both linguistic density and diversity. However, some notable differences between the estimated correlations for density and diversity were found. Overall, language density appears to be affected to a greater extent by environmental variables, including terrain elevation, seasonality, and temperature, than cultural factors, with the exception of political complexity. In particular, variable terrain elevation seems to have a more consistent effect on language density than diversity, suggesting that it facilitates diversification more than the maintenance of phylogenetic diversity over time. This provides some support for the idea that terrain elevation is a barrier to contact that leads to differentiation, although compared to some of the other variables in the model, the effect of terrain was found to be relatively weak. Cultural variables were found to impact linguistic diversity more consistently than environmental ones. These included political complexity, jurisdictional hierarchy within the local community, the level of dependence on agriculture as a subsistence strategy, and settlement strategy, ranging from nomadic to permanent settlements. The cultural organisation of neighbouring language communities was found to influence diversity the most, suggesting that politically complex, highly agricultural and sedentary language groups exert pressure on their neighbours to shift or converge to their language. This suggests that researchers like Comrie (2008) and Epps (2020) were correct in emphasising that the maintenance of co-territoriality involving many diverse linguistic communities over time is linked to the interplay of cultural attitudes and ways of life in those areas.

In Chapter 4, I presented a new version of a method called *multivAreate*, developed by Guzmán Naranjo and Mertner (2022) and Guzmán Naranjo, Mertner, and Urban (2024), with the goal of inferring variation in the diffusibility of different types of structural features. My focus was specifically on the range parameter which is

inferred by a GP, as this provides an estimate of the spatial extent within which features tend to form geographically contingent clusters which cannot be explained by other variables in the model. Gender and noun classes were found to be the least diffusible features in Africa, suggesting that if these features are impacted by contact, this tends to occur within a small area. The areal patterns which were visualised for this category of features supported this, as they were not characterised by an absence of spatial effects, but by few strong local clusters. Perhaps, in order for contact to affect gender/noun class systems, contact between speakers of languages with distinct systems should be very intense, which is how Thomason and Kaufman (1988) famously characterised the concept of feature borrowability: the more intensive the contact needed to cause changes to a domain, the less borrowable it is considered.

In this study, the inferred hierarchy of borrowability for the subset of features tested places gender and noun class systems at the low end along with nominal number marking; this is followed by bound verbal categories, word order, and tense/aspect markers which can be bound or free. A limitation of the model is that it sometimes infers a high degree of uncertainty in the range of certain features, which was the case for TAM marking and suggests that there is a high level of variation in the diffusibility of features belonging to that domain. Otherwise, the findings align with patterns of areality observed in Africa previously, with gender/noun class systems considered a highly stable feature within lineages while word order shows distinct areal patterns and inconsistency within lineages (Dimmendaal, 2008a; Güldemann, 2018b; Heine and Nurse, 2008; Van de Velde, 2019). As a secondary goal of the study, areal patterns could be visualised, some of which were discussed in detail in the paper with reference to prior findings; the rest are shown in Appendix B.

The final case study in Chapter 5 of this thesis found that the distribution of typological data on the structural features of language included in the most recent version of the World Atlas of Language Structures (Dryer and Haspelmath, 2022) is spatially and phylogenetically non-random, which provides a theoretical basis for the use of geographic and phylogenetic bias controls. Interestingly, the geographic bias was found to be greater than the phylogenetic bias, adding to a growing recognition of the importance of including controls for geography alongside language families in quantitative studies on linguistic typology, and perhaps casting some doubt on the results of previous studies which may not have included such controls (Guzmán Naranjo and Becker, 2021). This emphasises the importance of replication studies making use of new data and methods.

6.2 Future work

There is ample opportunity for future work in the intersection between Bayesian spatial modelling, phylogenetics, linguistic typology, and language evolution. These topics are growing in popularity among researchers, and many promising studies have recently advanced the field (such as those by Becker, Guzmán Naranjo, and Ochs (2023), Graff et al. (2024a,b), Guzmán Naranjo and Jäger (2022), Guzmán Naranjo and Mertner (2022), Neureiter et al. (2022), Ranacher et al. (2021), and Skirgård et al. (2023), and too many others to list here). There is still a great deal of variation in the methods used and in how distances between language locations are calculated, and while methodological diversity of this kind is clearly an asset, it can make the comparison of results and methods across studies more difficult. I will not argue for standardisation of the methods used; as Chapter 2 of this thesis shows, the suitability of different types of spatial models depends heavily on the type and amount of data which is available for a specific study. However, studies which focus explicitly on comparing the performance of methods or distance metrics, or on replicating previous studies using different methods (like Guzmán Naranjo and Jäger (2022) and Guzmán Naranjo, Mertner, and Urban (2024)), are crucial for the advancement of the field.

Bearing in mind that the addition of complexity to a statistical model should be done with careful consideration, the following sections will tackle some of the current issues in the field of spatial models for linguistic data. Many of these will touch upon the complexity of the real world and the ways in which this could be accounted for in future work. It should be considered that if we were to include all of the complexities of the real world in a single model, that model would never finish running (at least on the hardware which is currently available to researchers in this field). It would also be very hard to interpret its results, as the parameters of complex models are interdependent and should be interpreted in relation to each other. Thus, we cannot always assume that a more realistic model is necessarily going to be a better or more insightful one. At the same time, there may still be a great deal of important information which could be included in future spatial models in a considered way. In the following sections, I will focus on what I see as some of the most crucial considerations and avenues for advancement in the field as they relate to the study of linguistic typology, since there are too many to cover them all.

6.2.1 Water, areas, and other opportunities

Some of the results presented in this work support the idea that variable terrain is a barrier to contact between language groups. Distances across variable terrain can be

measured in a relatively simple way by using a topographic surface and calculating the shortest path between two points on that surface (Guzmán Naranjo and Jäger, 2022). However, the role of water in contact dynamics is less well-understood. In some cases, as when groups historically lacked sailing technology, the presence of a large body of water between them would preclude contact between them entirely. However, some groups developed technology which allowed them to cross bodies of water with relative ease. In other cases, water could actually facilitate contact. Rivers are a famous example of this, as linguistic groups in areas like the Amazon or in Egypt have historically settled near rivers and may have travelled faster by taking advantage of river systems (Bentz et al., 2018). These factors are particularly difficult to model, as quantifying how likely or how fast travel across water would be requires knowledge about the kind of technology that cultures in that area would have possessed. For instance, Kaiping (2022) models prehistoric migrations in South America, drawing on knowledge from archaeology, anthropology, and the ways in which river systems are utilised by present-day cultures in the region during travel. Synthesising available knowledge from different disciplines could be a promising approach to building a realistic model of interactions across water for different regions of the world.

Another common concern which was also raised in the introduction of this thesis relates to the variable sizes of different language territories. Even mapping a language community to a polygon with defined geographic boundaries represents an abstraction from reality; mapping such a community to a single coordinate, then, seems like an even more egregious one. In one of the chapters of this thesis, I used a polygon dataset created and curated by SIL (Eberhard and Fennig, 2023). However, as this dataset is not freely available, the search for alternatives is ongoing. Data collection is, of course, the ideal solution to this issue. Aside from that, the possibilities include such suggestions as tessellating points using Delaunay triangulation to create a Voronoi diagram (Aurenhammer and Klein, 2000). These methods suffer from a lack of proper testing, and since they disallow overlaps between language territories, the question must be asked whether these polygons would be a better representation of reality than points. A useful first step for future work could be to evaluate this in a systematic comparison between a spatial graph using point locations and distances against a binary or weighted Voronoi diagram. Another significant limitation related to the issue of polygons is that GP models, which I found to be the best for the dataset in my case study in Chapter 2, cannot be used with polygons at the moment of writing this. However, GPs for spatial statistics are rapidly gaining traction, so the possibilities for future development are promising (Gelfand and Schliep, 2016).

There is a related avenue of ongoing research in spatial statistics into methods which can adaptively infer the number of neighbours in a spatial graph (Levada,

Nielsen, and Haddad, 2024). This could be an alternative to polygon data with would account for differences in language density and deal with the common issue that a distance radius of 500 km includes a lot more languages in dense areas (which may not actually be in contact with groups that far away) while likely failing to capture contact between languages spoken in sparsely populated areas. Graphs like this can be used not just for the inference of spatial autocorrelation, but as the basis for simulations of evolving systems called agent-based models, which have already resulted in valuable insights into the interplay between innateness and culture in the evolution of language (Kirby, Dowman, and Griffiths, 2007) and historical linguistics (Hartmann, 2021).

6.2.2 Asymmetry in contact dynamics

Language contact occurs in a wide variety of sociolinguistic situations which can impact the structures, sound systems, and lexicons of the languages involved in diverse ways (Muysken, 2010). Some contact situations lack any obvious hierarchy between the languages involved, as in cases of egalitarian multilingualism (Lüpke, 2016; Pakendorf, Dobrushina, and Khanina, 2021). On the other hand, situations in which a politically powerful language, often with a larger number of speakers and/or a larger geographic area, come into contact with smaller, less politically powerful languages are well-attested historically and synchronically (Childs, 2010). These kinds of contact situations can lead to an asymmetric transfer of linguistic material, with the speakers belonging to the smaller language community converging and perhaps eventually shifting towards the other. The idea of asymmetry is built into the terminology commonly used to describe contact situations cross-linguistically, with one language described as the 'source language' and the other as the 'recipient language' (Matras and Sakel, 2007).

Despite asymmetry being a feature of many contact dynamics, its incorporation into computational methods remains an open issue. There are two main complications which need to be taken into account. The first is computational; the second is conceptual. As discussed in Chapter 2, most spatial modelling techniques require the underlying representation of the space between languages – whether that takes the form of raw distances or a neighbourhood matrix – to be symmetric. This excludes the possibility of explicitly building asymmetry into the spatial representations themselves for most spatial models. Fortunately, there is an exception to this rule, as the simultaneously autoregressive (SAR) model does not require a symmetric neighbourhood matrix (Whittle, 1954).

The directionality of contact effects does not always correlate with political power, prestige, or size (in terms of area or population). Sometimes, a dominant language adopts linguistic material from a less-dominant or even oppressed language. For example, when a community of speakers shifts from one language to another (often related to prestige or cultural or military dominance), they retain elements of their original language or choose the dominant-language variants which most closely match structures in their original language. These elements can then enter the dominant language as substrate effects (Muysken, 2010). This could be the mechanism by which labial-velar stops originated in central and West Africa (Bostoen and Donzo, 2013; Idiatov and Van de Velde, 2021).

A related issue is that hierarchies are temporally unstable. Based on current power relations in the region, one might assume that the borrowing of click consonants into Bantu languages represents a kind of transfer which defies hierarchical relations. Click consonants are found in the Khoisan languages spoken in the Kalahari desert and southern Africa, which comprise multiple distinct lineages, including Tuu and Kx'a. When Bantu speakers arrived in the region, they likely needed to rely on local knowledge for survival, and it is thought that clicks were adopted by these speakers during this time (Pakendorf et al., 2017). Later, speakers of Bantu languages would outnumber the hunter-gatherer communities. Cases such as this could create complications for detecting asymmetric contact based on synchronic data, as the relative sizes of linguistic communities shift over time. Hierarchical relationships are not temporally stable either and do not always correlate with community size. In the Rift valley in Tanzania, power relations have been unstable throughout history, which means that the languages in the region have likely all influenced each other at some point in time (Kießling, Mous, and Nurse, 2008).

The inclusion of social dynamics of language contact in a spatial model is a worthwhile pursuit for future research. A simple approach which relies on established methods could involve comparing a SAR model with an asymmetric neighbourhood graph with the same model with a symmetric graph. If the asymmetric model outperforms the symmetric model, this indicates that asymmetric relations play a role in shaping the distribution of linguistic features in the relevant area. On a related note, it is important to remember that the usefulness of statistical methods lies in their ability to detect underlying patterns in noisy data. Even if there are many exceptions to a rule, there is still value in detecting whether the rule holds in general for a given dataset. It is difficult to tell from a collection of real-world case studies whether the effects of contact are generally more likely to be asymmetric or bidirectional, and this is what a statistical model could do.

It would be even more useful if the model could detect asymmetry and the direction of influence (but see Dellert (2019) for a detailed discussion of the challenges associated with this and an algorithm for inferring the directionality of contact influence from lexical data). For spatial models, a relevant area of ongoing research is the development of methods by which a spatial weights matrix can be inferred from the data, so that it does not have to be specified by the researcher (Merk and Otto, 2022). Asymmetry in the weights matrix can also be inferred, such as when the spatial dataset comprises irregular polygons (which is the case for language polygon datasets) (Harke, Merk, and Otto, 2022). Future work could evaluate the suitability of these methods for linguistic data.

6.2.3 Time and space

The synthesis of what we know about the dynamics of language inheritance and contact is one of the main goals of statistical models of language change. In this thesis, the focus has been on the modelling of space, as this has received considerably less attention than the modelling of language history, despite the fact that it is arguably equally important. Furthering our understanding of language contact is essential to understanding how languages evolve and increasing the time depth at which we can reconstruct the history of languages. Because the field of spatial modelling in linguistics is newer than that of phylogenetics, there is still much to do in terms of evaluating how well these models can detect the processes we are interested in, even under the less-than-ideal conditions of data sparsity and uncertainty around the classification of languages. This thesis is one of several recent works attempting to make progress in this area.

Every chapter has included a control for phylogenetic relatedness, the most promising of which is phylogenetic regression, as this takes into account the entire structure of a language tree (see Grafen (1989) and Villemereuil and Nakagawa (2014) for an explanation of phylogenetic regression). However, there is a great deal of uncertainty around the relatedness of languages, especially in Africa, and especially at the level of subgroups like Bantu (Grollemund et al., 2015; Güldemann, 2018a). The inclusion of more detailed phylogenetic information can perhaps have the unintended side effect of building a level of certainty we do not have into the model architecture. Future work could tackle the issue of phylogenetic uncertainty within a Bayesian modelling framework (Villemereuil et al., 2012).

A method which combines the dimensions of time and space in a dynamic way includes phylogeographic models Koile et al. (2022) for reconstructing the history

of Bantu languages. However, this method still relied on the inference of a phylogeny prior to fitting the spatial model. Incorporating information about spatial relationships directly into a phylogenetic model could be an important step to reduce the interference of contact in establishing language histories. Future researchers could also adopt an approach which involves studying the spatial dynamics first, for example by examining the stability and diffusibility of the words and concepts which are used in phylogenetic inference. This is especially crucial for underdocumented language families like Bantu, which have a long history of both population movements and social interactions between groups that makes disentangling the effects of contact from those of inheritance particularly challenging. The effects of contact may have permeated the lexicons and structures of these languages more deeply than previously thought. It is also crucial to consider the temporal dimension of contact itself, particularly when migrations are a known part of the history of these groups. In the case of Bantu languages, we could draw on the wealth of studies which have inferred a homeland and a probable migration route to detect past contact influence across space (see Grollemund et al. (2015), Koile et al. (2022), and Pakendorf, Filippo, and Bostoen (2011), and references therein). This, in turn, would help us evaluate whether some of the proposed subgroups could be the result of contact between linguistic groups which are no longer geographically contingent.

6.2.4 Divergence

When related languages separate, a process of linguistic differentiation begins which leads to these languages gradually becoming more distinct over time. Differentiation, like convergence, can also happen for other reasons, such as evolutionary drift, functional or cognitive biases, and social pressures. Language contact can also lead to differentiation, in which case it is called divergence, referring specifically to "differentiation that is driven by language contact, rather than the absence of contact" (Mansfield, Leslie-O'Neill, and Li, 2023, p. 234).

In some cases, a desire for linguistic communities to distinguish themselves from their neighbours results in borrowing taboos, restricting the flow of lexical material between languages (Aikhenvald, 2002). However, these languages may still show convergence at the morphosyntactic level, meaning that such languages may still be susceptible the transfer of patterns without any transfer of forms (Aikhenvald, 2002). Drawing on data from a diverse sample of languages, Mansfield, Leslie-O'Neill, and Li (2023) find that divergence between dialects in close contact happens at the level of grammatical forms, but does not affect the presence or absence of forms expressing particular meanings. Thus, morphosyntactic variables of the kind included in

Grambank, all of which relate to patterns rather than forms and many of which relate to patterns of presence or absence, are not necessarily expected to show contact-induced divergence. However, as this is cross-linguistically understudied, we do not know whether divergence occurs at the pattern level in ways that could be detected across large geographic areas. As the study by Mansfield, Leslie-O'Neill, and Li (2023) focused on dialects or closely related varieties, it is not certain to what extent its findings can be generalised to more distantly related or unrelated languages.

It is worth remembering that language evolution is driven by speakers of diverse languages, and that these speakers can and do make conscious choices about the linguistic material they use and how they use it, and that these choices shape languages over time (Di Carlo and Good, 2023; Morrison, 2018; Mous, 2003). Under such a view of language evolution, we cannot exclude the possibility of divergence occurring at the level of language structure as well as surface forms, while noting that current research points to divergence being much less likely to occur in the domain of structural or conceptual categories (Graff et al., 2024b; Mansfield, Leslie-O'Neill, and Li, 2023). All this may be relevant for the kinds of spatial patterns detected by the model used in the present study, serving as a reminder not to exclude the possibility of contact-induced divergence or other kinds of differentiation which could occur between geographically proximate languages.

Detecting and incorporating linguistic divergence in quantitative models is a relatively recent goal for linguists studying variation. Much more attention has been devoted to convergence, perhaps for good reason. Convergence is well-attested at the macro-level. The WALS features examined by Murawaki and Yamauchi (2018) all showed positive (albeit, in some cases, weak) spatial autocorrelation. However, it should be noted that their model disallowed negative values, such that even if some features were negatively spatially autocorrelated, it would not be possible for the model to detect it.

6.3 Final remarks

In this thesis, I have demonstrated multiple new ways of modelling linguistic data in space and applied these predominantly to case studies of African languages. I hope I have shown how different Bayesian approaches can be applied to the study of linguistic typology and diversity, particularly as it relates to drawing inferences about the history and evolution of the languages under study. Uncertainty is a reality of this field on many levels; we deal with uncertainties in classification, in the locations of languages and their geographic extent, and in the data sources we work with. Uncertainty is not typically considered a positive thing, but under a

Bayesian framework, we utilise uncertainty as a way to ensure that our findings are robust, and to highlight when they are not. As such, it becomes a valuable source of information. In multiple chapters of this thesis, that information has played a key role in interpreting the results. It may also have left some with the impression that not much can be said about the topics covered in this thesis with any degree of certainty: from deciding whether hotbeds of linguistic diversity arose as a result of influences from the surrounding geographic environment or from the social configurations of the speaker groups within those hotbeds, to the question of whether some structural features are inherently more diffusible than others, about which I could only draw tentative conclusions with reference to a specific subset of features for a specific macroarea of the world.

The reality of spatial modelling is that it is fairly computationally intensive. As such, I have been faced with a series of trade-offs, as I discussed briefly earlier, which boil down to attempting to find the simplest possible model that is feasible to implement, run, and interpret. For the chapter on diffusibility, I chose to limit my dataset so that I could apply a model with a higher level of realism in its representation of space. For the chapter on data sparsity, I made the opposite choice. These and many other decisions have also been motivated by questions of practicality, as in the chapter on the correlates of linguistic diversity, where the decision of how to implement the method required a great deal of trial and error which will not be visible to the reader. I hope that any disagreements about the choices I have made here will spark interesting dialogues between researchers from different disciplines and fuel future work in this field.

Appendix A

Code availability

All the code for this thesis, including the Stan model files and R scripts for running the model and generating the plots, is available online. In this appendix, I will provide links to the repositories containing the code for each chapter.

The code is contained in a GitHub repository which can be accessed at https://github.com/anmiri/spatial-models-thesis. The code for each chapter is in a separate folder, and each folder can be navigated using the instructions given in the repository.

Appendix B

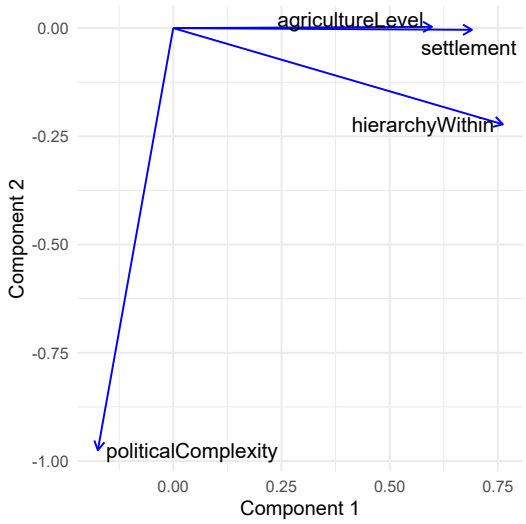
Data and intermediate results

B.1 Chapter 3

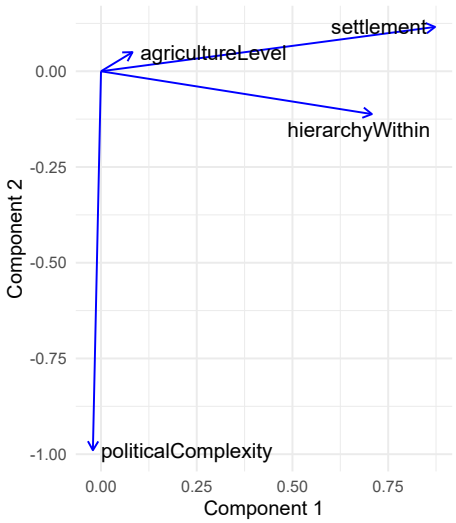
In Chapter 3, I transformed the D-PLACE variables (agriculture level, political complexity, settlement strategy, and community hierarchy) using Principal Components Analysis (PCA). The variables from D-PLACE are all ordinal, so these were transformed for use in the models without the phylogenetically decorrelated residuals using ordinal PCA as implemented in the R package **Gifi**. These PCAs were then transformed using Varimax rotation (Kaiser, 1958) The loadings plots from both of these transformations are shown in Figure B.1a and B.1b.

I applied the same method to the extracted non-phylogenetic residuals and phylogenetic latent variable z (which represents, on the same scale as the non-phylogenetic residuals, a continuous version of the original ordinal variables). For each plot, only the first two components are visualised. The three dimensions which resulted from each of these analyses was comparable, with one dimension defining the level of political complexity, another the level of dependence on agriculture and settlement strategy, and the third community hierarchy. These are shown in Figures B.2a and B.2b.

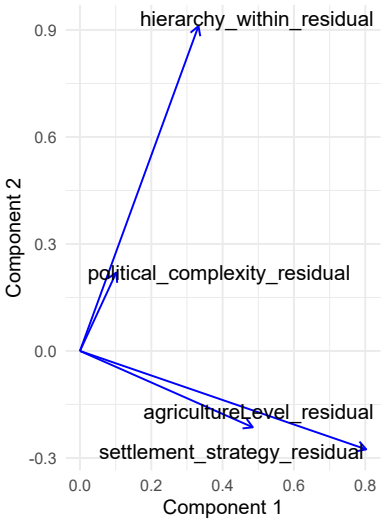
The climate variables were also transformed using a PCA, and the resulting loadings and screeplot (depicting how much variance is explained by each of the dimensions) are shown in Figure B.4a and B.4b.



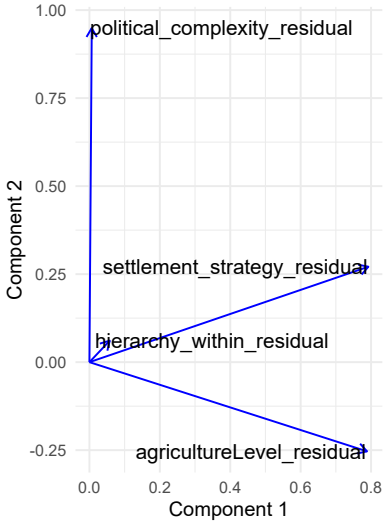
(A) Ordinal PCA loadings.



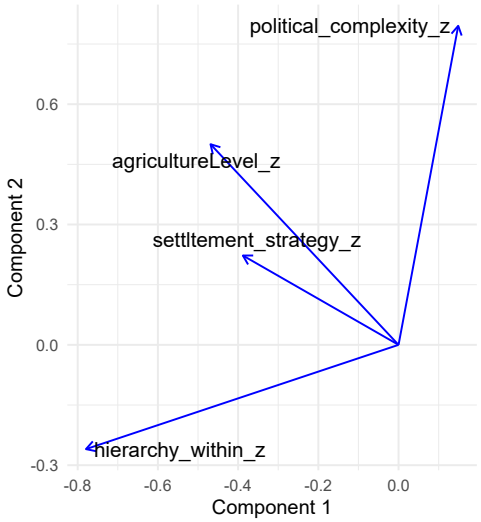
(B) Ordinal PCA loadings with Varimax rotation.



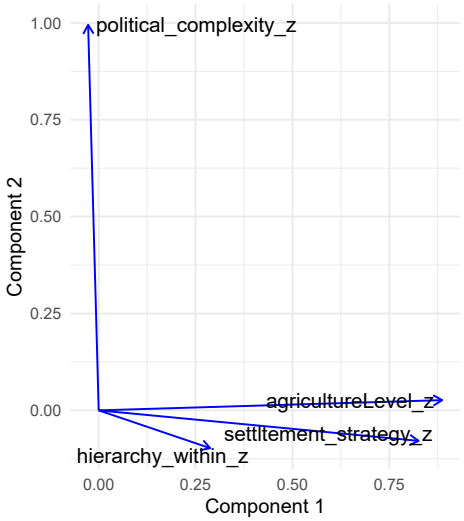
(A) Non-phylogenetic residuals: PCA loadings.



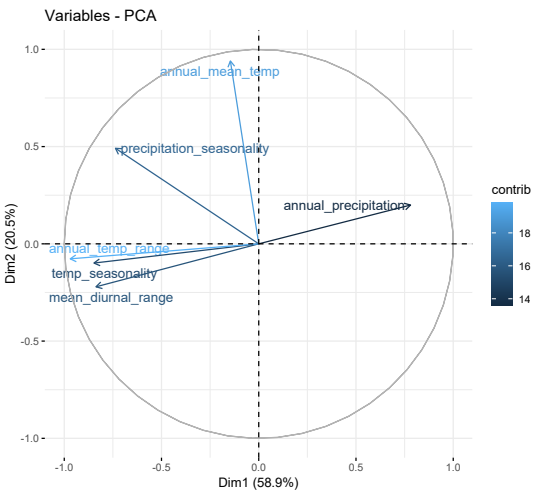
(B) Non-phylogenetic residuals: loadings with Varimax rotation.



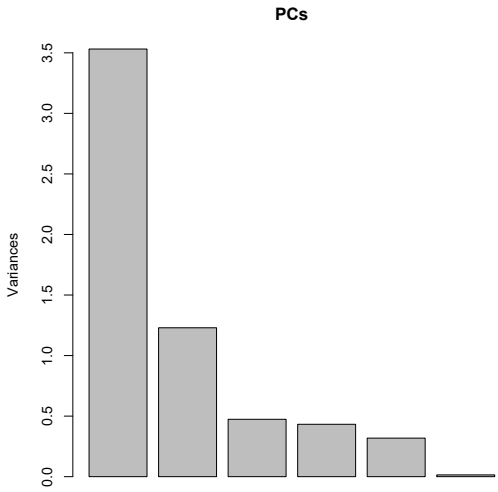
(A) Latent variable z: PCA loadings.



(B) Latent variable z: loadings with Varimax rotation.



(A) Loadings plot for climate PCs.



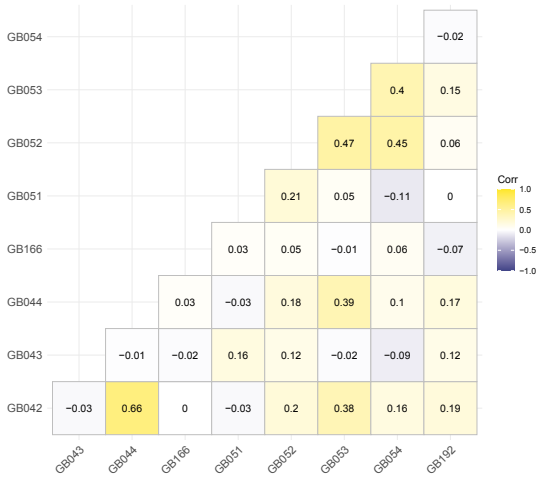
(B) Screeplot for climate variables.

B.2 Chapter 4

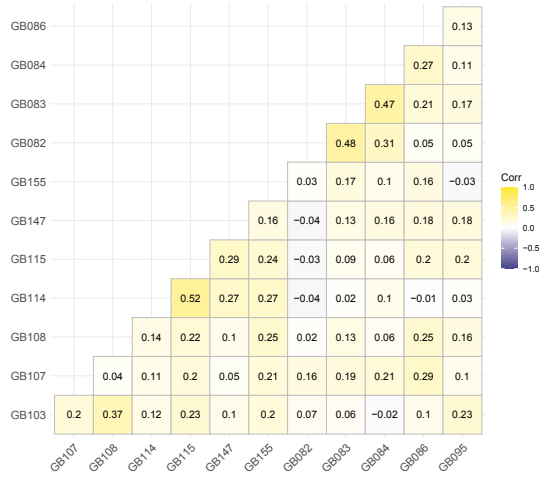
In this section, I will show some additional visualisations from the model presented in Chapter 4. The first, Figure B.5, shows the estimated inter-feature correlations. Low correlations are shaded in purple (dark) while high correlations are shaded in yellow (light). For ease of visualisation, these correlations are presented separately for the nominal domain (which includes nominal number and gender/noun classes), the verbal domain (which includes TAM and bound verbal categories), and word order. Otherwise, the correlations become very difficult to read.

The second set of plots shows the aggregated spatial effects which were not shown in the main text (Figure B.6). These are followed by the remaining individual spatial effects plots for all the features, sorted by domain, with nominal categories in Figure B.7, verbal categories in Figure B.8, and word order features in B.9. High-quality versions of these plots are also available in the code repository for this chapter.

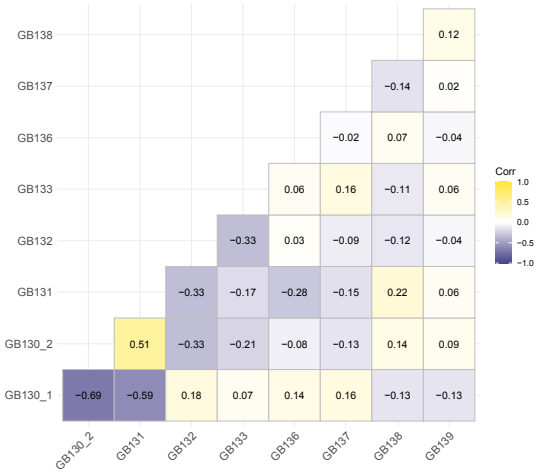
A set of the estimated phylogenetic intercepts is shown in B.10. Not all of them can be shown, as every language has its own intercept for every feature. However, these plots should illustrate how the intercepts vary along the lines of language families, with related languages having more similar intercepts.



(A) Correlations between nominal features.

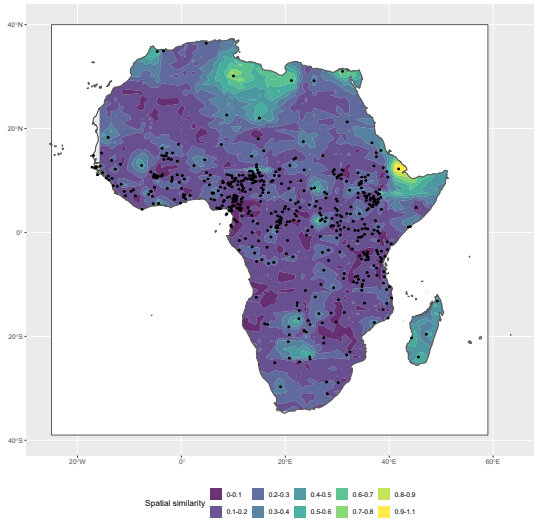


(B) Correlations between verbal features.

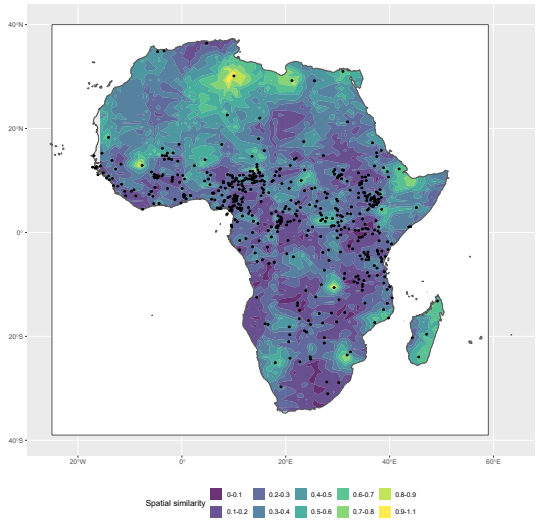


(C) Correlations between word order features.

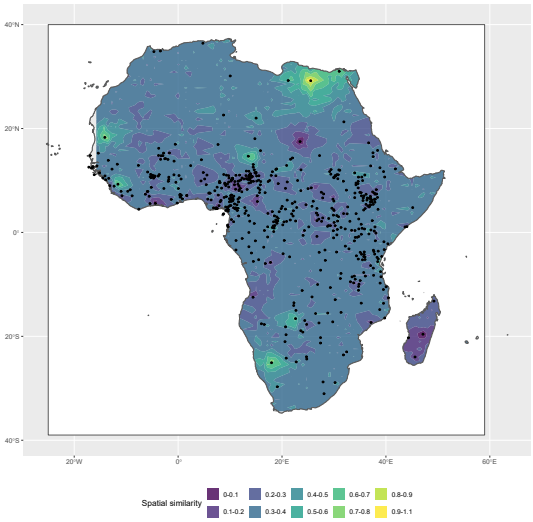
FIGURE B.5: Inter-feature correlations for features in the domains of nominal categories, verbal categories, and word order. Correlations across categories/domains may also be present, but they are presented separately here for ease of visualisation.



(A) Aggregated effects: word order.



(B) Aggregated spatial effects: TAM.



(C) Aggregated spatial effects: number.

FIGURE B.6: Aggregated spatial effects plots for word order, TAM, and nominal number.

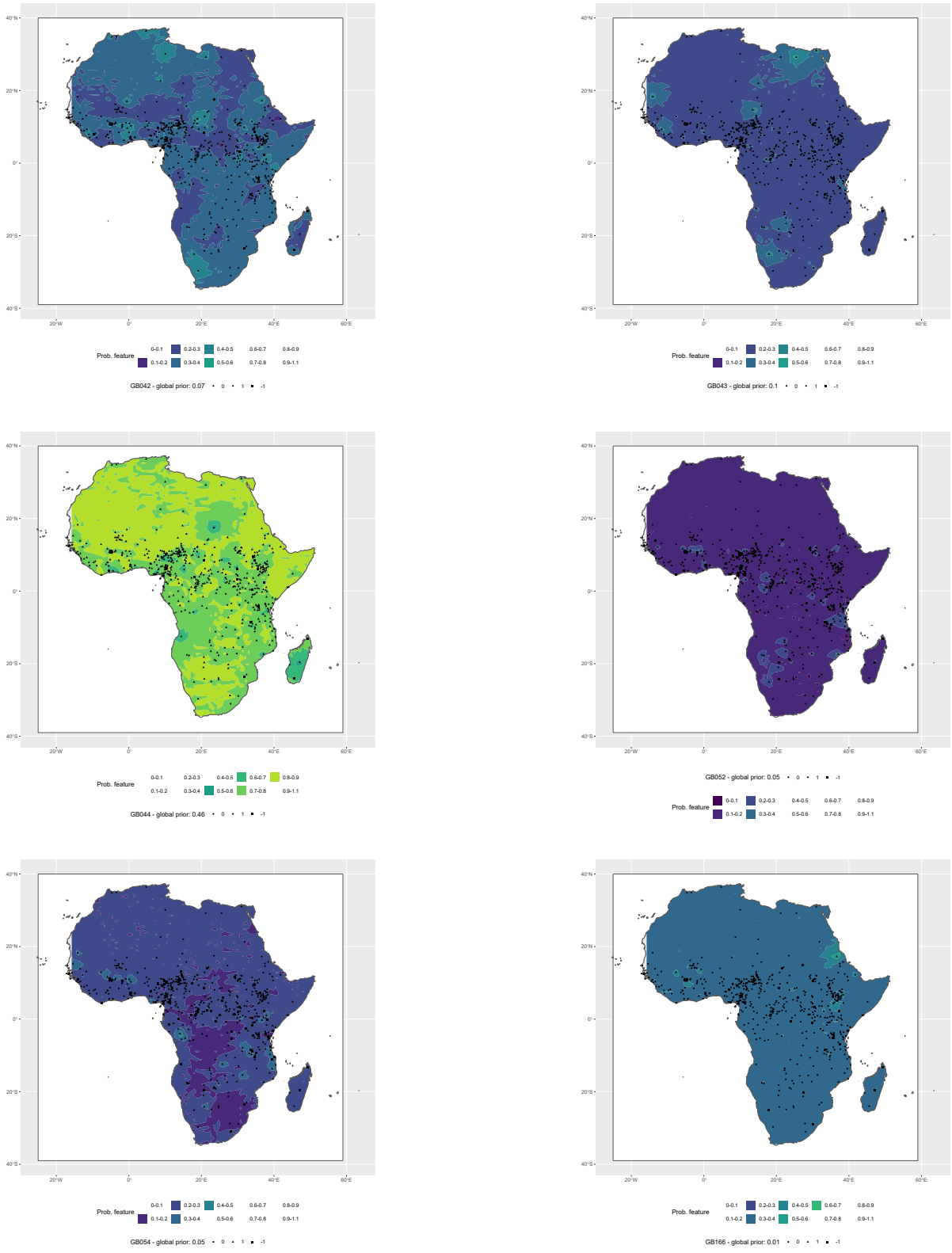


FIGURE B.7: Spatial predictions for the individual features (nominal domain).

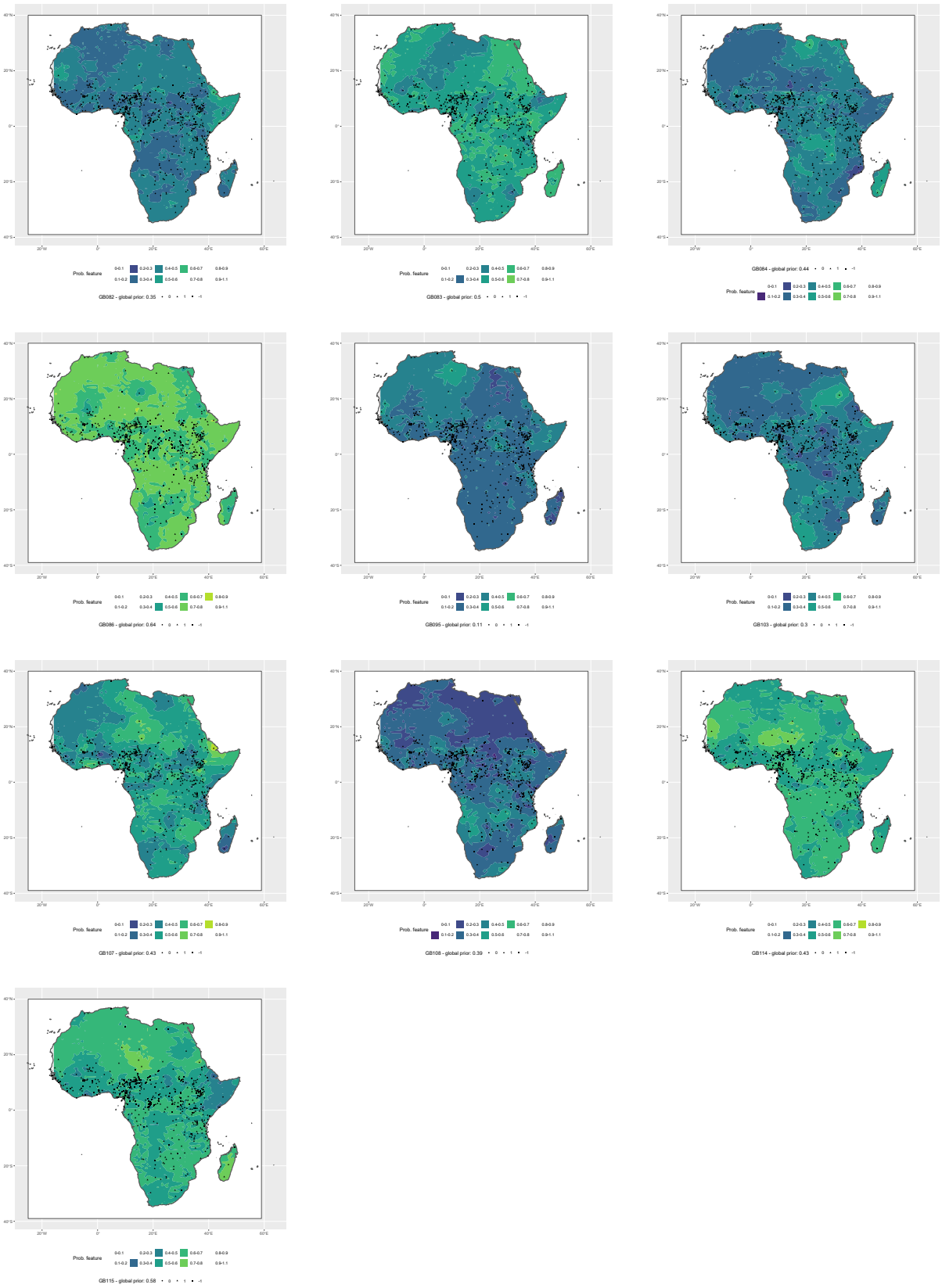


FIGURE B.8: Spatial effects plots (verbal domain).

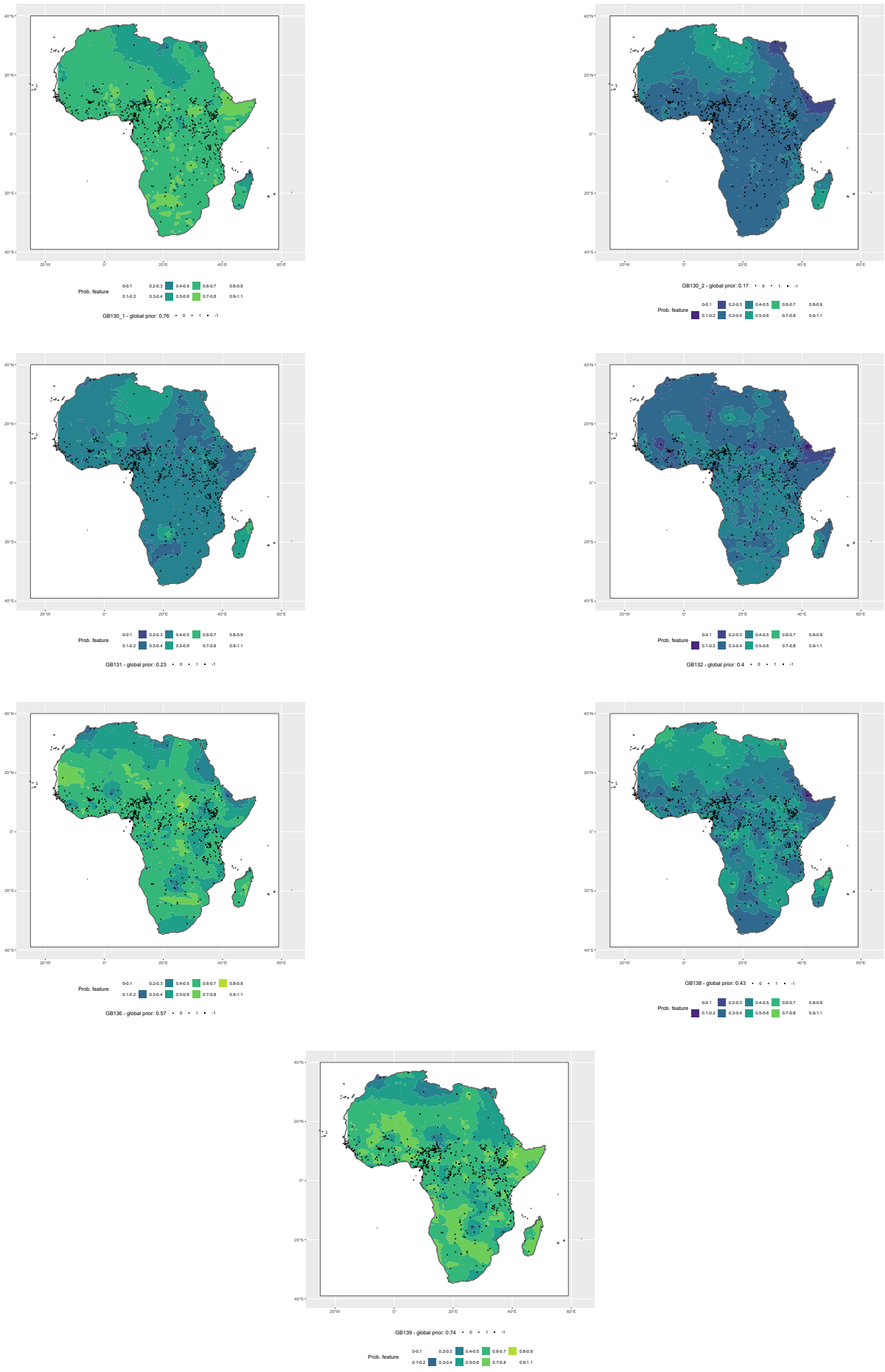
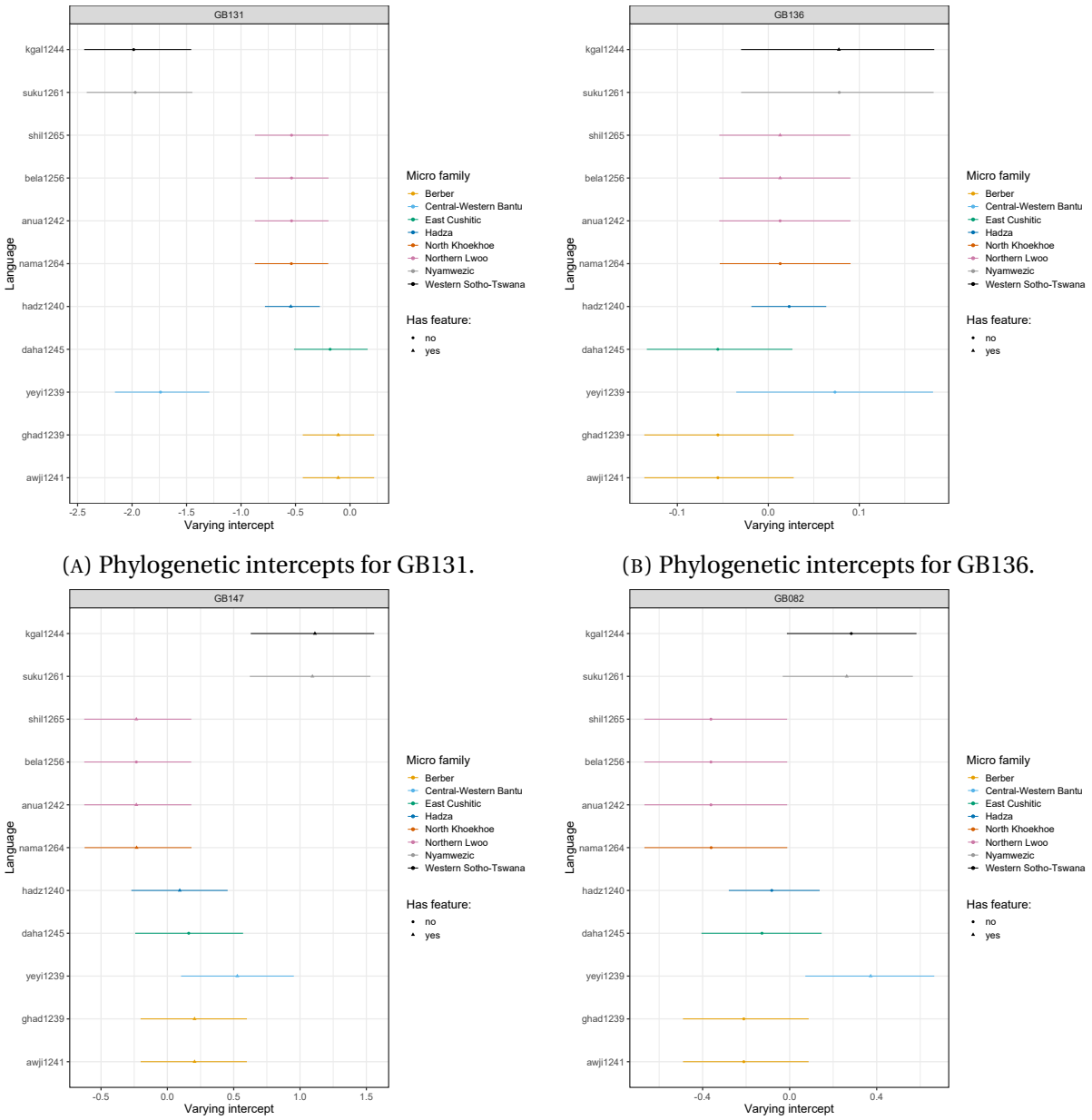


FIGURE B.9: Spatial predictions for the individual features (word order).



(A) Phylogenetic intercepts for GB131.

(B) Phylogenetic intercepts for GB136.

(C) Phylogenetic intercepts for GB147.

(D) Phylogenetic intercepts for GB095.

FIGURE B.10: A collection of phylogenetic intercepts for a sample of diverse languages.

B.3 Chapter 5

For Chapter 5, I did some preliminary analyses of missing values in Grambank (version 1.0) (Skirgård et al., 2023). Using the same method as in the chapter, I included all the languages in Glottolog and coded those languages which are in Grambank as 1 (present), and those which are not in Grambank as 0 (missing). Figure B.11 shows the proportion of missing values across macroareas.

As mentioned in Chapter 2, when we have binary data, the recommended approach to measuring spatial autocorrelation is to use **join-count statistics**, essentially counting how many spatial neighbours share the same value. The measure counts *BB* (a shared value of 1) and *WW* (a shared value of 0) against *BW* (neighbour pairs whose values do not match). These counts are compared to an expected number of matches given random chance. I performed a join-count test of spatial autocorrelation on the missing data in Grambank using a binary spatial weights matrix, with languages within 1000 km coded as neighbours. This allows for the creation of a fully connected graph. A positive z value of 16.6 and a p-value of 0.002 indicates significant positive spatial autocorrelation, i.e., missing languages in Grambank tend to cluster together spatially. This is robust to different neighbourhood specifications. An example of this is shown for the languages of Peru for neighbour pairs within a distance radius of 300 km (since higher distance thresholds make the visualisation becomes harder to interpret) in Figure B.12.

This suggests that spatial controls are important to reduce statistical bias induced by missing data in Grambank as well as WALS.

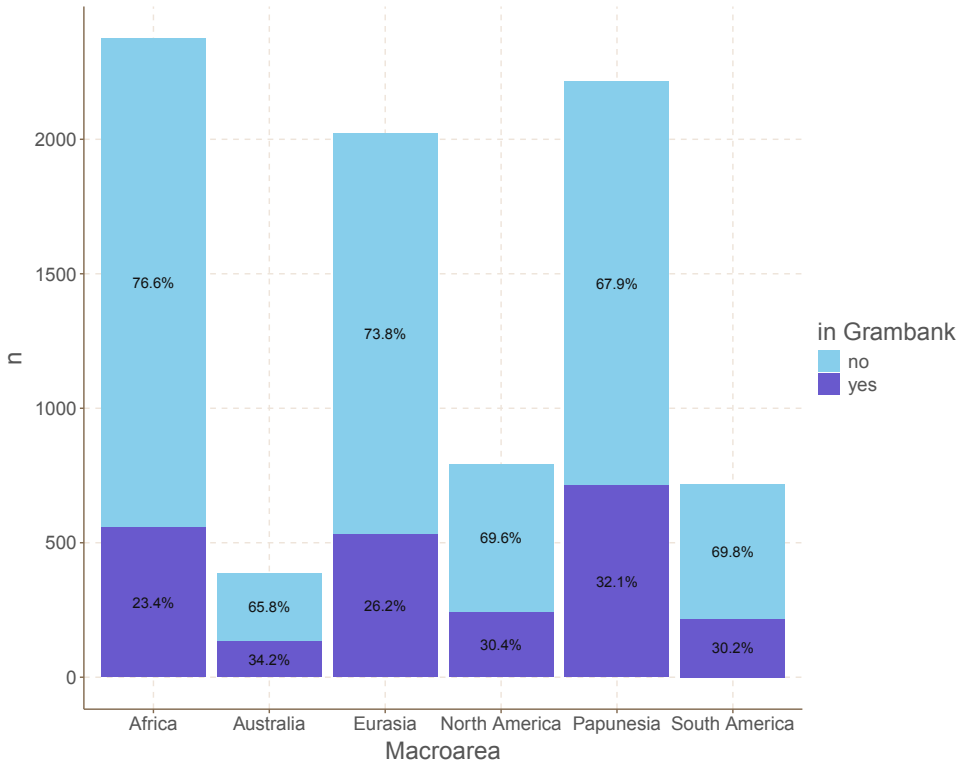


FIGURE B.11: Missing values in Grambank per macroarea.

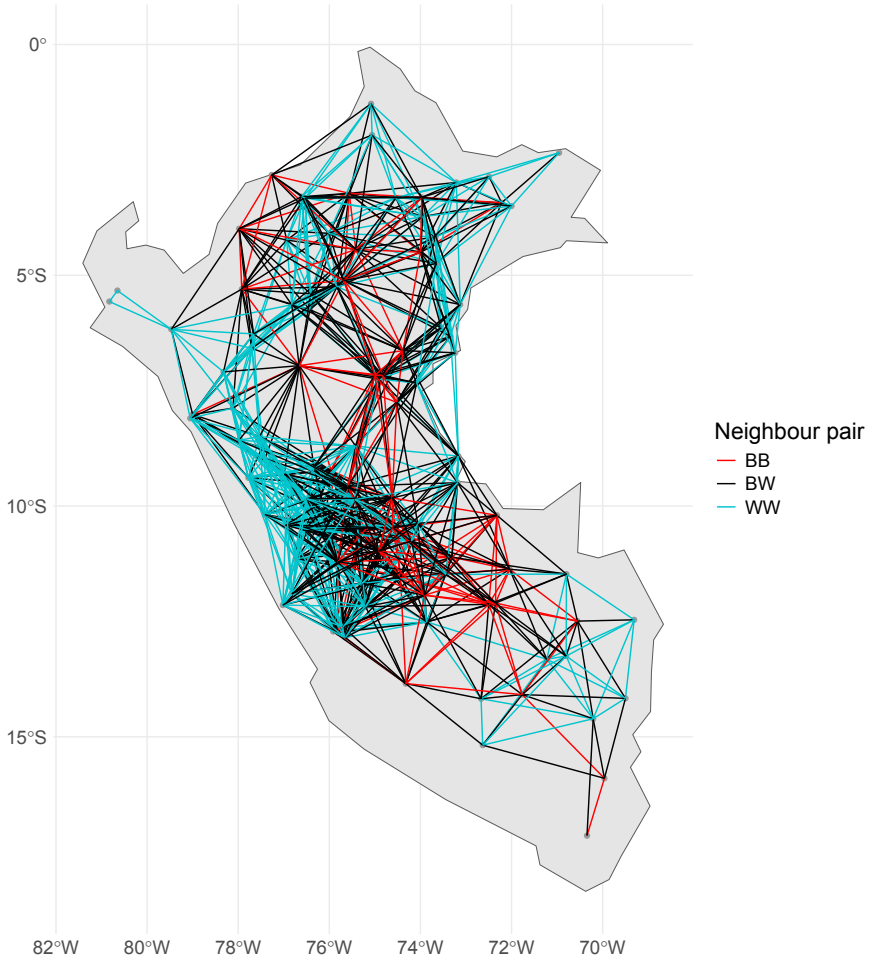


FIGURE B.12: Visualisation of join-count statistic for neighbour pairs in Peru.

Bibliography

Adelaar, Alexander (2010). "The amalgamation of Malagasy". In: *A Journey through Austronesian and Papuan Linguistic and Cultural Space: Papers in Honour of Andrew K Pawley*. Ed. by John Bowden, Nikolaus P. Himmelmann, and Malcolm Ross. Pacific Linguistics.

Aikhenvald, Alexandra (2002). *Language Contact in Amazonia*. New York: Oxford University Press. ISBN: 9780199257850.

Allassonnière-Tang, Marc, Olof Lundgren, Maja Robbers, Sandra Cronhamn, Filip Larsson, One-Soon Her, Harald Hammarström, and Gerd Carling (2021). "Expansion by migration and diffusion by contact is a source to the global diversity of linguistic nominal categorization systems". In: *Humanities and Social Sciences Communications* 8.1, p. 331. ISSN: 2662-9992. DOI: 10.1057/s41599-021-01003-5. URL: https://www.nature.com/articles/s41599-021-01003-5.

Anselin, Luc (1995). "Local Indicators of Spatial Association–Lisa". In: *Georgaphical Analysis* 27.2, pp. 93–115.

Atkinson, Quentin D. (2011). "Phonemic Diversity Supports a Serial Founder Effect Model of Language Expansion from Africa". In: *Science* 332.6027, pp. 346–349. DOI: 10.1126/science.1199295. URL: https://www.science.org/doi/abs/10.1126/science.1199295.

Aurenhammer, Franz and Rolf Klein (2000). *Voronoi Diagrams*. Ed. by J.-R. Sack and J. Urrutia. Amsterdam. DOI: https://doi.org/10.1016/B978-044482537-7/50006-1. URL: https://www.sciencedirect.com/science/article/pii/B9780444825377500061.

Bakker, Dik (2010). "Language sampling". In: *The Oxford handbook of linguistic typology*. Ed. by Jae Jung Song. Oxford: Oxford University Press, pp. 100–127.

Becker, Laura, Matías Guzmán Naranjo, and Samira Ochs (2023). "Socio-linguistic effects on conditional constructions: A quantitative typological study". In: *Sociolinguistic and Typological Perspectives on Language Variation*. Ed. by Silvia Ballarè and Guglielmo Inglese. Berlin, Boston: De Gruyter Mouton, pp. 121–154. ISBN: 9783110781168. DOI: doi:10.1515/9783110781168-005. URL: https://doi.org/10.1515/9783110781168-005.

Bell, Alan (1978). "Language samples". In: *Universals of human language I: Method and theory*. Ed. by Joseph H. Greenberg and Charles Albert Ferguson. Stanford: Stanford University Press, pp. 123–156.

Bentz, Christian, Dan Dediu, Annemarie Verkerk, and Gerhard Jäger (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. URL: http://www.nature.com/articles/s41562-018-0457-6.

Bentz, Christian, Annemarie Verkerk, Douwe Kiela, Felix Hill, and Paula Buttery (2015). "Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms". In: *PLOS ONE* 10.6, e0128254. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0128254.

Besag, Julian (1974). "Spatial Interaction and the Statistical Analysis of Lattice Systems". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 36.2, pp. 192–236. ISSN: 00359246. URL: http://www.jstor.org/stable/2984812.

Bickel, Balthasar (2007). "Typology in the 21st century: major current developments". In: *Linguistic Typology* 11, pp. 239–251.

— (2011). "Statistical modeling of language universals". In: *Linguistic Typology* 15.2, pp. 401–413. DOI: doi:10.1515/lity.2011.027. URL: https://doi.org/10.1515/lity.2011.027.

— (2013). "Distributional biases in language families". In: *Language Typology and Historical Contingency*. Ed. by Balthasar Bickel, Lenore A. Grenoble, David A. Peterson, and Alan Timberlake. John Benjamins Publishing Company, pp. 415–444. ISBN: 9789027270801. DOI: doi:10.1075/tsl.104.19bic. URL: https://doi.org/10.1075/tsl.104.19bic.

— (2017). "Areas and Universals". In: *The Cambridge Handbook of Areal Linguistics*. Ed. by Raymond Hickey. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, pp. 40–55.

— (2020). "Lange and ancient linguistic areas". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken. Oxford: Oxford University Press, pp. 78–93. ISBN: 9780198723813. DOI: 10.1093/oso/9780198723813.003.0007.

Blench, Roger M. (2013). "Why is Africa so linguistically undiverse? Exploring substrates and isolates". In: *Mother Tongue* 18, pp. 43–78.

— (2015). "The Bantoid Languages". In: *Oxford Handbook Topics in Linguistics*. Oxford University Press. ISBN: 9780199935345. DOI: 10.1093/oxfordhb/9780199935345.013.17.

Bostoen, Koen (2020). "The Bantu Expansion: Some facts and fiction". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken. Oxford: Oxford University Press, pp. 227–239. ISBN: 9780198723813. DOI: 10.1093/oso/9780198723813.003.0007.

Bostoen, Koen and Jean-Pierre Donzo (2013). "Bantu-Ubangi language contact and the origin of labial-velar stops in Lingombe (Bantu, C41, DRC)". In: *Diachronica* 30.4, pp. 435–468. DOI: 10.1075/dia.30.4.01bos.

Bouckaert, Remco, David Redding, Oliver Sheehan, Thanos Kyritsis, Russell Gray, Kate Jones, and Quentin Atkinson (2022). *Global language diversification is linked to socio-ecology and threat status*. DOI: 10.31235/osf.io/f8tr6.

Bowern, C. and Claire Atkinson (2012). "Computational phylogenetics and the internal structure of Pama-Nyungan". In: *Language* 88.4. DOI: 10.1353/LAN.2012.0081.

Bromham, Lindell, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua (2022). "Global predictors of language endangerment and the future of linguistic diversity". In: *Nat Ecol Evol* 6.2, pp. 163–173. ISSN: 2397-334x. DOI: 10.1038/s41559-021-01604-y.

Brown, Cecil H., Eric W. Holman, Søren Wichmann, and Viveka Velupillai (2008). "Automated classification of the worlds languages: a description of the method and preliminary results". In: *Language Typology and Universals* 61.4, pp. 285–308. DOI: doi:10.1524/stuf.2008.0026. URL: https://doi.org/10.1524/stuf.2008.0026.

Bürkner, Paul-Christian (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan". In: *Journal of Statistical Software* 80.1, pp. 1–28. DOI: 10.18637/jss.v080.i01.

Campbell, Lyle (2017). "Why is it so Hard to Define a Linguistic Area?" In: *The Cambridge handbook of areal linguistics*. Ed. by Raymond Hickey. Cambridge: Cambridge University Press, pp. 19–39.

Caron, Bernard and Petr Zima, eds. (2006a). *Sprachbund in the West African Sahel*. Afrique et Langage 11. Leeuven/Paris: Peeters.

"TAM verbal paradigms in the West African Sahel as areal (Sprachbund), genetic and sociolinguistic features (Where are we 75 years after Klingenheben?)" (2006b). In: *Sprachbund in the West African Sahel*. Ed. by Bernard Caron and Petr Zima. Afrique et Langage 11. Leeuven/Paris: Peeters, pp. 221–237.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). "Stan: A Probabilistic Programming Language". In: *Journal of Statistical*

Software, Articles 76.1, pp. 1–32. ISSN: 1548-7660. DOI: 10.18637/jss.v076.i01. URL: https://www.jstatsoft.org/v076/i01.

Cathcart, Chundra, Gerd Carling, Filip Larsson, Niklas Erben Johansson, and Erich Round (2018). "Areal pressure in grammatical evolution: An Indo-European case study". In: *Diachronica* 35, pp. 1–34. DOI: 10.1075/dia.16035.cat.

Childs, G. Tucker (2010). "Language Contact in Africa: A Selected Review". In: *The Handbook of Language Contact*. John Wiley & Sons, Ltd, pp. 695–713. ISBN: 9781444318159. DOI: 10.1002/9781444318159.ch34.

Chun, Yongwan and David A. Griffith (2013). *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Thousand Oaks, CA: Sage.

Clements, G. N. and Annie Rialland (2008). "Africa as a phonological area". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 3685.

Cliff, Andrew David and Keith J. Ord (1981). *Spatial processes: models and applications*. London: Taylor & Francis.

Cogneau, Denis and Yannick Dupraz (2014). "Questionable Inference on the Power of Pre-Colonial Institutions in Africa". In: URL: https://shs.hal.science/halshs-01018548.

Comrie, Bernard (2008). "Linguistic Diversity in the Caucasus". In: *Annual Review of Anthropology* 37, pp. 131–143. ISSN: 1545-4290. DOI: https://doi.org/10.1146/annurev.anthro.35.081705.123248. URL: https://www.annualreviews.org/content/journals/10.1146/annurev.anthro.35.081705.123248.

Crass, Joachim and Ronny Meyer (2007). "Ethiopia". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 228–250. ISBN: 978-0-521-87611-7. DOI: 10.1017/CBO9780511486272.008.

Creissels, Denis, Gerrit J. Dimmendaal, Zygmunt Frajzyngier, and Christa König (2008). "Africa as a morphosyntactic area". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 86150.

Cressie, Noel A. C. (1994). *Statistics for Spatial Data*. John Wiley & Sons. ISBN: 9780471002550.

Currie, T. E. and R. Mace (2009). "Political complexity predicts the spread of ethnolinguistic groups". In: *Proceedings of the National Academy of Sciences* 106.18, pp. 7339–7344. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0804698106. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0804698106.

Cysouw, Michael (2005). "Quantitative methods in typology". In: *Quantitative Linguistik: ein internationales Handbuch = Quantitative linguistics*. Ed. by Reinhard Koehler, Gabriel Altmann, and Rajmund G. Piotrowski. Berlin: De Gruyter, pp. 554–578.

— (2011). "Understanding transition probabilities". In: *Linguistic Typology* 15.2, pp. 415–431. DOI: https://doi.org/10.1515/lity.2011.028.

Dahl, Osten (2001). "Principles of areal typology". In: *Language Typology and Language Universals*. Ed. by Martin Haspelmath, König Ekkehard, Wulf Oesterreicher, and Wolfgang Raible. Vol. 2. Berlin, Boston: De Gruyter Mouton, pp. 1456–1470. ISBN: 9783110194265. DOI: doi:10.1515/9783110194265-042. URL: https://doi.org/10.1515/9783110194265-042.

Dahl, Otto Christian (1988). "Bantu substratum in Malagasy". In: *Études Océan Indien* 9, pp. 91–132.

Dediu, Dan (2011). "A Bayesian phylogenetic approach to estimating the stability of linguistic features and the genetic biasing of tone". In: *Proceedings of the Royal Society B: Biological Sciences* 278.1704, pp. 474–479. DOI: 10.1098/rspb.2010.1595. URL: https://royalsocietypublishing.org/doi/10.1098/rspb.2010.1595.

Dediu, Dan and Michael Cysouw (2013). "Some Structural Aspects of Language Are More Stable than Others: A Comparison of Seven Methods". In: *PLoS ONE* 8.1. Ed. by John P. Hart, e55009. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0055009. URL: https://dx.plos.org/10.1371/journal.pone.0055009.

Dediu, Dan and D. Robert Ladd (2007). "Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin". In: *Proceedings of the National Academy of Sciences* 104.26, pp. 10944–10949. DOI: 10.1073/pnas.0610848104. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0610848104.

Dediu, Dan, Scott R. Moisik, W. A. Baetsen, Abel Marinus Bosman, and Andrea L. Waters-Rist (2021). "The vocal tract as a time machine: inferences about past speech and language from the anatomy of the speech organs". en. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 376.1824, rstb.2020.0192, 20200192. ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2020.0192. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.2020.0192.

Dellert, Johannes (2019). *Information-theoretic causal inference of lexical flow*. Language Variation 4. Berlin: Language Science Press. DOI: 10.5281/zenodo.3247415.

Dellert, Johannes, Thora Daneyko, Alla Münch, Alina Ladygina, Armin Buch, Natalie Clarius, Ilja Grigorjew, Mohamed Balabel, Isabella Boga, Zalina Baysarova, Roland Mühlenbernd, Gerhard Jäger, and Johannes Wahle (2019). "NorthEuraLex 0.9". In:

Lang Resources & Evaluation. URL: https://doi.org/10.1007/s10579-019-09480-6.

Di Carlo, Pierpaolo, Angiachi Demetris Esene Agwara, and Rachel Ojong Diba (2020). "Multilingualism and the heteroglossia of ideologies in Lower Fungom (Cameroon)". In: *Sociolinguistic Studies* 14.3, pp. 321–345. DOI: 10.1558/sols.38799. eprint: https://doi.org/10.1558/sols.38799. URL: https://doi.org/10.1558/sols.38799.

Di Carlo, Pierpaolo and Jeff Good (2023). "Language contact or linguistic micro-engineering? Feature pools, social semiosis, and intentional language change in the Cameroonian Grassfields". In: *Linguistic Typology at the Crossroads*, 72–125 Pages. DOI: 10.6092/ISSN.2785-0943/17231. URL: https://typologyatcrossroads.unibo.it/article/view/17231.

Diamond, Jared and Peter Bellwood (2003). "Farmers and Their Languages: The First Expansions". In: *Science* 300.5619, pp. 597–603. DOI: doi:10.1126/science.1078208. URL: https://www.science.org/doi/abs/10.1126/science.1078208.

Dimmendaal, Gerrit J. (2001). "Language Shift and Morphological Convergence in the Nilotic Area". In: *Sprache und Geschichte in Afrika* 16, pp. 83–124.

— (2008a). "Africa's verb-final languages". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 272–308.

— (2008b). "Language Ecology and Linguistic Diversity on the African Continent". In: *Language and Linguistics Compass* 2, pp. 840–858. DOI: 10.1111/j.1749-818X.2008.00085.x.

— (2020). "Language diversification and contact in Africa". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken. Oxford: Oxford University Press, pp. 210–226. ISBN: 9780198723813. DOI: 10.1093/oso/9780198723813.003.0007.

— (2021). "The Comparative Method and Language Change in Accretion Zones: A View from the Nuba Mountains". In: *Historical Linguistics and Endangered Languages*. Ed. by Patience Epps, Danny Law, and Na'ama Pat-El. New York: Routledge. Chap. 7. DOI: https://doi.org/10.4324/9780429030390.

Dimmendaal, Gerrit J., Mily Crevels, and Peter Muysken (2020). "Patterns of dispersal and diversification in Africa". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken. Oxford: Oxford University Press, pp. 197–209. ISBN: 9780198723813. DOI: 10.1093/oso/9780198723813.003.0007.

Donegan, Connor (2022). "geostan: An R package for Bayesian spatial analysis". In: *The Journal of Open Source Software* 7.79, p. 4716. DOI: 10.21105/joss.04716.

Donegan, Connor, Yongwan Chun, and Daniel A. Griffith (2021). "Modeling community health with areal data: Bayesian inference with survey standard errors and spatial structure". In: *Int J Env Res Public Health* 18.13, p. 6856. DOI: 10.3390/ijerph18136856.

Dryer, Matthew and Martin Haspelmath (2022). *The World Atlas of Language Structures Online (v2020.3)*. URL: https://doi.org/10.5281/zenodo.7385533.

Dryer, Matthew S. (1989). "Large Linguistic Areas and Language Sampling". In: *Studies in Language. International Journal sponsored by the Foundation Foundations of Language* 13.2, pp. 257–292.

Dryer, Matthew S. and Martin Haspelmath (2013). *The World Atlas of Language Structures Online*. Munich: Max Planck Digital Library. URL: https://wals.info/.

Dunn, Michael, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray (2011). "Evolved structure of language shows lineage-specific trends in word-order universals". en. In: *Nature* 473.7345, pp. 79–82. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09923. URL: http://www.nature.com/articles/nature09923.

Eberhard David M., Gary F. Simons and Charles D. Fennig, eds. (2023). *Ethnologue: Languages of the World. Twenty-seventh edition*. Dallas, Texas: SIL International. URL: http://www.ethnologue.com.

Edelsten, Peter, Hannah Gibson, Rozenn Guérois, Gastor Mapunda, Lutz Marten, and Julius Taji (2022). "Morphosyntactic variation in Bantu: Focus on East Africa". In: *Journal of the Language Association of Eastern Africa* 1.1, pp. 1–22. URL: https://hal.science/hal-03924991v1.

Elhorst, Paul J. and Solmaria Halleck Vega (2017). "The SLX model: Extensions and the sensitivity of spatial spillovers to W (English translation)". In: *Papeles de Economía Española* 152, pp. 34 –50.

Epps, Patience (2020). "Amazonian Linguistic Diversity and Its Sociocultural Correlates". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken. Oxford: Oxford University Press, pp. 275–290.

Evans, Nicholas (2010). *Dying Words: Endangered Languages and What They Have to Tell Us*. Ed. by David Crystal and Nicholas Evans. Chichester, U.K.; Malden, MA: Wiley-Blackwell. ISBN: 9780631233053. DOI: 10.1002/9781444310450.

— (2017). "Did language evolve in multilingual settings?" In: *Biology & Philosophy* 32, pp. 905–933. DOI: 10.1007/s10539-018-9609-3.

— (2020). "Time, diversification, and dispersal on the Australian continent: Three enigmas of linguistic prehistory". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken. Oxford: Oxford University Press. ISBN: 9780198723813. DOI: 10.1093/oso/9780198723813.003.0007.

Evans, Nicholas and Stephen C. Levinson (2009). "The myth of language universals: Language diversity and its importance for cognitive science". In: *Behavioral and Brain Sciences* 32.5, pp. 429–448. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X0999094X. URL: https://www.cambridge.org/core/product/identifier/S0140525X0999094X/type/journal_article.

Everett, Caleb (2013). "Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives". In: *PLOS ONE* 8.6, pp. 1–10. DOI: 10.1371/journal.pone.0065275. URL: https://doi.org/10.1371/journal.pone.0065275.

Ferguson, Charles (1976). "The Ethiopian Language Area". In: *Language in Ethiopia*. Ed. by M. Lionel Bender, J. Donald Bowen, Robert Cooper, and Charles Ferguson. Oxford University Press, pp. 63–76.

Gelfand, Alan E. and Erin M. Schliep (2016). "Spatial statistics and Gaussian processes: A beautiful marriage". In: *Spatial Statistics* 18. Spatial Statistics Avignon: Emerging Patterns, pp. 86–104. ISSN: 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2016.03.006. URL: https://www.sciencedirect.com/science/article/pii/S2211675316300033.

Gerardi, Fabrício Ferraz, Stanislav Reichert, and Carolina Coelho Aragon (2021). "TuLeD (Tupían lexical database): introducing a database of a South American language family". In: *Language Resources and Evaluation* 55.4, pp. 997–1015. ISSN: 1574-0218. DOI: 10.1007/s10579-020-09521-5. URL: https://doi.org/10.1007/s10579-020-09521-5.

Getis, Arthur and J. K. Ord (1992). "The Analysis of Spatial Association by Use of Distance Statistics". In: *Geographical Analysis* 24.3, pp. 189–206. DOI: https://doi.org/10.1111/j.1538-4632.1992.tb00261.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1992.tb00261.x.

Gifi, Albert (1990). *Nonlinear Multivariate Analysis*. New York: Wiley.

Good, Jeff (2013). "A (micro-)accretion zone in a remnant zone? Lower Fungom in areal-historical perspective". In: *Language typology and historical contingency*. Ed. by Balthasar Bickel, Lenore A. Grenoble, David A. Peterson, and Alan Timberlake. Amsterdam: John Benjamins, pp. 265–282.

— (2025). "Historical morphosyntax and syntactic change". In: *The Oxford guide to the Bantu languages*. Ed. by Lutz Marten, Nancy C. Kula, Jochen Zeller, and Ellen Hurst. Oxford: Oxford University Press.

Grafen, A. (1989). "The phylogenetic regression". In: *Philosophical Transactions of the Royal society of London. Series B. Biological Sciences* 326, pp. 119–157.

Graff, Anna, Damián E. Blasi, Erik J. Ringen, Vladimir Bajić, Daphné Bavelier, Kentaro K. Shimizu, Brigitte Pakendorf, Chiara Barbieri, and Balthasar Bickel (2024a). "Global patterns of genetic admixture reveal effects of language contact

(preprint)". In: *bioRxiv*. DOI: 10.1101/2024.12.19.629340. eprint: https://www.biorxiv.org/content/early/2024/12/20/2024.12.19.629340.full.pdf. URL: https://www.biorxiv.org/content/early/2024/12/20/2024.12.19.629340.

Graff, Anna, Erik J. Ringen, Taras Zakharko, Mark Stoneking, Kentaro K. Shimizu, Balthasar Bickel, and Chiara Barbieri (2024b). "An inverse correlation between linguistic and genetic diversity (preprint)". In: *bioRxiv*. DOI: 10.1101/2024.12.18.628602. eprint: https://www.biorxiv.org/content/early/2024/12/20/2024.12.18.628602.full.pdf. URL: https://www.biorxiv.org/content/early/2024/12/20/2024.12.18.628602.

Gray, Joseph Patrick (1999). "A corrected ethnographic atlas". In: *World Cultures* 10, pp. 24–85.

Greenberg, Joseph H. (1959). "Africa as a linguistic area". In: *Continuity and change in African cultures*. Ed. by William R. Bascom and. Melville J. Herskovitz. Chicago: Chicago University Press, pp. 15–27.

— (1963). *The Languages of Africa*. The Hague: Mouton.

— (1966). *Universals of language*. Cambridge, MA: MIT press.

— (1983). "Some Areal Characteristics of African Languages". In: *Current Approaches to African Linguistics*. Ed. by Ivan R. Dihoff. Vol. 1. Berlin, Boston: De Gruyter, pp. 1–22. ISBN: 9783112420065. DOI: doi:10.1515/9783112420065-002. URL: https://doi.org/10.1515/9783112420065-002.

Greenhill, Simon J., Quentin D. Atkinson, Andrew Meade, and Russell D. Gray (2010). "The shape and tempo of language evolution". en. In: *Proceedings of the Royal Society B: Biological Sciences* 277.1693, pp. 2443–2450.

Grollemund, Rebecca, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel (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. DOI: 10.1073/pnas.1503793112. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1503793112.

Guzmán Naranjo, Matías and Laura Becker (2021). "Statistical Bias Control in Typology". In: *Linguistic Typology*. DOI: doi:10.1515/lingty-2021-0002. URL: https://doi.org/10.1515/lingty-2021-0002.

Guzmán Naranjo, Matías, Laura Becker, Miriam L. Schiele, and I-Ying Lin (2025). "Why modelling space is hard: no evidence for a serial founder effect in Polynesian phoneme inventories". In: *Linguistics*. DOI: doi:10.1515/ling-2024-0016. URL: https://doi.org/10.1515/ling-2024-0016.

Guzmán Naranjo, Matías and Gerhard Jäger (2022). *Euclide, the crow, the wolf and the pedestrian: distance methods for spatial typology*. Presentation at ALT 2022, Austin Texas.

Guzmán Naranjo, Matías and Miri Mertner (2022). "Estimating areal effects in typology: A case study of African phoneme inventories". In: *Linguistic Typology*. DOI: https://doi.org/10.1515/lingty-2022-0037.

Guzmán Naranjo, Matías, Miri Mertner, and Matthias Urban (2024). "Spatial effects with missing data". In: *Open Linguistics* 10.1, p. 20240032. DOI: doi:10.1515/opli-2024-0032. URL: https://doi.org/10.1515/opli-2024-0032.

Güldemann, Tom (1998). "The Kalahari Basin as an object of areal typology - a first approach". In: *Language, identity, and conceptualization among the Khoisan*. Ed. by Mathias Schladt. Quellen zur Khoisan-Forschung. Rüdiger Köppe, pp. 137–169.

— (2008). "The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 151–185.

— (2018a). "Historical linguistics and genealogical language classification in Africa". In: *The Languages and Linguistics of Africa*. Ed. by Tom Güldemann. Berlin, Boston: De Gruyter Mouton, pp. 58–444. ISBN: 9783110421668. DOI: doi:10.1515/9783110421668-002. URL: https://doi.org/10.1515/9783110421668-002.

— (2018b). "Language contact and areal linguistics in Africa". In: *The Languages and Linguistics of Africa*. Ed. by Tom Güldemann, pp. 445–545. ISBN: 9783110421668. DOI: 10.1515/9783110421668-003.

Güldemann, Tom and Edward D. Elderkin (2010). "On external genealogical relationships of the Khoe family". In: *Khoisan languages and linguistics: Proceedings of the 1st International Symposium, January 4-8, 2003, Riezlern/Kleinwalsertal*. Ed. by M. Brenzinger and C. König. Quellen zur Khoisan-Forschung. Rüdiger Köppe, pp. 15–52.

Güldemann, Tom and Anne-Maria Fehn, eds. (2014). *Beyond Khoisan: historical relations in the Kalahari Basin*. Current Issues in Linguistic Theory 330. Amsterdam: John Benjamins.

Güldemann, Tom and Rainer Vossen (2000). "Khoisan". In: *African languages: an introduction*. Ed. by Bernd Heine and Derek Nurse. Cambridge University Press, pp. 99 –122.

Hammarström, Harald (2018). "1. A survey of African languages". In: *The Languages and Linguistics of Africa*. Ed. by Tom Güldemann. Berlin, Boston: De Gruyter Mouton, pp. 1–57. ISBN: 9783110421668. DOI: doi:10.1515/9783110421668-001. URL: https://doi.org/10.1515/9783110421668-001.

Hammarström, Harald and Mark Donohue (2014). "Some Principles on the Use of Macro-Areas in Typological Comparison". In: *Language Dynamics and Change*

4.1, pp. 167–187. ISSN: 2210-5824, 2210-5832. DOI: 10.1163/22105832-00401001. URL: https://brill.com/view/journals/ldc/4/1/article-p167_5.xml.

Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank (2023). *Glottolog 4.8*. Leipzig: Max Planck Institute for Evolutionary Anthropology. DOI: https://doi.org/10.5281/zenodo.8131084. URL: http://glottolog.org.

Harke, Franz H., Miryam S. Merk, and Philipp Otto (2022). "Estimation of Asymmetric Spatial Autoregressive Dependence on Irregular Lattices". In: *Symmetry* 14.7. ISSN: 2073-8994. DOI: 10.3390/sym14071474. URL: https://www.mdpi.com/2073-8994/14/7/1474.

Hartmann, Frederik (2021). *Germanic phylogeny: a computational investigation using Bayesian inference and agent-based models (PhD dissertation)*. Universität Konstanz.

Haspelmath, Martin (2001). "The European linguistic area: Standard Average European". In: *Language Typology and Language Universals*. Ed. by Martin Haspelmath. Berlin: De Gruyter Mouton, pp. 1492–1510. ISBN: 978-3-11-019426-5. DOI: doi:10.1515/9783110194265-044. URL: 10.1515/9783110194265-044.

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

Haudricourt, André-George and Pascal Dibie (1987). *Les pieds sur terre*. Métailié.

Heggarty, Paul, Warren Maguire, and April McMahon (2010). "Splits or waves? Trees or webs? How divergence measures and network analysis can unravel language histories". In: *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 365.1559, pp. 38293843. DOI: https://doi.org/10.1098/rstb.2010.0099.

Heine, Bernd (1975). "Language typology and convergence areas in Africa". In: *Linguistics* 144, pp. 27–47.

— (1976). *A typology of African languages based on the order of meaningful elements*. Kölner Beiträge zur Afrikanistik. Berlin: Dietrich Reimer.

Heine, Bernd and Tania Kuteva (2003). "On contact-induced grammaticalization". In: *Studies in language: International Journal Sponsored by the Foundation "Foundations of Language"* 27, pp. 529–572. DOI: 10.1075/sl.27.3.04hei.

Heine, Bernd and Derek Nurse (2007). "Introduction". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 1–14. ISBN: 978-0-521-87611-7. DOI: 10.1017/CBO9780511486272.008.

— eds. (2008). *A Linguistic Geography of Africa*. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press.

Herbst, Jeffrey (2000). *States and Power in Africa: Comparative Lessons in Authority and Control*. STU - Student edition, 2. Princeton University Press. ISBN: 9780691164144. URL: http://www.jstor.org/stable/j.ctt9qh05m.

Hickey, Raymond (2010). "Language Contact: Reconsideration and Reassessment". In: *The Handbook of Language Contact*. John Wiley & Sons, Ltd, pp. 1–28. ISBN: 9781444318159. DOI: https://doi.org/10.1002/9781444318159.ch.

Holman, Eric W., Søren Wichmann, Cecil H. Brown, Viveka Velupillai, André Müller, and Dik Bakker (2008). "Explorations in automated language classification". In: *Folia Linguistica* 42.3-4, pp. 331–354. DOI: doi:10.1515/FLIN.2008.331. URL: https://doi.org/10.1515/FLIN.2008.331.

Hua, Xia, Simon J. Greenhill, Marcel Cardillo, Hilde Schneemann, and Lindell Bromham (2019). "The ecological drivers of variation in global language diversity". In: *Nature Communications* 10.1, p. 2047. ISSN: 2041-1723. DOI: 10.1038/s41467-019-09842-2. URL: http://www.nature.com/articles/s41467-019-09842-2.

Huisman, John L. A., Asifa Majid, and Roeland van Hout (2019). "The geographical configuration of a language area influences linguistic diversity". In: *PLOS ONE* 14.6. Ed. by Richard A. Blythe, e0217363. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0217363. URL: https://dx.plos.org/10.1371/journal.pone.0217363.

Idiatov, Dmitry (2018). "An areal typology of clause-final negation in Africa". In: *Aspects of linguistic variation*. Ed. by Daniël Van Olmen, Tanja Mortelmans, and Frank Brisard. De Gruyter Mouton, pp. 115–163. DOI: 10.1515/9783110607963-005.

Idiatov, Dmitry and Mark Van de Velde (2021). "The lexical distribution of labial-velar stops is a window into the linguistic prehistory of Northern Sub-Saharan Africa". In: *Language* 97.1, pp. 72–107. ISSN: 1535-0665. DOI: 10.1353/lan.2021.0002. URL: https://muse.jhu.edu/article/785540.

Jaeger, T. Florian, Peter Graff, William Croft, and Daniel Pontillo (2011). "Mixed Effect Models for Genetic and Areal Dependencies in Linguistic Typology". In: *Linguistic Typology* 15.2, pp. 281–319. URL: https://www.degruyter.com/view/journals/lity/15/2/article-p281.xml.

Joseph, Brian D. (2010). "Language Contact in the Balkans". In: *The Handbook of Language Contact*. Ed. by Raymond Hickey. John Wiley
Sons, Ltd, pp. 618–633. ISBN: 978-1-4443-1815-9. DOI: 10.1002/9781444318159.ch30. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444318159.ch30.

Josserand, Mathilde, E. Meeussen, A. Majid, and Dan Dediu (2021). "Environment and culture shape both the colour lexicon and the genetics of colour perception". In: *Scientific reports* 11.1. DOI: https://doi.org/10.1038/s41598-021-98550-3.

Jäger, Gerhard (2013). "Phylogenetic Inference from Word Lists Using Weighted Alignment with Empirically Determined Weights". In: *Language Dynamics and Change* 3, pp. 245–291. DOI: 10.1163/22105832-13030204.

— (2018). "Global-scale phylogenetic linguistic inference from lexical resources". In: *Scientific Data* 5.1, p. 180189. ISSN: 2052-4463. DOI: 10.1038/sdata.2018.189. URL: http://www.nature.com/articles/sdata2018189.

Jäger, Gerhard and Johannes Wahle (2021). "Phylogenetic Typology". In: *Frontiers in Psychology* 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.682132. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.682132.

Kaiping, Gereon (2022). "A network for simulating pre-colonial migration in the Americas". In: *GIScience 2021 Short Paper Proceedings: UC Santa Barbara, Center for Spatial Studies*. URL: http://dx.doi.org/10.25436/E21598.

Kaiser, Henry F. (1958). "The Varimax Criterion for Analytic Rotation in Factor Analysis". In: *Psychometrika* 23.3, pp. 187200. DOI: 10.1007/BF02289233.

Kaplan, Judith (2017). "From lexicostatistics to lexomics: Basic vocabulary and the study of language prehistory". In: *OSIRIS* 31.1, pp. 202–223. DOI: https://doi.org/10.1086/694093.

Kauhanen, Henri, Deepthi Gopal, Tobias Galla, and Ricardo Bermúdez-Otero (2018). "Geospatial distributions reflect rates of evolution of features of language". In: *arXiv:1801.09637 [cond-mat, physics:nlin, physics:physics]*. arXiv: 1801.09637. URL: http://arxiv.org/abs/1801.09637.

Kießling, Roland, Maarten Mous, and Derek Nurse (2008). "The Tanzanian Rift Valley area". In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 186–227.

Kirby, Kathryn R., Russell D. Gray, Simon J. Greenhill, Fiona M. Jordan, Stephanie Gomes-Ng, Hans-Jörg Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale, and Michael C. Gavin (2016). "D-PLACE: A Global Database of Cultural, Linguistic and Environmental Diversity". In: *PLOS ONE* 11.7. Publisher: Public Library of Science, pp. 1–14. DOI: 10.1371/journal.pone.0158391. URL: https://doi.org/10.1371/journal.pone.0158391.

Kirby, Simon, Mike Dowman, and Thomas L. Griffiths (2007). "Innateness and culture in the evolution of language". In: *Proceedings of the National Academy of Sciences* 104.12, pp. 5241–5245. DOI: 10.1073/pnas.0608222104. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0608222104. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0608222104.

Koile, Ezequiel, Simon J. Greenhill, Damián E. Blasi, Remco Bouckaert, and Russell D. Gray (2022). "Phylogeographic analysis of the Bantu language expansion supports a rainforest route". In: *Proceedings of the National Academy of Sciences* 119.32, e2112853119. DOI: 10.1073/pnas.2112853119. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2112853119. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2112853119.

Kälin, Fabiola (2017). "Global Analysis of the Influence of Geographical Factors on Contact-Induced Language Change". Zürich: Geographisches Institut der Universität Zürich.

Ladefoged, Peter and Ian Maddieson (1996). *The sounds of the worlds languages*. Oxford: Blackwell.

LeSage, James P. and R. Kelley Pace (2009). *Introduction to Spatial Econometrics*. New York: Chapman and Hall/CRC. ISBN: 9780429138089.

— (2019). "Interpreting Spatial Econometric Models". In: *Handbook of Regional Science*. Ed. by Manfred M. Fischer and Peter Nijkamp. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–18. ISBN: 978-3-642-36203-3. DOI: 10.1007/978-3-642-36203-3_91-1. URL: https://doi.org/10.1007/978-3-642-36203-3_91-1.

Levada, Alexandre Luís Magalhães, Frank Nielsen, and Michel Ferreira Cardia Haddad (2024). *Adaptive k-nearest neighbor classifier based on the local estimation of the shape operator*. arXiv: 2409.05084 [cs.LG]. URL: https://arxiv.org/abs/2409.05084.

Leyew, Zelealem (2008). "Is Africa a linguistic area?" In: *A Linguistic Geography of Africa*. Ed. by Bernd Heine and Derek Nurse. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press, pp. 15–35.

List, Johann-Mattis (2018). *Towards a history of concept list compilation in historical linguistics*. URL: https://hiphilangsci.net/2018/10/31/concept-list-compilation/.

List, Johann-Mattis, Robert Forkel, Simon J. Greenhill, Christoph Rzymski, Johannes Englisch, and Russell D. Gray (2022). "Lexibank, a public repository of standardized wordlists with computed phonological and lexical features". In: *Scientific Data* 9.1, p. 316. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01432-0. URL: https://doi.org/10.1038/s41597-022-01432-0.

Lupyan, Gary and Rick Dale (2010). "Language Structure Is Partly Determined by Social Structure". In: *PLOS ONE* 5.1, pp. 1–10. DOI: 10.1371/journal.pone.0008559. URL: https://doi.org/10.1371/journal.pone.0008559.

Lüpke, Friederike (2010). "Multilingualism and Language Contact in West Africa: Towards a holistic perspective". In: *Journal of Language Contact* 3.1, pp. 1–12. DOI:

10.1163/19552629-90000002. URL: https://brill.com/view/journals/jlc/3/1/article-p1_2.xml.

Lüpke, Friederike (2016). "Uncovering Small-Scale Multilingualism". In: *Critical Multilingualism Studies* 4.2, pp. 35–74. ISSN: 2325-2871.

Maho, Jouni (1999). *A comparative study of Bantu noun classes*. Göteburg: Acta Universitatis Gothoburgensis.

Mair, Patrick, Jan De Leeuw, and Patrick J. F. Groenen (2022). *Multivariate Analysis with Optimal Scaling*. DOI: 10.32614/CRAN.package.Gifi.

Manfredi, Stefano (2022). "An areal typology of kin terms in the Nuba Mountain languages". In: *Journal of African Languages and Linguistics* 43.2, pp. 199–247. DOI: doi:10.1515/jall-2022-8896. URL: https://doi.org/10.1515/jall-2022-8896.

Mansfield, John, Henry Leslie-O'Neill, and Haoyi Li (2023). "Dialect differences and linguistic divergence: A crosslinguistic survey of grammatical variation". English. In: *Language Dynamics and Change* 23.3. Publisher Copyright: © John Mansfield et al., 2023., pp. 1–45. ISSN: 2210-5824. DOI: 10.1163/22105832-bja10026.

Matras, Yaron (2007). "The borrowability of structural categories". In: *Grammatical Borrowing in Cross-Linguistic Perspective*. Ed. by Yaron Matras and Jeanette Sakel. Berlin, New York: De Gruyter Mouton, pp. 31–74. ISBN: 9783110199192. DOI: doi:10.1515/9783110199192.31. URL: https://doi.org/10.1515/9783110199192.31.

— (2010). "Contact, Convergence, and Typology". In: *The Handbook of Language Contact*. John Wiley & Sons, Ltd. Chap. 3, pp. 66–85. ISBN: 9781444318159. DOI: https://doi.org/10.1002/9781444318159.ch3. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444318159.ch3.

— (2011). "Explaining convergence and the formation of linguistic areas". In: *Geographical Typology and Linguistic Areas: With special reference to Africa*. Ed. by König Christa Hieda Osamu and Hirosi Nakagawa. Amsterdam: Benjamins, pp. 143–160.

Matras, Yaron and Jeanette Sakel (2007). "Introduction". In: *Grammatical Borrowing in Cross-Linguistic Perspective*. Ed. by Yaron Matras and Jeanette Sakel. Berlin, New York: De Gruyter Mouton, pp. 1–14. ISBN: 9783110199192. DOI: doi:10.1515/9783110199192.1. URL: https://doi.org/10.1515/9783110199192.1.

Matérn, Bertil (2013). *Spatial Variation*. Originally published by the Swedish National Institute for Forestry Research, 1960. Springer Science & Business Media.

McElreath, Richard (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, Florida: CRC press.

McIntosh, Susan Keech (1999). "Pathways to complexity: an African perspective". In: *Beyond Chiefdoms: Pathways to Complexity in Africa*. Ed. by Susan Keech McIntosh. New Directions in Archaeology. Cambridge: Cambridge University Press, pp. 1–30.

Merk, Miryam S. and Philipp Otto (2022). "Estimation of the spatial weighting matrix for regular lattice dataAn adaptive lasso approach with crosssectional resampling". In: *Environmetrics* 33.1, e2705. ISSN: 1180-4009, 1099-095X. DOI: 10.1002/env.2705. URL: https://onlinelibrary.wiley.com/doi/10.1002/env.2705.

Michalopoulos, Stelios and Elias Papaioannou (2013). "Pre-colonial Ethnic Institutions and Contemporary African Development". In: *Econometrica : journal of the Econometric Society* 81.1, pp. 113–152. ISSN: 0012-9682. DOI: 10.3982/ecta9613. URL: https://europepmc.org/articles/PMC4118452.

Miestamo, Matti (2008). *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: De Gruyter Mouton.

Miestamo, Matti, Dik Bakker, and Antti Arppe (2016). "Sampling for Variety". In: *Linguistic Typology* 20.02, pp. 233–296. DOI: 10.1515/lingty-2016-0006.

Moran, Steven and Daniel McCloy, eds. (2019). *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History. URL: https://phoible.org/.

Morrison, Michelle E. (2018). "Beyond derivation: Creative use of noun class prefixation for both semantic and reference tracking purposes". In: *Journal of Pragmatics* 123, pp. 38–56. ISSN: 0378-2166. DOI: https://doi.org/10.1016/j.pragma.2017.10.009. URL: https://www.sciencedirect.com/science/article/pii/S0378216617303983.

Mous, Maarten (2003). *The Making of a Mixed Language: The case of Ma'a/Mbugu*. Creole Language Library. Amsterdam: John Benjamins Publishing Company. ISBN: 9789027252487.

Mufwene, Salikoko S. (2001). *The Ecology of Language Evolution*. Cambridge Approaches to Language Contact. Cambridge: Cambridge University Press.

Mugane, John M. (2015). *The story of Swahili*. Athens: Ohio University Press. ISBN: 9780896804890.

Murawaki, Yugo and Kenji Yamauchi (2018). "A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features". In: *Journal of Language Evolution* 3.1, pp. 13–25. ISSN: 2058-4571, 2058-458X. DOI: 10.1093/jole/lzx022.

Murdock, George Peter (1967). *Ethnographic atlas*. Pittsburg: University of Pittsburgh Press.

Murdock G. P., R. Textor H. Barry III D. R. White J. P. Gray and W. T. Divale (1999). "Ethnographic Atlas (codebook)". In: *World Cultures* 10, pp. 24–136.

Muysken, Pieter (2010). "Scenarios for Language Contact". In: *The Handbook of Language Contact*. Ed. by Raymond Hickey. John Wiley & Sons, Ltd, pp. 265–281. ISBN: 9781444318159.

Nerbonne, John (2013). "How much does geography influence language variation?" In: *Space in Language and Linguistics*. Ed. by Peter Auer, Martin Hilpert, Anja Stukenbrock, and Benedikt Szmrecsanyi. Berlin, Boston: De Gruyter Mouton. ISBN: 978-3-11-031202-7. DOI: 10.1515/9783110312027.222. URL: https://www.degruyter.com/doi/10.1515/9783110312027.222.

Nerbonne, John and Wilbert Heeringa (2007). "Geographic distributions of linguistic variation reflect dynamics of differentiation". In: *Roots*. Ed. by Sam Featherston and Wolfgang Sternefeld. Berlin, New York: De Gruyter Mouton, pp. 267–298. ISBN: 978-3-11-019315-2. DOI: 10.1515/9783110198621.267. URL: https://www.degruyter.com/document/doi/10.1515/9783110198621.267/html.

Nettle, Daniel (1996). "Language Diversity in West Africa: An Ecological Approach". In: *Journal of Anthropological Archaeology* 15.4, pp. 403–438. ISSN: 02784165. DOI: 10.1006/jaar.1996.0015. URL: https://linkinghub.elsevier.com/retrieve/pii/S027841659690015X.

— (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. URL: https://linkinghub.elsevier.com/retrieve/pii/S0278416598903282.

Nettle, Daniel and Suzanne Romaine (2000). *Vanishing Voices*. Oxford: Oxford University Press.

Neureiter, Nico, Peter Ranacher, Nour Efrat-Kowalsky, Gereon A. Kaiping, Robert Weibel, Paul Widmer, and Remco R. Bouckaert (2022). "Detecting contact in language trees: a Bayesian phylogenetic model with horizontal transfer". In: *Humanities and Social Sciences Communications* 9.1, p. 205. ISSN: 2662-9992. DOI: 10.1057/s41599-022-01211-7. URL: https://www.nature.com/articles/s41599-022-01211-7.

Nichols, Johanna (1992). *Linguistic diversity in space and time*. Chicago: University of Chicago Press. ISBN: 978-0-226-58056-2.

— (2003). "Diversity and Stability in Language". In: *The Handbook of Historical Linguistics*. John Wiley & Sons, Ltd. Chap. 5, pp. 283–310. ISBN: 9781405166201. DOI: https://doi.org/10.1002/9781405166201.ch5. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405166201.ch5.

Nichols, Johanna (2020). "Dispersal patterns shape areal typology". In: *Language Dispersal, Diversification, and Contact*. Ed. by Mily Crevels and Pieter Muysken.

Oxford: Oxford University Press, pp. 25 –43. ISBN: 9780198723813. DOI: 10.1093/oso/9780198723813.003.0007.

Nikolaev, Dmitry (2019). "Areal dependency of consonant inventories". In: *Language Dynamics and Change* 9, pp. 104–126. DOI: 10.1163/22105832-00901001.

Nikolaev, Dmitry and Eitan Grossman (2018). "Areal Sound Change and the Distributional Typology of Affricate Richness in Eurasia". In: *Studies in Language* 42.3, pp. 562–599.

Nunn, Nathan (2008). "The Long Term Effects of Africa's Slave Trades". In: *Quarterly Journal of Economics* 123.1, pp. 139–176.

Nunn, Nathan and Diego Puga (2012). "Ruggedness: The Blessing of Bad Geography in Africa". In: *Review of Economics and Statistics* 94.1, pp. 20–36.

Nyst, Victoria (2010). "Sign languages in West Africa." In: *Sign Languages - A Cambridge language survey*. Ed. by D. Brentari. Cambridge: Cambridge University Press, pp. 405–432.

Odland, John (1988). *Spatial autocorrelation*. Newbury: Sage.

Ord, J. K. and Arthur Getis (1995). "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application". In: *Geographical Analysis* 27.4, pp. 286–306. DOI: https://doi.org/10.1111/j.1538-4632.1995.tb00912.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1995.tb00912.x.

Orsini, Luisa, Joost Vanoverbeke, Ine Swillen, Joachim Mergeay, and Luc De Meester (2013). "Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization". In: *Molecular Ecology* 22.24, pp. 5983–5999. DOI: https://doi.org/10.1111/mec.12561. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.12561. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12561.

Pakendorf, Brigitte, Nina Dobrushina, and Olesya Khanina (2021). "A typology of small-scale multilingualism". In: *International Journal of Bilingualism* 25.4, pp. 835–859. DOI: 10.1177/13670069211023137. eprint: https://doi.org/10.1177/13670069211023137. URL: https://doi.org/10.1177/13670069211023137.

Pakendorf, Brigitte, Cesare de Filippo, and Koen Bostoen (2011). "Molecular Perspectives on the Bantu Expansion: A Synthesis". In: *Language Dynamics and Change* 1.1, pp. 50 –88. DOI: 10.1163/221058211X570349. URL: https://brill.com/view/journals/ldc/1/1/article-p50_3.xml.

Pakendorf, Brigitte, Hilde Gunnink, Bonny Sands, and Koen Bostoen (2017). "Prehistoric Bantu-Khoisan language contact: A cross-disciplinary approach". In: *Language Dynamics and Change* 17, pp. 1–46.

Paradis, Emmanuel and Klaus Schliep (2019). "ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R". In: *Bioinformatics* 35, pp. 526–528. DOI: 10.1093/bioinformatics/bty633.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Manual. Vienna, Austria.

Ranacher, Peter, Nico Neureiter, Rik van Gijn, Barbara Sonnenhauser, Anastasia Escher, Robert Weibel, Pieter Muysken, and Balthasar Bickel (2021). "Contact-Tracing in Cultural Evolution: A Bayesian Mixture Model to Detect Geographic Areas of Language Contact". In: *Journal of the Royal Society Interface* 18.181, pp. 1–15.

Renfrew, Colin (1992). "Archaeology, Genetics and Linguistic Diversity". In: *Man* 27.3, pp. 445–478. ISSN: 00251496, 23972548. URL: http://www.jstor.org/stable/2803924.

Riley, Shawn, Stephen Degloria, and S.D. Elliot (1999). "A Terrain Ruggedness Index that Quantifies Topographic Heterogeneity". In: *International Journal of Science* 5, pp. 23–27.

Riutort-Mayol, Gabriel, Paul-Christian Bürkner, Michael R. Andersen, Arno Solin, and Aki Vehtari (2022). "Practical Hilbert space approximate Bayesian Gaussian processes for probabilistic programming". In: *Statistics and Computing* 33.17. ISSN: 1573-1375. DOI: 10.1007/s11222-022-10167-2. URL: https://doi.org/10.1007/s11222-022-10167-2.

Roger Bivand (2022). "R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data". In: *Geographical Analysis* 54.3, pp. 488–518. DOI: 10.1111/gean.12319.

Sakel, Jeanette (2007). "Types of loan: Matter and pattern". In: *Grammatical Borrowing in Cross-Linguistic Perspective*. Ed. by Yaron Matras and Jeanette Sakel. Berlin, New York: De Gruyter Mouton, pp. 15–30. ISBN: 9783110199192. DOI: doi:10.1515/9783110199192.15. URL: https://doi.org/10.1515/9783110199192.15.

Sands, Bonnie (1998). *Eastern and Southern African Khoisan: evaluating claims of distant linguistic relationships*. Vol. 14. Quellen zur Khoisan-Forschung. Rüdiger Köppe.

Sands, Bonny (2022). "Tracing Language Contact in Africas Past". In: *The Cambridge Handbook of Language Contact: Volume 1: Population Movement and Language Change*. Ed. by Salikoko S. Mufwene and Anna María Escobar. Cambridge Handbooks in Language and Linguistics. Cambridge: Cambridge University Press, pp. 84–121.

Sands, Bonny and Hilde Gunnink (2019). "Clicks on the fringes of the Kalahari Basin Area". In: *Theory and description in African Linguistics*. Ed. by Emily Clem, Peter

Jenks, and Hannah Sande. Berlin: Language Science Press, pp. 703–724. DOI: 10.5281/zenodo.3365789.

Segerer, Guillaume and Sébastien Flavier (2011-2021). *RefLex: Reference lexicon of Africa. Version 1.1*. URL: http://reflex.cnrs.fr/.

Segerer, Guillaume and Martine Vanhove (2022). "Areal patterns and colexifications of colour terms in the languages of Africa". In: *Linguistic Typology* 26.2, pp. 247–281. DOI: doi:10.1515/lingty-2021-2085. URL: https://doi.org/10.1515/lingty-2021-2085.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoğlu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray (2023). "Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss". In: *Science Advances* 9.16. DOI: 10.1126/sciadv.adg6175.

Skirgård, Hedvig (2021). *Multilevel dynamics of language diversity in Oceania (PhD dissertation)*. Canberra: Australian National University. DOI: 10.25911/WY2B-AQ30.

Skirgård, Hedvig, Hannah J. Haynie, Harald Hammarström, Damián E. Blasi, Jeremy Collins, Jay Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich,

Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Noor Karolin Abbas, Sunny Ananth, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Anina Bolls, Robert D. Borges, Mitchell Browen, Lennart Chevallier, Swintha Danielsen, Sinoël Dohlen, Luise Dorenbusch, Ella Dorn, Marie Duhamel, Farah El Haj Ali, John Elliott, Giada Falcone, Anna-Maria Fehn, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu H. Huntington-Rainey, Guglielmo Inglese, Jessica K. Ivani, Marilen Johns, Erika Just, Ivan Kapitonov, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Kate Lynn Lindsey, Nora L. M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Alexandra Marley, Tânia R. A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya, Michael Müller, Saliha Murado$\breve{g}$lu, HunterGatherer, David Nash, Kelsey Neely, Johanna Nickel, Miina Norvik, Bruno Olsson, Cheryl Akinyi Oluoch, David Osgarby, Jesse Peacock, India O.C. Pearey, Naomi Peck, Jana Peter, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W. P. Schmidt, Dineke Schokkin, Jeff Siegel, Amalia Skilton, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith VoSS, Daniel Wikalier Smith, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray (2023). *Grambank v1.0*. Version v1.0. Dataset. DOI: 10.5281/zenodo.7740140. URL: https://doi.org/10.5281/zenodo.7740140.

Solin, Arno and Simo Särkkä (2020). "Hilbert space methods for reduced-rank Gaussian process regression". In: *Statistics and Computing* 30.2, pp. 419–446. ISSN: 1573-1375. DOI: 10.1007/s11222-019-09886-w. URL: https://doi.org/10.1007/s11222-019-09886-w.

Song, Jae Jung, ed. (2013). *The Oxford Handbook of Linguistic Typology*. Oxford: Oxford University Press. ISBN: 9780199658404.

Stein, Michael L. (2012). *Interpolation of spatial data: some theory for kriging*. New York: Springer Science & Business Media.

Swadesh, Morris (1955). "Towards greater accuracy in lexicostatistic dating." In: *International Journal of American Linguistics* 21.2.

Thomason, Sarah (2010). "Contact Explanations in Linguistics". In: *The Handbook of Language Contact*. John Wiley & Sons, Ltd. Chap. 1, pp. 29–47. ISBN:

9781444318159. DOI: https://doi.org/10.1002/9781444318159.ch1. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444318159.ch1.

Thomason, Sarah G. (2001). *Language Contact: An Introduction*. Edinburgh: Edinburgh University Press.

Thomason, Sarah G. and Terrence Kaufman (1988). *Language Contact, Creolization and Genetic Linguistics*. Berkeley: University of California Press.

Tosco, M. (2000). "Is there an "Ethiopian language area"?" In: *Anthropological Linguistics* 42, pp. 329–365.

Towner, Mary C., Mark N. Grote, Jay Venti, and Monique Borgerhoff Mulder (2012). "Cultural Macroevolution on Neighbor Graphs: Vertical and Horizontal Transmission among Western North American Indian Societies". In: *Human Nature* 23.3, pp. 283–305. ISSN: 1045-6767, 1936-4776. DOI: 10.1007/s12110-012-9142-z. URL: http://link.springer.com/10.1007/s12110-012-9142-z.

Trudgill, Peter (2010). "Contact and Sociolinguistic Typology". In: *The Handbook of Language Contact*. Ed. by Raymond Hickey. Wiley, pp. 299–319. ISBN: 9781119485094.

— (2011). *Sociolinguistic Typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Urban, Matthias (2020). "Mountain linguistics". In: *Language and Linguistics Compass* 14.9, pp. 1–23. DOI: 10.1111/lnc3.12393. URL: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/lnc3.12393.

Urban, Matthias and Steven Moran (2021). "Altitude and the distributional typology of language structure: Ejectives and beyond". In: *Plos one* 16.2, e0245522.

Van de Velde, Mark (2019). "Nominal morphology and syntax". In: *The Bantu Languages 2nd Edition*. Ed. by Mark Van de Velde, Koen Bostoen, Derek Nurse, and Gérard Philippson. New York: Routledge, pp. 237–269.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". In: *Statistics and Computing* 27.5, pp. 1413–1432. ISSN: 1573-1375. DOI: 10.1007/s11222-016-9696-4. URL: https://doi.org/10.1007/s11222-016-9696-4.

Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry (2024). *Pareto Smoothed Importance Sampling*. arXiv: 1507.02646 [stat.CO]. URL: https://arxiv.org/abs/1507.02646.

Ver Hoef, Jay M., Ephraim M. Hanks, and Mevin B. Hooten (2018). "On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models". In: *Spatial Statistics* 25, pp. 68–85. ISSN: 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2018.04.006. URL: https://www.sciencedirect.com/science/article/pii/S2211675317302725.

Villemereuil, Pierre de and Shinichi Nakagawa (2014). *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology*. Berlin: Springer.

Villemereuil, Pierre de, Jessie A. Wells, Robert D. Edwards, and Simon P. Blomberg (2012). "Bayesian models for comparative analysis integrating phylogenetic uncertainty". In: *BMC Evolutionary Biology* 12.1, p. 102. ISSN: 1471-2148. DOI: 10.1186/1471-2148-12-102. URL: https://doi.org/10.1186/1471-2148-12-102.

Vossen, Rainer, ed. (2013). *The Khoesan Languages*. Routledge.

Wall, Melanie (2004). "A close look at the spatial structure implied by the CAR and SAR models". In: *Journal of Statistical Planning and Inference* 121, pp. 311–324. DOI: 10.1016/S0378-3758(03)00111-3.

Wang, Tianheng, Søren Wichmann, Quansheng Xia, and Qibin Ran (2023). "Temperature shapes language sonority: Revalidation from a large dataset". In: *PNAS Nexus* 2.12. Ed. by Emilio Moran, pp. 1–9. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgad384. URL: https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgad384/7457938.

Watanabe, Sumio (2010). "Asymptotic equivalence of Bayes cross validation and widely application information criterion in singular learning theory". In: *Journal of Machine Learning Research* 11, pp. 3571–3594.

White, Douglas R. and Lilyan A. Brudner-White (1988). "The Murdock Legacy: the Ethnographic Atlas and the Search for a Method". In: *Behavior Science Research* 22.1-4, pp. 59–81. ISSN: 0094-3673. DOI: 10.1177/106939718802200107. URL: http://journals.sagepub.com/doi/10.1177/106939718802200107.

Whittle, P (1954). "On Stationary Processes in the Plane". en. In: *Biometrika* 3/4, pp. 434–449. URL: https://www.jstor.org/stable/2332724.

Wichmann, Søren (2015). "Diachronic stability and typology". In: *Routledge Handbook of Historical Linguistics*. Ed. by Raymond Hickey. New York: Taylor & Francis, pp. 212–224. ISBN: 978-0-415-52789-7.

Wieling, Martijn, John Nerbonne, and R. Harald Baayen (2011). "Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially". In: *PLoS ONE* 6.9. Ed. by Matjaz Perc, e23613. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0023613. URL: https://dx.plos.org/10.1371/journal.pone.0023613.

Williams, Christopher KI and Carl Edward Rasmussen (2006). *Gaussian processes for machine learning*. Vol. 2. 3. Cambridge, MA: MIT Press.

Winter, Bodo and Martijn Wieling (2016). "How to analyze linguistic change using mixed models, Growth Curve Analysis and Generalized Additive Modeling". In: *Journal of Language Evolution*, pp. 7–18. DOI: https://doi.org/10.1093/jole/lzv003.

Wright, Sewall (1943). “Isolation by Distance”. In: *Genetics* 28.2, pp. 114–138. DOI: doi:10.1093/genetics/28.2.114.

Zeshan, Ulrika and Connie de Vos, eds. (2012). *Sign Languages in Village Communities: Anthropological and Linguistic Insights*. 1st ed. Berlin, Boston: De Gruyter Mouton. URL: http://www.jstor.org/stable/j.ctvbkjwzx.