Why modeling space is hard:
Polynesian phoneme inventories as a test case

Abstract

In recent years there has been an increased interest in computational modeling of spatial phenomena in typology (Guzmán Naranjo & Becker 2022, Guzmán Naranjo & Mertner 2023, Hartmann 2022, Hartmann & Jäger 2023, Ranacher et al. 2021, Urban & Moran 2021). While the main focus of most work so far has been on direct language contact, there are two different types of spatial dynamics of interest to typologists and areal linguists: language expansion, and unidirectional contact effects. In this paper we present three new statistical techniques to model expansion and unidirectional contact effects. We illustrate these techniques with a case study on Polynesian phoneme inventory sizes, and show that there is strong evidence for unidirectional effects, but little evidence for expansion effects (contra Atkinson 2011). We also argue that we are still far from having a complete understanding of how to model all spatial dynamics that can affect language contact, and that more attention should be paid to these issues.

1 Introduction

In recent years there has been an increased interest in computational modeling of spatial phenomena in typology (Guzmán Naranjo & Becker 2022, Guzmán Naranjo & Mertner 2023, Hartmann 2022, Hartmann & Jäger 2023, Ranacher et al. 2021, Urban & Moran 2021). While the main focus of most work so far has been on direct language contact, there are two different types of spatial dynamics of interest to typologists and areal linguists: language expansion, and unidirectional contact effects. Especially language expansion in the form of serial founder effects has received attention in typology in that serial founder effects have been argued to account for variation in phoneme inventory sizes (Trudgill 2004, Atkinson 2011b). A serial founder effect refers to a scenario in which populations expand from some point of origin in sequential migrations of small speaker communities. Such a scenario has been argued to lead to less complexity of certain linguistic structures, e.g. to smaller phoneme inventory sizes with increasing distance between a speech community and its origin.

Previous quantitative approaches to modeling spatial phenomena in general, and to serial founder effects in particular, have mostly not taken into account the complexities of spatial relations between languages. Spatial modeling has sometimes been presented as a sort of solved problem, or at least relatively easy to account for. We argue in this paper that space is neither, and that we are only beginning to understand how to take it into account in our models. Taking phoneme inventory sizes in Polynesian languages as a test case, this study proposes four different statistical techniques to model (i) symmetric contact between languages, (ii) spatial expansion from a common origin, (iii) unidirectional contact effects from one set of source languages to another set of potential target languages, and to (iv) estimate the likelihood of such unidirectional contact effects for a given potential target language.

Our choice of phoneme inventory sizes in Polynesian as a test case is a purely practical one. On the one hand, we have fairly detailed knowledge about the Polynesian expansion from a common Urheimat. This allows us to model spatial expansion in a relatively detailed way. On the other hand, serial found effects in expansion scenarios have been mainly discussed in relation to phoneme inventory sizes. Phoneme inventory size is also a variable that should be easily affected by contact through the borrowing of individual phonemes. In addition, Polynesian languages are also known to have had sustained contact with Non-Polynesian languages in Melanesia and, to a lesser extent, Micronesia.

This paper is structured as follows. Section 2 provides an overview of Polynesian languages and their expansion, as well as previous claims regarding phoneme inventory sizes and the serial founder effect. Section 3 describes the dataset used for the present study and our annotation choices. In Section 4, we give an overview of the general patterns concerning phoneme inventories in the Polynesian languages from our dataset in relation to evidence about diachronic developments and language contact from the literature. We then turn to modeling the association between spatial expansion, contact, and phoneme inventory sizes. Section 5 introduces the model components and Section 6 presents the model results. We discuss the implications thereof in Section 7; Section 8 concludes.

2 Polynesian phoneme inventories: A showcase of complex spatial relations

2.1 Polynesian languages

Polynesian languages belong to the Oceanic branch of the Austronesian language family. They are mainly situated in the Polynesian Triangle, which is a geographical region comprising Hawai'i as the north apex, New Zealand as the southernmost corner and Rapa Nui (Easter Island) as the easternmost point (Kurpa 1973: 1). Besides, about 20 so-called Polynesian Outlier languages are located on the periphery of eastern Micronesia and Melanesia (Pawley 1967: 260). Glottolog distinguishes 38 Polynesian varieties; our dataset includes 36 of them. Their location can be seen in Figure 1, with the core Polynesian languages plotted in dark blue and the Outlier languages in light blue.

Despite the fact that Polynesian has long been known to constitute a phylogenetic group within Oceanic (cf. Blust 2013: 118), the study of the internal phylogenetic structures be-

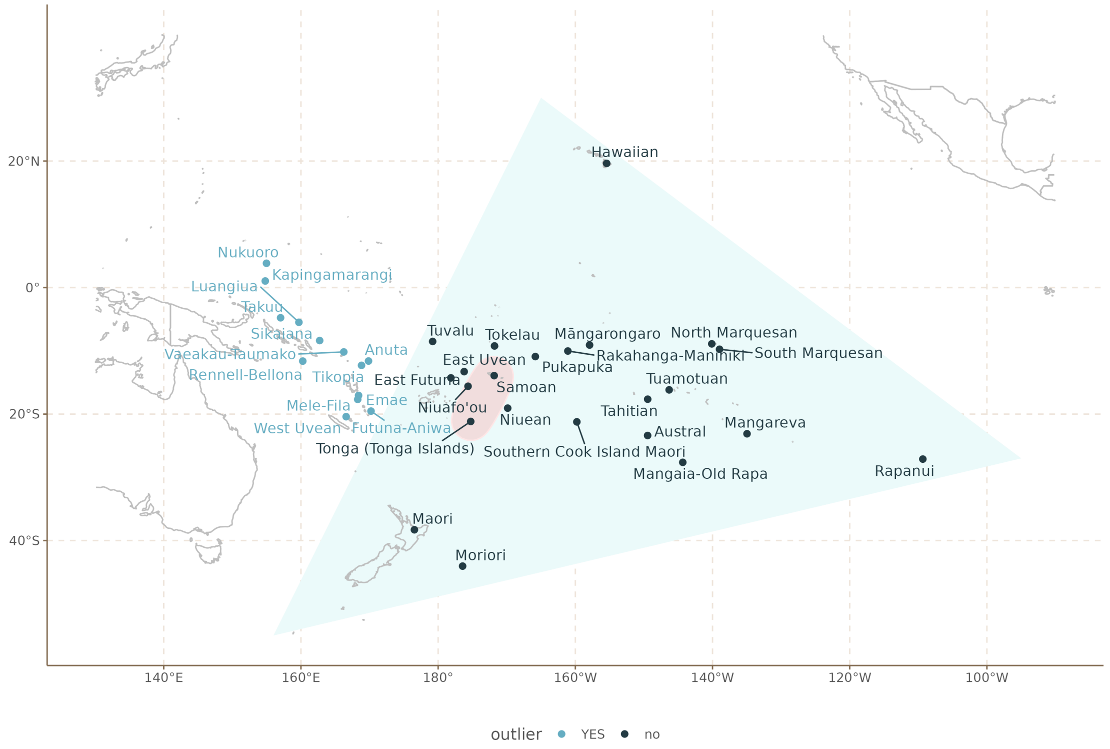

Figure 1: Location of the 36 Polynesian languages in the dataset

tween Polynesian languages only emerged in the 20th century. Elbert (1953) proposed an internal structure of Polynesian languages based on phonology, distinguishing between the two main branches of Nuclear Polynesian and Tongic (Elbert 1953: 169). Later, Pawley (1966: 39-40) proposed a more fine-grained subgrouping based on morphology and divided Nuclear Polynesian further into Samoic-Outliers and Eastern Polynesian. This division into Tongic, Nuclear Polynesian, several Outlier subbranches and Eastern Polynesian mainly corresponds to the phylogenetic relations distinguished today. Without going into further detail, we follow the internal phylogenetic structure of Polynesian as proposed by Glottolog.[1] The full phylogenetic tree of all 36 Polynesian languages in our dataset can be seen in Figure 2. As major groups, we thus distinguish Tongic vs. Nuclear Polynesian, the latter of which consists of several Outlier groups and East Polynesian (cf. also Biggs 1971).

It is widely accepted that the origin of the Polynesian homeland lies in an area around Tonga, Samoa, Futuna and ’Uvea (cf. Bellwood 1979, Geraghty 1983, Green 1981, Jennings 1979, Kirch 1984a, 1996, Kirch & Green 1992, 2001, Pawley & Green 1973, 1984). Archeological evidence ties the origin of Polynesians to the earlier Lapita culture in Melanesia, mostly characterized by complex ceramic potteries (cf. Kirch 1997, Pawley 2007). Based on archeological findings, we can assume that the Lapita culture spread from the Bismarck archipelago

[1] For more details about motivating phylogenetic groupings in Polynesian, see Biggs (1971), Green (1966), Greenhill & Clark (2011), Kirch (1984b), Marck (2000), Pawley (1966, 1967), Walworth (2014), Wilson (2012).

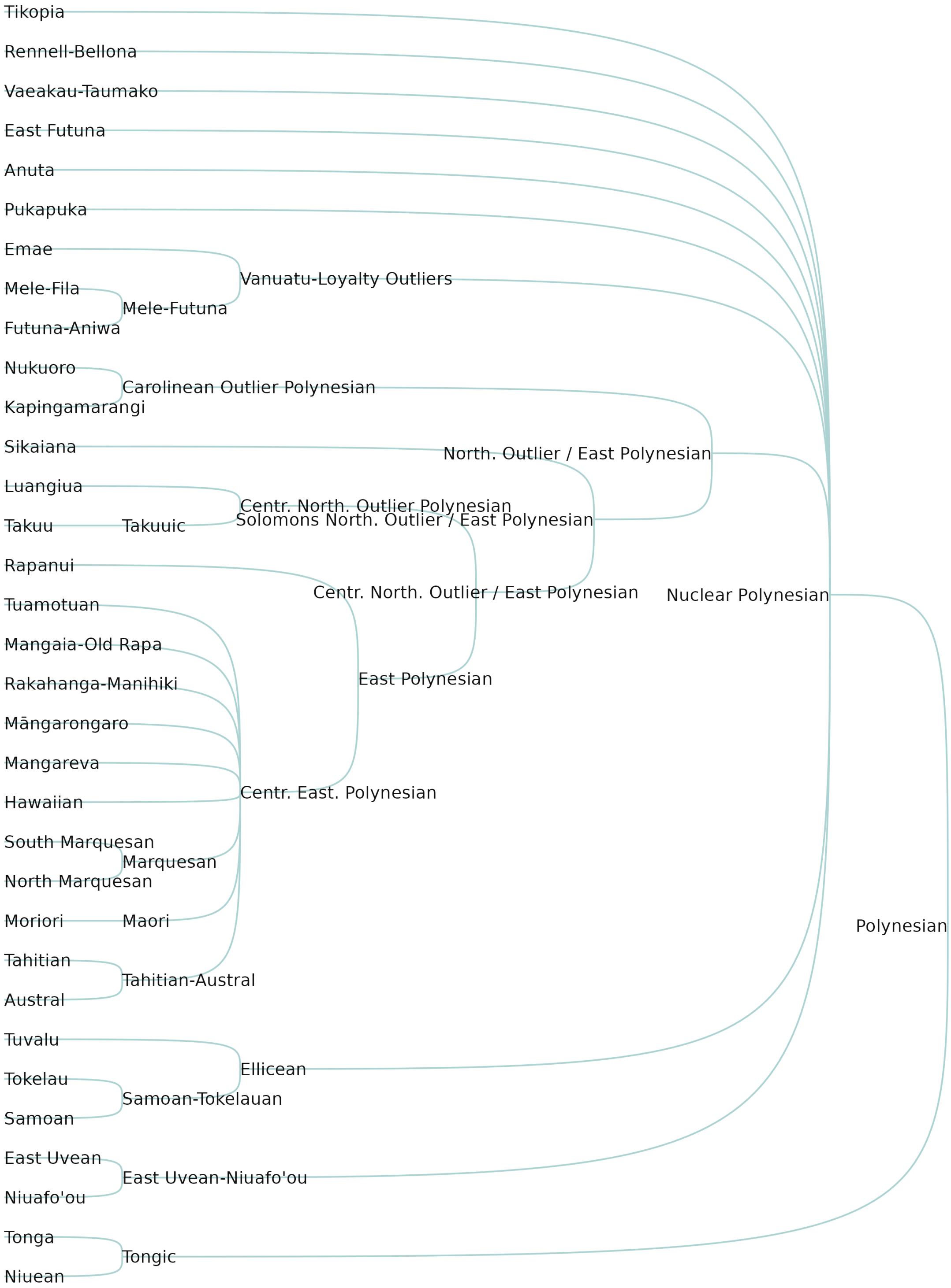

Figure 2: Phylogenetic structure of the Polynesian languages in the dataset

to the Solomon islands, Vanuatu, the Loyalites, Fiji to Tonga and Samoa. The beginning of the Lapita culture in Western Melanesia dates back to around 1500 B.C., and it persisted about 1000 years, until around 500 B.C. (Kirch 1996: 61-63). Early Polynesian culture gradually developed from the Lapita culture, which consisted of a chain of populations from Fiji to Samoa and Tonga. Over the course of time, Lapita populations in central Polynesia had less and less contact with Lapita populations in the West, which led to their isolation and thus to the development of a distinct, early Polynesian culture. Cultural artifacts suggests that this separation happened some time in the first milenium B.C. (Kirch 2017: 187-188).

Many details of the ensuing Polynesian expansion remain debated and unknown, especially regarding dates and time frames. There is, however, more consensus on the general spatial patterns of the expansion. We will give a brief overview here; see the map in Figure 4 in Section 3.3 for our operationalization of Polynesian migration paths.

The Polynesian expansion started in the Western Polynesian center around Tonga, Samoa, Futuna and 'Uvea. From there, early Polynesians sailed east, reaching the Southern Cook islands as the first area of expansion in Eastern Polynesia (Irwin 1994: 98). The Polynesians then traveled further to the Austral islands, Rapa island, the Society islands, the Tuamotu islands, the Marquesas islands, the Gambier islands and the Pitcairn islands. According to modern radiocarbon dating, Polynesians populated central and southern East Polynesia between 900 and 1200 A.D. (Kahn & Sinoto 2017, Kennett et al. 2012, Wilmshurst et al. 2011, cf. Kirch 2017: 198-200 for an overview). The most remote islands of Hawai'i, Rapa Nui and New Zealand were colonized somewhat later. The Hawai'i islands were likely settled from the Southern Cook islands along a chain of contact between the Society islands, the Tuamotu islands and the Marquesas islands (Kirch 2017: 200, Green & Weisler 2002: 234-235). The Polynesians reached Rapa Nui by moving eastward from Society islands via Mangareva island (Gambier) and Henderson island (Pitcairn) (Green & Weisler 2002: 234-235, Kieviet 2017: 2, Martinsson-Wallin & Crockford 2001). New Zealand was populated the latest from the Southern Cook islands, most likely in the 13th century (Goodwin, Browning & Anderson 2014, Wilmshurst et al. 2011, Kirch 2017: 200).

In addition to their eastward expansion, Polynesians also traveled west from central Polynesia (Ward, Webb & Levison 1973, Kirch 1984b) to Melanesia and Micronesia. There is evidence that some of this east-to-west expansion happened very early in the first milenium B.C., while other islands were settled only around 1500 A.D. (cf. Carson 2012, Kirch 1984b). By now, evidence points to two main origins of the Polynesian Outlier populations. Tuvalu is the most likely origin of the northern Outliers in Micronesia (Nukuoro and Kapingamarangi) as well as of Melanesian Outlier populations in Papua New Guinea (Takuu) and the Solomon Islands (Ontong Java, Sikaiana, Reef Islands, Duff Islands) (Kirch 2017: 161, Carson 2012: 28). 'Uvea most likely corresponds to the second point of origin for other Polynesian Outlier populations in Melanesia further south in the Solomon Islands (Anuta, Tikopia, Rennell, Bellona), Vanuatu (Futuna, Aniwa, Emae, Efate) and the Loyalty Islands (Ouvéa) (Carson 2012: 28,Kirch 1984b: 230).

2.2 Phoneme inventory sizes (PIS) and the serial founder effect

It has been proposed that language expansion leaves clear signatures on the grammar of the languages involved. Among these, the best known are probably serial founder effects (Atkinson 2011b, Deshpande et al. 2009, Fort & Pérez-Losada 2016, Pérez-Losada & Fort 2018). Originally, serial founder effects were proposed for population genetics (Betti et al. 2009, Manica et al. 2007, Pierce et al. 2014). The main idea is that when populations expand from a homeland in sequential migrations of small numbers of individuals, the genetic diversity decreases as the populations expand further and further. Sequential small migrations lead to genetic bottlenecks, which in turn decreases genetic diversity within groups.

Adapting this concept to linguistics, it has been proposed that serial founder effects can impact the structure of languages. This is due to a combination of socio-linguistic factors, namely the interaction of population sizes and linguistic complexity. Trudgill (2004) argues that if a linguistic community is isolated, small, or has minimal contact with other languages, there is a high probability that linguistic complexity in that language will persist and be transmitted to subsequent generations (Trudgill 2004: 306). He argues that, at the same time, languages with a low degree of contact and small populations can afford to develop smaller inventories, as opposed to languages with larger populations and more contact and therefore potential adult learners. The explanation given is that smaller inventories provide fewer possible segmental contrasts, which in turn lead to higher memory load. According to Trudgill (2004: 315-317), this is not a problem for small tight-knit communities with few L2 learners, while it does pose problems for looser-knit communities with more contact and L2 learners.[2]

In this vein, Trudgill (2004: 312), building on earlier ideas by Haudricourt (1961), suggests that "the two factors of isolation and small community size" may account for phoneme inventories of Polynesian languages, which are relatively small when compared to other Austronesian languages closer to Taiwan (cf. Trudgill 2004 and Trudgill 2011: 155-256).[3] The association between phoneme inventory sizes (PIS) and population sizes has since been tested by a number of other studies, with mixed results. Some studies found effects (e.g. Hay & Bauer 2007, Wichmann, Rama & Holman 2011), while others did not (e.g. Donohue & Nichols 2011, Moran, McCloy & Wright 2012, Pericliev 2004).

[2] See Bakker (2004) and Donohue & Nichols (2011) for a critique of this idea. See also Moran, McCloy & Wright (2012) for a more exhaustive discussion of the topic.

[3] See however, Rice (2004) on Athabaskan for an opposite result. Interestingly, Trudgill (2004: 313) mentions contact as an explanation for larger consonant inventories in Polynesian Outliers of Melanesia, but does not develop this point any further.

If, for the sake of argument, we assume that there is a real effect between PIS and population size, then the serial founder effect would lead to languages having gradually smaller PISs the further away from the homeland they are located, if the expansion happens in sequential migrations. This is what Atkinson (2011a) claims. Based on PIS, he argues that we can observe a serial founder effect in that languages have gradually smaller PISs with increasing distance to Africa, which is taken as a global point of origin for the spread of humans. Despite having received methodological and theoretical criticism (e.g. Cysouw, Dediu & Moran 2012, Wang et al. 2012), other recent works have claimed similar findings (e.g. Fenk-Oczlon & Pilz 2021, Fort & Pérez-Losada 2016, Pérez-Losada & Fort 2018).

The above-mentioned studies all take a global approach to the idea of a serial founder effect observable in PIS. Given the large scale and the variety of language-internal and language-external factors that can impact the PIS of a language, it is unclear whether the observed effects on a global scale can really be attributed to expansion. It is very likely that the effects are rather an artifact of different confounding factors.

Because of this, if we want to take the hypothesis of a serial founder effect seriously, it is important to test it in a more controlled setting and to control for other confounding factors. As far as we are aware, however, no study has looked at effects of geographical expansion on linguistic structure in detail for a small number of languages in a clearly defined region yet. In addition, although some of the more recent studies like Pérez-Losada & Fort (2018) or Fenk-Oczlon & Pilz (2021) make use of quantitative techniques, they do not try to fully control for genetic and other spatial confounds. Moreover, several other studies found no evidence that would support a potential serial founder effect with PISs (Creanza et al. 2015, Cysouw, Dediu & Moran 2012). This calls for a smaller-scale but more comprehensive quantitative approach to testing a potential serial founder effect on PISs, which this the aim of the present paper.

3 Dataset and annotation

3.1 Dataset

Our dataset includes 36 Polynesian languages whose spatial distribution was shown on the map in Figure 1.[4] As mentioned in Section 3.3, we can distinguish between core Polynesian languages (spoken within the Polynesian triangle) and Polynesian Outlier languages (spoken in Melanesia and Micronesia). In addition to the Polynesian languages, which are in the main focus of the present study, we included 124 non-Polynesian languages in our dataset. Those are languages from other Austronesian branches or other language families spoken in Melane-

[4] The full dataset including annotations can be found in the supplementary materials: `https://osf.io/eybv6/?view_only=4ec10a53fe5c4596b69de070750c89f5` (anonymized OSF link for review).

sia and Micronesia in the regions where the Polynesian Outlier languages are spoken. Figure 3 shows their geographic locations (blue dots) together with the 36 Polynesian languages (red triangles) from the dataset.

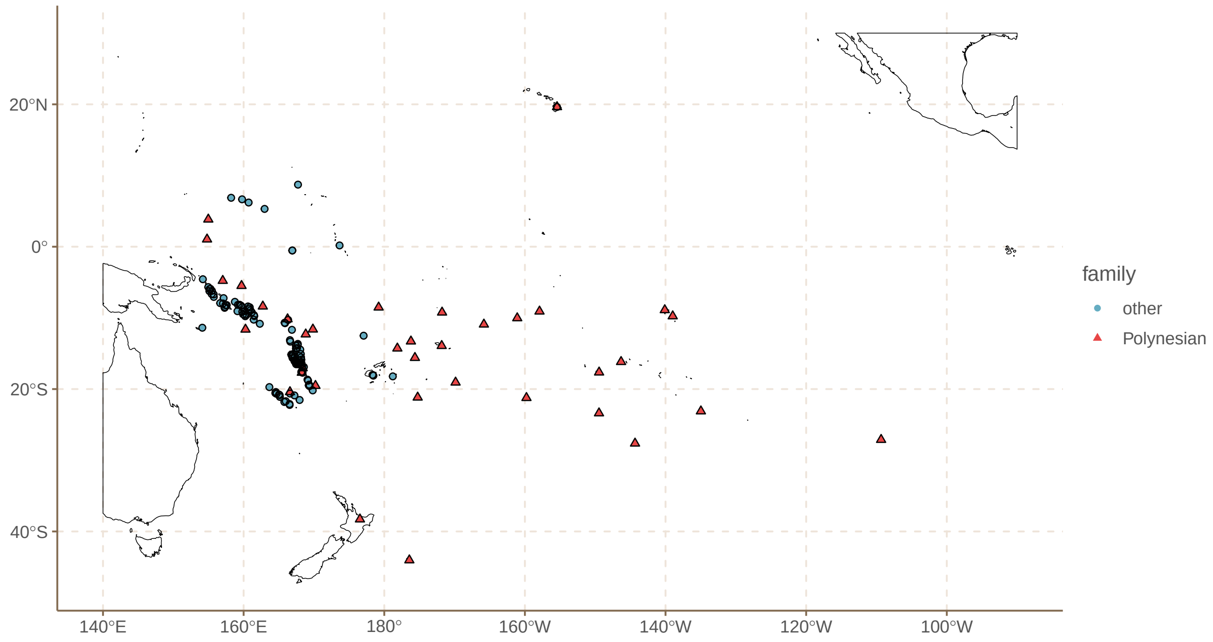

Figure 3: Map of the Polynesian and non-Polynesian languages in the dataset

3.2 Phoneme inventory annotation

All languages in the dataset (both Polynesian and non-Polynesian) were annotated for their phoneme inventories. We did this using the information provided in reference grammars and descriptions. Of course, determining which phones have phoneme status let alone determining the total number of phonemes a language has is not trivial and can be subject to biases on various levels. Therefore, we generally rely on the decisions in the grammatical descriptions made by experts on the respective languages.

The only systematic exception to this is length, which is treated differently across languages as well as descriptions and analyses. In some cases, segments that only differ in length with no other qualitative difference are treated as separate phonemes, whereas this is not done in other cases. Although there may be valid theoretical and/or language-internal reasons for either analysis, we systematically counted phonological length as an additional phonological feature. In other words, we treated /a/ and /a:/ as well as /l/ and /l:/ as separate phonemes. In most cases, either classification did not affect the final count of phonemes to a great extent, but in some cases it did. To give one example, Tryon (1995: 949-950) describes the phoneme inventory of Mele-Fila as having 15 consonants and 5 vowels. Yet, he notes that the language has distinctive length contrasts in all vowels and consonants. We therefore annotated the phoneme inventory size of Mele-Fila as 30+10=40.

Treating length distinctions this way in our dataset ensures systematicity in the annotation across languages, which follows the proposal of comparative concepts for typology (cf. Croft 2016, Haspelmath 2010, 2018). The main idea behind comparative concepts is to define a linguistic category that can be applied and used across languages for typological comparisons. Therefore, it must not rely on language-specific criteria and can differ from established definitions and classifications in linguistic traditions of single languages. We are aware of the fact that it may not be always useful or theoretically motivated to treat short and long segments phonological segments as separate phonemes.[5] For the purposes of this study, it is warranted for crosslinguistic comparability and systematicity. In principle, we could have also decided never to treat such segments as phonemes. This, however, would have led to fewer distinctions and the loss of important variation especially comparing the Outlier languages to core and Eastern Polynesian (cf. Section 4).

3.3 Polynesian expansion and distance to origin

To take the migration paths into account when modeling the Polynesian expansion, we require precise coordinates in order to calculate the distances between the current location of a language and the point of origin. Although there is a solid body of work from linguistics, archaeology and genetics related to the pathways of the Polynesian expansion, there is no straightforward model available that would meet our criteria. Therefore, we built a graph of migration paths based on what is known from the literature, assuming parsimony and simplifying certain aspects of the expansion for modeling purposes.[6]

As mentioned in Section 2.1, the most likely origin of the Polynesian expansion corresponds to the area between Samoa and Tonga. Given that we have to select single points in order to calculate distances, we calculated two sets of distances for each of the languages, taking Samoa or Tongan as the point of origin. This is a necessary simplification for modeling purposes; the early Polynesians most likely maintained a travel network between Samoa and Tonga rather than expanding from only one of the islands. Using both Samoa and Tonga as

[5] For instance, Anderson & Otsuka (2006), Taumoefolau (2002) and Rolle (2009) have argued that long vowels in Tongan and Niue correspond to the surface realization of two adjacent identical short vowels and should therefore not be treated as phonemic, but as a phonetic phenomenon. Other discussions of long vowels in Polynesian include Bauer (1993: 534-538) for Maori, Besnier (2000: 612-613) for Tuvaluan, Mosel & Hovdhaugen (1992: 28-31) for Samoan as well as Næss & Hovdhaugen (2011: 27-30) for Vaeakau-Taumako.

[6] See the file `polynesian-paths.csv` for more details on the islands, coordinates, and references of each of the migration paths.

points of origin for two alternative sets of distances, however, ensures that we can gauge the impact of these choices and check for potential differences in the results. For reasons of space, we only report results with Samoa as the point of origin, however, there were no noticeable differences in the models using Tonga as origin point.[7] Figure 4 shows our final expansion graph.

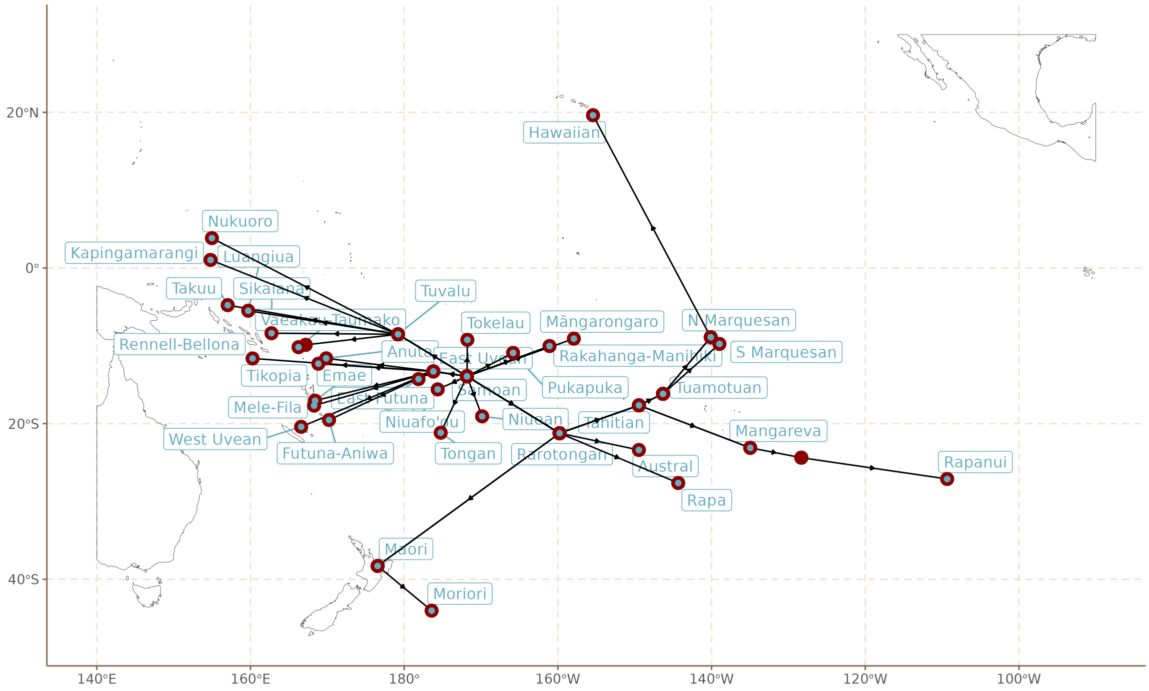

Figure 4: Graph of the Polynesian expansion with Samoa as the point of origin

For the purposes of this study, we assume a uni-directional expansion from one point to another. Despite the fact that the expansion involved more complex travel networks and maintained, bi-directional contact between different pairs of islands (cf. Irwin 1994), this is a necessary simplification in order to estimate the distances between current language locations and their point of origin.

For islands that are very close to the Tonga-Samoa area with no specific migration paths mentioned in the literature, we assume a direct path from Samoa (or Tonga) by parsimony. This is the case for Niue, Pukapuka, Niuafo'ou, Tuvalu, Tokelau and Tonga (Samoa).

Another simplification is that we use the coordinates provided by Glottolog as a proxy for islands/atolls and island/atoll groups. Often, the literature can only point to island groups, e.g. the Society islands, that have acted as important hubs and intermediate steps in the expansion. As mentioned before, we need precise coordinates for the calculation of distances, meaning that we have to select a point in that island group. This is not a trivial choice, especially

[7] See the supplementary materials for the corresponding plots and analyses with Tonga as the point of origin.

given that such island groups can span a territory of a thousand square kilometers. We can narrow down such island groups to those islands with archaeological findings which support long term presence of Polynesians. This still leaves us with a number of potential islands per island group in vairous cases. As a heuristic, we therefore selected islands that are also featured as the location of other Polynesian languages in our dataset. For instance, we use the Glottolog coordinates for Tahitian, spoken on Tahiti, to represent the Society islands as an intermediate step in the expansion from e.g. the Cook islands (Rarotonga Island) to the Gambier islands (Mangareva).

We included additional points that do not correspond to any language in our dataset in only two cases. The first is Henderson island as an intermediate step between Mangareva and Rapa Nui (cf. Green & Weisler 2002: 233). The second additional point with no associated Polynesian language from our sample is the island of Taumako. Because Glottolog locates Veakau-Taumako on the Reef islands, we added this additional step of Taumako between Tuvalu and the Reef islands (cf. Næss & Hovdhaugen 2011: 11).

For the migration to the northern Outliers, there are two possible paths, namely via both Tuvalu and Tokalau (cf. Carson 2012, Kirch 2017), that the methods described above could not resolve. In this case, we selected Tuvalu as the intermediate step due to its shorter absolute distance to the northern Outlier islands.

4 General trends and patterns

4.1 Consonants

The most common consonants in Polynesian are shown in Table 1. They all belong to the set of consonants reconstructed for Proto-Polynesian (cf. Biggs 1978: 708). The ones that are printed in black in Table 1 correspond to phonemes in at least 30 out of 35 languages (80%).

	labial	alveolar	velar	glottal
plosive	/p/	/t/	/k/	/ʔ/
fricative	/f/	/s/		/h/
nasal	/m/	/n/	/ŋ/	
liquid		/r/, /l/		
lateral				
glide				

Table 1: Common consonants (>80% in black, 50-80% in gray)

The only consonant that all 35 languages in the sample have is the plosive /p/; the other most common consonants are /k/ (32 languages), as well as the two nasals /m/ and /n/ (33 and 32 languages, respectively). This reflects earlier findings from historical linguistics, showing that the consonants /p/, /m/ and /n/ are the most diachronically stable ones in Polynesian with no or very few changes from Proto-Polynesian to the Polynesian languages spoken today (Marck 2000: 24). All other phonemes have undergone changes to varying degrees in a number of languages. This is reflected in Table 1, where the phonemes given in gray are present in 18-28 languages (50-80%). Out of all Proto-Polynesian phonemes, /ʔ/ has been shown to be the least stable in that it has been lost in comparatively many Polynesian languages (Marck 2000: 24-25).

Proto-Polynesian is reconstructed to have had both liquids /l/ and /r/, with /l/ developing into /r/ and vice versa in a number of Polynesian languages (Marck 2000: 52-57). This is why many Polynesian languages only feature one of the two glides: /r/ is found in 16 languages of our dataset, 19 have /l/, and only 7 languages have both. Five out of the languages with phonemic /l/ and /r/ are Outlier languages where /l/ is an innovation rather than a continuation from Proto-Polynesian. According to Elbert (1965: 435), Tikopia and Takuu split up /r/ into two distinct phonemes, whereas /l/ became a new phoneme in Mele-Fila, Emae, Futuna-Aniwa and Rennell-Bellona mainly through borrowing from Non-Polynesian languages.

Another unstable consonant in the history of Polynesian is the glide /w/, reconstructed for Proto-Polynesian (Marck 2000: 49). It developed into /v/ in a number of languages, which is why it is only part of the phoneme inventory in 6 languages from our dataset.

Voiced plosives are very rare in Polynesian languages. Out of all Polynesian languages in our dataset, only two languages have Rennell-Bellona and West Uvean have phonemic voiced plosives. Renell-Bellona features /b/ and /g/, while West Uvean has /b, g, d/ as well as the retroflex voiced plosives /ɖ/ and /ɟ/. In both languages, the phonemic status of these consonants is likely the result of language contact with speakers of Melanesian languages.

For West Uvean, there is clear evidence for contact with Non-Polynesian languages. The language is spoken on in the North and South of the Ouvéa island (Loyalities). It still is in direct contact with the Melanesian language Iaai, spoken in the center of the island (Ozanne-Rivierre 1994: 524). As will be discussed in more detail in Section 4.3, voiced plosives (among other phonemes) in West Uvean have been borrowed from Iaai (cf. also Sections 4.2 and 4.3). In addition, voiced plosives occurring in word-final position in West Uvean are the result of English and French loans rather than loans from Iaai (Ozanne-Rivierre 1994: 534-535).

Rennell-Bellona is currently not in contact with another Non-Polynesian language, but there is evidence for language contact in the (distant) past. Linguistic evidence for past language contact is a number of loan words of Melanesian origin in Rennell-Bellona (Carson 2012: 34). Long-term contact with Non-Polynesian Melanesian speaking communities (called *hiti*) is also plausible based on cultural and archaeological evidence (Elbert & Schütz 1988: 277-278). It is thus likely that when Polynesians settled in Renell and Bellona some time after

1000 AD, the islands were already inhabited by the *hiti* people, who then likely co-inhabited those islands with the Polynesians for some time (Carson 2012: 38).

Another group of consonants that are found in only some of the Polynesian languages are long consonants or geminates. Table 2 lists those long consonants together with the languages in our dataset that they occur in.

	/p:/	/t:/	/k:/	/f:/	/v:/	/s:/	/h:/	/m:/	/n:/	/ŋ:/	/l:/
Tuvalu	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Mele-Fila	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Sikaiana	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Takuu	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓
Nukuoro	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
Kapingamarangi	(p^{h})	(t^{h})	(k^{h})				✓	(m^{h})	(n^{h})	(ŋ^{h})	(l^{h})
West Uvean	(p^{h})	(t^{h})	(k^{h})					(m^{h})	(n^{h})	(ŋ^{h})	(l^{h})

Table 2: Phonematic long consonants

Except for Tuvalu, all languages with long consonants belong to the group of Outlier Polynesian languages, i.e. they are spoken in Melanesia and Micronesia to the West of the Polynesian triangle. Although Tuvalu is not classified as an Outlier language as such, it has been identified as one of the most likely origins or closest relative of some Outlier languages in the North (e.g. Carson 2012, Kirch 1984b, Pawley 1966). Among those Outliers are Nukuoru, Kapingamarangi, Takuu and Sikaiana, which also feature long consonants as shown in Table 2. Such long or geminated consonants have been argued to be a shared innovation in a number of Outlier languages (cf. Biggs 1978: 700-701 and Pawley 1967: 286-287). It is likely that they originate from earlier reduplication $C_1V_1C_2V_2$ sequences, in which V_1 was often unstressed and identical with V_2, which led to its loss. For instance, Proto-Polynesian **lelei* 'good' developed into Tuvaluan *llei* 'good' Pawley (1967: 286-287).

In addition to long consonants, Table 2 shows aspirated consonants in Kapingamarangi equivalent to the geminated ones in the other languages. It has been argued that aspirated consonants in Polynesian originate from the same $C_1V_1C_2V_2$ sequences as the geminated consonants. This development was explored in more detail by Milner (1958) for Tuvalu, providing convincing evidence that aspirated consonants are the result of reduplication with vowel elision.[8] Additionally, Milner (1958) argued that a similar process could have taken place in

[8] According to the description of Besnier (2000), we annotated Tuvalu as having long/geminated consonants rather than aspirated ones. However, Besnier (2000: 613) notes that "[g]eminated plosives/p t k/ are heavily aspirated." Thus, the phonetic properties of geminated consonants in Tuvalu provide more evidence for the fact that the developments of aspiration and gemination or long consonants are likely related.

Kapingamarangi as well. Næss & Hovdhaugen (2011) give a similar account of aspirated consonants for West Uvean; which we return to in Section 4.3.

According to Pawley (1967: 267), long consonants are the most important shared innovation for postulating a phylogenetic grouping which these languages all belong to. Still, he acknowledges that this shared feature may be the result of convergence, meaning that it was a parallel development facilitated by contact rather than an inherited property (Pawley 1967: 287). The remaining language with long consonants in Table 2 is Mele-Fila, another Outlier language spoken on the island of Efate and likely linked to East Uvean and East Futuna (Carson 2012) rather than Tuvalu. This does not fit in with the account of an inherited property in a very straight-forward way. A shared development due to contact, on the other hand, would be more apt to account for long consonants in Mele-Fila. Interestingly, we find a related phenomenon in Rennell-Bellona, another Outlier language from the southern group. Rennell-Bellona was described to have consonant clusters due to vowel loss in fast speech (Elbert & Schütz 1988: 17-19). It could well be that this reflects a similar development as the one leading to geminated (and aspirated) consonants, only that it stayed on the phonetic level without becoming integrated into the phonological system of the language.

4.2 Vowels

Polynesian vowel systems commonly distinguish between five vowel qualities. This can be seen in Table 3, showing the vowels that occur in >80% of the Polynesian languages in the dataset. These vowel qualities also correspond to the five reconstructed ones for Proto-Polynesian (Biggs 1978: 701). Reconstructing vowel length has proven to be difficult, as vowel length has long not been marked orthographically (Dempwolff 1929, Elbert 1953). While we do not have vowel length information for all reconstructed words, it is generally assumed that length distinctions for vowels were already part of Proto-Polynesian (Biggs 1971: 483).

	front		central		back	
high	/i/	/iː/			/u/	/uː/
mid	/e/	/eː/			/o/	/oː/
low			/a/	/aː/		

Table 3: Common vowels (>80%)

Compared to consonants, there is little variation or innovation in terms of vowel phonemes in Polynesian languages. There is one Outlier language in our dataset, West Uvean, that has a noteworthy innovation in terms of vowel phonemes.[9] West Uvean features the four additional

[9] Changes in vowel qualities have also been shown for Maori as a consequence of contact with New Zealand English during the last 100 years (Watson et al. 2016). Those changes in the vowel qualities differ from the

vowels /æ, ə, œ, y/. According to Ozanne-Rivierre (1994: 534), they have been borrowed from the Melanesian language Iaai due to long and intense language contact. Interestingly, in West Uvean the three vowels /œ/, /ə/ and /e/ are all used for /ə/ in Iaai loans, reflecting different stages of borrowing (Ozanne-Rivierre 1994: 540-541).

4.3 Inventory sizes

Figure 5 shows the geographic distribution of phoneme inventory sizes in Polynesian. The inventory sizes range from 13 in Tikopia to 40 phonemes in Mele-Fila, with a median of 20 phonemes. As can be seen in Figure 5, larger inventories tend to be located in the West in Melanesia and Micronesia, i.e. they tend to be found in Polynesian Outliers languages. Smaller inventories are more common in the Polynesian triangle, but they are certainly not confined to that area. Still, Figure 5 already suggests that language contact with non-Polynesian languages in Melanesia may be involved in explaining larger consonant inventories, especially for the seven largest inventories which are all situated in the west: Mela-Fila (40), West Uvean (36), Tuvalu (32) and Takuu (32), Sikaiana (30), Nukuoro (30), Vaeakau-Taumako (30) and Kapingamarangi (28).

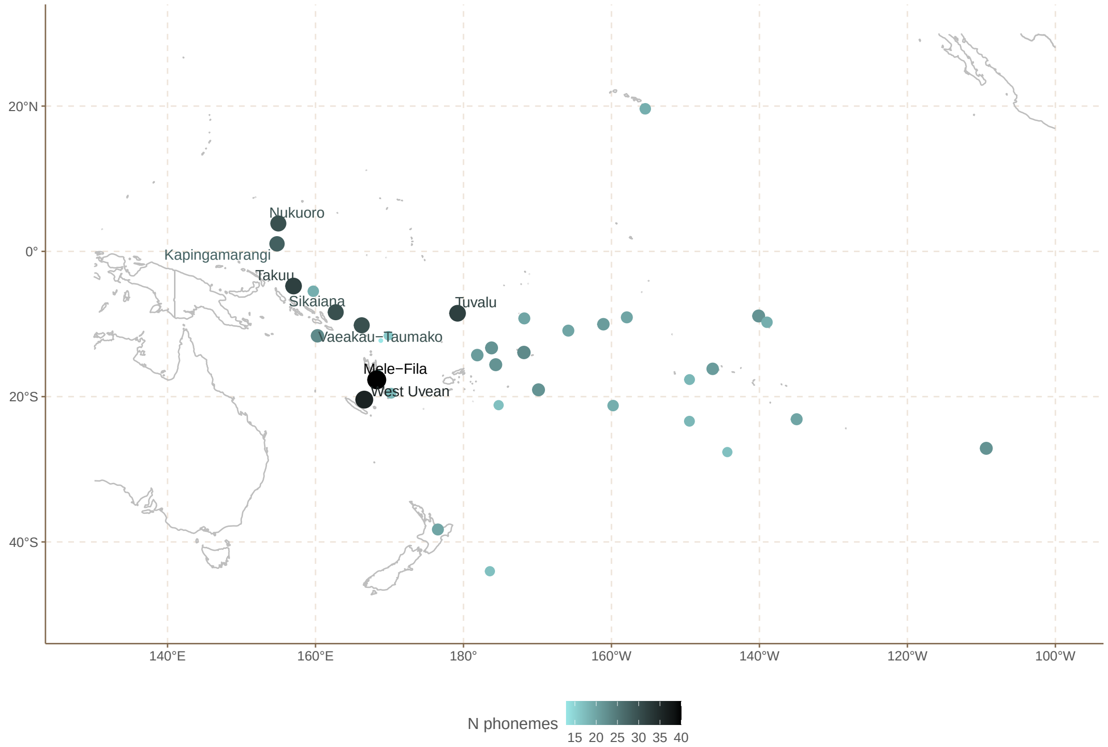

Figure 5: Distribution of phoneme inventory sizes

West Uvean situation, however, as they involve phonetic variation rather than the inclusion of additional vowel phonemes. Therefore, we do not discuss this in more detail here.

As was shown in Section 4.1, larger phoneme inventories are in part the consequence of the phonemic long consonants that have developed in a number of Outlier languages and in Tuvalu. Thus, except for West Uvean and Vaeakau-Taumako, the other languages with the largest phoneme inventories all have the phonemic opposition between long and short consonants (cf. Table 2). As was mentioned above, the development of long consonants is the result of vowel deletion between identical consonants. Yet, it is unclear to what extent this is necessarily a shared innovation in some common ancestor language or a parallel development in the single languages later on. The latter is not unlikely, given that we can assume extensive contact between the different Outlier languages (Clark 1994, Pawley 1967). Although not explicitly discussed in the literature, we cannot exclude contact with Non-Polynesian languages to have favored certain stress patterns that could have in turn facilitated the loss of unstressed vowels and thus the development of long consonants.

Related to that, Clark (1994: 117) describes a shift in word stress for Mele-Fila from the penultimate to the antepenultimate syllable. This shift is analyzed as the result of so-called intimate borrowing from Efate, which "requires prolonged intimacy between the two communities (such as frequent intermarriage over generations), affects all parts of linguistic structure, and in particular its lexical effects will not be localised but should pervade the lexicon as a whole" (Clark 1994: 113).

Geminated consonants are not the only extension that has led to larger phoneme inventory sizes in some of the Outliers, and it has long been noted that their phonological properties differ from Triangle Polynesian. Elbert (1965: 440-441) notes that West Uvean, Vaeakau-Taumako and Kapingamarangi show the greatest expansion in terms of phoneme inventories. He notes that both West Uvean and Vaeakau-Taumako are spoken in close proximity with other Non-Polynesian languages, whereas Kapingamarangi is fairly isolated linguistically. We will briefly discuss the phoneme inventory of these three languages here and relate them to their contact situation alluded to by Elbert (1965) as quoted above.

As mentioned, Kapingamarangi is fairly isolated, with the closest Non-Polynesian languages at about 450km (Clark 1994: 110). It is spoken on the Kapingamarangi atoll, which is the southern most point of Micronesia. The next closest atoll is Nukuoro (where another Polynesian Outlier language is spoken), at a distance of about 300km, which makes it one of the more isolated Polynesian languages. Its comparatively large phoneme inventory stems from the opposition between unaspirated and aspirated consonants as shown in Table 2 in Section 4.1. It is unclear whether this reflects a purely language-internal development or contact with other Polynesian languages.

One of the other languages that Elbert (1965) mentioned as having greatly increased their phoneme inventory is Vaeakau-Taumako. Vaeakau-Taumako is an Outlier language spoken the Reef Islands and Duff Islands in the eastern part of the Solomon Islands. There are many other Oceanic, Polynesian as well as Papuan languages spoken in the greater area of the Solomon Islands; the language spoken in closest vicinity to Vaeakau-Taumako is the Oceanic language Äiwoo.[10] Vaeakau-Taumako has indeed one of the largest phoneme inventories in our dataset, and similarly to Kapingamarangi, this is not due to phonemic long consonants either. One innovation is the phonemic status of the voiced plosives /b/ and /d/. Næss & Hovdhaugen (2011: 11) note that "it is possible that the presence of voiced oral stops in Vaeakau-Taumako, unusual for a Polynesian language, may reflect the presence of an earlier language community which shifted to the newly arrived language [...] However, as we have no knowledge of what such an earlier language may have been like, this cannot be verified; though it may be noted that the Main Reefs language Äiwoo, which might share an ancestor with this hypothetical original language if their speakers were part of the same Lapita expansion, has a full set of voiced oral stops."

Besides that, the large phoneme inventory of Vaeakau-Taumako is mainly due to a consistent distinction between unaspirated (/p, t, k, m, n, ŋ, l/) and aspirated (devoiced) consonants (/p^h, t^h, k^h, m^h, n^h, $ŋ^h$, l^h/). Although it is not entirely clear what the source of the aspirated consonant is for all cases, Næss & Hovdhaugen (2011: 37) mention two possible sources. For some cases, they show evidence that aspirated nasal is the result of vowel deletion between an unvoiced fricative and a unaspirated nasal. For other cases, Næss & Hovdhaugen (2011: 37) show that aspirated consonants have an origin in reduplication, similar to long or geminated consonants in a number of other Polynesian Outliers. They propose that in words with reduplicated material, the vowel of the reduplicated syllable was lost, and the resulting geminate consonant further developed into an aspirated consonant, as in *kai* (SG) > *kakai* (PL) > **kkai* > *khai* 'eat' (Næss & Hovdhaugen 2011: 37). In other words, aspirated, devoiced consonants in Vaeakau-Taumako may, at least in certain cases, represent a further development from geminated consonants. In Section 4.1, we showed that Kapingamarangi has aspirated variants of the same consonants that are commonly geminated in the other Outlier languages (cf. Table 2). Thus, while these developments are usually treated as a language-internal process, it is very likely that early contact between Outliers and other Polynesian as well as Non-Polynesian languages has spread or facilitated both developments. An opposition between unaspirated and aspirated plosives is not uncommon in the area and is found in other Melanesian languages. Tryon & Hackman (1983: 78) list a number of varieties spoken on Santa Isabel to "have developed a phonemic aspirated voiceless stop series". Other work on Oceanic languages of New Caledonia has shown correspondences between aspirated consonants in the North and high tones in the South, arguing for the same origin in reduplicated forms with vowel elision (Haudricourt 1968, Ozanne-Rivierre 1995, Rivierre 1993). To conclude, the development of

[10] The areas where Vaeakau-Taumako is spoken have experienced quite some migration over the last centuries, and still involves a high degree of mobility (Næss & Hovdhaugen 2011: 4, 11).

long and aspirated consonants has likely been facilitated by contact between Polynesian and other Non-Polynesian languages in Melanesia and West Polynesia.

The third language cited by Elbert (1965) for its large expansion of the phoneme inventory is West Uvean. As shown in Figure 5, West Uvean also has one of the largest phoneme inventories (36) in our dataset despite the lack of long consonants. Instead, West Uvean features a number of additional phonemes that are partially the result from phoneme splits, long-term contact with the Melanesian language Iaai and more recent contact with English and French. Although phoneme splits are traditionally seen as language-internal as opposed to language-external changes due to contact, certain language-internal developments related to voicing likely facilitated the borrowing of new phonemes from Iaai and their integration into West Uvean (Ozanne-Rivierre 1994: 538). Two important splits that occurred involve voicing. The two original voiceless plosives /k/ and /t/ became voiced /g/ and /d/, respectively, in word-initial and intervocalic positions. This contrast in voicing then also spread to bilabial plosives, resulting in the opposition between /p/ and /b/. On the one hand, the availability of a voicing opposition with plosives has allowed for word-final voiced plosives in later French and English loan words (Ozanne-Rivierre 1994: 534). On the other hand, it also likely facilitated maintaining a voicing contrast in the retroflex (/ʈ, ɖ/) and palatal plosives (/c, ɟ/) borrowed from Iaai. In addition, the Heo variety of West Uvean also retained some of the voicing opposition in nasals, laterals and glides from Iaai, which led to the additional voiceless /m̥ , n̥ , l̥ , w̥ / in the Heo variety of West Uvean. However, the Muli variety of West Uvean only uses their voiced counterparts /m, n, l, w/. Ozanne-Rivierre (1994: 540) relates this difference between varieties to differences in the degree of bilingualism due to their respective geographical locations. While the Heo variety is located in the North of the island of Ouvéa, the Muli variety is mostly spoken on Muli, a separate small island in the South West of Ouvéa. Ozanne-Rivierre (1994: 537) remarks that "[t]he Fagauvea [West Uvean] spoken in Heo (WUH), which is more influenced by Iaai because it is located on Ouvéa island itself, seems to have less of a tendency to 'nativise' than does the Fagauvea [West Uvean] spoken on the small island of Muli (WUM)." Thus, the Heo variety tends to conserve voiceless nasals borrowed from Iaai better than the Muli variety. Other West Uvean phonemes that are the result of language contact with Iaai are the consonants /θ, ʃ, w, ɲ/ and the vowels /æ, ə, œ, y/ (cf. Section 4.2). Interestingly, Ozanne-Rivierre (1994: 530-531) shows that /h/ in West Uvean is not simply a continuation of the reconstructed Proto-Polynesian **h*, but integrated into the language due to later loans from other Polynesian languages that West Uvean has been in contact with, possibly East Uvean or Tongan.

4.4 Taking stock

After the previous discussion, we come back to the questions we want to explore in this paper. First and center, we are interested in the distribution of PISs in Polynesian languages. This has been impacted by several factors: internal language change, contact between Polynesian languages during the expansion, and after, and contact to and borrowing from non-Polynesian languages. In order to explore the effect expansion has had on PIS, we need to take all these elements into account.

5 Statistical modeling of directed expansion and asymmetric contact

This section introduces the statistical techniques we use in this paper: phylogenetic regression to control for genetic biases (Section 5.1), a Stationary Gaussian Process to model symmetric contact between the Polynesian languages (Section 5.2), a Non-Stationary Gaussian Process to model the effect of the Polynesian expansion (Section 5.3), and what we call unidirectional contact effect estimation to estimate the effect of Non-Polynesian languages on the Polynesian ones (Section 5.4). For the implementation and model definition, see the supplementary materials.

5.1 Genetic bias: Phylogenetic regression

To control for genetic biases, we use phylogenetic regression (cf. Becker, Guzmán Naranjo & Ochs 2023, Bentz et al. 2015, Guzmán Naranjo & Becker 2022, Verkerk & Di Garbo 2022). We will not go into the details of phylogenetic regression in this paper, but the basic idea is that we add a group-level (random) effect for each language in our sample, and force the coefficient estimates of this term to be correlated between languages according to a pre-computed phylogenetic tree. This results in two languages which are more closely related to have more similar estimates than languages which are further apart in the phylogenetic tree. Instead of adding language families to the model (which are categorical), this approach captures that phylogenetic relations are gradual, languages can be more or less closely related and are not simply binned together in several categories. Therefore, phylogenetic regression also works well when all languages in a dataset come from the same phylogenetic unit such as Polynesian, as it can handle different degrees of relatedness as long we have information on that in the form of a phylogenetic tree. To build a phylogenetic tree, we used the phylogenetic information from Glottolog as shown in Figure 2.

5.2 Symmetric contact: Stationary Gaussian Process

To capture the fact that languages spoken in close proximity to each other have a greater potential to exhibit similar structures, we use a Gaussian Process (GP). GPs (Rasmussen 2003, 2004) were first introduced for use in typology by Guzmán Naranjo & Becker (2022) and Guzmán Naranjo & Mertner (2023), and their use has been further expanded by (Hartmann & Jäger 2023).[11] We will not discuss the mathematical details of GPs in this paper, but focus on high level explanations. The basic idea behind a GP is that we first calculate the distance between all observations in our dataset, and then use a kernel function to build a covariance matrix. This covariance matrix expresses the potential spatial correlation between observations. The key feature of a GP is that the spatial correlation between observation decays non-linearly with distance. Put differently, two languages spoken closely together will be highly correlated, but as the distance between languages increases, this correlation can drop to effectively 0.

One particular feature of the usual approach to GPs is that they are stationary. This means that the model effects are not dependent on the absolute location of the languages, but their locations in relation to each other. Consider the 1-dimensional toy example of a stationary GP in Figure 6. We see a stationary data-generating function (red line), data sampled from that function (light blue dots) and a corresponding 1-dimensional stationary GP fitted to the data (dark blue dots with uncertainty intervals).[12] In this type of scenario, the value of y_i does not depend on the absolute value of x_i, but on its distance to its neighbors and a general behavior of the function. The example in Figure 6 shows that a GP can track a non-linear function very well.

Since we deal with coordinates for the languages in the dataset, we need to build a two-dimensional distance matrix from the geodesic distance between all pairs of languages. We can then use the two-dimensional GP to estimate the spatial correlation between the PIS of Polynesian languages. In other words, the two-dimensional GP estimates how much of the variation in PIS can be attributed to the PIS of the languages spoken in proximity, i.e. contact.

5.3 Directed expansion: Non-stationary Gaussian Process

In addition to the stationary GP described above, we can use a different type of GPs to capture the potential effects of the Polynesian expansion from their historical point of origin. We can use a non-stationary GP to do that, because in contrast to stationary GPs, non-stationary GPs do consider absolute location in space. This is achieved by combining two kernels: a non-linear kernel like the one used in the stationary GP, and a linear kernel. The non-linear kernel

[11] There are other, similar approaches to modeling spatial relations such as splines. See Hartmann (2022) for a recent example.

[12] We sample from the data-generating function (a sine wave) by adding Gaussian noise to it.

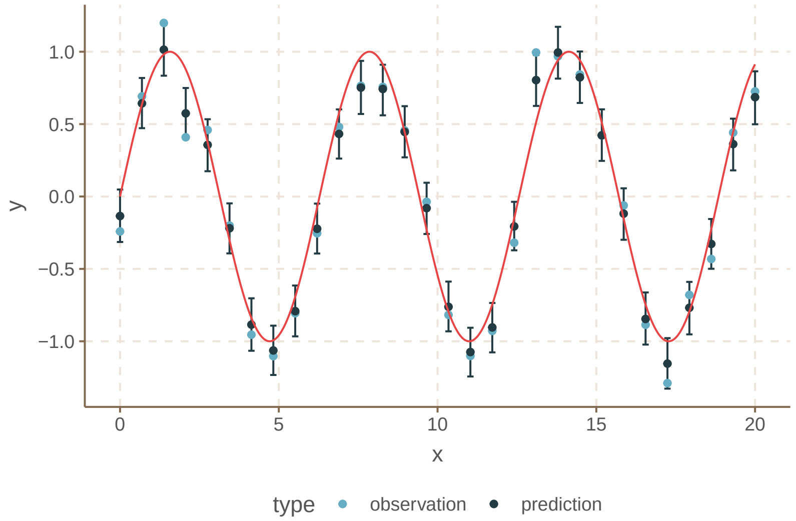

Figure 6: Stationary Gaussian Process

is responsible for capturing how observations influence each other, and the linear kernel is responsible for the absolute trend. To illustrate this, Figure 7 shows a non-stationary function (red line), the data sampled from that function (light blue dots), and a non-stationary GP fitted to the data (dark blue dots with uncertainty intervals).[13] In this example, the y value of observation i depends on both the absolute value of x for i, and the relative distance on x between i and its neighbors, and their values. While there are noticeable fluctuations in the value of y, overall, we can expect larger values of y for larger values of x. For this toy example in Figure 7, we see that the non-stationary GP is needed to capture the variation of y as a function of x.

To see how a stationary GP model performs when fitted to non-stationary data, compare Figures 8 and 9. The two figures show model performance when we fit the model to half of the data (x $<$ 10), and then predict the remaining half (x $>$ 10). The difference is clear: the stationary GP in Figure 9 fails to capture the general upward trend in the data and makes incorrect predictions for most of the test data. In contrast, the non-stationary GP in Figure 8 does a very good job of predicting the values of the test data including the overall upward trend.

This type of overall trend is what corresponds to the serial founder effect: as the distance between a language and the point of origin increases, the likelihood of a given event or feature systematically increases or decreases. Applied to the question of a potential founder effect for PIS, we would expect that phoneme inventories decrease in size with increasing overall distance to the point of origin.

While the previous example only contains 1-D data, we can expand this observation to

[13] In this case it is a sine wave added to a linear function of x.

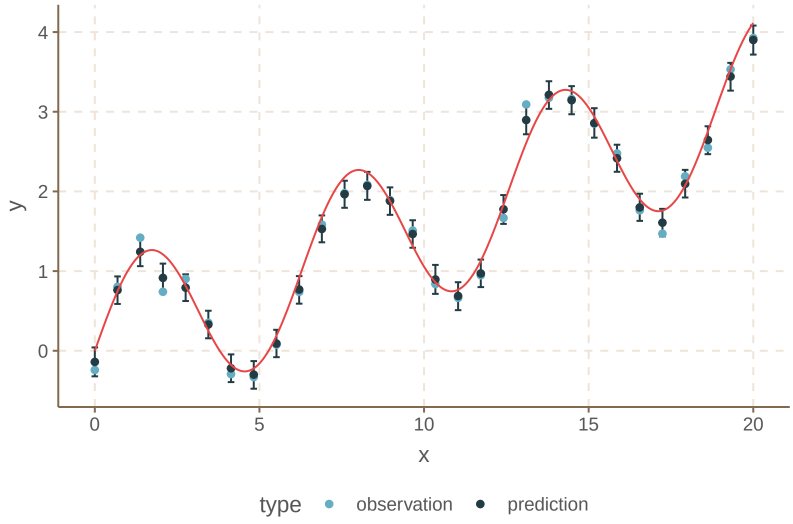

Figure 7: Non-stationary Gaussian Process

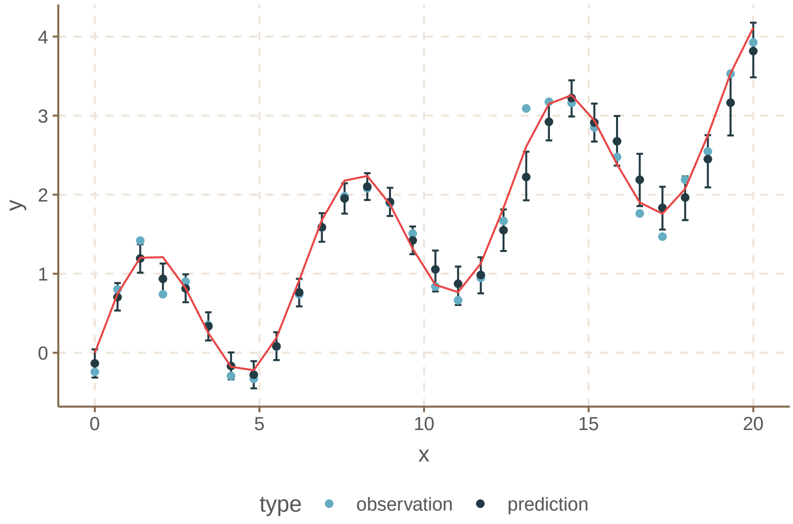

Figure 8: Non-stationary GP on non-stationary test data

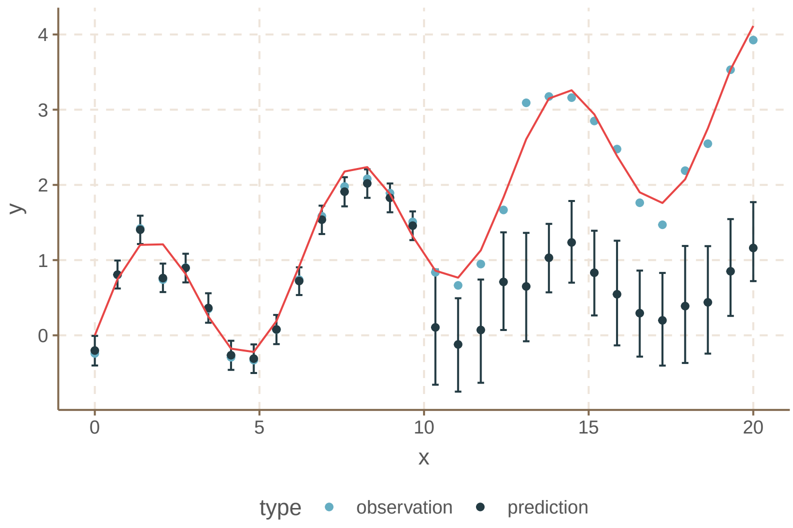

Figure 9: Stationary GP on non-stationary test data

a 2-dimensional scenario like before. The stationary component is built on top of distances between observations on a two dimensional plane, while the non-stationary component is built on top of the absolute distance of observations to a specific point on the plane. Thus, in addition to the two-dimensional stationary component described in Section 5.2, the two-dimensional non-stationary component of the GP can be used to model the potential effect of the absolute distance between the point of origin (Samoa) and the language on the PIS.

5.4 Asymmetric contact: unidirectional contact effect estimation

The final model component relates to the potential contact effect of Non-Polynesian languages on the Polynesian languages in our dataset. A simple solution would be to include all Non-Polynesian languages of the area as equal data points into our models. This solution, however, is not applicable to the scenario that we want to model for the purposes of the present study. There are three main reasons why Non-Polynesian languages cannot be included into the model in the same way as Polynesian languages.

First, the non-stationary GP estimates a linear and a non-linear component, but it can only do so jointly. Including the Non-Polynesian languages as equal data points would require us to apply the linear component to them as well. It is not clear how to do this, though, because we generally have much less information on potential migration paths, and because it is not even clear for some languages whether we should assume some form of migration in the first place. Second, for the purposes of the present study, the general properties of PISs in the region, as well as the effects Polynesian languages had on Non-Polynesian languages are not relevant.

Third, as this study aims to model variation in PISs, it would not be plausible to assume that Polynesian languages simply increase or decrease their number of phonemes in order to match the PIS of their neighboring Non-Polynesian languages. Instead, we expect contact to affect PIS in more indirect way in that it involves individual phonemes which then lead to changes in the PIS. Regardless the specific mechanisms involved, such types of contact effects cannot be captured by simply adding all Non-Polynesian languages to the model.

An alternative way to approach the effects of Non-Polynesian languages on the PIS of Polynesian languages is to model asymmetric contact effects. This corresponds to a scenario in which the borrowing of a new phoneme or a new contrastive feature (e.g. length) from Non-Polynesian leads to a change in the PIS of Polynesian languages.

To do so, we propose a new technique to capture this type of asymmetric contact influence situation. We call this method unidirectional contact effect estimation. The main idea is that we expect non-Polynesian languages to have an effect on Polynesian languages similarly to how we expect normal contact to happen, that is, languages closer to each other are expected to be able to influence each other more than languages further apart. In our context, we assume that Polynesian languages could borrow phonemes or features from non-Polynesian languages, and that this borrowing can lead to an increase on the PIS by direct incorporation of the feature. In general, as for the contact effects modeled by a stationary GP, we assume that languages spoken in closer proximity to each other can influence each other more than languages spoken further away. To capture this effect of Non-Polynesian on Polynesian, our method of unidirectional contact effect estimation includes three steps:

1. We fit a stationary GP to a model predicting the presence or absence of a phoneme or phonological feature p in Non-Polynesian languages. We do not include Polynesian languages in this part of the model. We also include a phylogenetic term for the non-Polynesian languages.
2. We predict the expected probability of p at the location of each Polynesian language L_i in our dataset. In other words, we do interpolation of probabilities based on the GP fitted to the Non-Polynesian languages for the locations of the Polynesian languages.
3. In a new model for the Polynesian languages, we use the estimated probability of p to predict the PIS of each L_i as a linear effect.

While we can expand this approach to an arbitrary number of phonemes, we need to consider that two phonemes can be correlated with each other, e.g [pʰ] and [kʰ]. Strong correlations between two phonemes can lead to multicolinearity issues in the estimates for the linear component. More importantly, strong correlations between multiple phonemes can lead to overly optimistic estimates if we model each phoneme independently. To account for this potential correlation, we do not fit independent (logistic) regression models to each phoneme. Instead, we fit a single multivariate probit model to all relevant phonemes (cf. Guzmán Naranjo &

Mertner 2023). Explained very briefly, multivariate probit regression models multiple binary outcomes as correlated by assuming an underlying multivariate normal distribution. For instance, this means that we do not model the probability of [p^h] separately from the probability of [k^h], but model both probabilities at the same time.

Another potential confound when estimating the asymmetric effects of Non-Polynesian languages on Polynesian languages comes from phylogenetic relatedness, since a number of them are Oceanic and thus related Austronesian languages. To control for this, we also include a phylogenetic term in the multivariate probit model.

The last remaining question regarding the effect of Non-Polynesian languages on Polynesian PISs is which phonemes should be included for the purposes of this study. We could technically include all phonemes found from all Non-Polynesian languages in our dataset, but this would cause difficulties with fitting the model. Therefore, we selected those phonemes and phonological features that cause the main amount of variation across Polynesian languages in terms of PIS. These are the presence / absence of:

- vowel length
- non-cardinal vowels
- consonant geminates
- voiced plosives
- long or aspirated plosives
- two or more liquids
- the phonemes /v/, /w/, /s/, /f/, /ŋ/, /ʔ/, /h/

It is important to note that we fit the unidirectional contact effects simultaneously with the rest of the model, not sequentially. We build a single model with multiple components, which we estimate jointly. That means all the components (phylogenetic, contact, expansion, unidirectional contact) are dependent on each other.

5.5 Estimating the likelihood of unidirectional contact: Mixture models

A further question that emerges from the discussion on how to capture unidirectional contact is that of detecting cases in which contact may have played a role and those in which it likely has not. In our particular case, we do not necessarily want to assume that all non-Polynesian languages have had an impact on all Polynesian languages. Inversely, we would like to know which Polynesian languages were likely most influence by non-Polynesian languages and which ones were not. This last point basically boils down to a question of probabilistic classification: how likely is each observation to belong to group A or group B.

To implement this idea we use what is called a mixture model[14] (Bradley, Fayyad & Reina 2000, Lindsay 1995, Rasmussen 1999). In a mixture model we assume that observations come from two different distributions. For example, Figure 10 shows the mixture of two normal distributions, one with mean = 0 and sd = 1 (group A), and the other with mean = 3 and sd = 2 (group B). In this examples, the observations come from two distinct groups (also called components) A and B. A mixture model can try to recover, for each observation, what the likelihood is of that observation belonging to either A or B.

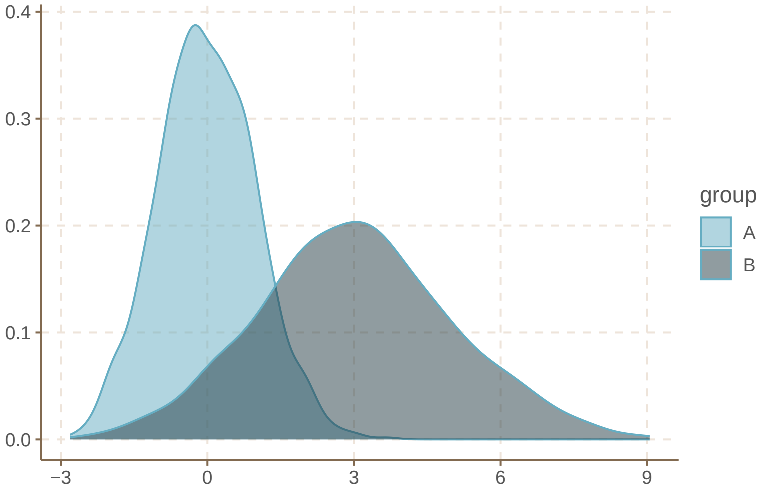

Figure 10: Example of mixture of two normal distributions

In our case, we assume that we have two distributions of observations: Polynesian languages not influenced by non-Polynesian languages, and Polynesian languages influenced by non-Polynesian languages. That is, one component is the model which only includes a stationary GP, and the other component is the model which includes a stationary GP and unidirectional contact.[15] It is more or less equivalent to using both models at the same time on the data, and trying to estimate which observations are likelier to be better explained by which model.

5.6 Taking stock

We thus have a more complex model structure that consists of two main components. The first one corresponds to the unidirectional contact effect estimation, where we use a multivariate probit model on the presence / absence of selected phonemes and phonological features in Non-Polynesian. This model includes a phylogenetic term (on each phoneme) and a stationary GP (on each phoneme.

[14] Not to be confused with a so-called mixed-effect model.

[15] While we could in theory use any combinations of the previous models and techniques, fitting mixture models can be very difficult computationally. We only explore this specific combination in this paper for reasons that will become apparent in the results section.

The second step corresponds to estimating PISs in Polynesian. To do so, we fit a series of Poisson models predicting the PISs of Polynesian languages. Here, we include a phylogenetic term, a GP (stationary or non-stationary) and a term based on the unidirectional contact effect estimation from step 1 to estimate the effect of Non-Polynesian on Polynesian. For this second step, we do not fit a single model including all terms. Instead, we fit a series of models including and excluding components in order to gauge their importance and effects on the final model. Table 4 shows a summary of all models fitted.

component 1	$\mathrm{m}_{\mathrm{lat}}$	unidirectional contact effect estimation
component 2	m_{s}	phylogenetic term + stationary GP
	m_{ns}	phylogenetic term + non-stationary GP
	$\mathrm{m}_{\mathrm{s\_uni}}$	phylogenetic term + stationary GP + unidirectional contact
	$\mathrm{m}_{\mathrm{ns\_uni}}$	phylogenetic term + non-stationary GP + unidirectional contact

Table 4: Summary of all models fitted

This model structure is more complex than that of more common linear or generalized models used in quantitative typology. This complexity is nevertheless necessary for a serious attempt to test the effect of spatial expansion (serial founder effect), including the effect of neighboring languages (asymmetric contact).

6 Model results

This section presents the model results. In Section 6.1, we analyze the fit of the models, which provides insights into how well they capture the observed data. Section 6.2 then examines model performance, comparing how well the models perform on new data. In Section 6.3, we zoom in on the model results for western Polynesian and Non-Polynesian contact effects. Finally, we explore the spatial patterns predicted by the models in Section 6.4.

6.1 Model fit and visual posterior evaluation

We begin by comparing how the four models that estimate PISs in Polynesian (m_{s}, m_{ns}, $\mathrm{m}_{\mathrm{s\_uni}}$, $\mathrm{m}_{\mathrm{ns\_uni}}$) fit the observed data. To do so, we can plot the posterior predictions of the models against the observed PISs. Figures 11 and 12 show the predictions vs. observations for the models m_{s} and $\mathrm{m}_{\mathrm{s\_uni}}$. Both models include a stationary GP to estimate the effects of contact within Polynesisan and, in the case of $\mathrm{m}_{\mathrm{s\_uni}}$, the effects of unidirectional contact from Non-Polynesian languages. The observed PISs are shown in red, with dots for core Polynesian and triangles for the Outlier languages. The predictions are shown in the form of box plots, with the central horizontal black line as the median estimate, the blue boxes correspond to the range that 50% of the predicted values falls into, and the whiskers indicate maximum and minimum values.

From the wide whiskers in Figures 11 and 12, we can observe that both models have a very large amount of uncertainty in the posterior predictions. At the same time, we see that the central mass of the predictions contains the observed values for most languages, as the observed values fall within the blue boxes. In Figure 11, which shows the results of m_s with no control for unidirectional contact with Non-Polynesian, there are seven languages for which the central mass of the predictions lies far from the real value. The languages are Anuta, Mele-Fila, Sikaiana, Takuu, Tikopia, Tuvalu and West-Uvean. All seven languages are all located in close proximity to Non-Polynesian languages; we can see from Figure 12 that the model m_{s_uni}, which accounts for unidirectional contact effects from Non-Polynesian languages, does better in predicting the PISs of those languages. We can thus conclude that Non-Polynesian languages likely have an impact on the PISs of a number of outlier Polynesian languages in Melanesia. Overall, the models fit the data relatively well, and most of the posterior mass is near the true value of the data. Mele-Fila is the only language in which the observed PIS lies outside of the central mass of the predictions, even when unidirectional contact effects from Non-Polynesian are included.

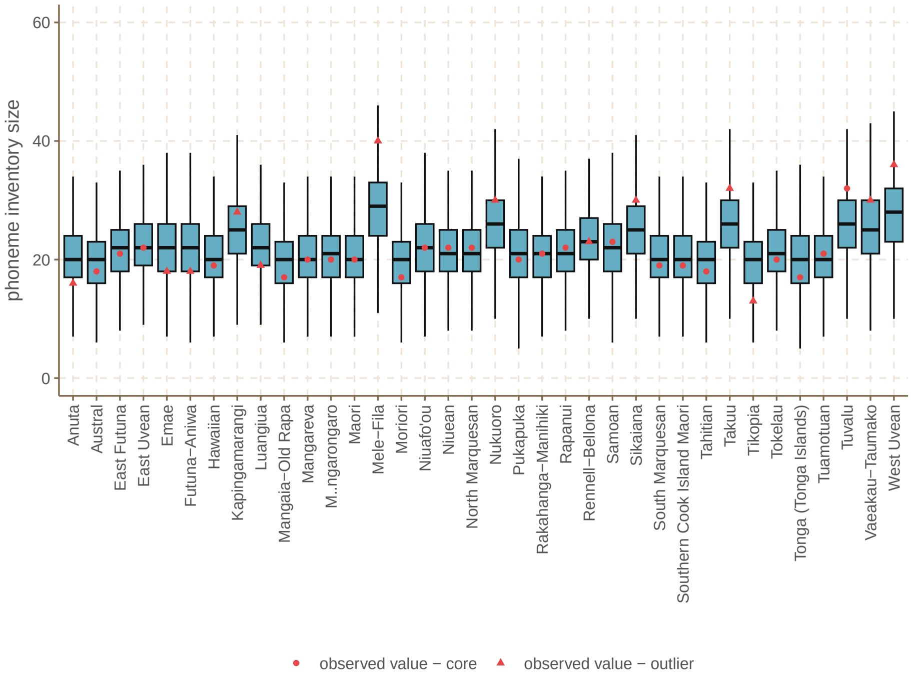

Figure 11: Predicted vs. observed PIS (m_s)

Figures 13 and 14 show the results for models m_{ns} and m_{ns_uni}, which include a non-stationary GP and the unidirectional contact effect component in the case of m_{ns_uni}. In contrast to the previous two models, these two thus estimate the effect of the distance between a

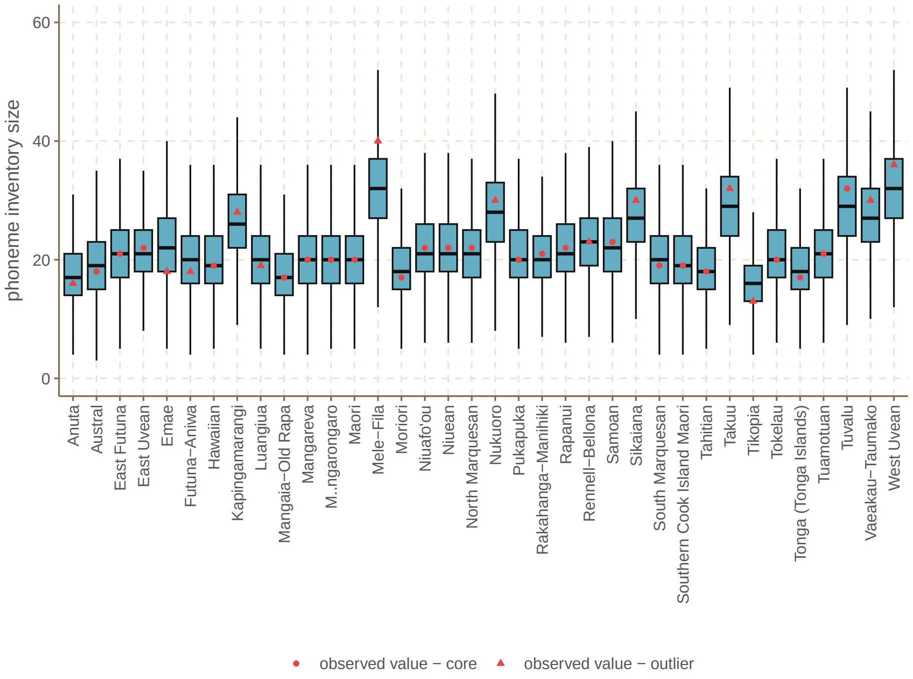

Figure 12: Predicted vs. observed PIS (m_{s_uni})

given Polynesian language and the Polynesian origin in Samoa.[16] In other words, Figures 13 and 14 show the results of models that estimate a serial founder effect for PISs in Polynesian. Comparing the model results to their respective non-stationary counterparts, i.e. Figures 13 and 11 as well as Figures 14 and 12, we see that the non-stationary GP component has little impact on model fit. Looking at the predictions from m_{ns}, which models the serial founder effect but no contact effects from Non-Polynesian, we see that the predictions for the seven languages mentioned above have not noticeably improved.

We can further explore model fit by comparing the root mean square error (RMSE) for the models. Simply put, RMSE values can range between zero and any positive value. The closer the value is to 0, the better is the model fit. Table 5 shows the RMSE values for the four models discussed, ranked according to their RMSE values from top to bottom. We can see a clear difference between the models including the unidirectional contact component and those that do not. The two models m_{ns_uni} and m_{s_uni} show a better model fit than the models m_{ns} and m_s. This confirms the impression from the model predictions discussed above that including unidirectional contact effects from Non-Polynesian languages is important to account for the variation in PISs in Polynesian. In addition, Table 5 suggests that there is very small advantage to include a non-stationary GP in terms of model fit.

Another way of assessing model fit is to visualize the mean error (ME) for each observation.

[16] We also tested Tonga as an origin for the expansion but saw no noticeable differences. Due to reasons of space we only report results for Samoa.

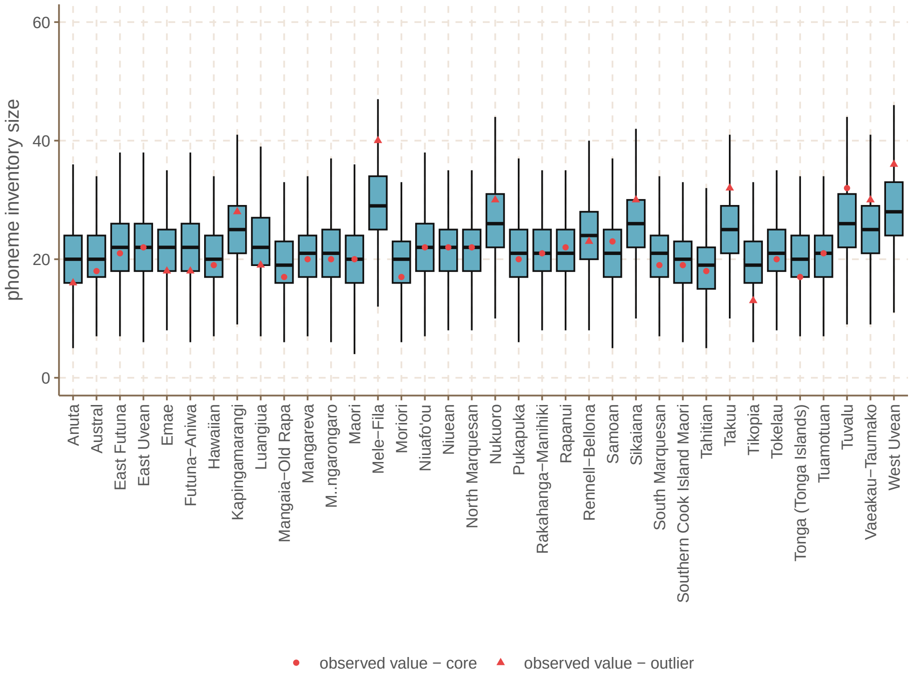

Figure 13: Predicted vs. observed PIS (m_{ns})

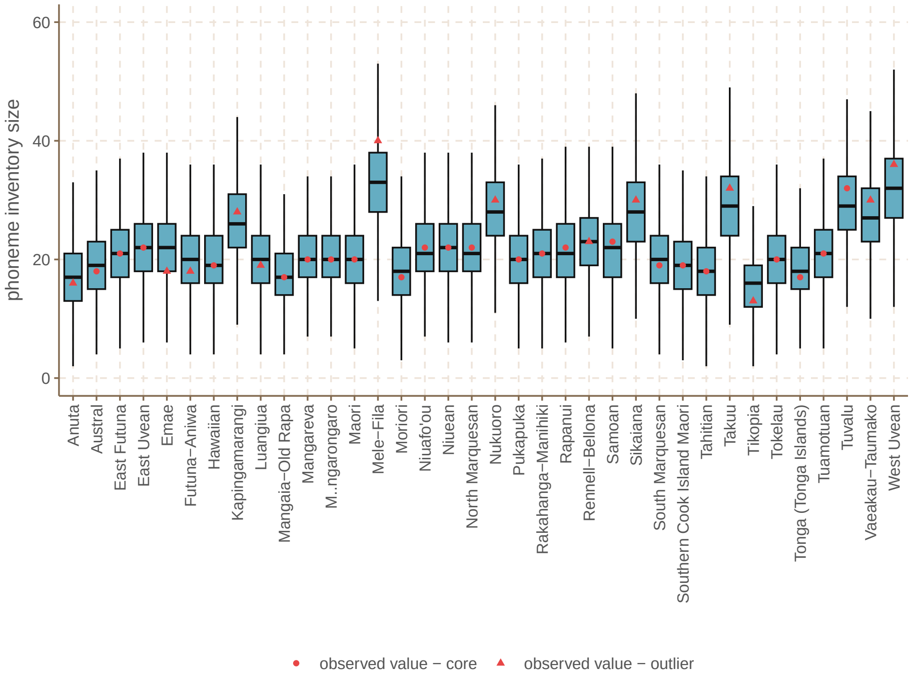

Figure 14: Predicted vs. observed PIS (m_{ns_uni})

model		RMSE
m_{ns_uni}	(non-stationary GP + unidirectional contact)	1.99
m_{s_uni}	(stationary GP + unidirectional contact)	2.15
m_{ns}	(non-stationary GP)	3.62
m_{s}	(stationary GP)	3.67

Table 5: RMSE for all six models.

We use ME values here instead of RMSE values, since ME values do not only indicate the magnitude of errors but indicate their direction. Positive values stand for over-preditions (too large predicted PIS), and negative ME values correspond to under-predictions (too small predicted PIS).

Figure 15 shows ME values for each observation for m_s, which captures neither unidirectional contact effects nor the expansion, and which had the worst model fit according to the RMSE scores.[17] Core Polynesian languages are represented by circles, and Outlier languages by triangles. Overpredictions are shown in read, and underpredictions in black. We observe that most of the errors occur in the western Polynesian languages, with the most extreme ME values for Takuu, Sikaiana, Tikopia, Mele-Fila and Tuvalu. Interestingly, we find both over-predictions (red) and under-predictions (black). Besides the Outlier languages, Tuvalu is one of the languages with larger magnitude ME.

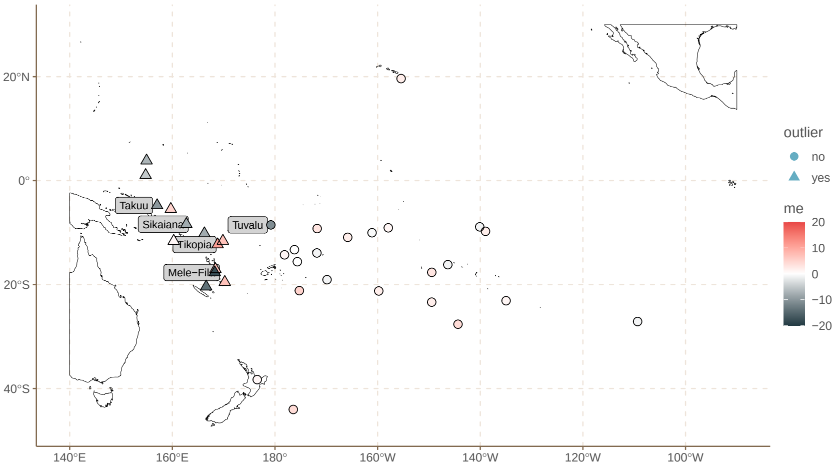

Figure 15: Mean errors (ME) for m_s

[17] The remaining plots based on the other three models can be found in the supplementary materials.

6.2 Model performance

In the previous section we looked at model fit, i.e. how well the models captured the data that they were trained on. This does not tell us how well the models perform on new data, though. To explore this question, we performed leave-one-out cross-validation (CV), i.e, we leave one observation out, fit the model on all other observations and try to predict the left-out observation. This is then repeated for all observations.

The results of the cross-validation are given in Table 6. The second column shows the difference in Expected Log Predictive Density (Δ ELPD) between the models. The absolute ELPD values are irrelevant; it is the relative difference between models that can be interpreted. The best-performing model is set to 0, and all other ELPD values are given in relation to the best-performing model. The larger the difference, the worse is the model's performance. To interpret the difference between ELPD values, the third column of Table 6 shows the standard error of the ELPD values. A usual threshold suggested for the standard error of Δ ELPD is that the latter should be four times as large as its standard error. Only then can we be certain about a real difference between the models and not just chance. Because ELPD values are hard to interpret, we also provide the RMSE of the model.

model		Δ ELPD	SE (Δ ELPD)	RMSE CV
m_s	(stationary GP)	0.0	0.0	6.10
m_{ns}	(non-stationary GP)	-0.5	0.5	6.07
m_{s_uni}	(stationary GP + unidirectional contact)	-5.8	2.2	6.23
m_{ns_uni}	(non-stationary GP + unidirectional contact)	-6.7	2.4	6.21

Table 6: Model performance

The ELPD results show that there is very little difference in terms of model performance between the 4 models. M_s and m_{ns} performs slightly better than the other two models including unidirectional contact from Non-Polynesisan languages. The differences are, however, rather small compared to their standard errors. Therefore, we do not have strong evidence for a substantial difference in model performance. With respect to the non-stationary GP that models the Polynesian expansion, this means that we do not have evidence supporting a serial founder scenario for PISs in Polynesian. Regarding unidirectional contact effects from Non-Polynesian, the ELPD results suggest that including those effects does not help with out-of-sample predictions, even though we saw in Section 6.1 that they improved model fit for the training data.

In terms of RMSE values, Table 6 shows that the model predictions on out-of-sample observations are about twice or three times worse than the predictions for observed data in Table 5. This means that although the models are able to track the spatial structures in the data, these spatial structures have relatively low predictive information for new observations.

6.3 Mixture model: Better representation of unidirectional contact

In the previous section, we have discussed how the unidirectional contact from Non-Polynesian languages ($\mathrm{m}_{\mathrm{s\_uni}}$ and $\mathrm{m}_{\mathrm{ns\_uni}}$) does not appear to improve model performance when predicting new observations. There are, however, a few important details regarding these contact effects that deserve further attention.

First, models $\mathrm{m}_{\mathrm{s\_uni}}$ and $\mathrm{m}_{\mathrm{ns\_uni}}$ assume that all Polynesian languages are subject to unidirectional contact effects from Non-Polynesian languages. As was briefly discussed in Section 4, however, we know from the literature that this mostly affects languages in Western Polynesia and especially the Outlier languages spoken in Melanesia and Micronesia. For languages spoken in central and eastern Polynesia, such contact is much less likely. In other words, we know that the amount of contact with Non-Polynesian languages is not identical for all Polynesian languages in the dataset.

We approach this issue with a so-called mixture model. The idea of a mixture model is that we assume for the response variable (PIS) to come from two independent distributions. In our case, one distribution contains information about unidirectional contact effects, while the other distribution does not contain any information thereof. The likelihood of each observation then results from the mixture of both distributions with differing proportions.

In theory, it is possible to build a model that knows exactly which observations belong the Outliers and which do not. This would provide the model with important information regarding which languages are highly impacted by unidirectional contact effects from Non-Polynesian and which are not.[18] However, this is not particularly useful. It is much more insightful to have the model estimate this information from the data itself and to compare the results with our previous assumptions based on evidence from the literature.

Figure 16 shows the estimated mixing proportions for unidirectional contact effects component. It is important to note that the mixture model has no information about which languages are Outliers (or any other classification to capture more and less contact with Non-Polynesian languages). The results are striking for how well they reflect our knowledge on how contact with Non-Polynesian languages affects different (groups of) Polynesian languages.

First, it is obvious that the higher mixing proportions for the unidirectional contact component are found among the Polynesian Outliers. The one exception to this overall picture is Tuvalu, which is not classified as an Outlier language but which is geographically (and genetically) very close to some of the Polynesian Outliers and likely serves as their origin (cf. Section 2.1). Second, the language with the highest mixing proportion is Mele-Fila, which had intense contact with non-Polynesian languages. Third, Nukuoro and Kapingamarangi are two Outlier languages spoken in Micronesia with relatively low mixing proportions. As was

[18] See the supplementary materials for a model that does exactly this.

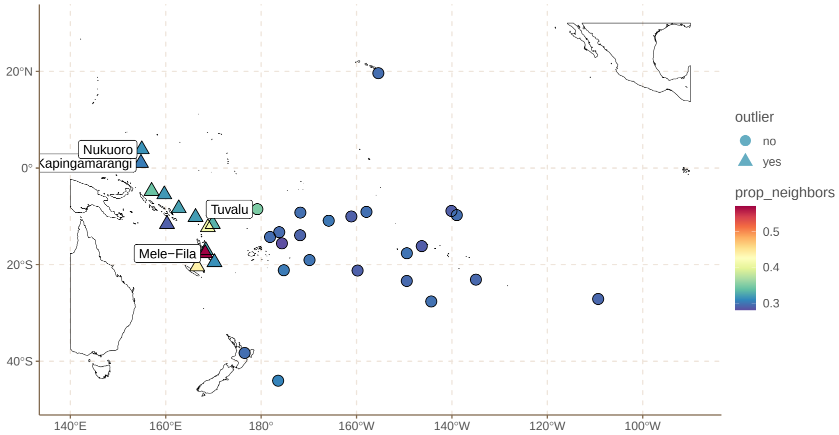

Figure 16: Mixing proportion for mixture model on neighborhood effects

mentioned in Section 4.3, this reflects their high degree of geographic isolation. According to Clark (1994: 110), the closest Non-Polynesian language is spoken on the Mortlock islands, which are at a distance of about 225km to Nukuoro and of about 475km to Kapingamarangi.

Although we only showed a brief example of using mixture models, it is an important approach to scenarios in which we need to identify the strength or relevance of contact between source languages and a number of potential target languages.

6.4 Predictions for spatial patterns

In this section, we explore the spatial effects that the different models predict. We use conditional effect plots to do so. A spatial conditional effect plot shows how the response variable (in this case PIS) changes across space, given the model parameters. The plot is created by defining a rectangular grid of regularly spaced of locations, and predicting the expected PIS at each location. These predictions do not take into account other aspects of the model (i.e. we fix the non-spatial parameters), and should not be interpreted in absolute terms. In other words, we can only interpret the relative differences between predictions across space, rather than absolute values of PIS. The predictions can be interpreted as the effect of the spatial component after controlling or accounting for all other predictors. Figures 17 to 19 all show the observed PIS values as red dots, with larger dots representing larger inventory sizes. The predictions are shown as colored areas ranging from dark blue / purple for relatively smaller PISs to orange for larger PISs.

We start by examining the conditional effects of m_s and m_{s_uni}, which have a stationary GP and control for unidirectional contact effects in the latter case. The conditional effects of those two models are shown in Figure 17. M_s (left plot) finds a relatively weak areal pattern.

Polynesian languages to the west (in Melanesia and central Polynesia) are estimated to have larger phoneme inventories than the languages in eastern Polynesia. This matches the fact that Melanesia contains a considerable number of non-Polynesian languages which have been in contact with the Polynesian languages spoken in this region.

The right plot in Figure 17 shows the spatial effects predicted by $\mathrm{m}_{\mathrm{s\_uni}}$, which includes a component to model unidirectional contact effects, finds a similar but weaker areal pattern. This is an important result. It means that a large portion of the spatial variance found with m_{s} can actually be captured by contact with Non-Polynesian languages and does not simply correspond to a spatial effect within Polynesian languages. In other words, this result of comparing the spatial effects of m_{s} and $\mathrm{m}_{\mathrm{s\_uni}}$ confirms that our technique to include unidirectional contact effects from Non-Polynesian languagse works as expected.

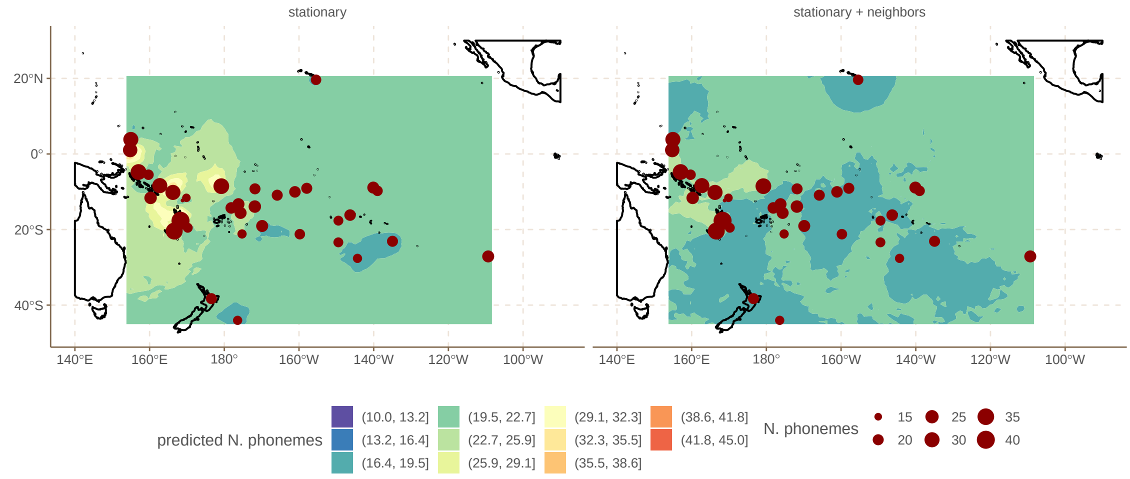

Figure 17: Spatial effects of m_{s} (left) and $\mathrm{m}_{\mathrm{s\_uni}}$ (right)

We turn to models m_{ns} and $\mathrm{m}_{\mathrm{ns\_uni}}$ next, which include a non-stationary GP and model unidirectional contact effects in the case of the latter. As described in Section 5.3, the non-stationary GP contains two components, namely a linear and a non-linear one. The linear component accounts for pontential expansion effects, while we use the non-linear one to capture contact. Figure 18 shows the linear GP component of both models without (m_{ns}) and with ($\mathrm{m}_{\mathrm{ns\_uni}}$) unidirectional contact effects.[19]

[19] The bubble shapes of the linear effects are due to how we calculated the distance from each point on the

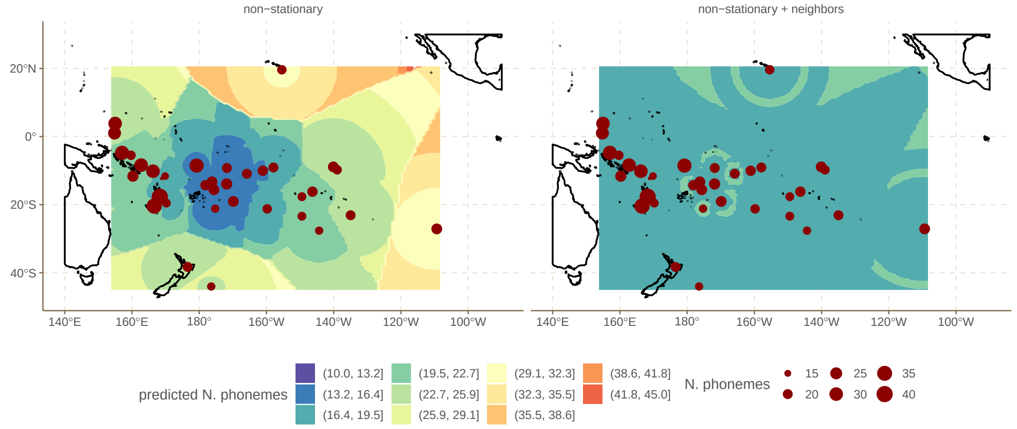

Figure 18: Spatial effects of the linear GP component of m_{ns} (left) and $\mathrm{m}_{\mathrm{ns\_uni}}$ (right)

Looking at the spatial effects of m_{ns} on the left of Figure 18, we see that PISs are predicted to be larger with larger distances to the origin in Samoa. This is exactly the opposite of what we would expect given the serial founder effect, which predicts that PISs become smaller with larger distances to the origin. However, once we control for unidirectional contact effects in $\mathrm{m}_{\mathrm{ns\_uni}}$ (right plot), the linear effect almost completely disappears. In other words, $\mathrm{m}_{\mathrm{ns\_uni}}$ no longer attributes most of the variation in PISs to the absolute locations of Polynesian languages but to an influence from neighboring Non-Polynesian languages. This shows that once we control for the effect of Non-Polynesian languages, we no longer find any evidence for an effect of the distance to Samoa as the Polynesian origin.[20] In other words, these results do not support the hypothesis of a serial founder effect in Polynesian PISs.

Figure 19 shows the combination of the linear and non-linear GP components of the non-stationary GP in the models m_{ns} (left) and $\mathrm{m}_{\mathrm{ns\_uni}}$ (right).[21] The spatial effects of the non-stationary GPs in m_{ns} and $\mathrm{m}_{\mathrm{ns\_uni}}$ shown here are much stronger than those of m_{s} and $\mathrm{m}_{\mathrm{s\_uni}}$, which only include stationary GPs (cf. Figure 17).

There are, however, two aspects to these patterns worth pointing out. First, since the non-

prediction grid to the origin. In order to calculate the distance from each location l on the grid to Samoa, we first calculate the geodesic distance from l to the nearest language in our dataset, and then we follow the graph from there to Samoa.

[20] Note that this is not due to the choice of Samoa as the origin of the Polynesian expansion. We see close to identical results for Tonga; see the supplementary materials for details.

[21] The spatial effects of the non-linear GP component of m_{ns} and $\mathrm{m}_{\mathrm{ns\_uni}}$ are not shown here, as they largely mirror the results shown in Figure 17. For more details, see the supplementary materials.

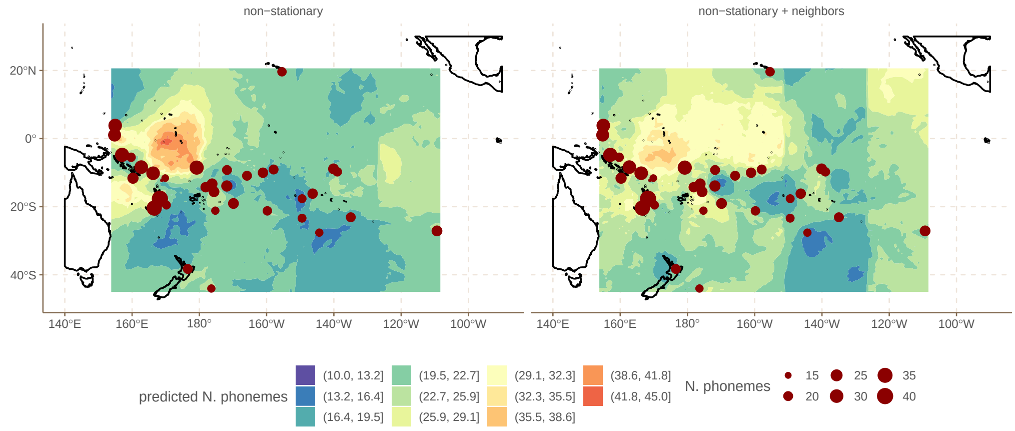

Figure 19: Linear and non-linear GP components of m_{ns} (left) and m_{ns_uni} (right)

stationary GP models m_{ns} and m_{ns_uni} did not perform noticeably better than the stationary GP models m_s and m_{s_uni}, it is likely that these more extreme areal patterns are over-fitting the data. Second, as we can see in both plots in Figure 19, the most extreme changes occur in the north-west, which is a region without any observations in our dataset. It is therefore likely that we deal with an artifact effect of the model, which means that it should be interpreted very carefully. The spatial effects of models m_s and m_{s_uni} with no component to model expansion (cf. Figure 17) did not show this hot spot.

7 Discussion

The results of this case study have several important consequences and implications regarding the hypothesis of a serial founder effect with PISs in particular, and the various aspects of modeling spatial and contact relations between languages in general. We first discuss the consequences of our results with respect to a potential serial founder effect in Section 7.1, the consequences for phoneme inventory sizes and contact in general in Section ?? and the broader implications for modeling space and languages in Section 7.3.

7.1 No evidence for a serial founder effects for PIS in Polynesian

The present study took as a point of departure the hypothesis of a serial founder effect for phoneme inventory sizes. Evidence for this hypothesis has been debated in the literature (e.g. Atkinson 2011a; Cysouw, Dediu & Moran 2012, Deshpande et al. 2009, Fort & Pérez-Losada 2016, Pérez-Losada & Fort 2018, Wang et al. 2012). The two most important methodological shortcomings of previous approaches that have argued for serial founder effects in the distribution of PISs are that they examine the effect on a global scale with many potential confounding factors, which are not sufficiently controlled for. In this study, we proposed a method to circumvent both shortcomings. We restricted our dataset to Polynesian languages, since their phylogenetic relations, spatial expansion from a common point of origin as well as contact with other languages is much better understood than it is for many other language families and areas of the world. In addition, we proposed advanced statistical methods to model spatial expansion (Section 5.3) while controlling for phylogenetic (Section 5.1) and contact relations between Polynesians (Section 5.2), as well as for unidirectional contact from Non-Polynesian languages on Polynesian (Section 5.4).

As was shown in Section 6, the results of our models not only show no support for a serial founder effect of PISs in Polynesia, but they show evidence against it. If unidirectional contact with Non-Polynesian languages is not taken into account, we do find an effect of expansion (cf. Figure 18). This effect, however, is the opposite to what has been argued for the serial founder effect. Instead of phoneme inventories to become smaller with increasing distance to the point of origin, we find that the models predict phoneme inventories to increase with increasing distance to the point of origin. We can say therefore say with a high degree of certainty that our results do not support the serial founder effect hypothesis with respect to PIS in Polynesian languages.

The results of this case study in the rather "controlled" context of Polynesian languages also has implications for previous global studies and their results. As we saw in Section 6, if we include contact with Non-Polynesian languages in the form of phoneme or feature borrowing into Polynesian, it can account for much of the variation of PISs in Polynesian across space. This suggests that language contact is an important confound not accounted for in previous quantitative accounts of serial founder effects with PISs. It is likely that previously found serial founder effects for PISs disappear once contact is more seriously controlled for in the models.

7.2 Linguistic implications for contact and phoneme inventory size

An important finding for Polynesian linguistics is a recurring pattern of Tuvalu in the models. Despite being a core Polynesian language, Tuvalu generally patterns with the Outlier languages with respect to its phoneme inventory (size). We observed this in error rates in predictions (Section 6.2) as well as in the mixing proportions of estimated indirect contact (Section 6.3). This result is in accordance with the literature about the special situation of Tuvalu with respect to Outlier Polynesian languages (cf. Sections 2.1 and 4).

Another important linguistic result is that we found strong evidence for phoneme inventory size to be affected by contact with neighboring languages. We clearly saw that Polynesian languages in Melanesia and western Polynesia with more contact to Non-Polynesian languages have larger phoneme inventories than languages further east. This finding confirms insights from the literature (cf. Matras 2009, Nichols 1992, Trudgill 2004) that long-term language contact with a high degree of bilingualism likely leads to the borrowing of phonemes and thus to an increase of PIS.

Nevertheless, we found that our models did not perform very well for predicting new data under cross-validation, despite fitting the data very well. This means that there is a portion of the variance in PIS in Polynesian languages which is not spatially or genetically conditioned.

7.3 Implications for modeling spatial relations between languages

In this study, we introduced four different modeling components to capture spatial relations: (i) a **stationary Gaussian Process** to model non-linear spatial correlations between observations, which can be used to capture symmetric contact effects between languages, (ii) a **non-stationary Gaussian Process** to model linear spatial correlations between observations, which can be used to capture the effects of spatial expansion from a point of origin, (iii) **unidirectional contact effect estimation** to model contact effects from one set of source languages on another set of potential target languages, and (iv) a **Mixture model** to estimate the likelihood of unidirectional contact effects in a set of potential target languages.

While the models with unidirectional contact effect estimation showed promising results, we can also observe that there were some aspects of the contact situation that our models did not capture. The models struggled most when predicting the PIS of Polynesian languages in close proximity to Non-Polynesian languages. This suggests that there are additional spatial and contact relations that would be integrated into our models.

We showed how unidirectional contact effect estimation can be implemented in statistical modeling. Our motivation for examining unidirectional contact effects was mainly a practical one, as our focus lied on modeling the variation in Polynesian phoneme inventory sizes and not in Non-Polynesian languages. There is, however, also linguistic evidence to support that contact effects between Melanesian languages and Polynesian Outliers have in fact been asymmetric, with a stronger influence from Melanesian languages on Polynesian than vice versa. For instance, Clark (1994) analyzes linguistic contact between the Non-Polynesian language Efate and the two Polynesian Outliers Mele-Fila and Emae in central Vanuatu. He finds a clear asymmetry in that Efate has influenced Mele-Fila and Emae to a larger extent on various linguistic levels than vice versa. To explain this, Clark (1994) distinguishes between "cultural" and "intimate" borrowing, with cultural borrowing being more superficial and intimate borrowing requiring a tighter and sustained socio-cultural interaction between the speaker populations. He concludes:

> The reason why Melanesian shows no signs of intimate borrowing from Polynesian in this case is probably to be explained simply on numerical grounds. We can safely assume that the first Polynesian-speaking immigrants were few in number and found an established Melanesian population. Even today, there are more than three times as many Melanesian as Polynesian speakers in the Efate region. If, as this suggests, Polynesians have always been a minority, they would, in establishing trade contacts or seeking spouses outside the village, have had to deal with Melanesian speakers more often than not, whereas Melanesian speakers, on average, would have had only a minority of Polynesian contacts. Melanesian wives, in particular, marrying into Polynesian villages, bearing and rearing children, speaking a Melanesian-influenced second-language variety, would have accomplished both the physical and the linguistic assimilation of the immigrants. (Clark 1986: 341)

This goes to show that our method of unidirectional contact effect estimation is useful to model cases of asymmetric language contact, where we have linguistic evidence for more influence from one language to the other.

As for the Mixture model technique, which we used to estimate the likelihood of unidirectional contact effects from Non-Polynesian to Polynesian languages, our results suggested that it worked as expected. For the most part, the results of this method were in accordance with the literature on the influence of Non-Polynesian languages on Polynesian languages. While further testing of this technique is still needed, it is a promising approach to detecting contact effects between groups of languages. The Mixture model method can be seen as a computational implementation of the idea proposed by Di Garbo & Napoleão de Souza (2023). Di Garbo & Napoleão de Souza (2023) suggest a method for finding contact which relies on looking at a target language (the language with potential contact effects), a neighbor language (from which the target language might have borrowed material), and a (set of) benchmark language (s) related to the target language. In their method, if the target and neighbor language share a feature not shared by the target and benchmark language, one can conclude borrowing between target and neighbor is likely. In our case, all Polynesian languages are target and benchmark at the same time, while the Non-Polynesian languages represent the neighbors.

Taking these aspects together, our case study showed that modeling spatial dynamics with relation to linguistic structure is not straightforward. Perhaps the main result is that there simply is no single best model. At most, one could argue that there is no strong justification for a model with an expansion component, but this is only so from a predictive perspective. Depending on the research question, it might make sense to include an expansion component even if it does not help predicting new observations. This would be the case, for instance, if the researcher wanted to "control for" all potential spatial confounds, including potential expansion effects.

One general implication is that there is no one silver bullet to "control for" contact or space. Our data is small enough that we are able to include much more complex spatial relations than adding some type of areas as group-level effect. However, the methods described in this study are not easy to scale up in order to handle large, global databases like WALS or Grambank. Our data is, by comparison to other global typological samples, fairly contained and manageable. Yet, we saw that there are difficult spatial relations which cannot be simply modeled using a single spatial component, but which require at least four different spatial components. And there are additional aspects that we would ideally want to capture better in a statistical model in order to represent spatial relation between in a more realistic way. For instance, we know that a more accurate spatial representation of single languages would not be a single point coordinate but a polygon that covers the entire area that this language is spoken in. Ideally, languages that are in contact should also be represented as overlapping polygons, given that contact stands for groups of speakers who use both languages.

To conclude, we showed that even a fairly contained area and research question can quickly lead to very complex statistical modeling if we want to represent the linguistic realities as accurately as possible. Other regions of the world may exhibit even more complex spatial dependencies. Researchers thus need to take special care when modeling spatial structures.

8 Concluding remarks

In this paper, we have proposed four different statistical modeling techniques to capture spatial relations between languages, using the distribution of phoneme inventory sizes in Polynesian as a test case. The four techniques were: (i) a **stationary Gaussian Process** to capture symmetric contact effects between languages, (ii) a **non-stationary Gaussian Process** to capture the effects of spatial expansion from a point of origin, (iii) **unidirectional contact effect estimation** to model contact effects from one set of source languages (Non-Polynesian) on another set of potential target languages (Polynesian), and (iv) a **Mixture model** to estimate the likelihood of unidirectional contact effects in a set of potential target languages. Overall, we have shown that there is no clear evidence for a serial founder effect with respect to phoneme inventory sizes in Polynesian. The method introduced in this paper could nevertheless be applied to other potential cases in which spatial expansion may play a more important role. We did find moderate evidence for unidirectional effects of Non-Polynesian languages on Polynesian in that the presence of certain phonemes or phonological features in Non-Polynesian languages led to larger phoneme inventories in neighboring Polynesian languages. Similarly, the method to model language unidirectional contact could be useful for other types of scenarios in which we expect some degree of asymmetry in contact, e.g. between languages with different degrees of prestige. Finally, the take-home message of this study is that modeling spatial relations between languages in a more realistic way is very complex. There is no single technique that will account or control for different types of contact and spatial phenomena, and it is necessary to test and combine different methods if we want to do the linguistic reality justice in statistical models.

References

Anderson, Victoria & Yuko Otsuka. 2006. The phonetics and phonology of ”definitive accent” in Tongan. *Oceanic Linguistics* 45(1). 21–42.

Atkinson, Quentin D. 2011a. Linking spatial patterns of language variation to ancient demography and population migrations. *Linguistic Typology* 15(2). 321–332.

Atkinson, Quentin D. 2011b. Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* 332. 346–349.

Bakker, Peter. 2004. Phoneme inventories, language contact, and grammatical complexity: a critique of trudgill. *Linguistic Typology* 8(3).

Bauer, Winifred. 1993. *Maori*. London: Routledge.

Becker, Laura, Matías Guzmán Naranjo & Samira Ochs. 2023. Socio-linguistic effects on conditional constructions: A quantitative typological study. In Silvia Ballarè & Guglielmo Inglese (eds.), *Sociolinguistic and typological perspectives on language variation*, 121–154. De Gruyter.

Bellwood, Peter. 1979. *Man's conquest of the Pacific*. New York: Oxford University Press.

Bentz, Christian et al. 2015. Adaptive communication: languages with more non-native speakers tend to have fewer word forms. *PLOS ONE* 10(6). e0128254.

Besnier, Niko. 2000. *Tuvaluan: a Polynesian language of the Central Pacific*. London: Routledge.

Betti, Lia et al. 2009. Distance from africa, not climate, explains within-population phenotypic diversity in humans. *Proceedings of the Royal Society B: Biological Sciences* 276(1658). 809–814.

Biggs, Bruce. 1971. The languages of Polynesia. In *Linguistics in Oceania*, 466–506. Berlin: De Gruyter.

Biggs, Bruce. 1978. The History of Polynesian Phonology. In *Second International Conference on Austronesian Linguistics: Proceedings*, 691–716. Canberra: Pacific Linguistics.

Blust, Robert. 2013. *The Austronesian languages*. Revised edition. Australia: Asia-Pacific Linguistics.

Bradley, Paul S, UM Fayyad & CA Reina. 2000. Clustering very large databases using em mixture models. In *Proceedings 15th international conference on pattern recognition. icpr-2000*, vol. 2, 76–80.

Carson, Mike. 2012. Recent developments in prehistory: Perspectives on settlement chronology, inter-community relations, and identity formation. In Richard Feinberg & Richard Scaglion (eds.), *Polynesian outliers: the state of the art* (Ethnology Monographs 21), 27–48. Pittsburgh, PA: University of Pittsburgh.

Clark, Ross. 1986. Linguistic convergence in Central Vanuatu. In Paul Geraghty & Louis Carrington (eds.), *FOCAL II : Papers from the Fourth International Conference on Austronesian Linguistics*, 333–342. Canberra: Pacific Linguistics.

Clark, Ross. 1994. The Polynesian outliers as a locus of language contact. In Tom Dutton & Darrell Tryon (eds.), *Language contact and change in the Austronesian world*, 109–140. Berlin: De Gruyter.

Creanza, Nicole et al. 2015. A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences* 112(5). 1265–1272.

Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). 377–393.

Cysouw, Michael, Dan Dediu & Steven Moran. 2012. Comment on "phonemic diversity supports a serial founder effect model of language expansion from africa". *Science* 335(6069). 657–657.

Dempwolff, Otto. 1929. Das austronesische Sprachgut in den polynesischen Sprachen. *Festbundel, uitgegeven door het Koninklijk Bataviaasch Genootschap van Künsten en Wetenschappen bij gelegenheid van zijn 150 jarig bestaan* 1. 62–86.

Deshpande, Omkar et al. 2009. A serial founder effect model for human settlement out of africa. *Proceedings of the Royal Society B: Biological Sciences* 276(1655). 291–300.

Di Garbo, Francesca & Ricardo Napoleão de Souza. 2023. A sampling technique for worldwide comparisons of language contact scenarios. *Linguistic Typology* 27(3).

Donohue, Mark & Johanna Nichols. 2011. Does phoneme inventory size correlate with population size? *Linguistic Typology* 15(2). 161–170.

Elbert, Samuel. 1953. Internal relationships of Polynesian languages and dialects. *Southwestern Journal of Anthropology* 9(2). 147–173.

Elbert, Samuel. 1965. Phonological expansion in outlier Polynesia. *Lingua* 14. 431–442.

Elbert, Samuel & Albert Schütz. 1988. *Echo of a culture: A grammar of Rennell and Bellona.* Honolulu, Hawaii: University of Hawaii Press.

Fenk-Oczlon, Gertraud & Jürgen Pilz. 2021. Linguistic complexity: relationships between phoneme inventory size, syllable complexity, word and clause length, and population size. *Frontiers in Communication* 6. Publisher: Frontiers Media SA, 626032.

Fort, Joaquim & Joaquim Pérez-Losada. 2016. Can a linguistic serial founder effect originating in africa explain the worldwide phonemic cline? *Journal of The Royal Society Interface* 13(117). 20160185.

Geraghty, Paul. 1983. *The history of the Fijian languages* (Oceanic Linguistics Special Publication 19). Honolulu: University of Hawaii Press.

Goodwin, Ian, Stuart Browning & Atholl Anderson. 2014. Climate windows for Polynesian voyaging to New Zealand and Easter Island. *Proceedings of the National Academy of Sciences* 111(41). 14716–14721.

Green, Robert & Marshall Weisler. 2002. The Mangarevan Sequence and Dating of the Geographic Expansion into Southeast Polynesia. *Asian Perspectives* 41(2). 213–241.

Green, Roger. 1966. Linguistic subgrouping within Polynesia: The implications for prehistoric settlement. *The Journal of the Polynesian Society* 75(1). 6–38.

Green, Roger. 1981. Location of the Polynesian homeland: A continuing problem. In Jim Hollyman & Andrew Pawley (eds.), *Studies in Pacific languages and cultures in honor of Bruce Biggs*, 133–158. Auckland: Linguistic Society of New Zealand.

Greenhill, Simon & Ross Clark. 2011. POLLEX-Online: The Polynesian Lexicon Project Online. *Oceanic Linguistics* 50(2). 551–559.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3).

Guzmán Naranjo, Matías & Miri Mertner. 2023. Estimating areal effects in typology: a case study of african phoneme inventories. *Linguistic Typology* 27(2).

Hartmann, Frederik. 2022. Methodological problems in quantitative research on environmental effects in phonology. *Journal of Language Evolution* 7(1). tex.eprint: https://academic.oup.com/jole/artic pdf/7/1/95/45052002/lzac003.pdf, 95–119.

Hartmann, Frederik & Gerhard Jäger. 2023. Gaussian process models for geographic controls in phylogenetic trees. *Open Research Europe* 3(57). 57.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different. In *Aspects of linguistic variation*, 83–114. De Gruyter.

Haudricourt, André. 1961. Richesse en phonèmes et richesse en locuteurs. *L'Homme* 1. 5–10.

Haudricourt, André. 1968. La langue de Gomen et la langue de Touho en Nouvelle-Calédonie. *Bulletin de la Société de Linguistique de Paris* 63(1). 218–235.

Hay, Jennifer & Laurie Bauer. 2007. Phoneme inventory size and population size. *Language* 83(2). 388–400.

Irwin, Geoffrey. 1994. *The prehistoric exploration and colonialisation of the Pacific*. Cambridge: Cambridge University Press.

Jennings, Jesse. 1979. *The prehistory of Polynesia*. Cambridge: Harvard University Press.

Kahn, Jennifer & Yosihiko Sinoto. 2017. Refining the Society Islands cultural sequence: Colonization phase and developmental phase coastal occupation on Mo'orea island. *Waka Kuaka* 126(1). 33–60.

Kennett, Douglas et al. 2012. A Bayesian AMS 14C chronology for the colonisation and fortification of Rapa Island. In Atholl Anderson & Douglas Kennett (eds.), *Taking the High Ground: The archaeology of Rapa, a fortified island in remote East Polynesia*, 189–202. Canberra: ANU Press.

Kieviet, Paulus. 2017. *A grammar of Rapa Nui* (Studies in Diversity Linguistics 12). Berlin: Language Science Press.

Kirch, Patrick. 1984a. *The evolution of the Polynesian chiefdoms.* Cambridge: Cambridge University Press.

Kirch, Patrick. 1984b. The polynesian outliers: Continuity, change, and replacement. *Journal of Pacific History* 19. 224–238.

Kirch, Patrick. 1996. Lapita and its aftermath: The Austronesian settlement of Oceania. *Transactions of the American Philosophical Society* 86(5). 57–70.

Kirch, Patrick. 1997. *The Lapita peoples: Ancestors of the Oceanic world.* Oxford: Blackwell.

Kirch, Patrick. 2017. *On the ROAD of the WINDS: An archaeological history of the Pacific Islands before European contact.* Oakland, CA: University of California Press.

Kirch, Patrick & Roger Green. 1992. History, phylogeny, and evolution in Polynesia. *Current Anthropology* 33(1). 161–186.

Kirch, Patrick & Roger Green. 2001. *Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology.* Cambridge: Cambridge University Press.

Kurpa, Viktor. 1973. *Polynesian Languages: A Survey of Research.* The Hauge/Paris: Mouton.

Lindsay, Bruce G. 1995. *Front cover image for mixture models : theory, geometry, and applications mixture models : theory, geometry, and applications.* Hayward, Calif.: Institute of Mathematical Statistics.

Manica, Andrea et al. 2007. The effect of ancient population bottlenecks on human phenotypic variation. *Nature* 448(7151). 346–348.

Marck, Jeff. 2000. *Topics in Polynesian language and culture history.* Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Martinsson-Wallin, Helène & Susan J. Crockford. 2001. Early Settlement of Rapa Nui (Easter Island). *Asian Perspectives* 40(2). 244–278.

Matras, Yaron (ed.). 2009. *Language contact.* Cambridge: Cambridge University Press.

Milner, G.B. 1958. Aspiration in two Polynesian languages. *Bulletin of the School of Oriental and African Studies, University of London* 21(1). 368–375.

Moran, Steven, Daniel McCloy & Richard Wright. 2012. Revisiting population size vs. phoneme inventory size. *Language* 88(4). 877–893.

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan reference grammar.* Oslo: Scandinavian University Press.

Næss, Åshild & Even Hovdhaugen. 2011. *A grammar of Vaeakau-Taumako.* Berlin: De Gruyter.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: The University of Chicago Press.

Ozanne-Rivierre, Françoise. 1994. Laai loanwords and phonemic changes in Fagauvea. In Tom Dutton & Darrell Tryon (eds.), *Language contact and change in the Austronesian world*, 523–549. Berlin: De Gruyter.

Ozanne-Rivierre, Françoise. 1995. Structural Changes in the Languages of Northern New Caledonia. *Oceanic Linguistics* 34(1). 45–72.

Pawley, Andrew. 1966. Polynesian languages: A subgrouping based on shared innovations in morphology. *The Journal of the Polynesian Society* 75(1). 39–64.

Pawley, Andrew. 1967. The Relationships of Polynesian Outlier Languages. *The Journal of the Polynesian Society* 76(3). 259–296.

Pawley, Andrew. 2007. The origins of early Lapita Culture: The testimony of historical linguistics. In Stuart Bedford, Christophe Sand & Sean Connaughton (eds.), *Oceanic explorations: Lapita and western Pacific settlement*. Canberra: ANU Press.

Pawley, Andrew & Roger Green. 1973. Dating the Dispersal of the Oceanic Languages. *Oceanic Linguistics* 12(1/2). 1–67.

Pawley, Andrew & Roger Green. 1984. The Proto-Oceanic language community. *The Journal of Pacific History* 19(3). 123–146.

Pérez-Losada, Joaquim & Joaquim Fort. 2018. A serial founder effect model of phonemic diversity based on phonemic loss in low-density populations. *PLOS ONE* 13(6). e0198346.

Pericliev, Vladimir. 2004. There is no correlation between the size of a community speaking a language and the size of the phonological inventory of that language. *Linguistic Typology* 8(3). 376–383.

Pierce, Amanda A. et al. 2014. Serial founder effects and genetic differentiation during worldwide range expansion of monarch butterflies. *Proc Biol Sci.* 281(1797).

Ranacher, Peter et al. 2021. Contact-tracing in cultural evolution: a Bayesian mixture model to detect geographic areas of language contact. *Journal of the Royal Society Interface* 18(181). 1–15.

Rasmussen, Carl. 1999. The infinite gaussian mixture model. *Advances in neural information processing systems* 12.

Rasmussen, Carl. 2003. Gaussian processes in machine learning. In *Summer school on machine learning*, 63–71. Springer.

Rasmussen, Carl. 2004. Gaussian processes in machine learning. In Olivier Bousquet, Ulrike von Luxburg & Gunnar Rätsch (eds.), *Advanced lectures on machine learning*, 63–71. Berlin: Springer.

Rice, Keren. 2004. Language contact, phonemic inventories, and the Athapaskan language family. *Linguistic Typology* 8(3). 321–343.

Rivierre, Jean-Claude. 1993. Tonogenesis in New Caledonia. *Oceanic Linguistics Special Publications* 24. 155–173.

Rolle, Nicholas. 2009. The Phonetic Nature of Niuean Vowel Length. *Toronto Working Papers in Linguistics* 31.

Taumoefolau, Melanaite. 2002. Stress in Tongan. *MIT Working Papers in Linguistics* 44. 341–354.

Trudgill, Peter. 2004. Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8(3). 305–320.

Trudgill, Peter. 2011. Social structure and phoneme inventories. *Linguistic Typology* 15(2). 155–160.

Tryon, Darrell & B.D. Hackman. 1983. *Solomon Islands languages: An internal classification.* Canberra: Pacific Linguistics.

Tryon, Darrell T. (ed.). 1995. *Comparative austronesian dictionary: an introduction to austronesian studies.* Berlin, New York: De Gruyter Mouton.

Urban, Matthias & Steven Moran. 2021. Altitude and the distributional typology of language structure: ejectives and beyond. *Plos one* 16(2). Publisher: Public Library of Science San Francisco, CA USA, e0245522.

Verkerk, Annemarie & Francesca Di Garbo. 2022. Sociogeographic correlates of typological variation in northwestern bantu gender systems. *Language Dynamics and Change* 1. Publisher: Brill, 1–69.

Walworth, Mary. 2014. Eastern Polynesian: The Linguistic Evidence Revisited. *Oceanic Linguistics* 53(2). 256–272.

Wang, Chuan-Chao et al. 2012. Comment on "Phonemic diversity supports a serial founder effect model of language expansion from Africa". *Science* 335(6069). 657–657.

Ward, Gerard, John Webb & M. Levison. 1973. The Settlement of the Polynesian Outliers: A Computer Simulation. *The Journal of the Polynesian Society* 82(4). 330–342.

Watson, Catherine et al. 2016. Sound change in Māori and the influence of New Zealand English. *Journal of the International Phonetic Association* 46(2). 185–218.

Wichmann, Søren, Taraka Rama & Eric W. Holman. 2011. Phonological diversity, word length, and population sizes across languages: The ASJP evidence. *Linguistic Typology* 15(2). 177–197.

Wilmshurst, Janet M et al. 2011. High-precision radiocarbon dating shows recent and rapid initial human colonization of east polynesia. *Proceedings of the National Academy of Sciences* 108(5). 1815–1820.

Wilson, William H. 2012. Whence the East Polynesians? Further Linguistic Evidence for a Northern Outlier Source. *Oceanic Linguistics* 51(2). 289–359.