Journal of Language Evolution, N/A, 1–98
doi: N/A
Advance Access Publication Date: N/A
Paper

A Sea of Languages: predicting the number of languages per island group in Remote Oceania

Anonymous Author[1,*]

[1]Department, Organization, Street, Postcode, State, Country

*Corresponding author. email-id.com

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

There are more than 7,000 languages on our planet today, but they are not evenly distributed. There are over 100 languages in Vanuatu, but only one in Sāmoa. This is in spite of the populations being similarly sized, having similar histories etc. Why might this be? This paper explores this question for one particular region: Remote Oceania. Remote Oceania comprises the eastern Solomon Islands (Temotu), Vanuatu, New Caledonia, Fiji, Polynesia and Micronesia. The region was settled more recently than the rest of Oceania and initially entirely by Austronesian speakers. Through studies in archaeology, history and linguistics, as well as community knowledge, much is known about migration patterns and societal organization of this part of the world. The comparatively shallow time depth, the relatedness of the languages and established knowledge about the history make it an ideal base for understanding language diversification. One particular theory has been proposed to explain the discrepancy in language richness over Remote Oceania — societal organisation. In this study, we test if this factor remains an important predictor of the number of languages also when other relevant factors are taken into account, such as time-depth and environment, using a Bayesian Regression model approach. The results show that the theory does hold in terms of statistical power, but that there are important caveats such as the instantiation of the variable "political complexity" and its relationship with contact with Oceanic neighbours in the New Guinea region.

Key words: languages, Oceania, predicting, political complexity

Introduction

The number of languages in the world is not equally distributed over area and population. In the modern country of Papua New Guinea, there are 835 languages over a population of 9.7 million people (United Nations, Department of Economic and Social Affairs, Population Division, 2022; Hammarström et al., 2017). However, in South Korea, there are only two indigenous spoken languages in a population of 52 million people. How can this be?

Understanding the mechanics of language change and diversification is essential to understanding human history. Language is interlinked with cultural identity, learning about the historical dynamics of language informs our understanding of our story generally. The question of language splitting in particular is interesting as it also implies community splitting. What causes us to lose touch, to diverge and go separate ways? This research topic has received a lot of attention in recent years (c.f. Gavin et al. (2017); Greenhill (2015); Coelho et al. (2019); Hua et al. (2019)). In this study, we will take a closer look specifically at the region of Remote Oceania.

Fig. 1 shows the language richness of the islands of Oceania. Languages are points, coloured by language family. In the Polynesian islands (Sāmoa, Tonga, Rarotonga, Tokelau etc), there is generally one language per island group. However, in central west Remote Oceania (Vanuatu, Temotu, Fiji and Kanaky) there are many more languages — sometimes up to 20 languages on the very same island. The archipelagos of Chuuk, Wa'ab and Pohnpei also feature more languages than the Polynesian islands and atolls. What explains this difference?

Let us first consider some plausible factors in language diversification. Firstly, diversification takes time. We would expect a higher probability of language diversification in areas where people have lived for a longer time simply due to natural drift as communities spread out and form separate clusters. All of Remote Oceania was settled in approximately the last 3,500 years but at different stages. A large part of Remote Oceania was settled for the first time by Austronesian-speaking people in a rapid expansion 3,600 - 2,800 BP known as the "Lapita expansion" (Bellwood (2006, 106-7); Rieth & Cochrane (2018, 137)). This first area covers all islands of Vanuatu, Temotu, New Caledonia, Fiji and parts of Western Polynesia. The dates of the first settlement for these islands are relatively similar, but there still is a large discrepancy in the number of languages (see Fig 1). This makes for a convenient case for comparison. If we include other regions of Oceania, such as Australia or the New Guinea region (Ross, 2017), the comparison becomes more complex as they have been settled for a much much longer time-span — possibly up to 60,000 years. Here we focus specifically on Remote Oceania in order to test specific theories about the diversification there that have been presented by anthropologists and linguists.

Let us compare two island groups in this region as an example: the Sāmoan archipelago in Western Polynesia and the island of Malakula in Vanuatu. There is one indigenous language spoken in Sāmoa — Sāmoan — and in

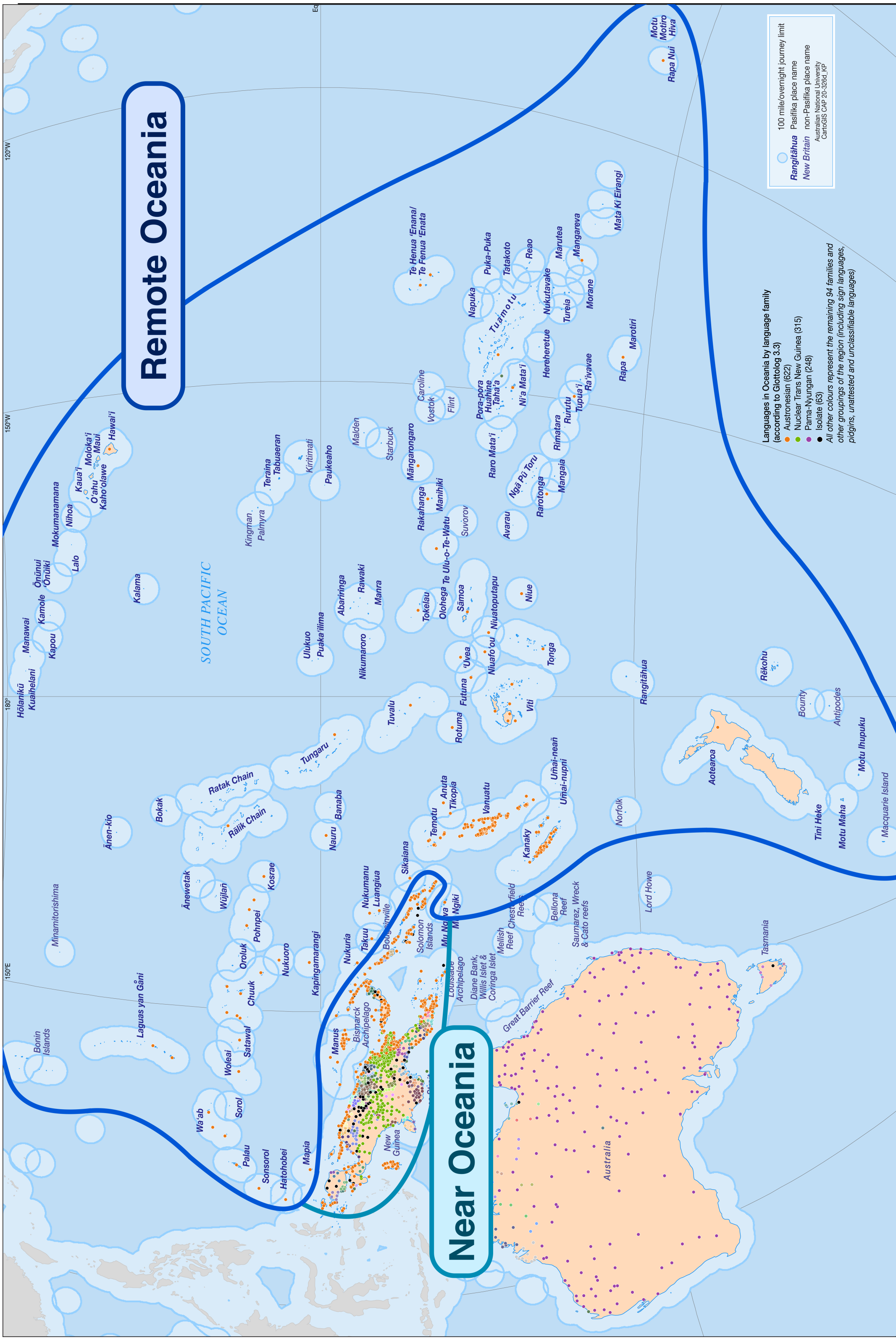

Fig. 1. Map of Oceania, with the regions of Near and Remote Oceania marked out (first labelled by Pawley & Green (1973)). Points represent languages, based on data Hammarström et al. (2017) with a few small adjustments by author.

Malakula there are 33 distinct languages spoken (Hammarström et al., 2017). Sāmoa was settled approximately 400 years after Malakula (Rieth & Cochrane, 2018, 137-8). This gives Malakula a bit of a "head-start", but most likely not enough to explain the vast difference.

Besides time depth, we might also expect that more space entails greater language diversity. A larger area makes it possible to spread out communities, which can entail language splitting. However, the shoreline[1] of Malakula is 67% of that of Sāmoa. Sāmoa is larger than Malakula. What other factors are there that can aid our understanding here?

Turner (1884) and Pawley (1981, 2007b) have suggested that the discrepancy between the number of languages in different parts of Remote Oceania can be related to societal structure. Societies in Vanuatu and New Caledonia tend to be less hierarchical than societies in Polynesia. Societies with fewer vertical levels of political structure are associated with more linguistic and cultural diversity. Conversely, more hierarchically complex societies are associated with linguistic and cultural homogeneity.

This study tests the hypothesis that more vertical levels of political organisation are meaningfully connected to the number of languages, also when we incorporate other possible factors like settlement time, size and ecological environment (c.f. Nettle (1998), Gavin & Sibanda (2012) and Hua et al. (2019)). If the results show that political complexity has a significant correlation with the number of languages per island group when other relevant factors are controlled for, this lends support to Turner and Pawley's hypothesis.

Previous research

Linguists, anthropologists and scholars of other fields have long remarked on the link between linguistic diversity and political structure in Remote Oceania. One of the oldest examples of this observation is the European Turner who travelled in the region 1861-1884 and hypothesized that the lack of regional variation *within* Sāmoan entailed that they had long had a centralized government (Turner, 1884, 172). Linguists have linked types of societal organization to different diversification processes. Language change is embedded in social structure (c.f. (Weinreich et al., 1968)), or as Grace (1992, 124) writes: "linguistic similarities and differences [are] reflections of community structures". As community ties weaken and what was once one social network splits in twain, we expect that this is reflected in language variation as well. In this way, language studies provide insights to neighbouring fields such as archaeology, anthropology etc.

Pawley (1981, 2007b) discusses the specific hypothesis that while much of Remote Oceania was settled in similar ways by Austronesian-speaking people, and at comparable time depths, and developed similarly

[1] Since Austronesian settlements are mainly found on the coast, we are comparing shorelines instead of land area.

immediately after the first settlement, the differences between them stem from the rise of powerful chiefs and maintenance of long-distance voyaging in the island groups of Polynesia. This theory suggests that the steps of the diversification process were more or less the same, but that the *rate* by which it progressed differed and that this is the cause for the variation we see today.[2] The societal structure in Polynesia was such that the process of cultural diversification was slower than in Vanuatu and New Caledonia.

The following quote from Pawley (2007b) summarises a few of these points and links them to the political organisation:

> *When they found a new island group the Lapita colonists rapidly explored and lightly settled it. This initial colonising phase was typically followed by a sequence of demographic, economic and technological developments. Population growth led to a denser distribution of settlements, more intensive agriculture, more local marriage and trade, weakening of kin ties with distant communities, and linguistic divergence. This sequence of events happened in both Southern Melanesia and West Polynesia and Fiji but it happened more slowly in West Polynesia/Fiji. In [Fiji and West Polynesia] the maintenance of long-distance voyaging, both within island groups and between neighbouring island groups, can be attributed in large part to the rise of powerful chiefs. These chiefs had political, economic and social motives for maintaining long distance connections and were able to use their authority to drive the production of a food surplus which in turn could be used to support specialist craftsmen who could, among other things, build and sail large ocean-going canoes.* (Pawley, 2007b, 28)

This proposal by Pawley (1981, 2007b) can be expressed in a Directed Acyclic Graph (DAG, c.f. Pearl (1995) and McElreath (2020). This lets us view the chain of argumentation in a stylized and formal visualisation that can lead directly to statistical modelling and comparison between proposals (c.f. Roberts et al. (2020)). Fig. 2 contains such a graph, based on the author's interpretation of Pawley's proposal in more detail. The connection between the rise of powerful chiefs and language richness is not direct, but mediated through a number of other factors.

Fig. 19 in H contains a larger version that incorporates more of the suggested linkages beyond those suggested by Pawley, for example, that political complexity may affect the modularity and interconnectedness of social networks and this, in turn, affects language splitting.

[2] Lynch (1981) calls this "the diversification cycle theory".

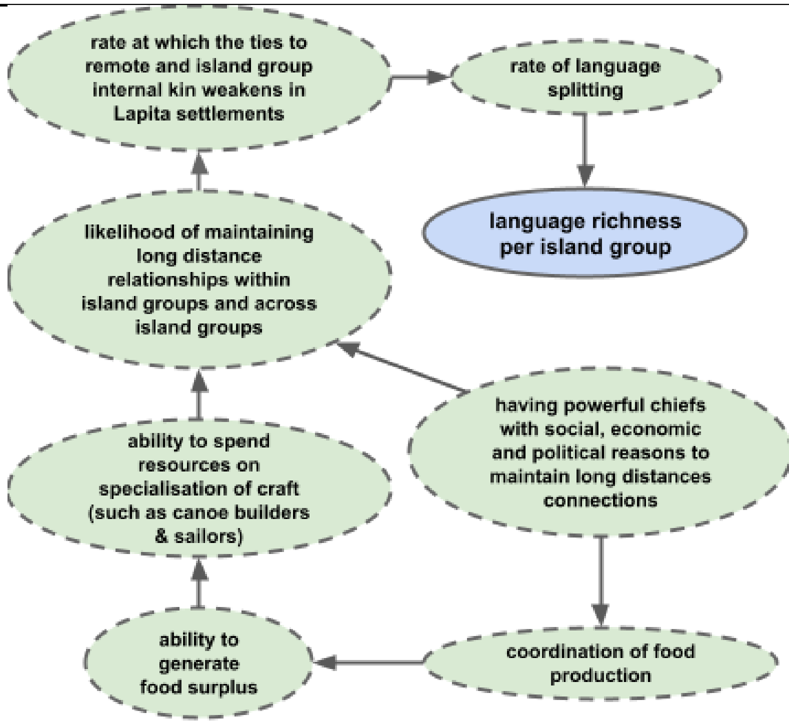

Fig. 2: Directed Acyclic Graph of the argument presented in Pawley (2007b). Blue = variable to be predicted (response).

It is possible that individual leaders in more stratified societies do not have a direct effect on the nature of the language of the community by means of their personality, conscious policies or directives. It is possible the political structure itself is indicative of a certain community network structure and attitude.

In a similar vein to Pawley, Currie & Mace (2009) suggests that more stratified societies are correlated with larger language areas in Africa and Eurasia. If one language covers a large area, then the number of languages per square kilometre is lower than otherwise — the language density is lower. Their research can be interpreted to mean that more stratified societal structures are more capable of sustaining linguistic homogeneity over a larger geographical area (or possibly reducing diversity by cultural dominance and warfare). Greater maintenance of cultural homogeneity would result in less language splitting over larger areas, and therefore a larger language area correlates with greater stratification. The approach is different from Pawley (1981, 2007b), but the scientific line of argumentation is similar.

Returning to language richness in Remote Oceania, a significant factor to consider besides those already discussed is contact — in particular with people from the New Guinea region (c.f. Ross (2017)). Prior to European colonisation, all indigenous communities of Remote Oceania spoke the languages of the Austronesian family (see

Fig 1). This family has its ultimate origins in Taiwan, and spread into the Remote Oceania approximately 3-4,000 years ago (c.f. Fig. 3). The peopling of New Guinea, the Bismarck archipelago and much of the Solomon Islands however is older than that, and there is a great amount of language and cultural diversity there that may have influenced Austronesian people during and/or after their journeys east.

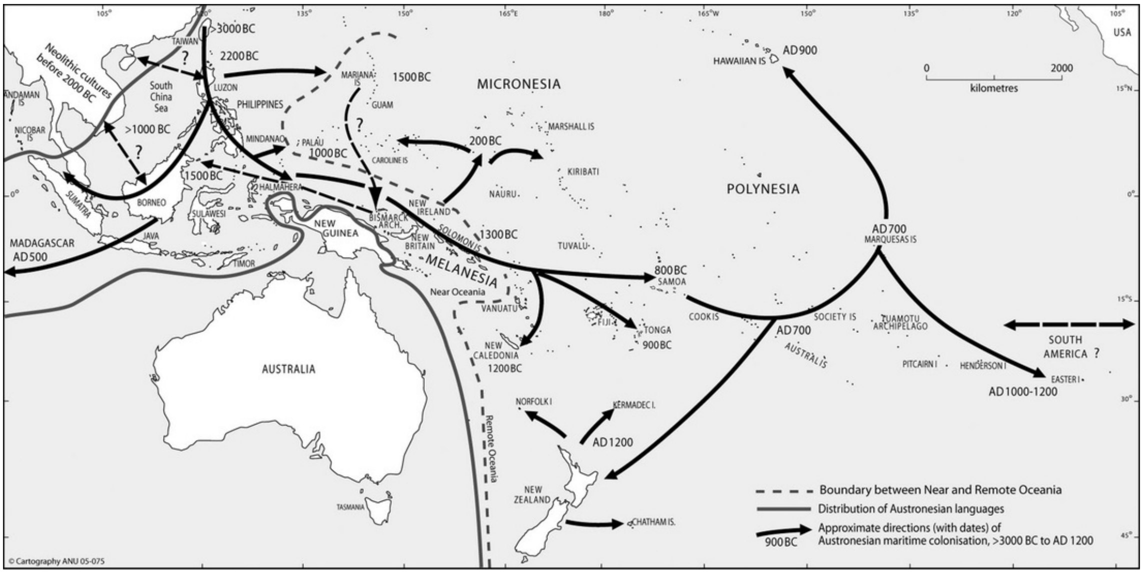

Fig. 3: Map of the expansion of the Austronesian language family, based on work by Peter Bellwood. Map made by ANU CartoGIS and Peter Bellwood.

Recent studies of ancient and modern DNA (Lipson et al., 2018; Posth et al., 2018) have shown that there is evidence of significant levels of genes that are associated with origins in non-Austronesian speaking communities from New Guinea and Solomons among Vanuatu people. The suggestion is that while the initial settlement of Remote Oceania was likely to be wholly Austronesian (i.e. genetically similar to East Asian populations), there have since been more people coming who have deeper ancestry from the New Guinea region. Interestingly, Posth et al. (2018) highlight that this had little effect on the languages of the region, communities in Vanuatu still speak languages that are clearly Austronesian in origin. This could be because the New Guinea people already spoke Austronesian languages at the point of migration or due to other reasons (c.f. Barlow (2023, 288, 321); Bedford et al. (2018, 209-210)).

Likewise, further studies such as Liu et al. (2022) show similar genetic influence from New Guinea and Solomons northwards in Palau, Chuuk and Pohnpei island groups. It is becoming increasingly clear in the overall scientific literature on the ancient history of Oceania that there were a lot more connections and interactions than previously postulated — including to island groups of Near Oceania. It is not the case that people travelled to one place, and then stayed put there with little influence from other people of Oceania. Oceania is an interconnected place today, and it was historically too.

It is possible that these interactions, which were more frequent in Vanuatu than in Polynesia, have implications for many domains of human life — including political organisation and language diversity. Linguist and anthropologist Lynch (1981, 104) has stressed the relevance of contact with non-Austronesian communities from New Guinea, Bismarck, Bougainville and the Solomon Islands as a key factor in explaining why there are more languages in Vanuatu compared to the rest of Oceania [3]. Lynch does not offer a precise account but outlines several different potential scenarios whereby contact with non-Austronesian groups caused greater diversification among Oceanic languages.

Unfortunately, it is difficult to incorporate data on contact with non-Austronesian speakers explicitly in the models of this study. This is primarily due to lack of information on this variable in island groups such as Kanaky (New Caledonia), Ratak & Ralik (Marshall Islands), Marquesas, Aotearoa and other island groups. However, this factor is very important and we will return to it in the discussions.

Finally, there has been a growing interest in environmental effects on language diversification[4]. One of the oldest and most influential bodies of work in this vein is by Nettle. Nettle (1998) showed that there is a correlation between number of languages per country and ecological risk (as measured by mean growing season). Nettle suggests that high-risk environments encourage wider social networks and therefore reduce the number of languages in that area. Hua et al. (2019) expanded on this idea and explains that “Smaller social groups are presumed to be more likely to be stable and self-sufficient in areas with a more abundant and reliable year-round food supply.” Hua et al. (2019) constructed a more complex and fine-grained model of how environmental factors influence language diversity. They found that factors associated with risk (precipitation and temperature seasonality and season length) did indeed predict much of the distribution of languages in the world. Similarly, Gavin et al. (2017) showed that a simple model that takes into account rainfall and an upper bound of population size can to a large extent predict the distribution of indigenous languages in Australia. Given the success of these models, the amount and seasonality of rainfall and temperature variables are also included in our model predicting language diversity in Remote Oceania.

[3] Lynch (1981) uses the terms “Papuan” and contrasts Melanesia and Polynesia. I have chosen to summarise him here in more precise terms in relation to how his work impacts the current study. He writes specifically about the influence of languages of New Britain, Bougainville, central Solomon Islands, Manus and the southeast of New Guinea mainland (Oro province, Milne Bay province and National Capital District) on Oceanic languages of those regions and in Temotu.

[4] For a longer summary of recent studies of language diversification, see Gavin et al. (2013) and Greenhill (2015).

In their study of global environmental predictors of language diversification, Hua et al. (2019) noted that there were certain areas that remained unexplained (New Guinea, Himalayas, West Africa and Mesoamerica). This may be an indication that the model needs to be adjusted to account for more variables or that the variables do not have the same effect globally. A recent study of language diversification in North America (Coelho et al., 2019) showed that the best predictors of language change may vary from place to place and that the factors are interlinked causally in a complex manner. Their study included environmental variables and data on the population. The authors note that political complexity might be an important variable too, but they were, unfortunately, unable to include it (Coelho et al., 2019, 7).

This study focuses on a particular hypothesis concerning a particular region of the world. It is possible that these models perform less well globally and the results should primarily be compared against the specific theories of Lynch (1981) and Pawley (1981, 2007b) about Remote Oceania. As one of the few studies to include data on both environmental and socio-cultural variables, it is able to contribute a valuable perspective to the field of language diversity mechanics.

Materials and Method

The approach of this paper is to gather relevant data on island groups and their societies and then fit a statistical model which aims to predict the number of languages given all the variables. From the composition of variables in the resulting model we can infer which have the greatest predicting power when they are co-estimated.

The unit of analysis is "island group". We cannot use modern political boundaries ("French Polynesia", "Cook Islands" etc) because they are often historically and culturally inappropriate. As we want to capture something about the communities at large we are not satisfied with using individual islands. For these reasons, we use two different ways of grouping islands: overnight sailing distances and shared language (see appendix B.3 for details).

A simple test of Spearman's Rank Order correlation reveals that the number of languages and political complexity per island group are indeed moderately correlated (rho = -0.4, p value < 0.001 for islands grouped for shared language). However, this is not sufficient. We want to investigate whether this correlation still holds once we explicitly incorporate other variables that we expect might explain language diversity as well, such as time-depth and environmental factors.

Variables were included in the model if they meet all three of the following criteria: a) have been suggested as relevant in previous literature, b) there is enough data on the variable in a published source and c) if they relate to the response variable (language richness) in a non-mediated manner (c.f. Pearl (1995)). This results in the following variables:

- **response variable**:
  - language classification, based on Hammarström et al. (2019) and other sources (see Appendix B.2).
- **predicting variables:**
  - number of vertical levels of jurisdictional hierarchy per society [5], based on the Ethnographic Atlas (Gray, 1998; Kirby et al., 2018), Sheehan et al. (2018) and own modifications. Also known as "political complexity" and "EA033" (see appendix B.4 for details on definition and coding).
  - archaeological dates, grouped into waves (Intoh & Ono, 2007; Intoh, 2008; Rieth & Cochrane, 2018; Levin et al., 2019; Feinberg & Scaglion, 2012; Napolitano et al., 2019). (See appendix B.5 for details.)
  - historical environmental data from Lima-Ribeiro et al. (2015): mean temperature, temperature seasonality, mean rainfall and rainfall seasonality. (See appendix B.6 for details.)
  - environmental data from NASA satellites (Running & Zhao, 2021a,b): Net Primary Productivity. See appendix B.6 for details.)
  - island geography (latitude and shoreline) from the Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG) (Wessel & Smith, 1996). (See appendix B.6 for details.)

The variables Net Primary Production (NPP; the amount of biomass or carbon produced), temperature & rainfall (mean and seasonality) and absolute latitude are all associated with the suitability of a place for life and they correlate with each other, at least to some degree (c.f. Figs 25 and 24). In this study, they are all tracking a similar concept: the suitability for human populations to thrive in smaller groups. Because of this, we are running a Principal Components Analysis on all seven variables, this allows us to extract a variable that explains most of the information. We ran a Non-Graphical Cattel's Scree Test (Cattell, 1966; Raiche & Magis, 2022) to determine the number of optimal components (overnight sailing distance island groups = 2 PCs, shared language island groups = 3 PCs). For more details, see Appendix B.6.

In order to create a formula for the statistical model, we consider the suggested causal relationships based on previous research (see section 2). Besides the effect of the environmental variables on language richness, we also hypothesise that there can be a joint effect with the size of the island. Similarly, the time-depth and size of islands may also interact.

[5] Nota Bene that there are other kinds of political complexity besides vertical, for example, horizontal complex exchange networks. See Appendix B.4 for more.

The relationship between variables can be represented in a formalised manner in a Directed Acyclic Graph (DAG; Pearl (1995) and McElreath (2020)), see Fig. 4. For the full DAG, including excluded variables and variables with missing data, see appendix H. Interactions are denoted with a deterministic node labelled with $*$.

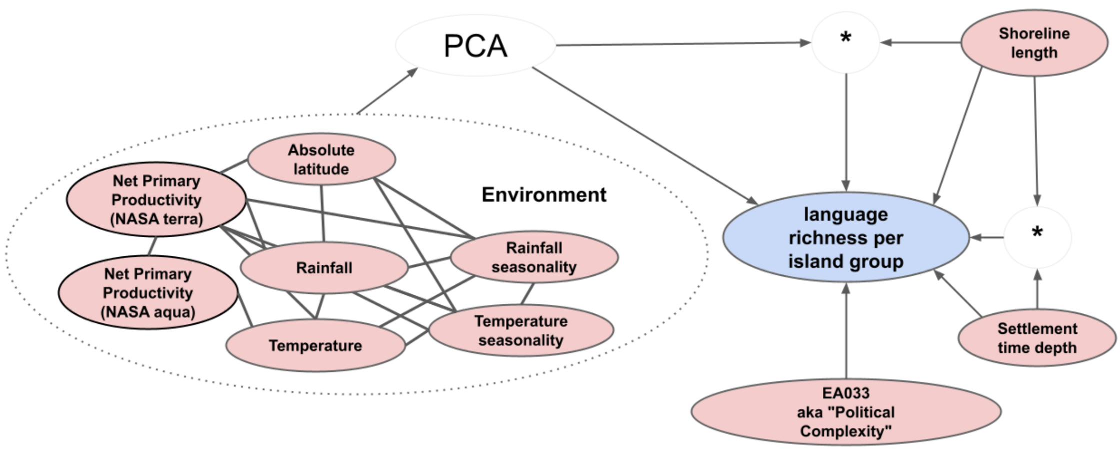

Fig. 4: Directed Acyclic Graph of the variables in this study (DAG, c.f. McElreath (2020)). Blue = variable to be predicted (response), red = predictors. Asterisk nodes represent variable interactions, a type of deterministic DAG-node. PCA represents Principal Components Analysis. PCA was run on the incoming variables, and the optimal number of components (given a Non-Graphical Cattel's Scree Test) were carried forward. Terra and Aqua are the names of the NASA satellites.

This results in the following model formula[6]:

```
Number of languages ~ Political complexity (mode) +
      Environmental PC1 * Shoreline (log10 +
      Environmental PC2 * Shoreline (log10 +
      (Environmental PC3 * Shoreline (log10) +
      Settlement order * Shoreline (log10)
```

Shoreline was log-10-transformed. All predicting variables were then also normalized in the conventional manner by subtracting the mean and dividing by the standard deviation.

To test the impact of several variables predicting the same response variable we can use a Bayesian Regression Model (Gelman & Hill, 2006; Bürkner, 2017). The distribution of languages per island group is over-dispersed.

[6] Settlement order was reversed from Fig. 15 such that high numbers represent further back in time.

There are many island groups with one language and only a few with higher numbers (see Figs. 10 and 11 in Appendix B.3). For the overnight-sailing island groups, the group that contains Vanuatu and Temotu has 130 languages, whereas most island groups in Polynesia have 1 or 2. This skewed distribution necessitates a Poisson distribution for the response variable in the model, which is able to adjust the relationship between the mean and the variance appropriately[7]. For further details on the model set-up, see appendix I.

Due to missing data, we are testing our models over 65 out of 104 shared language groups and 56 out of 67 overnight distance groups. This sample still contains some of our more extreme sites, like Santo, Kanaky (New Caledonia mainland) and Malakula. Variables that did not cover Vanuatu and New Caledonia with enough specific data-points are not included in this study, for example historic population data, non-Austronesian gene flow or jurisdictional hierarchy within local communities.

To test the robustness of the conclusions over all observations, we drop one observation out at a time and calculate the predictor effects and predictions of the number of languages. This allows us to asses which island groups represent the largest challenge to fitting the model.

The results were calculated in R 3.6.0 (R Core Team, 2019) using the function `brms::brm()` (Bürkner, 2017), tidyverse suite (Wickham et al., 2019) and the modEvA-package (Barbosa et al., 2016). For a full list of packages used, see appendix L.

Results

We will first go through the outcomes for the shared language island groups (i.e. groups of islands that share at least one language), and secondly the overnight distance groups. In common for both is that political complexity is a relevant predicting factor, with a negative effect (the more political vertical levels, the lower the number of languages). This lends support to the proposal by Pawley (1981) and Pawley (2007b). However, there are details of the result that suggest that the effect may be due to non-Austronesian contact rather than political complexity *per se*.

Fig. 5 shows the actual observed language counts (purple triangles) and predictions of this model (green boxplots). The model does predict that Santo, Kanaky and Malakula have the highest language counts in the dataset. Our observed language counts for Santo, Kanky and Malakula are 24, 29 and 33 respectively. The model is very close for Santo and Kanaky for its average prediction value but underestimates Malakula by 17 languages. The absolute average difference between the mean predicted values and the observed is 1.3.

[7] This issue was also faced by Gavin & Sibanda (2012) in their study of environmental factors in language diversification in the Pacific. They chose another approach, a reciprocal transformation, based on model fitness compared to two other kinds of transformations of the response variable.

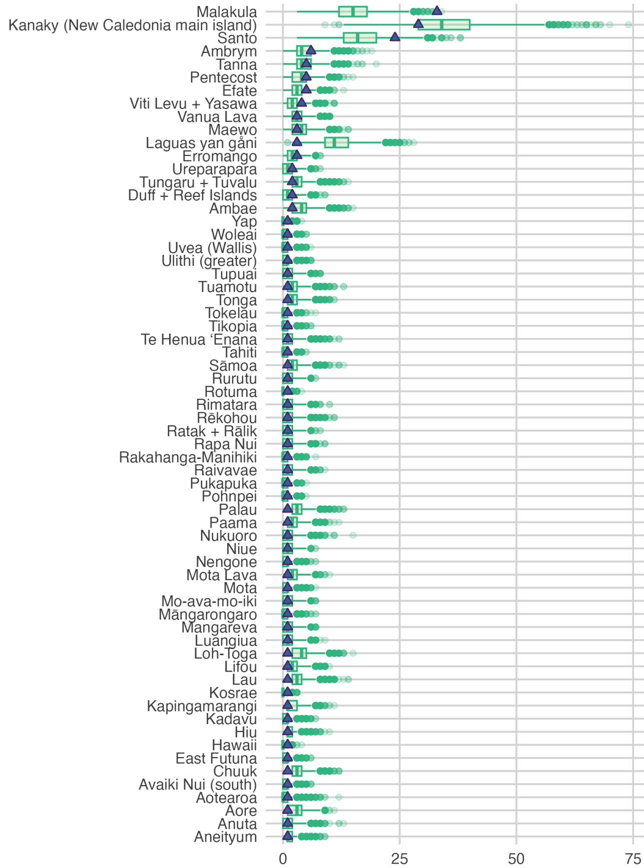

Fig. 5: Prediction from the model with island grouped for shared language. Number of languages is on the x-axis, observations on the Y. The dark blue triangles represent the observed number of languages, the green boxplot the predictions of the model. The average absolute difference between the mean of the predictions and the observed value is 1.3.

For the shared language island groups, the variables that have a 95% confidence interval which are not different from zero are: political complexity (EA033), environmental PC1, Shoreline, time depth and shoreline in interaction with time depth. Fig 20 in Appendix K shows the full distributions.

Most of the island groups can be removed with little effect on the overall accuracy of predictions (see Fig. 23 in K), however, when Malakula (central Vanuatu), is left out, the difference average difference between predicted and observed becomes lower (1.3 $\rightarrow$ 0.9). Malakula is the island group with the highest number of languages (see Fig 11 in appendix B.3). Political complexity remains an important predictor regardless of which island group is dropped.

For the overnight sailing distances, several island groups are much larger than the previous grouping. The islands of Vanuatu (including Anuta and Tikopia) and Temotu are all within the same group (see Fig 10 in appendix B.3; this is similar to the groups considered in Pawley (2007b)). This also leads to a larger discrepancy between the number of languages, with Vanuatu and Temotu having 130 and Sāmoa still only one. The average absolute difference between the predicted and observed number of languages is somewhat greater than for the previous grouping — 1.6. Fig 6 shows the prediction vs observed per island group.

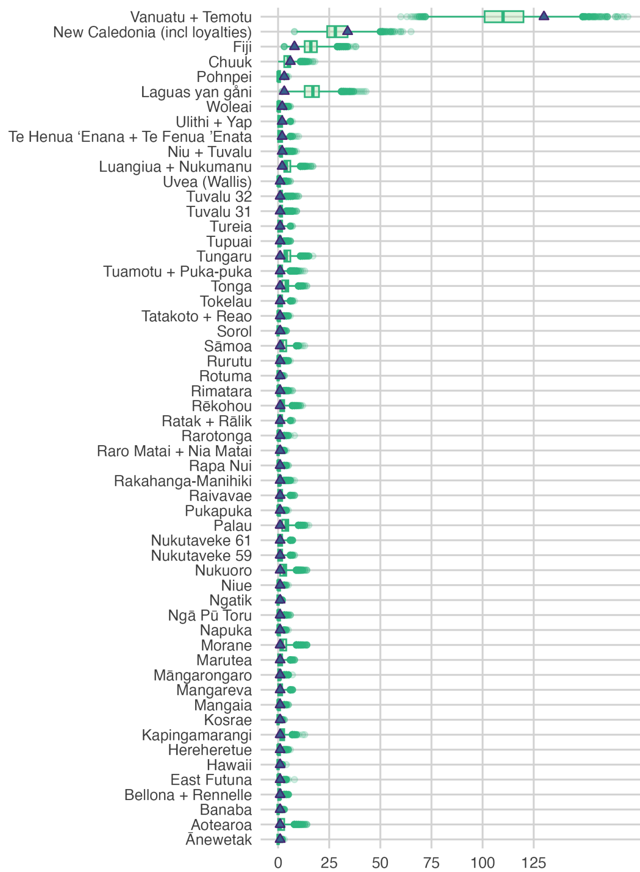

Fig. 6: Prediction from the model with island grouped for overnight sailing distances. The dark blue triangles represent the observed number of languages, the green boxplot the predictions of the model. The average absolute difference between the mean of the predictions and the observed value is 1.6.

The most important variables in this model (i.e. 95% confidence interval of coefficient distribution does not overlap 0) were: political complexity, environmental PC2, shoreline, time depth and time-depth in interaction with shoreline (see Fig 20). However, if we drop out Vanuatu + Temotu this no longer holds. When Vanuatu

+ Temotu are left out, the coefficient distribution of political complexity overlaps zero. Instead, Shoreline in interaction with time depth and environmental PC1 in interaction with Shoreline are relevant predictors. When Vanuatu + Temotu are dropped out, the overall difference between predicted and observed values goes down the most (from 1.6 to 0.7, see Fig. 22). When Vanuatu + Temotu are excluded, the model is more accurate. If we ask this model that has been trained *without* the Vanuatu + Temotu data to predict how many languages there are there, it grossly underestimates and proposes 7 languages (observed = 130). If we drop out any other island group (including Kanaky), political complexity remains an important predictor.

For tables of the predictor Estimates, Rhat and ESS of both models, please see appendix J.

Discussion

The results indicate that political complexity could be relevant for understanding the dynamics of language diversification in Remote Oceania, but further studies are needed — especially regarding Vanuatu.

When the island group Vanuatu + Temotu was dropped, the model for overnight sailing island groups no longer drew on political complexity as a relevant predictor of language richness. This may indicate that it primarily made use of the variable to tell apart Vanuatu + Temotu from all other island groups. Vanuatu + Temotu is by far the island group with the most languages, 130, and it is dominated by societies that are classed as level 1 on the scale of vertical political complexity used by the Ethnographic Atlas (Gray, 1998) and other studies (Watts et al., 2018), see Fig. 14 in appendix E).

Furthermore, if we tease apart the argument concerning political complexity (for example in the full DAG in Fig. 19 in appendix H or Fig. 2 in section 2), we find that conceptually the link between the number of languages and political complexity is not direct. It could be mediated through other factors such as network modularity, the propensity to maintain long-distance contact etc.

Watts et al. (2018) found that politically complex societies in Oceania became Christian at a faster speed. This may indicate that changes propagate through a population quicker if the society is not egalitarian. If Christianity spreads faster in societies with higher complexity, it is possible to reason that other changes also might spread faster. If changes propagate quickly throughout a population, that leaves less possibility for diversification through natural drift or groups in the periphery staying more archaic compared to the centre. This interpretation relies on the understanding that the variable for political complexity (EA033) reveals something not only about jurisdictional power relationships in a given society, but also captures something about the social network of *all* of its members (not just the leaders). Under this theory, people living in societies with higher political complexity would be able to travel further and have wider-ranging social networks, which in turn would lead to greater internal cultural homogeneity.

Maintaining long-distance contact can indeed be a function of shared cultural identity. Skirgård (2020, 218-291) notes that there are not only fewer languages in the island groups of Polynesia, but they are also more similar to each other lexically than languages in Vanuatu are. This relationship does not hold for grammatical features however, languages in Polynesia are similarly different from each other grammatically as languages of Vanuatu and other island groups in Remote Oceania are (as measured by Grambank features, (Skirgård et al., 2023)). It has been suggested, by Silverstein (1981), François (2011) and others (c.f. Mansfield et al. (2023)), that communities pay special attention to words and sound differences as emblematic of shared ties — but grammar tends to fly under the radar in this respect. This difference between Vanuatu and Polynesia in terms of lexical dissimilarity and language richness, but the similarity in terms of grammatical diversification may indicate that the cause does indeed lie in attitudes and social network organisation rather than merely time and environmental factors.

Currie & Mace (2009) argue that the reason behind their results (higher political complexity = larger language areas) is that societies with higher political complexity are more likely to be able to replace existing groups in an area, or otherwise incorporate (and possibly assimilate) them. It may also be the case that diversity in Polynesia was reduced by later expansionist efforts by societies with higher political complexity. For example, the island of Niuafo'ou which lies between Tonga, 'Uvea and Sāmoa is home to a language that is most closely related to 'Uvean. This language has however become *more* similar to Tongan during the period of Tongan domination in the area 1000-1300's (see Aswani & Graves (1998) and (Tsukamoto, 1988, 2-9)). It may be that politically complex societies do not only *maintain* homogeneity by continuing to be in contact after first settlement, but can also reduce diversity which emerges after settlement through incorporation and cultural dominance later.

Whatever the political complexity variable is tracking, it is clear that it is one of the variables that most clearly picks out Vanuatu and that Vanuatu is the place with the most languages. Where did this difference in societal organisation come from? As previous studies have indicated (e.g. Lynch (1981)), the answer may lie in contact with people of non-East Asian origin further west. It is possible that contact with communities from New Guinea and Solomons (in Bismarck before migration to Vanuatu or in Vanuatu after initial settlement) brought with it societal changes to Vanuatu that caused it to be a different place from islands further east (c.f. Pawley's idea of the same process at different speeds). Besides the possible cultural influence on attitudes and societal organisation, they may also have brought with them their own cultural heritage which would add to the diversity in itself.

Similarly to Gavin & Sibanda (2012), Hua et al. (2019) and others, the results of this study indicate that environmental factors, as encapsulated by the principal components, play some in explaining language diversity in Remote Oceania. However, size, time depth and political complexity are clearly necessary to take into account for this region.

Conclusions

This study investigated the hypothesis that besides time depth and environmental factors (c.f. Currie & Mace (2009); Gavin & Sibanda (2012); Hua et al. (2019) and Coelho et al. (2019)), language richness in Remote Oceania is also significantly influenced by interaction patterns that can be measured by political complexity (c.f. Pawley (1981, 2007b)). The results lend support to this thesis, with some caveats.

Political complexity is, perhaps unsurprisingly, a complex variable. There are a number of phenomena in our social world that it could track, including network dynamics and in the case of Remote Oceania specifically — deep contact with communities in New Guinea and Solomon Islands. It is clear that in terms of constructing predictive models of the number of languages per island group, it contributes significant power to the case of Remote Oceania. However, in order to understand the situation we need to deconstruct this concept further and conduct studies that tease apart the specific circumstances — especially in Vanuatu.

References

Ager, Simon. 2020. Tuvaluan (Te ’gana Tūvalu). `https://www.omniglot.com/writing/tuvaluan.htm`.

Aitken, Robert T. 1971[1930]. Ethnology of tubuai, bernice p. *Bernice Pauahi Bishop Museum Bulletin* 70. 14.

Alkire, William H. 1991. Woleai. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 382–384. Boston, Massachusetts: G.K. Hall and co.

Ammar, Ron. 2019. *randomcolor: Generate attractive random colors*. `https://CRAN.R-project.org/package=randomcoloR`. R package version 1.1.0.1.

Arnold, Jeffrey B. 2021. *ggthemes: Extra themes, scales and geoms for ggplot2*. `https://github.com/jrnold/ggthemes`. R package version 4.2.4.

Aswani, Shankar & Michael W Graves. 1998. The Tongan Maritime Expansion: A Case in the Evolutionary Ecology of Social Complexity. *Asian Perspectives* 135–164.

Athens, J Stephen. 2007. Prehistoric population growth on kosrae, eastern caroline islands. In Rallu J Kirch PV (ed.), *The growth and collapse of pacific island societies: Archaeological and demographic perspectives.*, 257–277. University of Hawai’i Press.

Barbosa, AM, JA Brown, A Jimenez-Valverde & R Real. 2016. Package ‘modEvA’. `https://cran.r-project.org/web/packages/modEvA/`.

Barlow, Russell. 2023. Papuan-austronesian contact and the spread of numeral systems in melanesia. *Diachronica* 40(3). 287–340.

Bayliss-Smith, Tim. 1974. Constraints on population growth: The case of the polynesian outlier atolls in the precontact period. *Human Ecology* 2. 259–295.

Beaglehole, E. & P. Beaglehole. 1938. Ethnology of pukapuka. *Bernice Pauahi Bishop Museum Bulletin* 110. 1–419.

Bedford, Stuart, Robert Blust, David V Burley, Murray Cox, Patrick V Kirch, Elizabeth Matisoo-Smith, Ashild Naess, Andrew Pawley, Christophe Sand & Peter Sheppard. 2018. Ancient dna and its contribution to understanding the human history of the pacific islands. *Archaeology in Oceania* 53(3). 205–219.

Bedford, Stuart & Matthew Spriggs. 2008. Northern vanuatu as a pacific crossroads: The archaeology of discovery, interaction, and the emergence of the” ethnographic present”. *Asian Perspectives: Special Issue: Maritime Migration and Colonization in Indo-Pacific prehistory* 47(1). 95–120. `https://www.jstor.org/stable/42928734`.

Bellwood, Peter. 2006. Austronesian Prehistory in Southeast Asia: Homeland, Expansion and Transformation. In Peter Bellwood, James J Fox & Darrell Tryon (eds.), *The Austronesians: Historical and Comparative Perspectives*, ANU E Press.

Bellwood, Peter S. 1971. Varieties of ecological adaptation in the southern cook islands. *Archaeology & Physical Anthropology in Oceania* 6(2). 145–169.

Bergen, Benjamin K. 2001. Nativization Processes in L1 Esperanto. *Journal of Child Language* 28(3). 575–595.

Birket-Smith, K. 1956. An ethnological sketch of rennell island. *Kong. Danske Videnskab. Selskab, Hist. -Filol. Medd.* 35. iii, 1–222.

Bivand, Roger S., Edzer Pebesma & Virgilio Gomez-Rubio. 2013. *Applied spatial data analysis with R, second edition*. Springer, NY. `https://asdar-book.org/`.

Blommaert, Jan. 2008. Artefactual Ideologies and the Textual Production of African Languages. *Language & Communication* 28(4). 291–307.

Bollt, Robert. 2008. Excavations in peva valley, rurutu, austral islands (east polynesia). *Asian perspectives* 47. 156–187.

Bolton, Lissant. 1998. Chief Willie Bongmatur Maldo and the Role of Chiefs in Vanuatu. *The Journal of Pacific History* 33(2). 179–195.

Bonnemaison, Joël. 1972. Système de Grades et Différences Régionales En Aoba (Nouvelles Hébrides). *Cahiers ORSTOM. Série Sciences Humaines* 9(1). 87–108.

Bonnemaison, Joël. 1996a. The art of power. In *Arts of vanuatu*, University of Hawaii Press.

Bonnemaison, Joël. 1996b. Graded Societies and Societies Based on Title: Forms and Rites of Traditional Power in Vanuatu. In *Arts of Vanuatu*, 200–216. University of Hawaii Press.

Bright, William. 1997. Notes. *Language in Society* 26(3). 469–470.

Buck, P. H. 1930. Samoan material culture. *Bernice Pauahi Bishop Museum Bulletin* 75. 1–724.

Buck, P. H. 1932a. Ethnology of manihiki and rakahanga. *Bernice Pauahi Bishop Museum Bulletin* 99. 1–238.

Buck, P. H. 1932b. Ethnology of tongareva. *Bernice Pauahi Bishop Museum Bulletin* 92. 1–225.

Buck, P. H. 1934. Mangaian society. *Bernice Pauahi Bishop Museum Bulletin* 122. 1–207.

Buck, P. H. 1950. *The material culture of kapingamarangi*, vol. 200. Bernice Pauahi Bishop Museum Press.

Buck, P. H. 1952. *The coming of the maori*. Wellington: Maori Purposes Fund Board; [distributed by] Whitcombe and Tombs.

Buck, P. H. 1971[1938]. Ethnology of mangareva. *Bernice Pauahi Bishop Museum Bulletin* 117. 1–519.

Bürkner, Paul-Christian. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software* 80. 1–28.

Burrows, E. G. 1971[1937]. Ethnology of uvea. *Bernice Pauahi Bishop Museum Bulletin* 145. 1–176.

Burrows, E. G. & M. E. Spiro. 1953. *An atoll culture: Ethnography of ifaluk in the central carolines*. New Haven: Human Relations Area Files.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). 1–28. doi:10.18637/jss.v080.i01.

Bürkner, Paul-Christian. 2018. Advanced Bayesian multilevel modeling with the R package brms. *The R Journal* 10(1). 395–411. doi:10.32614/RJ-2018-017.

Bürkner, Paul-Christian. 2021. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software* 100(5). 1–54. doi:10.18637/jss.v100.i05.

Bürkner, Paul-Christian. 2023. *brms: Bayesian regression models using stan*. `https://CRAN.R-project.org/package=brms`. R package version 2.19.0.

Capell, Arthur. 1958. *The culture and language of futuna and aniwa, new hebrides*. Sydney, Australia: University of Sydney.

Carroll, Vern. 1966. *Nukuoro kinship*. University of Chicago, Department of Anthropology. dissertation.

Carroll, Vern. 1975. Demographic concepts and techniques for the study of small populations. In Vern Carroll (ed.), *Pacific atoll populations*, University of Hawaii Press.

Carson, Mike T. 2012. Recent Developments in Prehistory: Perspectives on Settlement Chronology, Inter-Community Relations, and Identity Formation. In Richard Feinberg & Richard Scaglion (eds.), *The Polynesian Outliers: State of the Art* (Ethnology Monographs 21), 27–48. United States of America: Department of Anthropology, University of Pittsburgh.

Carucci, Laurence M. 1991. Marshall islands. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 191–194. Boston, Massachusetts: G.K. Hall and co.

Cattell, Raymond B. 1966. The scree test for the number of factors. *Multivariate behavioral research* 1(2). 245–276.

Charpentier, Jean-Michel & Alexandre François. 2012. *Linguistic Atlas of French Polynesia-Atlas Linguistique de La Polynésie Française*. de Gruyter Mouton.

Coelho, Marco Túlio Pacheco, Elisa Barreto Pereira, Hannah J. Haynie, Thiago F. Rangel, Patrick H Kavanagh, Kathryn R. Kirby, Simon J. Greenhill, Claire Bowern, Russell D. Gray, Robert K. Colwell, Nicholas Evans & Michael C. Gavin. 2019. Drivers of Geographical Patterns of North American Language Diversity. *Proceedings of the Royal Society B: Biological Sciences* 286(1899). 20190242. doi:10.1098/rspb.2019.0242.

Conte, Eric & Patrick Vinton Kirch. 2004. *Archaeological investigations in the mangareva islands (gambier archipelago), french polynesia*. Berkeley, CA: University of California.

Cooper, Nicholas. 2022. *Ncmisc: Miscellaneous functions for creating adaptive functions and scripts*. `https://CRAN.R-project.org/package=NCmisc`. R package version 1.2.0.

Cordy, Ross. 1983. Social stratification in the mariana islands. *Oceania* 53(3). 272–276.

Crocombe, R. G. 1967. From ascendancy to dependency: The politics of atiu. *The Journal of Pacific History* 2. 97–111.

Cummins, HG. 1977. Tongan society at the time of european contact. In N Rutherford (ed.), *Friendly islands: A history of tonga*, 63–89. Melbourne, Australia: John Sands Ltd.

Currie, T.E. & R. Mace. 2009. Political Complexity Predicts the Spread of Ethnolinguistic Groups. *PNAS - Proceedings of the National Academy of Sciences* 106(18). 7339–7344.

Currie, Thomas E, Simon J Greenhill, Russell D Gray, Toshikazu Hasegawa & Ruth Mace. 2010. Rise and Fall of Political Complexity in Island South-East Asia and the Pacific. *Nature* 467(7317). 801.

Dahl, David B., David Scott, Charles Roosen, Arni Magnusson & Jonathan Swinton. 2019. *xtable: Export tables to latex or html*. `http://xtable.r-forge.r-project.org/`. R package version 1.8-4.

Damon, Frederick H. 2002. Kula valuables: The problem of value and the production of names. *L'homme* 2(162). 107–136.

Davenport, WH. 1969. Social organization notes on the northern santa cruz islands: the main reef islands. *Baessler-Archiv, Neue Folge* 17(1). 151–243.

Deacon, A. B. & Camilla H. Hildegarde Wedgwood. 1934. *Malekula: a vanishing people in the new hebrides*. London: George Routledge and Sons, Ltd.

Donner, William W. 1991. Ontong java. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 253–255. Boston, Massachusetts: G.K. Hall and co.

Dowle, Matt & Arun Srinivasan. 2023. *data.table: Extension of 'data.frame'*. `https://CRAN.R-project.org/package=data.table`. R package version 1.14.8.

Dubois, Marie-Joseph. 1984. *Gens de maré: Nouvelle-calédonie*. Paris, France: éditions Anthropos.

Eberhard, David M, Gary F. Simons & Charles D. Fennig (eds.). 2019a. *Ethnologue: Languages of the World*. Dallas, Texas: SIL international twenty-second edn. `https://www.ethnologue.com`.

Eberhard, David M, Gary F. Simons & Charles D. Fennig. 2019b. The Problem of Language Identification. In David M Eberhard, Gary F. Simons & Charles D. Fennig (eds.), *Ethnologue: Languages of the World*, Dallas, Texas: SIL international twenty-second edn. `https://www.ethnologue.com/about/problem-language-identification`.

Edwards, Edmundo. 2003. Archaeological survey of ra'ivavae.

Edwards, Stefan McKinnon. 2022. *lemon: Freshing up your ggplot2 plots*. `https://github.com/stefanedwards/lemon`. R package version 0.4.6.

Eilers, A. 1934. *Islands around ponape : Kapingamarangi, nukuoro, ngatik, mokil, pingelap*. Hamburg, Germany: Friederichsen, De Gruyter and Co.

Ellis, S James. 2012. *Saipan Carolinian, One Chuukic Language Blended from Many*. University of Hawaii at Manoa. PhD Thesis.

Emory, Kenneth P. 1975. *Material culture of the tuamotu archipelago* 22. Honolulu: Department of Anthropology, Bernice Pauahi Bishop Museum.

Erdland, August & Richard. Neuse. 1961[1914]. The marshall islanders: life and customs, thought and religion of a south seas people. In *Human relations area files*, New Haven, CT: Sine nomine.

Evans, Nicholas. 2017. Did Language Evolve in Multilingual Settings? *Biology & Philosophy* 32(6). 905–933.

Faaniu, Simati. 1983. *Tuvalu: A History*. University of South Pacific.

Facey, EE. 1981. Hereditary chiefship in nguna. In Michael Allen (ed.), *Vanuatu: Politics, economics and ritual in island melanesia*, Sydney, Australia: M. Allen Sydney Academic Press.

Feinberg, Richard. 1988. Socio-spatial symbolism and the logic of rank on two polynesian outliers. *Ethnology* 27(3). 291–310.

Feinberg, Richard. 1991. Anuta. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 13–16. Boston, Massachusetts: G.K. Hall and co.

Feinberg, Richard & Richard Scaglion (eds.). 2012. *Polynesian Outliers — The State of the Art* (Ethnology Monographs 21). United States of America: Department of Anthropology, University of Pittsburgh.

Ferdon, Edwin N. 1987. *Early tonga*. Tucson, AZ.: University of Arizona Press.

Firth, Raymond. 1939. *Primitive polynesian economy*. London, UK: George Routledge and sons.

Firth, Raymond. 1959. *Social change in tikopia: re-study of a polynesian community after a generation*. London, UK: Allen and Unwin.

Firth, Raymond. 1991. Tikopia. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 324–327. Boston, Massachusetts: G.K. Hall and co.

Force, Roland W. 1960. Leadership and cultural change in palau. *Fieldiana. Anthropology* 50. 1–211.

Foucault, Michel. 2007. *Security, territory, population: lectures at the collège de france, 1977-78.* Springer.

François, A, M. Franjieh, S. Lacrampe & S. Schnell. 2015. The Exceptional Linguistic Density of Vanuatu. In A. François, S. Lacrampe, M. Franjieh & S. Schnell (eds.), *The Languages of Vanuatu: Unity and Diversity* (Studies in the Languages of Island Melanesia 5), 1–21. Canberra: Asia Pacific Linguistics Open Access.

François, Alexandre. 2011. Social Ecology and Language History in the Northern Vanuatu Linkage - A Tale of Divergence and Convergence. *Journal of Historical Linguistics* 1(2). 175–246.

Furrer, Reinhard, Florian Gerber & Roman Flury. 2022. *spam: Sparse matrix.* `https://www.math.uzh.ch/pages/spam/`. R package version 2.9-1.

Furrer, Reinhard & Stephan R. Sain. 2010. spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. *Journal of Statistical Software* 36(10). 1–25. doi:10.18637/jss.v036.i10.

Gabry, Jonah & Rok Češnovar. 2022. *cmdstanr: R interface to cmdstan.* Https://mc-stan.org/cmdstanr/.

Gardiner, J Stanley. 1898. The natives of rotuma. *The Journal of the Anthropological Institute of Great Britain and Ireland* 27. 396–435.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Antônio Pedro, Sciaini, Marco, Scherer & Cédric. 2023. *viridis(Lite) - colorblind-friendly color maps for r.* doi:10.5281/zenodo.4679424. `https://sjmgarnier.github.io/viridis/`. Viridis package version 0.6.3.

Garnier, Simon. 2023. *viridis: Colorblind-friendly color maps for r.* `https://CRAN.R-project.org/package=viridis`. R package version 0.6.3.

Gavin, Michael C, Carlos A Botero, Claire Bowern, Robert K Colwell, Michael Dunn, Robert R Dunn, Russell D Gray, Kathryn R Kirby, Joe McCarter, Adam Powell, Thiago F. Rangel, John R. Stepp, Michelle Trautwein, Jennifer L. Verdolin & Gregor Yanega. 2013. Toward a Mechanistic Understanding of Linguistic Diversity. *BioScience* 63(7). 524–535.

Gavin, Michael C, Thiago F Rangel, Claire Bowern, Robert K Colwell, Kathryn R Kirby, Carlos A Botero, Michael Dunn, Robert R Dunn, Joe McCarter, Marco Túlio Pacheco Coelho & Russell D. Gray. 2017. Process-Based Modelling Shows How Climate and Demography Shape Language Diversity. *Global Ecology and Biogeography* 26(5). 584–591.

Gavin, Michael C & Nokuthaba Sibanda. 2012. The Island Biogeography of Languages. *Global Ecology and Biogeography* 21(10). 958–967.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

Gerber, Florian & Reinhard Furrer. 2015. Pitfalls in the implementation of Bayesian hierarchical modeling of areal count data: An illustration using BYM and Leroux models. *Journal of Statistical Software, Code Snippets* 63(1). 1–32. doi:10.18637/jss.v063.c01.

Gerber, Florian, Kaspar Moesinger & Reinhard Furrer. 2017. Extending R packages to support 64-bit compiled code: An illustration with spam64 and GIMMS NDVI3g data. *Computer & Geoscience* 104. 109–119. doi:10.1016/j.cageo.2016.11.015.

Gietzelt, Dale. 1989. The indonesianization of west papua. *Oceania* 59(3). 201–221.

Giuliano, Paola & Nathan Nunn. 2018. Ancestral Characteristics of Modern Populations. *Economic History of Developing Regions* 33(1). 1–17.

Gladwin, Thomas & Thomas Gladwin. 2009. *East is a big bird: Navigation and logic on puluwat atoll*. Harvard University Press.

Goldsmith, Michael. 1991. Tuvalu. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 354–357. Boston, Massachusetts: G.K. Hall and co.

Goodenough, Ward H. 1991. Truk. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 351–354. Boston, Massachusetts: G.K. Hall and co.

Goodenough, Ward Hunt. 2002. *Under heaven's brow: Pre-christian religious tradition in chuuk*, vol. 246. Philadelphia, PA: American Philosophical Society.

Gooskens, Charlotte & Vincent J van Heuven. 2017. Measuring Cross-Linguistic Intelligibility in the Germanic, Romance and Slavic Language Groups. *Speech Communication* 89. 25–36.

Grace, George W. 1992. How Do Languages Change? (More on "Aberrant" Languages). *Oceanic Linguistics* 31(1). 115–130. `http://www.jstor.org/stable/3622968`.

Graves, Michael W. 1986. Lare prehistoric complexity on lelū: alternatives to cordy's model. *The Journal of the Polynesian Society* 95(4). 479–489.

Gray, J Patrick. 1998. Ethnographic Atlas Codebook. *World Cultures* 10(1). 86–136.

Green, Roger. 1991. Near and Remote Oceania: Disestablishing "Melanesia" in Culture History. In A Pawley (ed.), *Man and a Half: Essays in Pacific Anthropology and Ethnobotany in honour of Ralph Bulmer*, 491–502. The Polynesian Society.

Green, Roger C & Marshall I Weisler. 2000. *Mangarevan archaeology: Interpretations using new data and 40 year old excavations to establish a sequence from 1200 to 1900 ad*. Otago, New Zealand: Department of Anthropology & Archaeology; University of Otago.

Greenhill, Simon. 2015. Demographic Correlates of Language Diversity. In Claire Bowern & Bethwyn Evans (eds.), *The Routledge Handbook of Historical Linguistics*, 557–578. Abingdon, UK and New York, USA: Routledge Taylor & Francis Group.

Guiart, Jean. 1952. *L'organisation sociale et politique traditionnelle à maré (iles loyalty)*. Nouméa, New Caledonia: Institut Français d'Oceanie.

Guo, Jiqiang, Jonah Gabry, Ben Goodrich & Sebastian Weber. 2023. *rstan: R interface to stan*. `https://CRAN.R-project.org/package=rstan`. R package version 2.21.8.

Hadfield, E. 1920. *Among the natives of the loyalty group*. London: Macmillan.

Hammarström, Harald, Thom Castermans, Robert Forkel, Kevin Verbeek, Michel A Westenberg & Bettina Speckmann. 2018. Simultaneous Visualization of Language Endangerment and Language Description. *Language Documentation & Conservation* 12. 359–392.

Hammarström, Harald, Robert Forkel & Martin Haspelmath (eds.). 2017. *Glottolog 3.0*. Jena: Max Planck Institute for the Science of Human History.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2019. Glottolog/Glottolog: Glottolog Database 4.0. `https://doi.org/10.5281/zenodo.3260726`.

Hammarström, Harald, Robert Forkel, Martin Haspelmath & Sebastian Bank. 2020. About Languoids. In Harald Hammarström, Robert Forkel, Martin Haspelmath & Sebastian Bank (eds.), *Glottolog 4.1 (Website)*, Jena: Max Planck Institute for the Science of Human History. doi:10.5281/zenodo.3260726.

Hanlon, David L. 2019. *Upon a stone altar: A history of the island of pohnpei to 1890*. Honolulu, HI: University of Hawaii Press.

Harlow, Ray B. 1973. Regional Variation in Maori. *New Zealand Journal of Archaeology* 1. 123–138.

Haun, Alan Edward. 1984. *Prehistoric subsistence, population, and sociopolitical evolution on ponape, micronesia*. University of Oregon. dissertation.

Hau'ofa, Epeli. 1993. Our Sea of Islands. In Eric Waddell, Vijay Naidu & Epeli Hau'ofa (eds.), *A New Oceania: Rediscovering Our Sea of Islands*, 2–14. Suva: School of Social and Economic Development, The University of the South Pacific.

Hayes, Geoffrey R. 1981. *The cook islands 1820-1950*. Wellington, New Zealand: Victoria University Press.

Hedvig Skirgård, Michael Zurmühle, Ursin Maurus Beeli & Peter Ranacher. 2023. Grouping islands in the pacific based on sailing distances makes sense for understanding the region before colonialism. Talk given at *Annual conference of the new zealand studies association*, University of Stockholm and University of Turku.

Hijmans, Robert J. 2023. *raster: Geographic data analysis and modeling*. `https://rspatial.org/raster`. R package version 3.6-23.

Hocart, A.M. 1929. Lau islands, fiji. *Bernice Pauahi Bishop Museum Bulletin* 62. 1–230.

Hooper, Antony & Judith Huntsman. 1973. A demographic history of the tokelau islands. *The Journal of the Polynesian Society* 82(4). 366–411.

Howard, Alan. 1963. Conservatism and non-traditional leadership in rotuma. *The Journal of the Polynesian Society* 65–77.

Howard, Alan. 1991. Rotuma. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 280–283. Boston, Massachusetts: G.K. Hall and co.

Hua, Xia, Simon J Greenhill, Marcel Cardillo, Hilde Schneemann & Lindell Bromham. 2019. The Ecological Drivers of Variation in Global Language Diversity. *Nature communications* 10(1). 2047. doi:10.1038/s41467-019-09842-2.

Huffman, Kirk. 1996. Trading, Cultural Exchange and Copyright: Important Aspects of Vanuatu Arts. In *Arts of Vanuatu*, 182–194. University of Hawaii Press.

Humphreys, C. B. 1926. The southern new hebrides: An ethnological record. CUP Archive.

Hunt, Edward Eyre, David Murray Schneider, Nathaniel R Kidder & William D Stevens. 1949. *The micronesians of yap and their depopulation: Report of the peabody museum expedition to yap island, micronesia, 1947-1948*. Washington: Peabody Museum, Harvard University.

Ihaka, Ross, Paul Murrell, Kurt Hornik, Jason C. Fisher, Reto Stauffer, Claus O. Wilke, Claire D. McWhite & Achim Zeileis. 2023. *colorspace: A toolbox for manipulating and assessing colors and palettes*. `https://CRAN.R-project.org/package=colorspace`. R package version 2.1-0.

Intoh, Michiko. 2008. Ongoing Archaeological Research on Fais Island, Micronesia. *Asian Perspectives* 47(1). 121–138.

Intoh, Michiko & Rintaro Ono. 2007. Reconnaissance Archaeological Research on Tobi Island, Palau. *People and Culture in Oceania* 22. 53–83.

Joseph, A. & V. F. Murray. 1951. *Chamorros and carolinians of saipan: personality tests with an analysis of the bender gestalt test by lauretta bender*. Cambridge: Harvard University Press.

Kassambara, Alboukadel. 2023. *ggpubr: ggplot2 based publication ready plots*. `https://rpkgs.datanovia.com/ggpubr/`. R package version 0.6.0.

Keesing, F. M. 1934. *Modern samoa:its government and changing life*. London: G. Allen & Unwin ltd.

Keesing, Roger M & Andrew J. Strathern. 1981. *Cultural Anthropology. A Contemporary Perspective*. Fort Worth: Harcourt Brace College Publishers.

Kirby, KR, R. D. Gray, S. J. Greenhill, F. M. Jordan, S. Gomes-Ng, H-J Bibiko, Damián E. Blasi, Carlos A. Botero, Claire Bowern, Carol R. Ember, Dan Leehr, Bobbi S. Low, Joe McCarter, William Divale & Michael C. Gavin. 2018. D-PLACE/dplace-data: D-PLACE – the Database of Places, Language, Culture and Environment (Version v2.0.1) [Data set]. `http://doi.org/10.5281/zenodo.1466634`.

Kirch, Patrick V. 2002. Te kai paka-anuta: food in a polynesian outlier society. doi:https://doi.org/10.4000/jso.1404.

Kirch, Patrick Vinton. 1984. *The evolution of the polynesian chiefdoms*. Cambridge University Press.

Kirch, Patrick Vinton. 1994. *The wet and the dry: irrigation and agricultural intensification in polynesia*. Chicago, IL: University of Chicago Press.

Kirch, Patrick Vinton. 2010. *How chiefs became kings: Divine kingship and the rise of archaic states in ancient hawai'i*. Oakland, CA: Univ of California Press.

Kirch, Patrick Vinton. 2017. *On the Road of the Winds: An Archaeological History of the Pacific Islands before European Contact*. Univ of California Press.

Kirch, PV. 2012. Baseline Prehistory. The Polynesian Outliers: Continuity, Change, and Replacement. In Richard Feinberg & Richard Scaglion (eds.), *The Polynesian Outliers: State of the Art* (Ethnology Monographs 21), 17–26. United States of America: Department of Anthropology, University of Pittsburgh.

Kislev, Mordechai E, Anat Hartmann & Ofer Bar-Yosef. 2006. Early domesticated fig in the jordan valley. *Science* 312(5778). 1372–1374.

Kuhlken, Robert. 2002. Intensive agricultural landscapes of oceania. *Journal of Cultural Geography* 19(2). 161–195.

Lambert, B. 1966. The economic activities of a gilbertese chief. In Tuden A. M. J. Schwartz, Turner VW (ed.), *Political anthropology*, 155–172. Chicago: Aldine Publishing Company.

Lambert, Bernd. 1975. Makin and the outside world. In Vern Carroll (ed.), *Pacific atoll populations*, 212–285. University Press of Hawaii.

Lambert, Bernd. 1991. Kiribati. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 120–124. Boston, Massachusetts: G.K. Hall and co.

Lane, R. B. 1956. The heathen communities of southeast pentecost. *Journal de la Sociéte des Oceanistes* 12. 139–180.

Lessa, W. A. 1950. The ethnography of ulithi atoll. Unpublished Manuscript Ulithi (Micronesia).

Lessa, William Armand. 1966. *Ulithi: A micronesian design for living*. New York: Holt, Rinehart and Winston.

Levin, Maureece J, Katherine Seikel & Aimee Miles. 2019. A Partial Chronological Sequence of Human Habitation for Pingelap Atoll (Pohnpei State, Federated States of Micronesia). *Radiocarbon* 61(3). 765–776. doi:10.1017/RDC.2019.30.

Lima-Ribeiro, Matheus S., Sara Varela, Javier González-Hernández, Guilherme de Oliveira, José Alexandre F Diniz-Filho & Levo Caroma Terribile. 2015. ecoClimate: A Database of Climate Data from Multiple Models for Past, Present, and Future for Macroecologists and Biogeographers. *Biodiversity Informatics* 10. 1–21.

Lindstrom, Lamont. 1991. Tanna. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 314. Boston, Massachusetts: G.K. Hall and co.

LINGUIST List. 2014. Multitree: A Digital Library of Language Relationships. Department of Linguistics, The LINGUIST List, Indiana University. `http://multitree.org/`.

Lipson, Mark, Pontus Skoglund, Matthew Spriggs, Frederique Valentin, Stuart Bedford, Richard Shing, Hallie Buckley, Iarawai Phillip, Graeme K. Ward, Swapan Mallick, Nadin Rohland, Nasreen Broomandkhoshbacht, Olivia Cheronet, Matthew Ferry, Thomas K. Harper, Megan Michel, Jonas Oppenheimer, Kendra Sirak, Kristin Stewardson, Kathryn Auckland, Adrian V.S. Hill, Kathryn Maitland, Stephen J. Oppenheimer, Tom Parks, Kathryn Robson, Thomas N. Williams, Douglas J. Kennett, Alexander J. Mentzer, Ron Pinhasi & David Reich. 2018. Population Turnover in Remote Oceania Shortly after Initial Settlement. *Current Biology* 28(7). 1157–1165.

Liu, Yue-Chen, Rosalind Hunter-Anderson, Olivia Cheronet, Joanne Eakin, Frank Camacho, Michael Pietrusewsky, Nadin Rohland, Alexander Ioannidis, J Stephen Athens, Michele Toomay Douglas et al. 2022. Ancient dna reveals five streams of migration into micronesia and matrilocality in early pacific seafarers. *Science* 377(6601). 72–79.

Loeb, E. M. 1978. *History and traditions of niue*, vol. 32. Honolulu: Bernice P Bishop Museum.

Lüpke, F. & A Storch. 2013. *Repertoires and Choices in African Languages*. Berlin, Boston: De Gruyter Mouton.

Lynch, John. 1981. Melanesian Diversity and Polynesian Homogeneity: The Other Side of the Coin. *Oceanic Linguistics* 20(2). 95–129.

Macdonald, Barrie. 1982. *Cinderellas of the empire: Towards a history of kiribati and tuvalu*. Canberra, Australia: Australian National University Press.

Macdonald, Barrie K. 2020. Tuvalu. *Encyclopædia Britannica* `https://www.britannica.com/place/Tuvalu`.

Macgregor, G. 1935. Notes on the ethnology of pukapuka. *Bernice P Bishop Museum Occasional Papers* 11. vi, 1–52.

Macgregor, G. 1937. *Ethnology of tokelau islands*, vol. 146. Honolulu: Bernice P Bishop Museum.

Mahony, Frank J. 1960. Taro cultivation in truk. In John E. deYoung (ed.), *Taro cultivation practices and beliefs: the eastern carolines and the marshall islands* (Anthropological Working Papers 6), 69–98. Guam, Micronesian Islands: Office of the Staff Anthropologist.

of Hawaii at Manoa, University. 2023. Catalogue of endangered languages. `http://www.endangeredlanguages.com`.

Mansfield, John, Henry Leslie-O'Neill & Haoyi Li. 2023. Dialect differences and linguistic divergence: A crosslinguistic survey of grammatical variation. *Language Dynamics and Change* 1(aop). 1–45.

Marck, Jeffrey C. 1986. Micronesian Dialects and the Overnight Voyage. *Journal of the Polynesian Society* 95(2). 253–258.

Marck, Jeffrey C. 2000. *Topics in Polynesian Language and Culture History*. Canberra: Pacific Linguistics.

Marean, Curtis W, Miryam Bar-Matthews, Jocelyn Bernatchez, Erich Fisher, Paul Goldberg, Andy IR Herries, Zenobia Jacobs, Antonieta Jerardino, Panagiotis Karkanas, Tom Minichillo et al. 2007. Early human use of marine resources and pigment in south africa during the middle pleistocene. *Nature* 449(7164). 905–908.

Marlow, Karina. 2016. Australia day, invasion day, survival day: What's in a name? Published 21 January 2016, Retrieved 30 July 2016. `https://www.sbs.com.au/nitv/article/australia-day-invasion-day-survival-day-whats-in-a-name/a348kzmlw`.

McElreath, Richard. 2020. *Statistical rethinking: A bayesian course with examples in r and stan*. CRC press 2nd edn.

Meleiseā, Malama. 1995. To Whom God and Men Crowded: Chieftainship and Hierarchy in in Ancient Samoa. In Judith Huntsman (ed.), *Tonga and Samoa: Images of Gender and Polity*, Christchurch: Macmillan Brown Centre for Pacific Studies.

Miscevic, Nenad. 2020. Nationalism. In Edward N. Zalta (ed.), *The stanford encyclopedia of philosophy*, Metaphysics Research Lab, Stanford University (fall 2020 edition) edn. `https://plato.stanford.edu/archives/fall2020/entries/nationalism/`.

Monberg, Torben. 1991. *Bellona island beliefs and rituals*. University of Hawaii Press.

Muller, W. 1917. Yap. In G. Thilenius (ed.), *Ergebnisse der südsee-expedition 1908-1910*, vol. 2, B, iii, 1–380. Hamburg: Friederichsen.

Murdock, G. P. 1962. *Ethnographic Atlas*, vol. I-XXVII. University of Pittsburgh Press.

Murdock, G. P., C. S. Ford & J. W. M. Whiting. 1944. West caroline islands.

Murdock, G.P., A.E. Hudson C.S. Ford, R. Kennedy, L.W. Simmons & J.W. M. Whiting. 1945. *Outline of Cultural Materials II*. Department of Anthropology, Yale University.

Métraux, A. 1971. *Ethnology of easter island*. Honolulu, HI: Bernice P Bishop Muesum.

Müller, Kirill & Hadley Wickham. 2023. *tibble: Simple data frames*. `https://CRAN.R-project.org/package=tibble`. R package version 3.2.1.

Napolitano, Matthew F., Scott M. Fitzpatrick, Geoffrey Clark & Jessica H. Stone. 2019. New Investigations of Early Prehistoric Settlement on Yap, Western Caroline Islands. *The Journal of Island and Coastal Archaeology* 14(1). 101–107. doi:10.1080/15564894.2017.1335661.

Nettle, Daniel. 1998. Explaining Global Patterns of Language Diversity. *Journal of Anthropological Archaeology* 17(4). 354–374. `https://doi.org/10.1006/jaar.1998.0328`.

Nordhoff, Sebastian & Harald Hammarström. 2011. Glottolog/Langdoc: Defining Dialects, Languages, and Language Families as Collections of Resources. Talk given at *First International Workshop on Linked Science 2011-In Conjunction with the International Semantic Web Conference (ISWC 2011)*, Koblenz, Germany.

Nychka, Douglas, Reinhard Furrer, John Paige, Stephan Sain, Florian Gerber & Matthew Iverson. 2022. *fields: Tools for spatial data*. `https://github.com/dnychka/fieldsRPackage`. R package version 14.1.

Oakes, Leigh. 2001. *Language and national identity: Comparing france and sweden*, vol. 13. John Benjamins Publishing.

Oliver, Douglas L. 2019. *Ancient tahitian society*, vol. 2: Social Relations. University of Hawaii Press.

Ooms, Jeroen. 2014. The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv:1403.2805 [stat.CO]* https://arxiv.org/abs/1403.2805.

Ooms, Jeroen. 2023. *jsonlite: A simple and robust json parser and generator for r.* https://jeroen.r-universe.dev/jsonlitehttps://arxiv.org/abs/1403.2805. R package version 1.8.7.

Ottolinger, Philipp. 2019. *bib2df: Parse a bibtex file to a data frame.* https://github.com/ropensci/bib2df. R package version 1.1.1.

Pawley, Andrew. 1981. Melanesian Diversity and Polynesian Homogeneity - a Unified Explanation for Language. In Jim Hollyman & Andrew Pawley (eds.), *Studies in Pacific Languages & Cultures in Honour of Bruce Biggs*, 269–309. Auckland: Linguistic Society of New Zealand.

Pawley, Andrew. 2007a. Locating Proto Oceanic. In Malcolm Ross, Andrew Pawley & Meredith Osmond (eds.), *The Physical Environment*, vol. 3 The Lexicon of Proto Oceanic: The Culture and Environment of Ancestral Oceanic Society, 17–34. Canberra: ANU Electronic Press.

Pawley, Andrew. 2007b. Why Do Polynesian Island Groups Have One Language and Melanesian Island Groups Have Many? Patterns of Interaction and Diversification in the Austronesian Colonisation in Remote Oceania. Talk given at *Workshop on Migration*, Ile de Porquerolles, France.

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

Pearl, Judea. 1995. Causal diagrams for empirical research. *Biometrika* 82(4). 669–688.

Pebesma, Edzer & Roger Bivand. 2023. *sp: Classes and methods for spatial data.* https://CRAN.R-project.org/package=sp. R package version 2.0-0.

Pebesma, Edzer J. & Roger Bivand. 2005. Classes and methods for spatial data in R. *R News* 5(2). 9–13. https://CRAN.R-project.org/doc/Rnews/.

Peoples, James G. 1991. Kosrae. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 128–131. Boston, Massachusetts: G.K. Hall and co.

Pollock, Nancy J. 1995. The power of kava in futuna and uvea/wallis. *Canberra anthropology* 18(1-2). 136–165.

Posth, Cosimo, Kathrin Nägele, Heidi Colleran, Frédérique Valentin, Stuart Bedford, Kaitip W. Kami, Richard Shing, Hallie Buckley, Rebecca Kinaston, Mary Walworth, Geoffrey R. Clark, Christian Reepmeyer, James Flexner, Tamara Maric, Johannes Moser, Julia Gresky, Lawrence Kiko, Kathryn J. Robson, Kathryn Auckland, Stephen J. Oppenheimer, Adrian V. S. Hill, Alexander J. Mentzer, Jana Zech, Fiona Petchey, Patrick Roberts, Choongwon Jeong, Russell D. Gray, Johannes Krause & Adam Powell. 2018. Language Continuity despite Population Replacement in Remote Oceania. *Nature Ecology & Evolution* 2. 731–740.

Powell, Adam, Stephen Shennan & Mark G Thomas. 2009. Late pleistocene demography and the appearance of modern human behavior. *Science* 324(5932). 1298–1301.

Quackenbush, Edward Miller. 1968. *From Sonsorol to Truk: A Dialect Chain.* University of Michigan. PhD Thesis.

Quain, Buell H. 1948. *Fijian village.* Chicago: University oF Chicago Press.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org/.

R Core Team. 2023. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing Vienna, Austria. `https://www.R-project.org/`.

Raiche, Gilles & David Magis. 2022. *nfactors: Parallel analysis and other non graphical solutions to the cattell scree test.* `https://CRAN.R-project.org/package=nFactors`. R package version 2.4.1.1.

Raviv, Limor, Antje Meyer & Shiri Lev-Ari. 2019. Larger Communities Create More Systematic Languages. *Proceedings of the Royal Society B* 286(1907). 20191262.

Ray, Sidney H. 1917. The people and language of lifu, loyalty islands. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* 47. 239–322.

Raynor, WC & JH Fownes. 1991. Indigenous agroforestry of pohnpei: 1. *Agroforestry Systems* 16. 139–157.

Rensch, Calvin R & Eugene H. Casad. 1992. Calculating Lexical Similarity. In *Windows on bilingualism*, 13–15. Summer Institute of Linguistics.

Revelle, William. 2023. *psych: Procedures for psychological, psychometric, and personality research.* `https://personality-project.org/r/psych/https://personality-project.org/r/psych-manual.pdf`. R package version 2.3.6.

Riesenberg, Saul. 1968. *The native polity of ponape. washington.* Smithsonian Institution Press.

Rieth, Timothy M. & Ethan E. Cochrane. 2018. The Chronology of Colonization in Remote Oceania. In Ethan E. Cochrane & Terry L. Hunt (eds.), *The Oxford Handbook of Prehistoric Oceania*, Oxford: Oxford University Press. `https://doi.org/10.1093/oxfordhb/9780199925070.013.010`.

Ripley, Brian. 2023. *Mass: Support functions and datasets for venables and ripley's mass.* `http://www.stats.ox.ac.uk/pub/MASS4/`. R package version 7.3-60.

Roberts, Seán G, Anton Killin, Angarika Deb, Catherine Sheard, Simon J Greenhill, Kaius Sinnemäki, José Segovia-Martín, Jonas Nölle, Aleksandrs Berdicevskis, Archie Humphreys-Balkwill, Hannah Little, Christopher Opie, Guillaume Jacques, Lindell Bromham, Peeter Tinits, Robert M Ross, Sean Lee, Emily Gasser, Jasmine Calladine, Matthew Spike, Stephen Francis Mann, Olena Shcherbakova, Ruth Singer, Shuya Zhang, Antonio Benítez-Burraco, Christian Kliesch, Ewan Thomas-Colquhoun, Hedvig Skirgård, Monica Tamariz, Sam Passmore, Thomas Pellard & Fiona Jordan. 2020. Chield: the causal hypotheses in evolutionary linguistics database. *Journal of Language Evolution* 5(2). 101–120.

Roberts, Seán Geraint. 2013. *An Evolutionary Approach to Bilingualism.* The University of Edinburgh. PhD Thesis.

Rolett, Barry & Jared Diamond. 2004. Environmental Predictors of Pre-European Deforestation on Pacific Islands. *Nature* 431(7007). 443.

Rolett, Barry V. 2002. Voyaging and interaction in ancient east polynesia. *Asian Perspectives* 182–194.

Ross, Malcolm. 2017. Languages of the new guinea region. In Raymond Hickey (ed.), *The cambridge handbook of areal linguistics*, Cambridge University Press.

Rudis, Bob, Ben Bolker & Jan Schulz. 2017. *ggalt: Extra coordinate systems, geoms, statistical transformations, scales and fonts for ggplot2.* `https://github.com/hrbrmstr/ggalt`. R package version 0.4.0.

Running, Steven & Maosheng Zhao. 2021a. Modis/aqua net primary production gap-filled yearly l4 global 500m sin grid v061. NASA EOSDIS Land Processes DAAC, accessed via the Application for Extracting and Exploring Analysis Ready Samples (AppEEARS). Accessed 2023-05-24 from `https://doi.org/10.5067/MODIS/MYD17A3HGF.061`.

Running, Steven & Maosheng Zhao. 2021b. MODIS/Terra Net Primary Production Gap-Filled Yearly L4 Global 500m SIN Grid V061. NASA EOSDIS Land Processes DAAC, accessed via the Application for Extracting and Exploring Analysis Ready Samples (AppEEARS). Accessed 2023-05-24 from `https://doi.org/10.5067/MODIS/MYD17A3HGF.061`.

Running, Steven W & Maosheng Zhao. 2015. Mod17 user's guide. daily gpp and annual npp (mod17a2/a3) products nasa earth observing system modis land algorithm. Tech. rep. Moderate Resolution Imaging Spectroradiometer,The United States of America's National Aeronautics and Space Administration. `https://www.umt.edu/numerical-terradynamic-simulation-group/files/modis/MOD17UsersGuide2015_v3.pdf`.

Sacson, Iris. 2018. Change Request Number: 2018-090. `https://iso639-3.sil.org/request/2018-090`.

Sahlins, Marshall D. 1963. Poor Man, Rich Man, Big-Man, Chief: Political Types in Melanesia and Polynesia. *Comparative Studies in Society and History* 5(3). 285–303.

Sahlins, Marshall David. 1958. *Social stratification in polynesia.* Seattle, WA: University Washington Press.

Salesius. 1907. Die karolinen-insel yap. fin beitrag zur kentnis von land und leuten in unseren deutschen südsee-kolonien [english trans by hraf: The caroline island of yap. a contribution towards an understanding our german south sea colonies].

Sarkar, Deepayan. 2008. *Lattice: Multivariate data visualization with r.* New York: Springer. `http://lmdvr.r-forge.r-project.org`.

Sarkar, Deepayan. 2023. *lattice: Trellis graphics for r.* `https://lattice.r-forge.r-project.org/`. R package version 0.21-8.

Scarr, Deryck. 1984. *Fiji: a short history.* Sydney Australia: George Allen and Unwin.

Schneider, D. M. 1962. Double descent on yap. *Journal of the Polynesian Society* 71. 1–24.

Schneider, David M. 1953. Yap kinship terminology and kin groups. *American Anthropologist* 55. 215–236.

Schneider, David M. 1957. Political organization, supernatural sanctions and the punishment for incest on yap. *American Anthropologist* 59. 791–800.

Schoeffel, Penelope. 1987. Rank, Gender and Politics in Ancient Sāmoa: The Genealogy of Salamāsina o Le Tafaifā. *Journal of Pacific History* 22(4). 174–194.

Schutten, Gerrit-Jan, Chung hong Chan, Thomas J. Leeper & Detlef Steuer. 2023. *readods: Read and write ods files.* `https://github.com/ropensci/readODS`. R package version 1.8.0.

Schwartz, Theodore. 1978. Where Is the Culture? Personality as the Distributive Locus of Culture. *The making of psychological anthropology* 419–441.

Sheehan, Oliver, Joseph Watts, Russell D Gray & Quentin D Atkinson. 2018. Coevolution of Landesque Capital Intensive Agriculture and Sociopolitical Hierarchy. *Proceedings of the National Academy of Sciences* 115(14). 3628–3633.

Silverstein, Michael. 1981. *The Limits of Awareness.* Southwest Educational Development Laboratory Austin, TX.

Skirgård, Hedvig. 2020. *Multilevel dynamics of language diversity in oceania.* Australian National University. dissertation. `https://openresearch-repository.anu.edu.au/handle/1885/218982`.

Skirgård, Hedvig, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoglu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye , Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson & Russell D. Gray. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances* 9. doi:10.1126/sciadv.adg6175.

Slowikowski, Kamil. 2023. *ggrepel: Automatically position non-overlapping text labels with ggplot2*. `https://github.com/slowkow/ggrepel`. R package version 0.9.3.

Smith, S et al. 1983[1902-1903]. *Niue the island and its people*. Suva, Fiji: The Polynesian Society.

Spehr, A. 1954. Saipan. *Fieldiana: Anth.* 41. 1–383.

Spriggs, Matthew. 1982. Taro cropping systems in the southeast asian-pacific region: Archaeological evidence. *Archaeology in Oceania* 17(1). 7–15.

Spriggs, Matthew. 1986. Landscape, land use and political transformation in southern melanesia. In PV Kirch (ed.), *Island societies: Archaeological approaches to evolution and transformation*, 6–19. Cambridge University Press.

Spriggs, Matthew & Stephen Wickler. 1989. Archaeological research on erromango: Recent data on southern melanesian prehistory. *Bulletin of the Indo-Pacific Prehistory Association* 9. 68–91.

Stauffer, Reto, Georg J. Mayr, Markus Dabernig & Achim Zeileis. 2009. Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations. *Bulletin of the American Meteorological Society* 96(2). 203–216. doi:10.1175/BAMS-D-13-00155.1.

Swadesh, Morris. 1954. Perspectives and problems of Amerindian comparative linguistics. *Word-journal of The International Linguistic Association* 10(2-3). 306–332.

Tetens, A. 1958. Among the savages of the south seas. (Trans. F. M. Spoehr).

Tetens, A. & J. Kubary. 1873. Die carolineninsel yap. *Journal des Museum Godeffroy* 1. 84–120.

Thompson, L. 1940a. Southern lau, fiji. *Bernice Pauahi Bishop Museum Bulletin* 162.

Thompson, L. 1971[1945]. The native culture of the marianas islands. *Bernice Pauahi Bishop Museum Bulletin* .

Thompson, Laura. 1940b. *Fijian frontier* (Studies of the Pacific 4). San Francisco: Institute of Pacific Relations.

Tierney, Nicholas, Di Cook, Miles McBain & Colin Fay. 2023. *naniar: Data structures, summaries, and visualisations for missing data*. `https://github.com/njtierney/naniar`. R package version 1.0.0.

Tierney, Nicholas & Dianne Cook. 2023. Expanding tidy data principles to facilitate missing data exploration, visualization and assessment of imputations. *Journal of Statistical Software* 105(7). 1–31. doi:10.18637/jss.v105.i07.

Tonkinson, R. 1981. Church and kastom in southeast ambrym in vanuatu: Politics, economics and ritual in island melanesia. In M Allen (ed.), *Vanuatu: Politics, economics and ritual in island melanesia*, 237–267. Sydney, Australia: Academic Press.

Tsukamoto, Akihisa. 1988. *The Language of Niuafo'ou Island*. Australian National University. PhD Thesis.

Turner, George. 1884. *Samoa: A Hundred Years Ago and Long before; Together with Notes on the Cults and Customs of Twenty-Three Other Islands in the Pacific*. London: Macmillan and Co.

United Nations, Department of Economic and Social Affairs, Population Division. 2022. World population prospects 2022. Online Edition. `https://population.un.org/wpp/Download/Standard/MostUsed/`.

Van Meijl, Toon. 1995. Maori socio-political organization in pre-and proto-history: On the evolution of post-colonial constructs. *Oceania* 65(4). 304–322.

Venables, W. N. & B. D. Ripley. 2002. *Modern applied statistics with s*. New York: Springer 4th edn. `https://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

Walter, Michael AHB. 1978. An examination of hierarchical notions in fijian society: A test case for the applicability of the term" chief". *Oceania* 49(1). 1–19.

Walter, R. 1996. Settlement pattern archaeology in the southern cook islands: a review. *Journal of the Polynesian Society* 105. 63–99.

Walter, R & A Anderson. 1995. Archaeology of niue island: Initial results. *Journal of the Polynesian Society* 104. 471–481.

Walworth, Mary. 2015. *The Language of Rapa Iti: Description of a Language in Change*. University of Hawaii at Mānoa. PhD Thesis.

Watters, R.F. 1958. Cultivation in old samoa. *Economic Geography* 43. 338–351.

Watts, Joseph, Oliver Sheehan, Joseph Bulbulia, Russell Gray & Quentin Atkinson. 2018. Christianity Spread Faster in Small, Politically Structured Societies. *Nature Human Behaviour* 2. 559–564. doi:10.1038/s41562-018-0379-3.

Weinreich, U., W. Labov & M.I. Herzog. 1968. Empirical Foundations for a Theory of Language Change. In W. Lehmann & Y. Malkiel (eds.), *Directions for Historical Linguistics: A Symposium*, Austin, TX: University of Texas Press.

Wessel, På l & Walter HF Smith. 1996. A Global, Self-Consistent, Hierarchical, High-Resolution Shoreline Database. *Journal of Geophysical Research: Solid Earth* 101(B4). 8741–8743. Data used is from version 2.3.7.

Wickham, Hadley. 2007. Reshaping data with the reshape package. *Journal of Statistical Software* 21(12). 1–20. `http://www.jstatsoft.org/v21/i12/`.

Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. `https://ggplot2.tidyverse.org`.

Wickham, Hadley. 2020. *reshape2: Flexibly reshape data: A reboot of the reshape package*. `https://github.com/hadley/reshape`. R package version 1.4.4.

Wickham, Hadley. 2022. *stringr: Simple, consistent wrappers for common string operations*. `https://CRAN.R-project.org/package=stringr`. R package version 1.5.0.

Wickham, Hadley. 2023. *forcats: Tools for working with categorical variables (factors)*. https://CRAN.R-project.org/package=forcats. R package version 1.0.0.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo & Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43). 1686. doi:10.21105/joss.01686.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani & Dewey Dunnington. 2023a. *ggplot2: Create elegant data visualisations using the grammar of graphics*. https://CRAN.R-project.org/package=ggplot2. R package version 3.4.2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller & Davis Vaughan. 2023b. *dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr. R package version 1.1.2.

Wickham, Hadley & Lionel Henry. 2023. *purrr: Functional programming tools*. https://CRAN.R-project.org/package=purrr. R package version 1.0.1.

Wickham, Hadley, Jim Hester & Jennifer Bryan. 2023c. *readr: Read rectangular text data*. https://CRAN.R-project.org/package=readr. R package version 2.1.4.

Wickham, Hadley, Jim Hester, Winston Chang & Jennifer Bryan. 2022. *devtools: Tools to make developing r packages easier*. https://CRAN.R-project.org/package=devtools. R package version 2.4.5.

Wickham, Hadley & Dana Seidel. 2022. *scales: Scale functions for visualization*. https://CRAN.R-project.org/package=scales. R package version 1.2.1.

Wickham, Hadley, Davis Vaughan & Maximilian Girlich. 2023d. *tidyr: Tidy messy data*. https://CRAN.R-project.org/package=tidyr. R package version 1.3.0.

Williamson, Sabath MD I. 1982. Island population, land area, and climate: a case study of the marshall islands. *Human Ecology* 10. 71–84.

Winslow, Donna. 1991. Ajië. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 9. Boston, Massachusetts: G.K. Hall and co.

Xie, Yihui. 2014. knitr: A comprehensive tool for reproducible research in R. In Victoria Stodden, Friedrich Leisch & Roger D. Peng (eds.), *Implementing reproducible computational research*, Chapman and Hall/CRC. ISBN 978-1466561595.

Xie, Yihui. 2015. *Dynamic documents with R and knitr*. Boca Raton, Florida: Chapman and Hall/CRC 2nd edn. https://yihui.org/knitr/. ISBN 978-1498716963.

Xie, Yihui. 2023. *knitr: A general-purpose package for dynamic report generation in r*. https://yihui.org/knitr/. R package version 1.43.

Young, Michael. 1991. Dobu; goodenough island. In Terence E. Hays (ed.), *Encyclopedia of world cultures: Ii oceania*, 85–88. Boston, Massachusetts: G.K. Hall and co.

Zeileis, Achim, Jason C. Fisher, Kurt Hornik, Ross Ihaka, Claire D. McWhite, Paul Murrell, Reto Stauffer & Claus O. Wilke. 2020. colorspace: A toolbox for manipulating and assessing colors and palettes. *Journal of Statistical Software* 96(1). 1–49. doi:10.18637/jss.v096.i01.

Zeileis, Achim, Kurt Hornik & Paul Murrell. 2009. Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics & Data Analysis* 53(9). 3259–3270. doi:10.1016/j.csda.2008.11.033.

Data and code availability

All data in this study is available freely. The data has been gathered from multiple sources and is aggregated and will be published along with the code for the study on OSF/GitHub/Zenodo. The code is also made available, from gathering data to conducting analysis and preparing plots and tables.

Anonymised OSF project for double-blind peer review: `https://osf.io/amnh9/?view_only=403d57d715254539a3a99c798b4`

Information on latitude and shoreline is taken from the Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG) (Wessel & Smith, 1996). Rainfall and temperature information is taken from the ecoClimate database (Lima-Ribeiro et al., 2015). Data on Net Primary Production comes from the United States of America's National Aeronautics and Space Administration's Moderate Resolution Imaging Spectroradiometer (MODIS, Running & Zhao (2021b,a)). Language classification is taken from Glottolog (version 3.0), with some modification (see section B.2).

GitHub locations:

- predicting_number_lgs_in_remote_oceania withheld while in review process
- glottolog-cldf (v3) `https://github.com/glottolog/glottolog/tree/v3.0`

Zenodo locations:

- predicting_number_lgs_in_remote_oceania TBA
- glottolog-cldf (v3) `https://doi.org/10.5281/zenodo.437430`

In the scripts, overnight-sailing distances island groups are labelled "SBZR" (initials of authors of Hedvig Skirgård & Ranacher (2023)) and shared-language-groups "medium".

Definitions of variables

Defining Remote Oceania

The distinction of Near and Remote Oceania was first suggested by Pawley & Green (1973)[8] and was further elaborated on in Green (1991). This distinction is based on geology, flora and fauna. Near Oceania consists of chains of inter-visible large islands and archipelagos, whereas the islands and atolls of Remote Oceania are separated by much greater distances. This also coincides with human settlement patterns, with Remote Oceania only being colonized in the last 3,000 years but Near Oceania long before that.

There is a marked difference in the biodiversity of Near and Remote Oceania, due to the greater distances between islands in Remote Oceania. This has consequences for both flora and fauna, as Green (1991) and Pawley (2007a) write:

> *all terrestrial mammals other than rats and mice or those which accompanied people reach their eastward limit in the Solomons. The same applies to all fresh-water mussels, and most of the Palaeo-Oriental land-snail fauna. Thirty Papuan and Malayan genera of birds find their eastern limits here, as do 162 genera of seed-plants, about 24% of the total.*
>
> Green (1991, 495)

> *Even in marine life the difference is marked. The reefs of the Bismarck and Solomons show a much richer diversity of fish, molluscs, echinoderms, crustacea, seaweeds, and other edible life than those of Remote Oceanic [sic].*
>
> Pawley (2007a, 19)

The greater distances between the islands of Remote Oceania is also the reason why they were settled by people much later than Near Oceania was (3,000 years versus 20,000-60,000 years ago). It was necessary to develop particular sailing techniques to traverse these greater distances in order to settle the region.

[8] This is the first occurrence of the two terms, although Green wrote earlier papers which included some of the ideas.

In this study, the region of Remote Oceania is defined slightly differently from the maps in Green (1991) and Pawley (2007a). All Northern Polynesian Outliers in Papua New Guinea and Solomons are incorporated, as is Mapia in northwest Papua New Guinea. The lines in the maps of Green (1991) and Pawley (2007a) appear not to include these in Remote Oceania. I have included them in part to complete the Polynesian subgroup of languages and the languages of the Micronesian archipelagos, but also because it appears that these were largely not inhabited prior to the Austronesian expansion into the rest of Remote Oceania (c.f. Kirch (2012, 23)) which suggests that they are sufficiently distant from the larger islands of Near Oceania to be counted as "Remote".

Regarding the terms "Near" and "Remote", there are two interpretations of what they are referring to:

- islands and atolls of "Near Oceania" are nearer *to each other*
- islands and atolls of "Near Oceania" are nearer to large landmasses west-wards, such as Australia.

The difference between these two interpretations has no direct bearing on the study at hand, but the first one appears more fitting given the ecological framing of distance for people, flora and fauna to travel to colonise the new land. The islands of the Pacific are often construed as "islands in a vast sea" (Hau'ofa, 1993) and a reading of "Remote" as "far away from places of power/important places" may enforce that understanding which is unnecessary and detrimental to our understanding of the history of the region (especially considering the importance of Sāmoa and Tahiti as centres of power).

Defining "language"

This study investigates factors influencing language diversification. The most central part of our data is the number of languages per island group; this is the response variable of our models. It is notoriously difficult to distinguish dialects, languages and language groups and there exist many different competing standards for defining what a language is. Linguists and others have debated these issues for a long time. Given their importance and centrality to this research, this section will give an introduction to the language identification standards adopted in two of the most commonly used resources: SIL International's ISO 639-3 codes for language names (which is implemented in Ethnologue (Eberhard et al., 2019a)) and Glottolog's glottocodes Hammarström et al. (2017).

It should be noted that while different language identification standards do differ, their findings do not vary dramatically. For example, SIL's Ethnologue states that there are 7,111 living languages today, Glottolog reports that there are 6,989[9]. The difference between these two total counts is "only" 122 — which is quite small considering how controversial language identification standards can be. This is in line with Nettle's observation that while linguists often disagree on how precisely to define a language, there is often considerable agreement in practice when it comes to categorising specific language varieties (Nettle, 1998, 356).

The Summer Institute of Linguistics International (SIL) is the official Registration Authority of the International Organization for Standardisation (ISO) standard for codes to represent language names — ISO 639-3. There are other ISO standards for languages, but ISO 639-3 is the most comprehensive and widely used. Most students and scholars are familiar with this code set from the SIL publication Ethnologue (which in 2016 became fully accessible only to paying subscribers, contributors and users from developing countries). However, it should be noted that the code set is also available independently of Ethnologue at https://iso639-3.sil.org/. The ISO 639-3 is technically separate from Ethnologue, and maintained by separate staff of SIL International.

The purpose of ISO 639-3 is to coordinate language names. There can be many different names for the same language, for example, Armenian is known both as "Haieren" and "Ermenice" in academic literature (LINGUIST List, 2014). A code standard is needed in order to coordinate work within language technology, libraries and scholarly research. The ISO published the code standard 639-3 in 2007, but it was based on classifications in editions of Ethnologue published as early as 1984.

How then does ISO 639-3 classify languages? Under the heading "The Problem of Language Identification" the editors of Ethnologue outline their approach (Eberhard et al., 2019b). While they discuss the inherent complexities with defining languages in a universal standard, and even borrow metaphors from quantum physics ("Language as particle, wave, and field") they also provide specific criteria in the form of a list:

[9] This the number of languoids in glottolog classifies as "language" and not labelled as "extinct".

- *Two related varieties are normally considered varieties of the same language if speakers of each variety have inherent understanding of the other variety at a functional level (that is, can understand based on knowledge of their own variety without needing to learn the other variety).*
- *Where spoken intelligibility between varieties is marginal, the existence of a common literature or of a common ethnolinguistic identity with a central variety that both understand can be a strong indicator that they should nevertheless be considered varieties of the same language.*
- *Where there is enough intelligibility between varieties to enable communication, the existence of long-standing distinctly named ethnolinguistic identities coupled with well-developed standardization and literature that are distinct can be treated as an indicator that they should nevertheless be considered to be different languages.*

These criteria probably correspond rather well to the definition of “a language” by non-academics. However, they can be tricky to apply, and they are also very dependent on the particular cultural history and literary development of the communities in question. It is often said that language classification is as much a political issue as it is a matter for scholarly classification. Language is tied to cultural identity and that identity is necessarily defined in opposition to other communities. Self reported group identity is rarely generalisable and possible to distil into a neat universal standard. It has been said that *a language is a dialect with an army and a navy*[10], meaning that what is a language versus a dialect may be a product of political power structures in a given region.

The SIL in their definition of language vs dialect also refer to “intelligibility”, but it is unclear how this is defined and measured. On the page “Language information” in the online version of Ethnologue, we learn that there are particular cut-off scores for lexical similarity and intelligibility (85%):

> *Intelligibility and dialect relations. A measure of inherent intelligibility with other varieties is given by percent. Values of less than 85% are likely to signal difficulty in comprehension of the indicated language.*
>
> ..
>
> *Intelligibility may not be reciprocal or mutual, thus the wording of the intelligibility description may indicate the direction of the intelligibility*
>
> ..
>
> *Lexical similarity. The percentage of lexical similarity between two linguistic varieties is determined by comparing a set of standardized wordlists and counting those forms that show similarity in both form and meaning. Percentages higher than 85% usually indicate a speech variant that is likely a dialect of the language with which it is being compared. Unlike intelligibility, lexical similarity is bidirectional or reciprocal.*
>
> Eberhard et al. (2019b)

The word lists used are most likely standardised lists of basic vocabulary similar to Swadesh-lists. The particular method of calculating lexical similarity is based on a standardised procedure by Rensch & Casad (1992). Swadesh (1954, 326) has suggested a cut-off of lexical similarity as 81% instead for deeming two varieties to be of the same language.

It is still unclear how intelligibility is defined and measured by SIL. Mutual intelligibility can vary with context and need not be symmetrical (Nettle, 1998, 356). In their study of mutual intelligibility between pairs of different European languages, Gooskens & van Heuven (2017) found that Dutch speakers had a higher than 85% success rate at a spoken picture task in German[11]. Their study involved many different kinds of tasks, and depending

[10] The exact origin of this phrase is not known. It is most often attributed to Max Weinreich, but it may also have been first stated by Joshua Fishman (a student of Weinreich’s) or be an expansion of something declared by Antoine Meillet (Bright, 1997, 469). Weinreich himself has also said that he got it from an anonymous member of a lecture audience during 1943-4.

[11] For more information on how this study took into account schooling and passive knowledge of the target language by their participants, please see Gooskens & van Heuven (2017).

on which one was used certain language pairs would sometimes be considered as the same language by the ISO 639-3 criteria, and sometimes not.

The ISO 639-3 code standard also recognises so-called "macrolanguages", for example "Arabic" which covers 30 separate languages. Besides the ISO 639-3, there are also other ISO standards for languages and in some of these "Chinese" is counted as one language[12]. The concept of "macrolanguage" aids in linking ISO 639-3 to another set of language codes (ISO 639-2). "Macrolanguages" also provide an identifier for sets of languages that could be recognised as one language based on shared literature and ethnolinguistic identity, but they are not sufficiently mutually intelligible to qualify as one language by the 639-3 standard. Languages in this category are for example "Akan", which is broken into Fanti and Twi in 639-3 and "Chinese" which covers 14 languages. This gives us an insight into borderline cases, linguistic entities that by some of the criteria are the same language but which fail on the crucial test of enough mutual intelligibility and/or lexical similarity.

SIL International welcomes change requests and are continuously updating their classification. Since 2007, they have adopted 88 requests for splitting languages and 152 for merging. They have also rejected 22 requests for splitting and 11 for merging. Looking through a few of these requests, many make explicit reference to Bible translation projects. For example, change request 2018-090 which seeks to merge [nns] into [nbr] is based on a dialect survey which has as its primary aim "to determine indicators of Bible translation needs for Luke Initiative for Scripture Translation (LIST)" (Sacson, 2018), in order to optimise the ongoing Bible translation work with speakers there.

SIL International is a "faith-based" organisation with its roots in Evangelical Christianity. They conduct a lot of their work together with their sister organisation Wycliffe Bible Translations, which is explicitly a missionary organisation. This has led some to ask whether or not the goal of spreading the word of God has influenced the scholarly work of SIL and Ethnologue. For example, the above criteria of ISO 639-3 (mutual intelligibility, shared literature and distinct ethnolinguistic identities) may in practice be heavily influenced by the work of coordinating and optimising Bible translation. Inspired by this observation, Lüpke & Storch (2013) and Blommaert (2008) reformulated the famous statement about a language being "a dialect with an army and a navy" as:

> *a language is a dialect with a missionary and a dictionary.*
>
> Lüpke & Storch (2013) and Blommaert (2008)

Alongside SIL International and the ISO 639-3 code standard for language names, an alternative code set has emerged and is gaining in popularity, namely Glottolog's "glottocodes". Glottolog started out from the work of Nordhoff & Hammarström (2011) and the online databse is now on its 4th edition (Hammarström et al., 2019). The aim of the project is to provide stable identifiers for languages (as well as dialects and families) and coordinate bibliographic information on the world's languages. Similarly to Ethnologue, Glottolog also contains information on language genealogy and endangerment levels (Hammarström et al., 2018). If the aims of the SIL and Ethnologue may be influenced by spreading the Bible, it could be said that Glottolog is dominated more by the aim of coordinating existing bibliographic resources.

Glottolog identifies three different kinds of linguistic entities: families, languages and dialects. These are collectively known as "languoids". Families are any entity above a "language" (highest order genealogical unit as well as lower level branches) and dialects are anything below "language". All of these entities are assigned a unique stable identifier, a "glottocode". We will first discuss their definition of "language" and then the advantages of providing stable identifiers for dialects as well as languages.

Unlike SIL's ISO 639-3, Glottolog does not discuss shared cultural identity or literature but focuses solely on mutual intelligibility when defining what is a language and what is a dialect (Hammarström et al., 2020). On the website, the editors outline their process for language classification as a decision tree (see Fig. 7), with the first question asked being "Is the putative language assertably distinct from all other known languages?".

> *By distinct, we mean not mutually intelligible with any other language. In principle, any convincing evidence to this effect is sufficient. For example, direct comparison of language data or testimonies of non-intelligibility to all neighbouring languages is the most straightforward kind of evidence. But also, various types of evidence*

[12] Many of the other ISO standards for language names are primarily focussed on library use, as evidenced by the fact that the registration authority for ISO 639-2 is the American Library of Congress.

for isolation from all other humans for a long time could make a convincing case that a language is indeed distinct from all others.

(Hammarström et al., 2020)

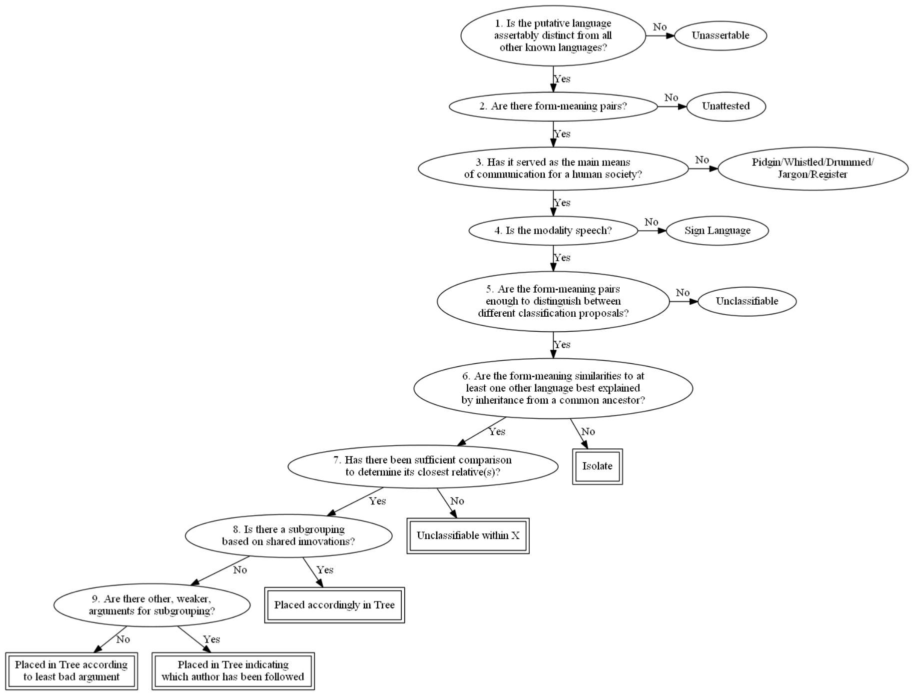

Fig. 7: Decision tree for language identification and classification in Glottolog (Hammarström et al., 2020).

In the quote above from Glottolog on "distinctiveness", the editors state that one can use language data and/or reports on mutual intelligibility. When empirical data on mutual intelligibility is missing, "an approximate minimal requirement is 50 items or so of basic vocabulary" needed for determining the distinctness of a language and where it should be placed in a genealogical tree (Hammarström et al. (2020) and Hammarström p.c.). This is similar to the lexical similarity measurements used by the SIL.

One important difference between Glottolog and SIL International lies in how the evidence is documented. Glottolog provides references for every language identification and tree, making it possible for other researchers to examine the evidence on their own. SIL International does not consistently provide this information. While change requests to ISO 639-3 will often refer to reports etc and many language entries have references for at least part of the information provided, there are many instances where the information is not accessible to readers. While the Ethnologue website states that "sources used for classifications are available on request by contacting the Editor" (Eberhard et al., 2019b), personal correspondence with the editors revealed that the sources are incomplete and cannot be requested in full.

Glottolog also considers whether or not the putative language has "served as the main means of communication for a human society" (Hammarström et al., 2020). This disqualifies most artificial languages, whistle registers, ritual speech registers and pidgins. SIL International does not discuss this particular criteria explicitly. However, among the artificial languages Esperanto is included in the catalogue but Klingon, High Valyrian and Angosey are not. Unlike most other artificial languages, Esperanto does have a few native speakers (Bergen, 2001), which is probably why it is included in Ethnologue. This suggests that SIL does consider something similar to Glottolog's "main means of communication" criteria even if they do not spell it out explicitly. Glottolog does catalogue constructed languages, pidgins etc, placing them in so-called "non-genealogical trees"[13]. These groups

[13] The list of non-genealogical trees in Glottolog are: Sign Language, Unclassifiable, Pidgin, Unattested (data missing), Mixed Language, Artificial Language and Speech Register.

of languages are included in the catalogue and given glottocodes, but the editors note that they are not subject to the same process of language identification and classification as the rest.

The fact that Glottolog provides unique stable identifiers for dialects and families, as well as languages, makes it possible to represent data in finer detail than previously possible. It also allows individual scholars to make different decisions on what is and what is not a language from Glottolog, while maintaining comparability. For example, there are several language varieties spoken on the Austral islands of the Pacific. Glottolog has classified them as being one language (Austral [aust1304]) with four dialects (Ra'ivavae [raiv1237], Rimatara [rima1237], Rurutu [ruru1237] and Tubuai [tubu1240]). However, other linguists disagree and have argued that these should be viewed as four different languages due to wordlists indicating they were not sufficiently mutually intelligible at the time of the arrival of Europeans (Walworth & François p.c.). In this dataset, these are counted as four separate languages because it appears to be more in line with the state of the world before the arrival of Europeans. It is easy to adjust for this in the dataset (and simple to undo) because the Glottolog dialect codes are linked to the language Austral [aust1304].

Being able to refer to dialects and families with a stable identifier is useful and an important difference between the SIL International ISO 639-3 code standard and Glottolog's glottocodes. When working with cross-linguistic and cross-cultural databases, it is often useful to be able to refer to a language variety in a more granular detail than the ISO 639-3 set allows for. Nordhoff & Hammarström (2011) point out that there may be differences between dialects that are crucial to certain research that would be confused if it was only possible to refer in a stable manner to the language level and not also to specific dialects. For example, certain dialects of German make a distinction between /eː/ and /æː/ whilst others don't. If references were not accurately tied to a language variety, this would be confusing as "German" would have different numbers of vowels in different sources. By providing stable identifiers for dialects as well as languages and providing a controlled hierarchy, it is possible for researchers to document their data accurately while still making it possible to aggregate up to the "language-level" when it is desirable. For example, Grambank and the Standard Cross Cultural Sample of D-PLACE both contain entries for specific languages and dialects. There are 158 one-to-one matches between the glottocodes of these two databases. In some cases though, the databases contain different dialects of the same language. If we were to lump dialects of the same language together in each of the datasets, the overlap increases to 183.

The definitions of language in ISO 639-3 (as described in Ethnologue (Eberhard et al., 2019b)) and Glottolog both focus on mutual intelligibility. Ethnologue notes that intelligibility does have to be symmetric, but neither of the resources discuss multilingualism, or multidialectalism in any detail. It is possible that the world used to be much more multilingual than it is today (Evans, 2017), and there are still many places where communities are highly competent in many different language varieties. Ethnologue state that they focus on *inherent* intelligibility, as opposed to *acquired*, but in situations where languages are not formally taught but rather present in the community continuously it can be difficult to draw the boundary between inherent and acquired. The underestimation of multilingualism has significant consequences for studies of the evolution of language (Roberts, 2013). However, unless two languages are exclusively spoken in a fully multilingual setting and never outside of it, the total count of languages should remain the same.

The language classification standards for Glottolog and SIL International do differ conceptually, and many of their differences may reflect their different aims (facilitating academic research by making references more accessible and integrated as compared to facilitating description and translation of the Bible into the world's languages). However, the counts for languages in the world that they produce are not that different in the end.

I will be using the classification of Glottolog, with a few adjustments, since it is more transparent and better referenced. It should be noted that using the ISO 639-3 results in the language count for the Vanuatu being 108, as opposed to 105 in Glottolog, which is not a huge difference.

The aim of this study is to explore language diversification in Remote Oceania prior to the arrival of European people. Because of this, I have made a few adjustments to the Glottolog classifications of languages in the relevant region. Languages that have come into the region directly because of the arrival of Europeans have been excluded (Bislama, English, French etc). Furthermore, I have made three other changes from Glottolog in the classification of indigenous languages of the region based on other accounts of their mutual intelligibility at the relevant point in time[14]. Firstly: in Fiji, Glottolog's Eastern Fijian [fiji1243] is split into three languages: Southeast Fijian [sout2864], Northeast Fijian [nort2842] and Kadavu [kada1285] based on personal correspondence with Andrew Pawley and Paul Geraghty. Secondly, Māori [maor1246] is split into two languages: Morori [mori1267] on Rēkohou (Chatham Islands) and Māori [maor1246] on mainland Aotearoa (New Zealand). This classification is based on Harlow (1973) and personal advice from Andrew Pawley. Third and finally, as was previously mentioned, the languages of the Austral Islands have been divided from one to four based on advice from Mary Walworth and Alexandre François.

The splitting of Eastern Fijian [fiji1243] into three languages has as consequence that the total language count for Fiji as one island group based on overnight sailing distances is 8 instead of 5. When the island groups are defined in relation to whether they share at least one language, the great island of Viti Levu is paired with Yasawa into one group sporting 4 languages instead of 2 and Kadavu separates out as a separate island group with 1 language. For Māori [maor1246] the difference in language classification makes less difference. Aotearoa and Rēkohou were always separated as different overnight distance island groups, each sporting one language (even though it is the same language under Glottolog's definition). For the island groups based on shared language they continue to be separated under the new classification, meaning in practice that we have yet another small Polynesian island group with one language. The situation is similar for the Austral islands. The new classification makes no difference for the island groups as defined by overnight distances since all four are separated out from each other already. For the island groups based on shared language, they continue to be separated out as different island groups. Overall the new classification results in more languages in Fiji, and more small Polynesian island groups with one language each.

Grouping Islands and atolls

For this study, we need a good way of grouping islands and atolls in Remote Oceania (5,525 landmasses) that is not dependent on modern politics (e.g. nation-states), but that reflects possible meaningful networks in Oceania prior to colonization. For this purpose, we will use two distinct approaches. The first is based on voyaging distances by canoe, and the second is based on whether the islands have a language in common or not. The assumption is that a group of islands is a meaningful unit for comparison if it forms a significant *network of contact* and that this can be approximated by shorter sailing distances or evidenced by the maintenance of a shared language.

These two principles, overnight sailing distances and sharing a language, have been applied consistently over all islands and all languages using data available in the Ethnologue (Eberhard et al., 2019a), Glottolog (Hammarström et al., 2017) and in some cases specialist literature (Faaniu (1983); Charpentier & François (2012); François et al. (2015); Macdonald (2020); Ager (2020) and a map of indigenous languages of New Caledonia published by CNRS-LACITO). For more on the definition of languages, see section B.2.

Grouping islands based on voyaging distances. Marck (1986, 2000) has suggested that islands within 100 miles[15] are reachable by overnight voyages in a traditional canoe (24h). This is taken to represent a good approximation of island connections of frequent contact and is supported by other accounts, such as (Gladwin & Gladwin, 2009, 38) who notes a higher frequency of travel between islands within 100 miles of each other. This is a distance people may travel for tribute offerings, trade and community events, which strengthens cultural ties. Further away, and it may not be so easy to get together as often. Fig. 8 shows this principle applied to central East Polynesia.

This idea proved useful also in terms of understanding language communities. Marck (1986, 2000) demonstrates that this way of grouping islands often conforms neatly to linguistic boundaries — especially in the north and east Pacific. For example, in Fig 8 the boundaries mostly line up well with well-established

[14] Tahitianization and dialect/language levelling has occurred, and therefore languages that used to be more different from each other are now more similar.

[15] Note that Marck uses *land* miles, not *sea* miles (Marck personal correspondence).

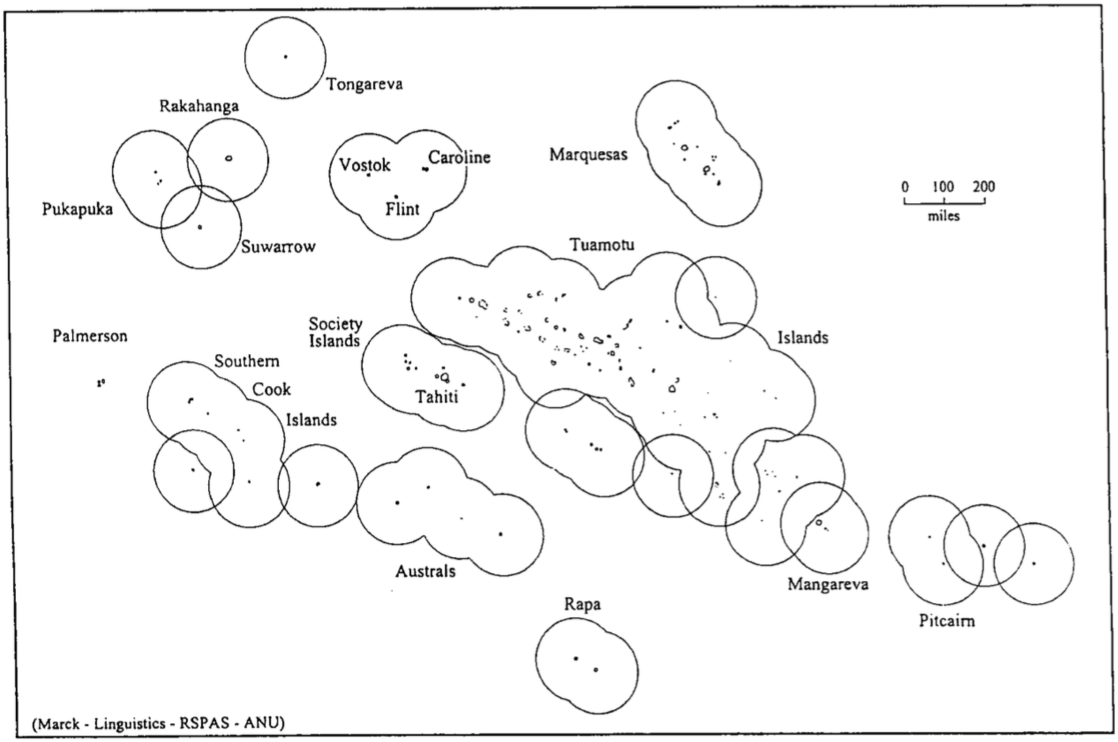

Fig. 8: Map of central East Polynesia from Marck (2000), showing 100 mile radii.

linguistic boundaries (with some exceptions, such as Tuamotus being split Australs merged etc). These island groups also line up somewhat with the "interaction spheres" in East Polynesia identified by Rolett (2002)[16].

However, this way of grouping islands does not take into account currents and winds - they are "just" a simple 100-mile distance straight from the coastline. Hedvig Skirgård & Ranacher (2023) improved upon this initial idea by incorporating a cost-surface that accounts for wind, currents and canoe profile shapes. The canoe that was used to model the distances was a v-shaped outrigger, see Fig. 9. The results were largely similar, strengthening the assumption that 100 miles is a decent approximation. It is this latter island grouping that we will use in this study. We will refer to these island groups as *overnight distance groups*. Fig. 10 shows this grouping, along with the number of languages per group. Out of the 56 island groups of this kind that were possible to include in the analysis, 45 have only one language.

Grouping islands based on (at least one) shared language. If two islands share a language, they are classified as of the same island group. For example, all islands of the Tuamotu archipelago (where Tuamotuan [tuam1242] is spoken) are merged, even though they are grouped into several separate island groups by the overnight distance approach. Vanuatu, being a land of great language diversity, has a different group for almost every island. Another example is the Tungaru islands (western islands of modern-day Kiribati) and Tuvalu which actually share a language. Most people of the country of Tuvalu speak Tuvaluan, but there is one island, Nui, where Tungaru (Gilbertese [gilb1244]) is also spoken (Faaniu, 1983; Macdonald, 2020; Ager, 2020). We will call these island groups *shared language groups*. Fig 11 shows these groups, and their language counts. Out of the 65 island groups that were included in the modelling, 21 had only one language.

[16] Rolett & Diamond (2004) consider islands within 50 km of each other as a meaningful unit, but do not extrapolate on why this is.

Fig. 9: The painting “A Fishing Canoe off North Kona” by Herbert Kawainui Kāne. Illustration of an outrigger canoe with v-shaped hull. Copyright Herbert K. Kāne, LLC.

Finally, these island groups should be taken with a grain of salt. They are most likely better than using modern nation-states and they are more explicitly and principally defined than many other island groups in other studies which rely on conventional definitions of well-known island groups or consider each island separately. However, the task of finding definite meaningful social geographic units in Oceania prior to colonization is difficult. Most likely the parameters vary over the space, for example with access to different kinds of sailing technology or different customs. Furthermore, from the initial settlement of Remote Oceania approximately 3,000 years ago until Western colonisation, there must have been a large amount of variation. (Rolett, 2002) for example discusses the decline in inter-island group voyaging in East Polynesia after 1450 CE. That said, these two approaches offer two different ways of viewing Oceania as a socially interconnected space — one more granular than the other — and are the best available island groupings for historical research known to the author.

Each island group is associated with the languages spoken there, based on information in the various sources cited earlier for defining groups by a shared language. Naturally, some islands will have many languages assigned to them, most notably Santo and Malakula in Vanuatu with 24 and 33 languages respectively. Other languages are spread out over many islands and atolls. Tuamotuan [tuam1242] is for example assigned to 1,302 distinct landmasses (in this dataset the parts of atolls and reefs that are above water appear as several small landmasses). For shared language groups, these landmasses are all joined under the label ‘Tuamotu’. However, as defined by overnight sailing distances this archipelago is split into several different island groups each with one language. The fact that it is the same language is not accounted for. Gavin & Sibanda (2012) also arrange their data similarly.

In the scripts, overnight-sailing distances island groups are labelled “SBZR” (initials of authors of Hedvig Skirgård & Ranacher (2023)) and shared-language-groups “medium”.

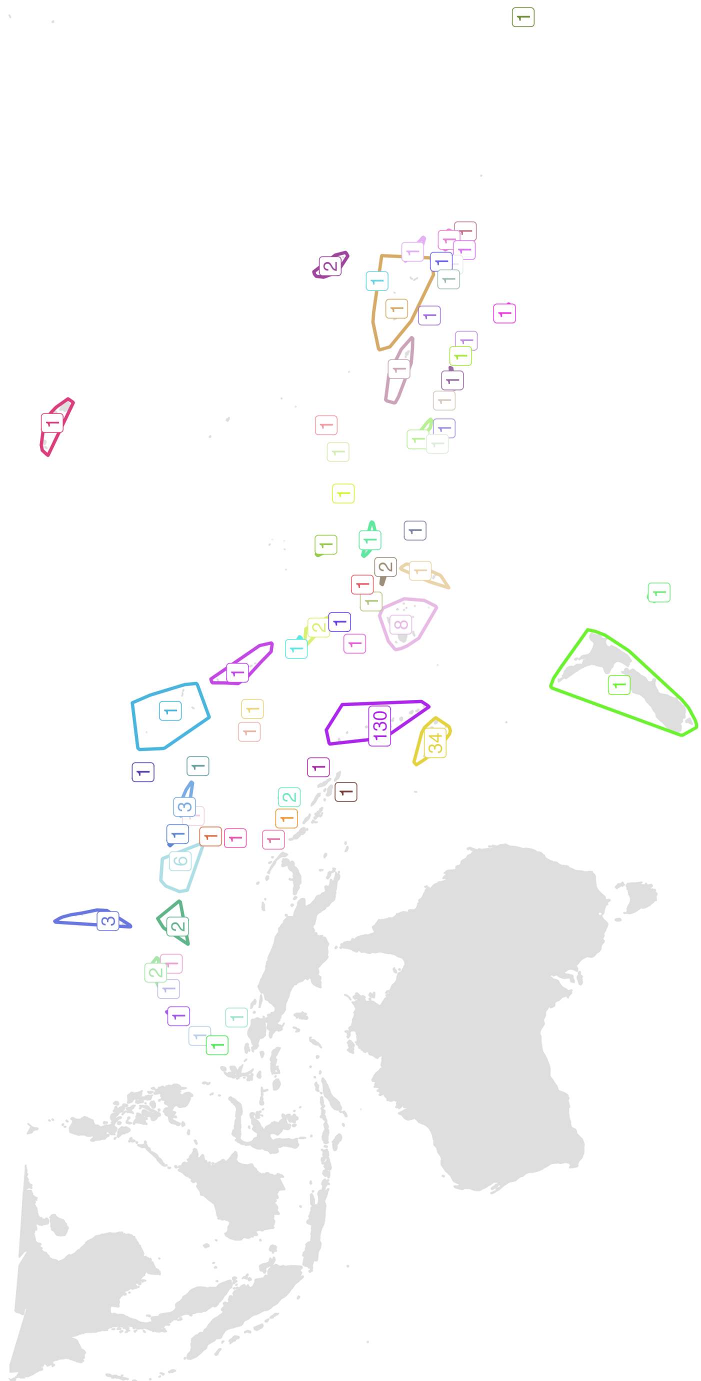

Fig. 10. Distribution of languages per overnight sailing distance island group.

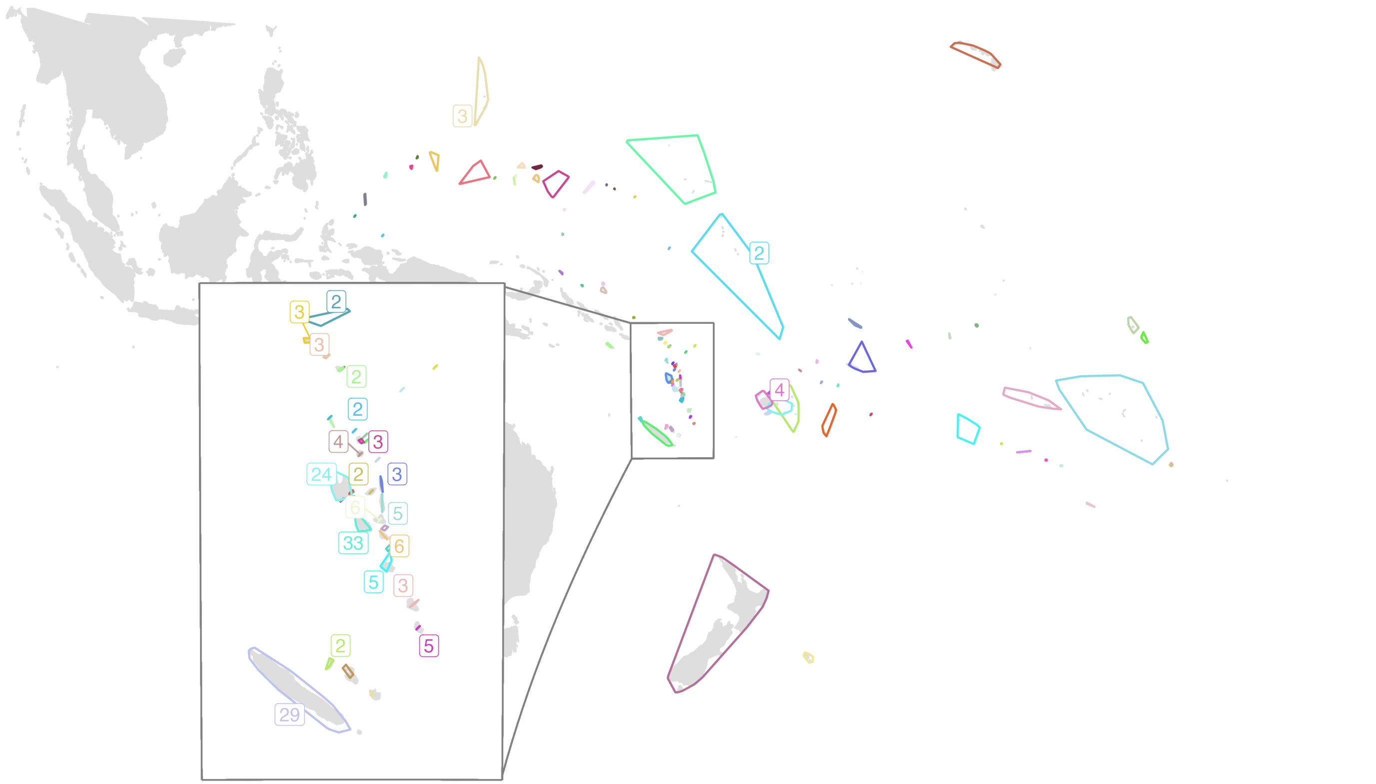

Fig. 11. Distribution of languages per shared language island group. Island groups with only one language do not have a label because otherwise the map is too difficult to read. All island groups without a label have 1 language.

"Political complexity"

The hypothesis tested in this study relies on a systematic measure of political structure across societies. The most widely used variable in this regard is variable 33 of the Ethnographic Atlas (Murdock, 1962): "jurisdictional hierarchy beyond local community" (EA033). This section gives an overview of how this variable is defined and its distribution in Remote Oceania.

The Ethnographic Atlas was first published in 1962 and is one of the data-sets included in D-PLACE (Kirby et al., 2018). The variable EA033 is widely used to represent "political complexity" in studies of cultural evolution. This variable has been used to study the relationships between language area, subsistence strategies and political organisation (Currie & Mace, 2009); processes of rise and fall of "political complexity" in South-east Asia and the Pacific (Currie et al., 2010); the spread of Christianity in the Pacific (Watts et al., 2018); and co-evolution of intensive use of natural resources and political structure (Sheehan et al., 2018). In this study, we will also refer to this variable as "political complexity".

In all of the studies referenced above, EA033 is used as a way of quantifying vertical political structure within a given society beyond the local community. "Local community" is in the ethnographic/anthropological literature defined as the "maximal group of persons who normally reside together in face-to-face association" (Murdock et al., 1945). The size and distribution of local community may vary greatly. For Remote Oceania, the Ethnographic Atlas records entries ranging from less than 50 (Ponape) to between 400-1000 (Tikopia).

"Society" is not defined explicitly in most of the literature, but it is sometimes used interchangeably with "ethnic group" or "culture". Keesing & Strathern (1981) write:

> *[A]ll the communities that are connected politically and economically (and hence comprise a kind of total social system) can be taken as comprising a* ***society****. Characteristically, a society comprises a total social system whose members share a common language and cultural tradition.*
>
> (Keesing & Strathern, 1981, 22) [17]

It is not possible to give more detail to the definition of "society" as employed in the Ethnographic Atlas dataset, but it does seem to correspond to "language community" to a great extent. In cases where societies are multilingual (c.f. Evans (2017)), it is less likely that the definition relies on at least one shared language even if not all are shared. It is, however, unclear how such cases are represented in the Ethnographic Dataset.

In some instances, it may be that "local community" and "society" are one and the same. For example, the total population for the "ethnic group" (variable EA202) of the society Tikopia is 1,300 which is not that much larger than the mean size of the local community (EA031) noted above. On the other hand, the total population of the ethnic group of Ponape is 8,000, which would result in an estimate of 160 local communities given the value of mean size of local community (EA031).

Political complexity (EA033) as a variable is observed per society, i.e. it does *not* capture relationships *between* distinct societies. It also does not track the political structure *within* the local community[18] or *horizontal* political relationships within a society (e.g. political relationships between equal genealogical groups). It specifically measures *vertical* jurisdictional levels *within* a given society, *beyond* the local community. Furthermore, it targets "jurisdictional" authority — not just the existence of rank. In other words, the levels of authority need to be tied to some kind of jurisdictional power, most likely the power to exert punishment for transgressions and the possibility to declare war upon other societies.

Specifically, it is necessary to underline that despite the common name for this variable — "Political complexity" — it does not include politically complex systems such as the Kula exchange ringDamon (2002), see Fig 12. Exchange systems of trade/debt are definitely complex and political, but they are not included in EA033 as they do not track *vertical* complexity but rather horizontal.

[17] Ironically, later in this section Keesing & Strathern (1981, 23) quote Schwartz (1978, 422) in saying that the "Manus people" as a society are easily linguistically bounded, when linguists have counted up to 19 languages on the island (Hammarström et al., 2019). In D-PLACE (Kirby et al., 2018) the society "Manus" in the Ethnographic Atlas is specifically linked to one language out of these 19, Titan [tita1241].

[18] Ethnographic Atlas variable 32 captures jurisdictional hierarchies *within* local communities.

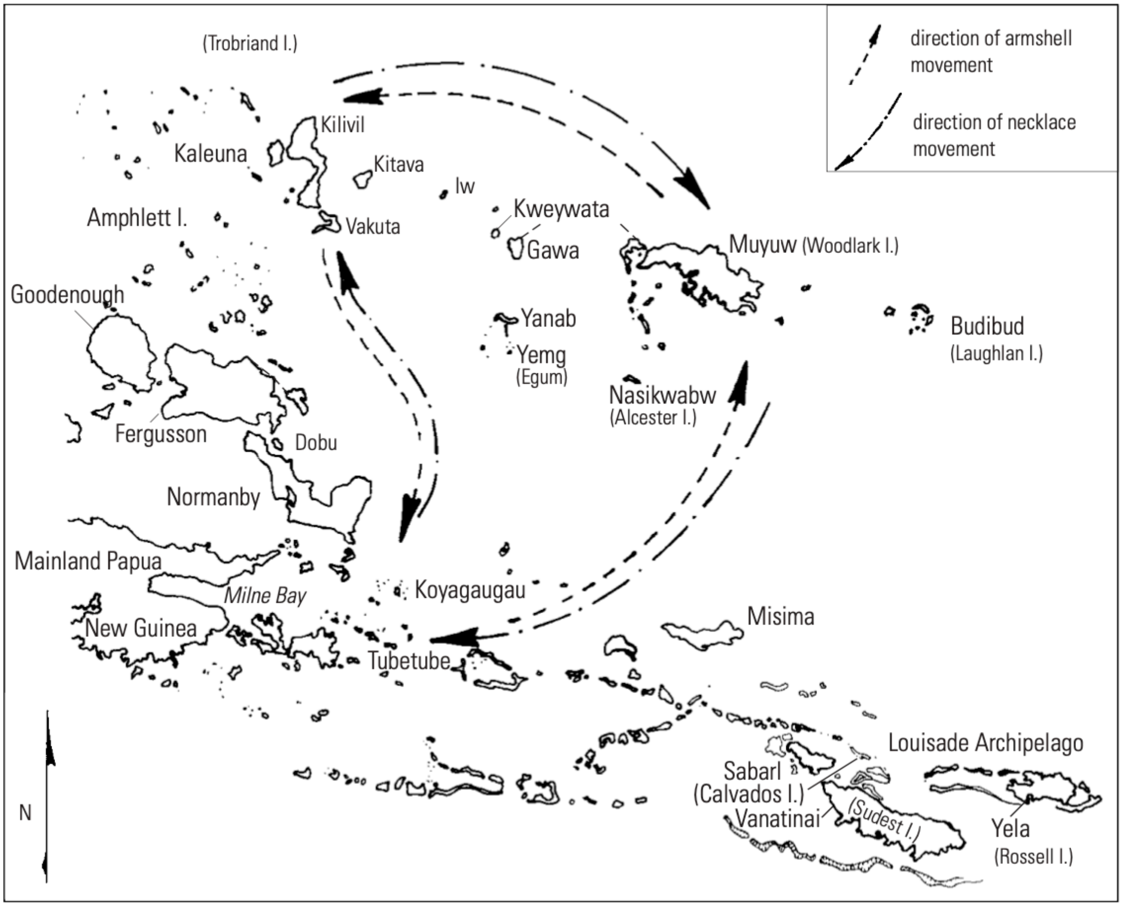

Fig. 12: Kula exchange ring, from Damon (2002).

Complex socio-political horizontal networks can also be found in Vanuatu, for example on the island of Tanna. Lindstrom (1991, 314) even remarks on the relationship with cultural diversity, writing: “Kava-drinking grounds across the island are linked by a complex system of ”roads” along which men exchange messages, goods, and spouses. This road network, by which each Tannese village is linked to all others, has produced cultural homogeneity across the island, despite linguistic diversity.”

When interpreting the effects of a variable like EA033, it is necessary to keep the above caveats in mind.

The possible values societies can take for political complexity (EA033) are:

1. No levels (no political authority beyond community)
2. One level (e.g., petty chiefdoms)
3. Two levels (e.g., larger chiefdoms)
4. Three levels (e.g., states)
5. Four levels (e.g., large states)

Giuliano & Nunn (2018) write:

> *[I]f the local village chief is the highest level of authority, and he or she does not answer to anyone above them, then the variable would take on a value of [1]. If above the chief, there was a district leader, then above*

this, there was a territory leader, and above this a provincial leader, and above this the paramount chief, then this variable would take on the value of [5].

Giuliano & Nunn (2018, 9)

This way of characterising political structure aligns well with ethnographic accounts of political structures in Western Polynesia, as we have for example seen in descriptions by Sahlins (1963). Consider Fig. 13 (Meleiseā, 1995, 22) for example, which illustrates the ideal characterisation of political power in Sāmoa. The village level (*matai* titles) corresponds to value 1 for EA033, subdistrict (*ali'i* titles) level 2, district level (*ali'i pa'ia / ao* titles) 3 and nation (*tafa'ifā / tupu* titles) corresponds to level 4. Sāmoa is however coded as level 3 since paramount titles are very unstable in the archipelago, even though they theoretically do exist.

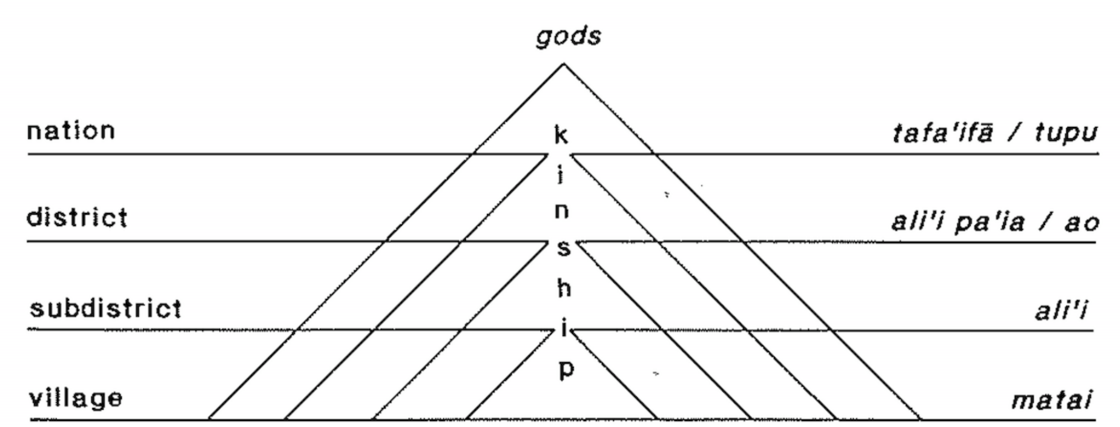

Fig. 13: Illustration of the traditional political organisation of Sāmoa from Meleiseā (1995, 22).

However, this framework is less easy to apply to the Melanesian islands of Remote Oceania where political structure often is described more in terms of autonomous and equal groups organised into horizontal structures such as exchange networks — both within and between societies (c.f. Bonnemaison (1996b) and Huffman (1996)). Bolton (1998) points out that the very concept of "chief" is new to Vanuatu, writing that chiefs were in many places first introduced by Europeans to fill the role of "representative for a community towards outsiders" and then transformed into "representative for traditional culture (kastom)" (Bolton, 1998, 185). In this study, political complexity could also be a proxy measurement of social network cohesiveness (c.f. Grace (1992)). Differences between modern Vanuatu "chiefs" and pre-European political structures are therefore unlikely to be large enough to give rise to dramatically different scores in this metric. As has been shown in previous studies, while this metric may be coarse-grained and not capture all the nuances of political life in the region, it appears to still be useful in testing hypotheses such as the one investigated in this study.

Given our language classification earlier, there are 233 language communities in our maximal sample. The ethnographic data is collected per society, and as was mentioned earlier it is possible for different societies to have the same language. Societies with the same language have been merged and our smallest cultural units of analysis are the language-society[19]. Each language is associated with the islands and atolls where it is described as spoken. The full geographical dataset contains 5,525 unique landmass polygons[20] which have been grouped

[19] American Sāmoans and Upolu Sāmoans were the only societies that needed merging in this fashion.

[20] The total number of polygons in the geographical dataset is 9,750, but this includes New Guinea and uninhabited islands (Phoenix Islands for example). 5,525 is the number of unique polygons that can be associated with a community of Remote Oceania.

into our smallest geographical unit of analysis: 104 shared language island groups. In the analysis, we will also be aggregating the islands into 67 overnight distance groups.

D-PLACE has a value recorded for political complexity for 44 “language-societies” in Remote Oceania. There have also been more recent studies where other authors have added more entries. We will be using additional data-points from a separate publication by Sheehan et al. (2018), resulting in another 27 data-points. There was an overlap between these two datasets, 26 language-socities having a coding both in the original Ethnographic Atlas dataset in D-PLACE and in Sheehan et al. Of these, 11 societies (i.e. 43%) were coded exactly the same. Upon closer investigation of the instances where the coding differed, I found that Sheehan et al often had access to *more* data, more *comprehensive* data and more *recent* data. Each of these instances was re-evaluated, and most often the coding from Sheehan et al was kept[21]. The data was also further supplemented with information from Bonnemaison (1996b, 201), resulting in 66 more data-points in Vanuatu. See appendix E for a table of the coding of this variable.

It should be noted that this data does not represent the state of these societies for the entire period between first settlement and European arrival. As Meleiseā (1995) writes, anthropology of the Pacific often depicts an unrealistic “ethnographic present”. For example, Schoeffel (1987, 185) writes about the shift in the political world of Sāmoa under the rule of Salamāsina which gave rise to a greater distinction between two kinds of chiefs: *tulāfale* and *ali'i*. Kirch (2017, 249) also mentions this distinction but notes that it mainly came to prominence in the 1500's. As a way of avoiding this problem, the D-PLACE database (Kirby et al., 2018) contains information on the focal years of the description. The data for the variable on political complexity in Remote Oceania varies from 1800 (Hawai'i) and 1940 (Vanua Levu). This study assumes that also more recent data may sufficiently accurately reflect the state of past societies, such that it remains useful for testing the central hypothesis.

Fig. 14 represents the coding of political complexity for the relevant societies. A more detailed table can be found in appendix E. While it is indeed the case that societies of Melanesian Remote Oceania tend to have lower levels of political complexity, it should be noted that many societies in Polynesia and Micronesia also score a value of 1.

For the calculations of this study, the mode political complexity per island group was used.

[21] I am grateful to Sheehan for very helpful personal correspondence during this process.

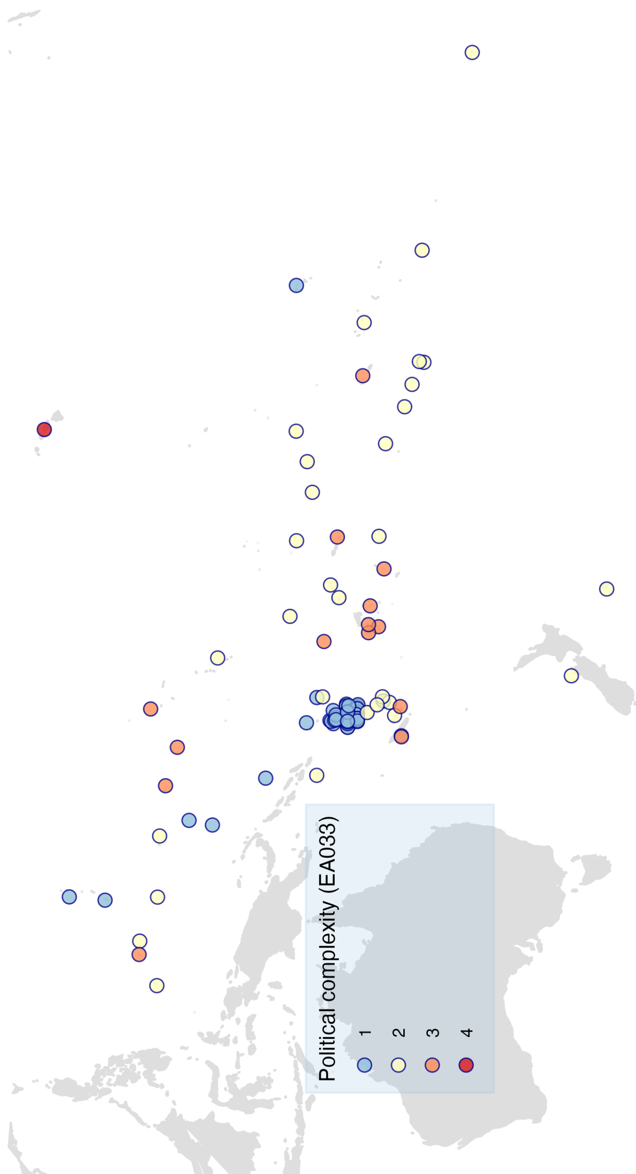

Fig. 14. Map of Remote Oceania with associated values for political complexity (EA033).

Waves of settlement

In order to test whether political complexity is driving language diversification in Remote Oceania, we need to control for other relevant variables — in particular time depth of human settlement. The time of settlement indicates how long a community is likely to have been in a certain place.

For most islands in our dataset, there is at least one archaeological study indicating time of first settlement. The archaeological data here is mainly based on an overview of the literature by Rieth & Cochrane (2018), supplemented by the following studies: Intoh & Ono (2007); Intoh (2008); Carson (2012); Kirch (2012); Napolitano et al. (2019); Ellis (2012) and Levin et al. (2019). See appendix F for a table of dating data.

For most of the island groups, the labels provided in Rieth & Cochrane (2018) or the other sources neatly corresponded to the island groups in our data (e.g. Mangareva = Mangareva). However, sometimes the label refers to a larger area, this is the case of "Austral Islands" which in our data-set is broken down into Rimatara, Tupuai, Ra'ivavae and Rurutu respectively. In such cases it is assumed that the time can be generalised over the smaller island groups (this is also the case for the political complexity metric). Furthermore, some island groups have not been subject to archaeological research, but they are known to have been settled in association with another place that has been studied. For example, while there have been no archaeological excavations on the Sorol atoll of Micronesia, it is known to be closely related to Ulithi (Quackenbush, 1968, 23) and it is likely that it was settled at a similar time. In such cases, the time depth is inferred based on information from nearby islands. For every island, the table in the appendix F lists the label of the island group in the source and whether the settlement timing has been inferred based on evidence from nearby islands or not.

Archaeological dates are based on radiocarbon findings which are calibrated based on conditions at the excavation site and other finds there. There exist different calibration methods[22]. In order to make the dates from different sites directly comparable they need to be re-calibrated in the same way. It was not possible to access all the necessary information on each publication and re-calibrate the dates appropriately. Instead of using the raw dates from the different studies, the islands have been sorted into a settlement order with twelve waves based on the dates in the literature and descriptions of what was settled in the same wave. Fig. 15 illustrates the settlement wave order per island. Exact dates and references are found in appendix F.

For comparison across island groups, the oldest settlement order per group was used. Unfortunately, this means that settlement data for so called Polynesian outlier communities in areas that were already inhabited (c.f. Southern Vanuatu), who arrived much more recently, is ignored.

[22] Radiocarbon calibration is the process by which radiocarbon years are converted into calendar years. Because the ratio of atmospheric $^{14}C/^{12}C$, which is a key element in this process, has not been stable historically, different methods exist and these may produce different results.

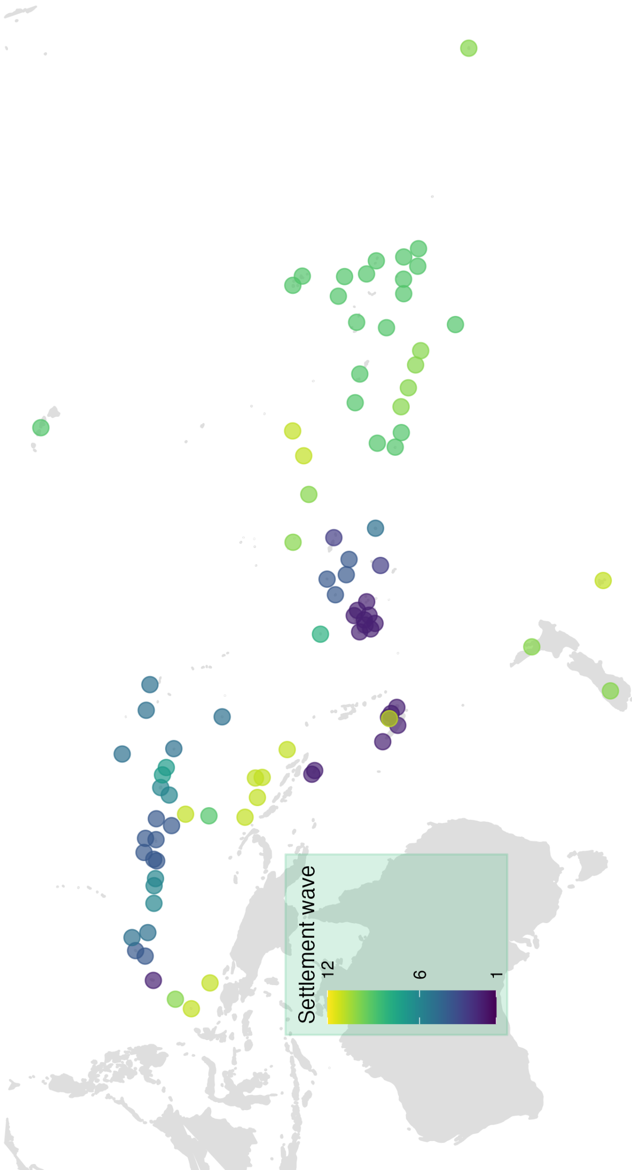

Fig. 15. Map of island groups in Remote Oceania labelled with settlement order.

Environmental factors

As previous studies of language diversification have shown (e.g. Greenhill (2015); Gavin et al. (2017); Coelho et al. (2019); Hua et al. (2019)), environmental factors such as latitude, rainfall etc can be useful in estimating language richness. This is due to the connection between these variables and the ability to sustain several self-sufficient smaller groups.

We are using measurements of absolute latitude, Net Primary Production, Precipitation (rainfall), temperature and size of islands to estimate the environments ability to sustain small groups. The data comes from Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG) (Wessel & Smith, 1996), the ecoClimate database (Lima-Ribeiro et al., 2015) and the Moderate Resolution Imaging Spectroradiometer (MODIS, Running & Zhao (2021b,a)). The coordinates of the centroids of each land mass from GSHHG was used to extract the information from ecoClimate and MODIS. The ecoClimate data used are simulations of the pre-industrial period (∼1760).

The Net Primary Production (NPP) data from MODIS is from 2001, 2002, 2003 and 2004 (these are the oldest available dates). NPP reflects the amount of carbon captured by plants in an ecosystem, after accounting for losses due to respiration. The data is gathered by two space satellites (Terra and Aqua) of the United States of America's National Aeronautics and Space Administration (NASA). They collect spectral indices of land-based vegetation, which are then used to infer NPP which is accumulated into a grid where each cell is 500 meters by 500 meters. For more details, see Running & Zhao (2015) [23].

In total, the following environmental variables are included in the study:

- Global Self-consistent, Hierarchical, High-resolution Geography Database (GSHHG)
  - Absolute latitude
  - Shoreline
- ecoClimate
  - Bio1: Annual mean temperature (Celsius)
  - Bio4: Temperature seasonality (standard deviation *100)
  - Bio12: Annual precipitation (mm/m2)
  - Bio15: Precipitation seasonality (coefficient of variation)
- Moderate Resolution Imaging Spectroradiometer (MODIS)
  - MOD17A3HGF.061: NPP data from the satellite Terra
  - MYD17A3HGF.061: NPP data from the satellite Aqua

The variables Net Primary Production (NPP; the amount of biomass or carbon produced), temperature & rainfall (mean and seasonality) and absolute latitude are all associated with the suitability of a place for life and they correlate with each other, at least to some degree (c.f. Figs 25 and 24). In this study, they are all tracking a similar concept: the suitability for human populations to thrive in smaller groups. Because of this, we are running a Principal Components Analysis on all seven variables, this allows us to extract a variable that explains most of the information. We ran a Non-Graphical Cattel's Scree Test (Cattell, 1966; Raiche & Magis, 2022) to determine the number of optimal components (overnight sailing distance island groups = 2 PCs, shared language island groups = 3 PCs).

Caveat regarding time-span

To understand the vast majority of human history and diversification, we should consider the state of societies as they were before the introduction of Western colonisation, industrialisation, globalization etc as these have only been around in the last 500 years. Behaviourally modern humans emerged between 90,000 and 160,000 years ago

[23] The manual is for a previous release, where the grid cells were 1km by 1km. The more recent release is 500m by 500m.

(Powell et al. (2009): Marean et al. (2007)). The last 500 years represent between 0.6% and 0.3% of our history. If we consider only the time period of more sedentary living and agriculture, i.e. approximately the last 11,000 years (c.f. Kislev et al. (2006)), the proportion of time since Columbus landing in the Caribbean is 4.5%.

The aforementioned processes, while recent, have resulted in mass immigration/invasion (c.f. Marlow (2016)) to places like the Americas, South Africa, Australia and New Zealand and oppressive policies that reduce language diversity. They have had a massive impact in the comparatively short time span that they have been active and are not only perpetuated by Western forces but also other powerful groups (see for example the case of Indonesian power in Western New Guinea (Gietzelt, 1989)). Another driving force that is relatively recent and related is the emergence of modern nation-states and their ideologies (c.f. Foucault (2007), Oakes (2001, 21-22)). In connection with the aforementioned forces, the formation of modern nation-states has further reduced diversity that was present within groups (through means like ideological movements (e.g.the *Volkgeist*-movement), nationally standardized mass education, mass media, print-media etc). The creation of nation-states and their continued development is often tied to ideas of one sole homogeneous culture (c.f. Miscevic (2020)), which results in the disappearance of regional variation, minority languages etc within what we today conceive of as “countries”/“nation-states”.

The massive effect of these forces in the last 500 years means that for historical scientists, there is a big difference between considering the human landscape before and after this point — especially in terms of cultural history and diversity. Currie & Mace (2009, 7340-7341) find that the data available for language territories in Australia and the Americas is so affected by colonial history and population replacement to make it impossible to include it in the analysis alongside the other regions.

There is also a problem with information pertaining to history before these massive forces step onto the stage. Ethnographic accounts are often “snapshots” with unclear time-depth (Bedford & Spriggs, 2008, 113). Human culture everywhere is dynamic and changing. Meleiseā (1995) has also written about how anthropology of the Pacific often depicts an unrealistic “ethnographic present”. This is one of the reasons why the ethnographic database D-PLACE (Kirby et al., 2018) includes “focal year” for each variable.

Through studies like this, we can model and discuss what the past may have been like and what could be driving certain changes, such as language splitting, but we must always be wary of the shortcomings of our data and methods. While the coding of political complexity and environmental data may not perfectly represent all the time between the settlement of Remote Oceania 3,000 years ago until now, it may be a decent proxy of large patterns and conclusions drawn may still be informative. For a more detailed understanding, deeper engagement with local communities and agent-based modelling is desirable.

Regarding population and the number of languages

While it is true that the number of languages is not entirely evenly spread out over the population, it is still the case that there is a general trend connecting a greater number of people with a greater number of languages (as can be seen in Figs 16 and 17). In addition, Raviv et al. (2019) and other studies have shown that community size may influence language diversification independently from network structure.

However, due to difficulties with data, this is hard to study historically. In this particular study, population numbers could not be included due to a lack of detailed and reliable data on enough island groups prior to Western colonization. While resources like of Hawaii at Manoa (2023) have population numbers for many places, it does not sufficiently cover Oceania to allow the variable to be included. This is similar to Currie & Mace (2009, 7340-7341) who found that the data available for language territories in Australia and the Americas is so affected by colonial history and population replacement to make it impossible to include it in the analysis alongside the other regions.

Figs 16 and 17 show the relationship between the number of languages per modern country/territory (Hammarström et al., 2017) and its population in 1950 (United Nations, Department of Economic and Social Affairs, Population Division, 2022), with Remote Oceanic countries and territories highlighted. Modern nation-states are not historically relevant cultural units in many places in the world, include the Pacific. For example, Cook Islands (also known as Avaiki Nui) and the Federated States of Micronesia encompasses many island groups with diverse histories (c.f. Fig 1 and Appendix B.3). However, even with this noise, there is a relationship — more people → more languages.

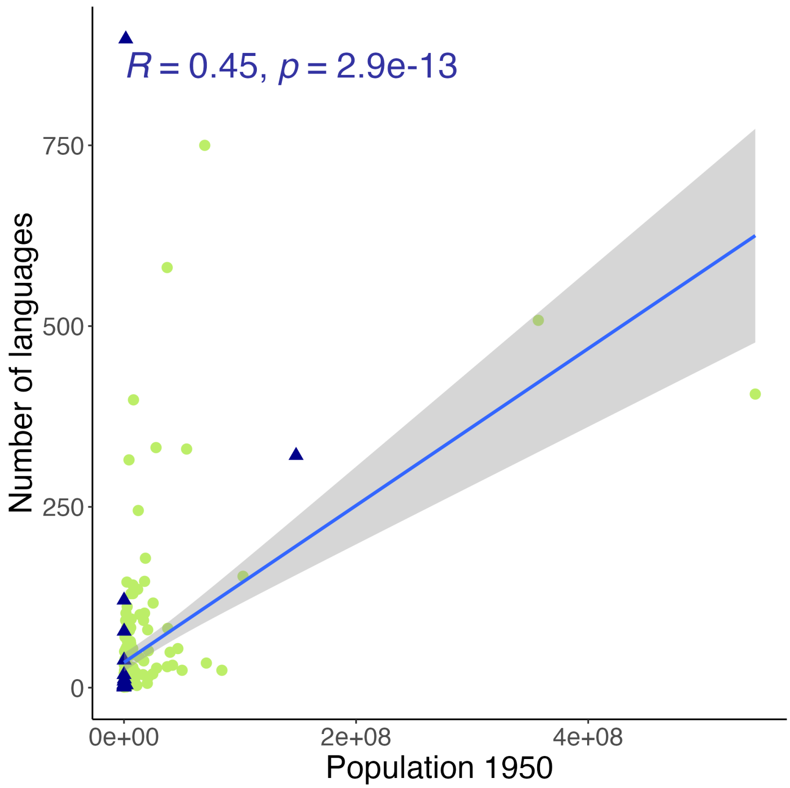

Fig. 16: Population in 1950 per country (x-axis) and number of languages per country (y-axis). Dark blue triangles = countries with at least some landmasses in Remote Oceania, green circles = all other countries.

While it is not possible to include historic population size as a variable, it is possible that the environmental variables are relevant here as they may suggest island groups that could be home to larger populations. Fig. 18 shows the relationship between the population data and language count data above to the variables of this study. The first four rows represent the data summarised per country (i.e. from United Nations, Department of Economic and Social Affairs, Population Division (2022) and Hammarström et al. (2017)) and the rest are per island group in Remote Oceania. This means there is a mismatch of units of analysis, for example: Hawai'i is a part of the United States of America and therefore is included in the data point for the entire USA. The "lg count (island group)" contains only the number of languages in Hawai'i in the island groups included in this study, i.e. 1.

There is a negative correlation of -0.56 between the population (log10) per country and the environmental PC1 per island group. This may indicate that the environmental PCs are picking up on information relevant to population size.

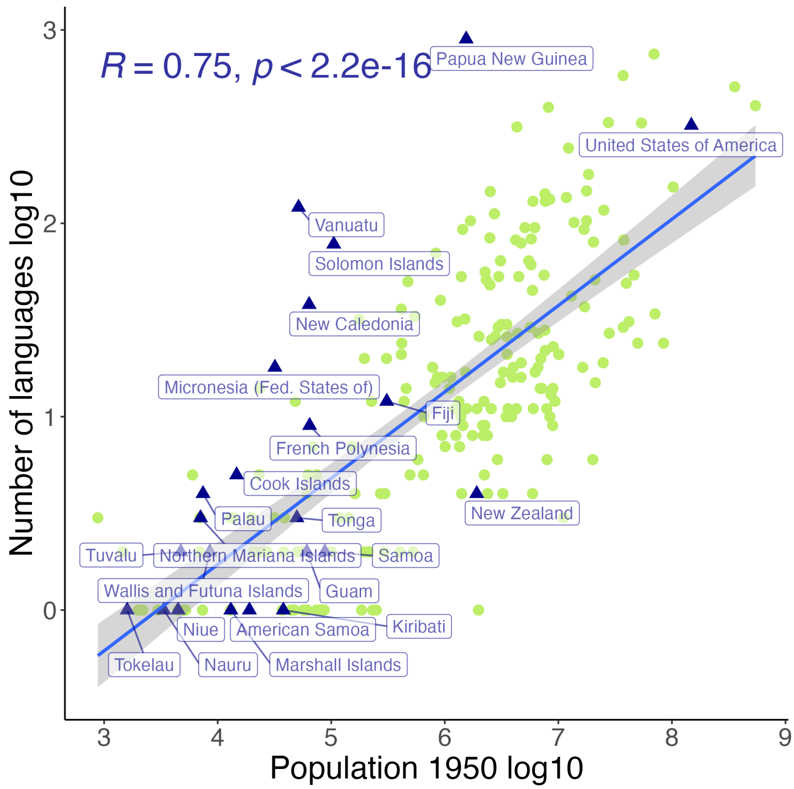

Fig. 17: Population in 1950 per country (x-axis) and number of languages per country (y-axis) - log-10-transformed. Dark blue triangles = countries with at least some landmasses in Remote Oceania, green circles = all other countries.

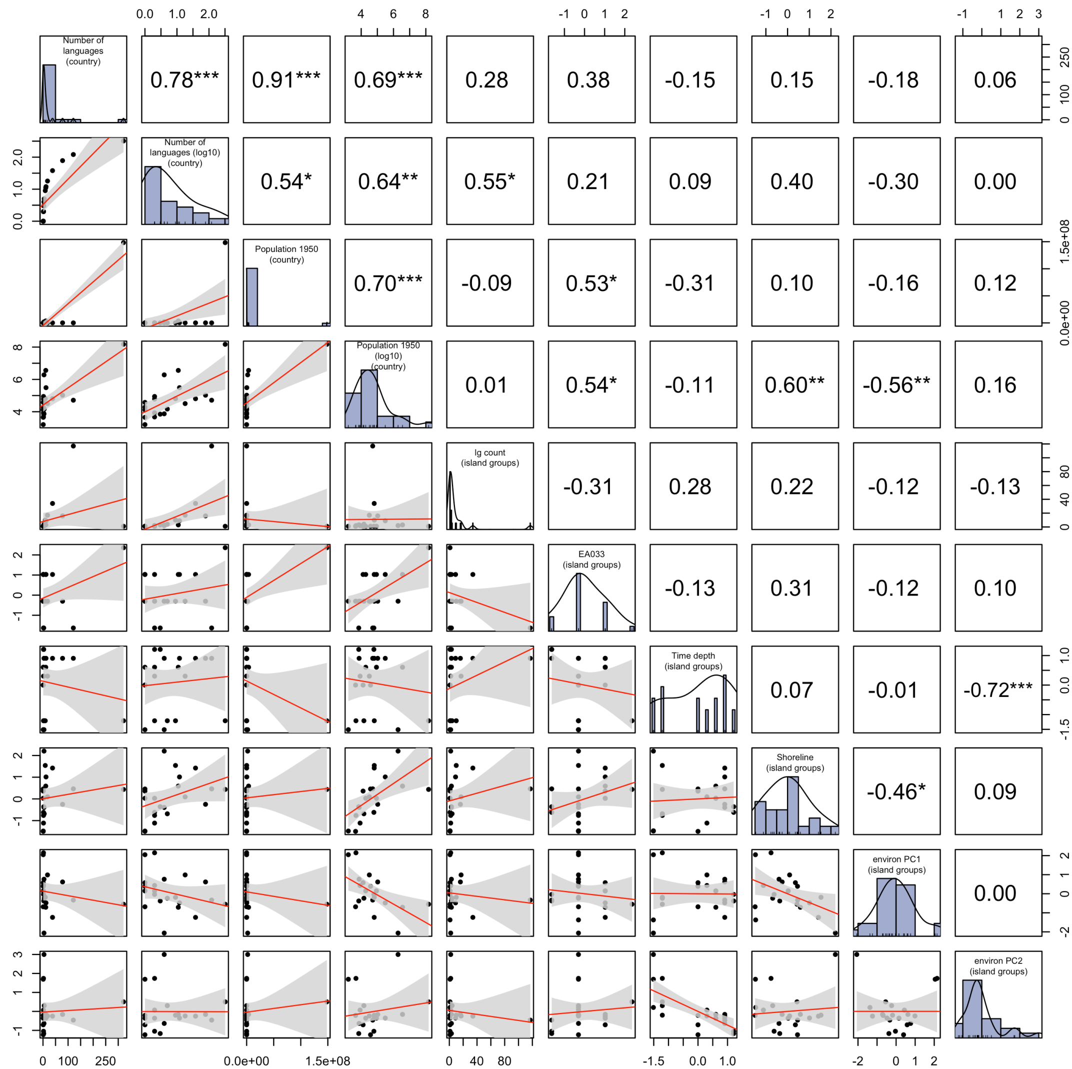

Fig. 18: Scatterplot matrix of all variables available, over countries. Note that the population data is taken from United Nations, Department of Economic and Social Affairs, Population Division (2022) and refers to the whole country, i.e. the population of the entire USA — not just Hawai'i..

Table of political complexity per island group

All code and data will also be available on OSF/GitHub/Zenodo.

Anonymised OSF project for double-blind peer review: `https://osf.io/amnh9/?view_only=403d57d715254539a3a99c798b4`

Island group (overnight-sailing)	Island group (shared language)	Political complexity (EA033)	glottocodes	References
Aotearoa	Aotearoa	2	maor1246	Sahlins (1958), Buck (1952), Kirch (1984), Van Meijl (1995)
Banaba	Tungaru + Tuvalu	2	gilb1244	Lambert (1966), Lambert (1975), Lambert (1991), Macdonald (1982)
Bellona + Rennelle	Mo-ava-mo-iki	2	renn1242	Birket-Smith (1956), Monberg (1991)
Chuuk	Chuuk	2	chuu1238	Goodenough (1991), Goodenough (2002), Mahony (1960)
East Futuna	East Futuna	2	east2447	Kirch (1994)
Fiji	Kadavu	3	kada1285	Kuhlken (2002), Scarr (1984), Walter (1978)
Fiji	Lau	3	laua1243	Hocart (1929), Quain (1948), Thompson (1940a), Thompson (1940b)
Fiji	Viti Levu + Yasawa	3	sout2864, nort2843	Kuhlken (2002), Scarr (1984), Walter (1978)
Hawaii	Hawaii	4	hawa1245	Kirch (1994), Kirch (2010)
Hereheretue	Tuamotu	2	tuam1242	Emory (1975)
Kanaky	Kanaky (New Caledonia main island)	1	ajie1238	Winslow (1991)
Kanaky	Kanaky (New Caledonia main island)	3	xara1244	Young (1991)
Kanaky	Lifou	2	dehu1237	Hadfield (1920), Ray (1917)
Kanaky	Nengone	3	neng1238	Dubois (1984), Guiart (1952)
Kapingamarangi	Kapingamarangi	1	kapi1249	Buck (1950)
Kosrae	Kosrae	3	kosr1238	Athens (2007), Graves (1986), Peoples (1991)

Laguas yan gåni	Laguas yan gåni	1	cham1312, caro1242	Cordy (1983), Thompson (1971[1945]), Joseph & Murray (1951), Spehr (1954)
Luangiua + Nukumanu	Luangiua	1	onto1237	Sahlins (1958), Bayliss-Smith (1974), Donner (1991)
Mangaia	Avaiki Nui (south)	2	raro1241	Bellwood (1971), Buck (1934), Crocombe (1967), Hayes (1981), Walter (1996)
Mangareva	Mangareva	2	mang1401	Buck (1971[1938]), Conte & Kirch (2004), Green & Weisler (2000)
Marutea	Tuamotu	2	tuam1242	Emory (1975)
Morane	Tuamotu	2	tuam1242	Emory (1975)
Māngarongaro	Māngarongaro	2	penr1237	Buck (1932b)
Napuka	Tuamotu	2	tuam1242	Emory (1975)
Ngatik	Pohnpei	3	pohn1238	Hanlon (2019), Haun (1984), Raynor & Fownes (1991), Riesenberg (1968)
Ngā Pū Toru	Avaiki Nui (south)	2	raro1241	Bellwood (1971), Buck (1934), Crocombe (1967), Hayes (1981), Walter (1996)
Niu + Tuvalu	Tungaru + Tuvalu	2	gilb1244, tuva1244	Lambert (1966), Lambert (1975), Lambert (1991), Macdonald (1982), Goldsmith (1991)
Niue	Niue	2	niue1239	Loeb (1978), Smith et al. (1983[1902-1903]), Walter & Anderson (1995)
Nukuoro	Nukuoro	1	nuku1260	Carroll (1966), Carroll (1975), Eilers (1934)
Nukutaveke 59	Tuamotu	2	tuam1242	Emory (1975)
Nukutaveke 61	Tuamotu	2	tuam1242	Emory (1975)
Palau	Palau	2	pala1344	Force (1960)
Pohnpei	Pohnpei	3	pohn1238	Hanlon (2019), Haun (1984), Raynor & Fownes (1991), Riesenberg (1968)

Pukapuka	Pukapuka	2	puka1242	Beaglehole & Beaglehole. (1938), Macgregor (1935)
Raivavae	Raivavae	2	raiv1237	Aitken (1971[1930]), Bollt (2008), Edwards (2003)
Rakahanga-Manihiki	Rakahanga-Manihiki	2	raka1237	Buck (1932a)
Rapa Nui	Rapa Nui	2	rapa1244	Sahlins (1958), Kirch (1984), Métraux (1971)
Raro Matai + Nia Matai	Tahiti	3	tahi1242	Oliver (2019)
Rarotonga	Avaiki Nui (south)	2	raro1241	Bellwood (1971), Buck (1934), Crocombe (1967), Hayes (1981), Walter (1996)
Ratak + Rālik	Ratak + Rālik	3	mars1254	Carucci (1991), Erdland & Neuse (1961[1914]), Williamson (1982)
Rotuma	Rotuma	3	rotu1241	Gardiner (1898), Howard (1963), Howard (1991)
Rurutu	Rurutu	2	ruru1237	Aitken (1971[1930]), Bollt (2008), Edwards (2003)
Rēkohou	Rēkohou	2	mori1267	Sahlins (1958), Buck (1952), Kirch (1984), Van Meijl (1995)
Sorol	Ulithi (greater)	2	ulit1238	Lessa (1950), Lessa (1966)
Sāmoa	Sāmoa	3	samo1305	Sahlins (1958), Buck (1930), Keesing (1934), Watters (1958)
Tatakoto + Reao	Tuamotu	2	tuam1242	Emory (1975)
Te Henua ‘Enana + Te Fenua ’Enata	Te Henua ‘Enana	1	nort2845	Sahlins (1958)
Tokelau	Tokelau	2	toke1240	Hooper & Huntsman (1973), Macgregor (1937)
Tonga	Tonga	3	tong1325	Kirch (1984), Cummins (1977), Ferdon (1987)
Tuamotu + Puka-puka	Tuamotu	2	tuam1242	Emory (1975)

Tungaru	Tungaru + Tuvalu	2	gilb1244	Lambert (1966), Lambert (1975), Lambert (1991), Macdonald (1982)
Tupuai	Tupuai	2	tubu1240	Aitken (1971[1930]), Bollt (2008), Edwards (2003)
Tureia	Tuamotu	2	tuam1242	Emory (1975)
Tuvalu 31	Tungaru + Tuvalu	2	tuva1244	Macdonald (1982), Goldsmith (1991)
Tuvalu 32	Tungaru + Tuvalu	2	tuva1244	Macdonald (1982), Goldsmith (1991)
Ulithi + Yap	Ulithi (greater)	2	ulit1238	Lessa (1950), Lessa (1966)
Ulithi + Yap	Yap	3	yape1248	Hunt et al. (1949), Muller (1917), Murdock et al. (1944), Salesius (1907), Schneider (1953), Schneider (1957), Schneider (1962), Tetens (1958), Tetens & Kubary (1873)
Uvea (Wallis)	Uvea (Wallis)	2	wall1257	Burrows (1971[1937]), Pollock (1995)
Vanuatu + Temotu	Ambae	1	east2443, west2513	Bonnemaison (1972)
Vanuatu + Temotu	Ambrym	1	sout2859	Tonkinson (1981)
Vanuatu + Temotu	Aneityum	2	anei1239	Humphreys (1926), Spriggs (1982), Spriggs (1986)
Vanuatu + Temotu	Anuta	1	anut1237	Feinberg (1988), Feinberg (1991), Kirch (2002)
Vanuatu + Temotu	Aore	1	aore1237	Bonnemaison (1996a)
Vanuatu + Temotu	Duff + Reef Islands	1	ayiw1239	Davenport (1969)
Vanuatu + Temotu	Efate	2	nort2836	Facey (1981)
Vanuatu + Temotu	Erromango	2	siee1239	Humphreys (1926), Spriggs & Wickler (1989)
Vanuatu + Temotu	Futuna + Aniwa	2	futu1245	Capell (1958)
Vanuatu + Temotu	Hiu	1	hiww1237	Bonnemaison (1996a)
Vanuatu + Temotu	Loh-Toga	1	loto1240	Bonnemaison (1996a)

Vanuatu + Temotu	Maewo	1	baet1237, cent2058, mari1426	Bonnemaison (1996a)
Vanuatu + Temotu	Malakula	1	aulu1238, avok1244, axam1237, bign1238, burm1263, bwen1239, dixo1238, katb1237, labo1244, lare1249, lete1241, ling1265, litz1237, maee1241, malf1237, malu1245, mara1399, mask1242, mpot1241, nese1235, nisv1234, niti1249, port1285, rere1240, unua1237, urip1239, vinm1237, vivt1234, sout2857	Bonnemaison (1996a), Deacon & Wedgwood (1934)
Vanuatu + Temotu	Mota	1	mota1237	Bonnemaison (1996a)
Vanuatu + Temotu	Mota Lava	1	motl1237	Bonnemaison (1996a)
Vanuatu + Temotu	Paama	1	paam1238	Bonnemaison (1996a)
Vanuatu + Temotu	Pentecost	1	saaa1241	Lane (1956)
Vanuatu + Temotu	Santo	1	butm1237, polo1242, akei1237, ambl1237, fort1240, lore1244, mere1242, moro1286, noku1237, piam1242, rori1237, saka1289, shar1244, tamb1253, tasm1246, tial1239, tolo1255, valp1237, wail1242, wusi1237, tang1347	Bonnemaison (1996a)
Vanuatu + Temotu	Tanna	1	sout2869	Lindstrom (1991)
Vanuatu + Temotu	Tikopia	2	tiko1237	Kirch (1994), Sahlins (1958), Firth (1939), Firth (1959), Firth (1991)
Vanuatu + Temotu	Ureparapara	1	leha1243, leha1244	Bonnemaison (1996a)
Vanuatu + Temotu	Vanua Lava	1	leme1238, vera1241, vure1239	Bonnemaison (1996a)
Woleai	Woleai	2	wole1240	Alkire (1991), Burrows & Spiro. (1953)

Ānewetak	Ratak + Rālik	3	mars1254	Carucci (1991), Erdland & Neuse (1961[1914]), Williamson (1982)

Table 1: Table of political complexity values (EA033).

Table of settlement date per island group based on archaeology

All code and data will also be available on OSF/GitHub/Zenodo.

Anonymised OSF project for double-blind peer review: `https://osf.io/amnh9/?view_only=403d57d715254539a3a99c798b4`

Table 2 shows data on settlement dates from several different archaeological studies.

The column "Source" contains the relevant reference, and "Source (meta)" the citation for the source that is citing the reference in "Source", where that is useful. In cases where it was reasonable to infer a settlement date based on a nearby island, this is indicated in the column "Based on inference from neighbouring island?".

In the summary table in Rieth & Cochrane (2018), the Vanuatu islands Santa Cruz and Reef islands are mentioned specifically on a separate row to "Vanuatu". Therefore, we take "Vanuatu" to mean islands in Vanutu excluding Santa Cruz and Reef islands

In the case of the so-called "Polynesian Outliers", i.e. communities who speak languages that belong to the Polynesian linguistic group but are living in islands outside of the traditional "Polynesian triangle", where possible we used the settlement dates of the first human settlement even if this is different from the Polynesian outlier socities living there today. For example, Carson (2012) describes non-Polynesian horizons in Mo-ava-mo-iki (a.k.a Renelle and Bellona). For Taumako, Carson describes the initial Lapita settlement as 3,000 years ago and the Polynesian outlier settlement 1000 years ago.

The column "Time depth settlement group" contains the grouping of islands into settlement waves, which is what is used in the models. For each of the island groups, the oldest settlement group was used.

Island group (finest)	Date ranges	Island group (overnight-sailing)	Island group (shared language)	Time depth settlement group	Based on inference from neighbouring island?	Name in source	Source	Source (meta)
Guam	3500, 2950	Laguas yan gåni	Laguas yan gåni	1		Mariana Islands	Carson (2014) and Athens et al (2004)	Rieth & Cochrane (2018)
Saipan	3500, 2950	Laguas yan gåni	Laguas yan gåni	1		Mariana Islands	Carson (2014) and Athens et al (2004)	Rieth & Cochrane (2018)
Ambae	3200, 3000	Vanuatu + Temotu	Ambae	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Ambrym	3200, 3000	Vanuatu + Temotu	Ambrym	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Aneityum	3200, 3000	Vanuatu + Temotu	Aneityum	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Anuta	2950	Vanuatu + Temotu	Anuta	2		Anuta	Carson (2012)	
Aore	3200, 3000	Vanuatu + Temotu	Aore	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Araki	3200, 3000	Vanuatu + Temotu	Araki	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Belep	3050, 3000	Kanaky	Belep	2		New Caledonia	Sand (2001)	Rieth & Cochrane (2018)
Bellona	3100, 2900	Bellona + Rennelle	Mo-ava-mo-iki	2		Bellona	Carson (2012)	
Duff Islands	3000	Vanuatu + Temotu	Duff + Reef Islands	2		Taumako	Leach and Davidson (2008)	Carson (2012)

Efate	3200, 3000	Vanuatu + Temotu	Efate	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Epi	3200, 3000	Vanuatu + Temotu	Epi	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Erromango	3200, 3000	Vanuatu + Temotu	Erromango	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Etarik	3200, 3000	Vanuatu + Temotu	Efate	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Gaua	3200, 3000	Vanuatu + Temotu	Gaua	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Hiu	3200, 3000	Vanuatu + Temotu	Hiu	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Kadavu	3130, 2870	Fiji	Kadavu	2		Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Kanaky (New Caledonia main island)	3050, 3000	Kanaky	Kanaky (New Caledonia main island)	2		New Caledonia	Sand (2001)	Rieth & Cochrane (2018)
Lau	3130, 2870	Fiji	Lau	2		Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)

Leveuka (Lomaviti)	3130, 2870	Fiji	Lomaviti	2	Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Lifou	3050, 3000	Kanaky	Lifou	2	New Caledonia	Sand (2001)	Rieth & Cochrane (2018)
Loh-Toga	3200, 3000	Vanuatu + Temotu	Loh-Toga	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Lomaviti	3130, 2870	Fiji	Lomaviti	2	Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Lomlom	3185, 2639	Vanuatu + Temotu	Duff + Reef Islands	2	Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)
Lopevi	3200, 3000	Vanuatu + Temotu	Paama	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Maewo	3200, 3000	Vanuatu + Temotu	Maewo	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Mafea	3200, 3000	Vanuatu + Temotu	Mafea	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Malakula	3200, 3000	Vanuatu + Temotu	Malakula	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)

Mataso	3200, 3000	Vanuatu + Temotu	Efate	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Merelava group	3200, 3000	Vanuatu + Temotu	Merelava group	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Mota	3200, 3000	Vanuatu + Temotu	Mota	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Mota Lava	3200, 3000	Vanuatu + Temotu	Mota Lava	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Nendö	3185, 2639	Vanuatu + Temotu	Nendö	2	Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)
Nengone	3050, 3000	Kanaky	Nengone	2	New Caledonia	Sand (2001)	Rieth & Cochrane (2018)
Ouvea	3050, 3000	Kanaky	Ouvea	2	New Caledonia	Sand (2001)	Rieth & Cochrane (2018)
Ouvea (Iaai)	3050, 3000	Kanaky	Ouvea	2	Ouvea	Sand (2001)	Rieth & Cochrane (2018)
Paama	3200, 3000	Vanuatu + Temotu	Paama	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Palau	3100, 2900	Palau	Palau	2	Palau	Athens and Ward (2001), Clark et al (2006) and Liston (2005, 2013)	Rieth & Cochrane (2018)

Pentecost	3200, 3000	Vanuatu + Temotu	Pentecost	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Reef Islands	3185, 2639	Vanuatu + Temotu	Duff + Reef Islands	2	Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)
Rennelle	3100, 2900	Bellona + Rennelle	Mo-ava-mo-iki	2	Rennelle	Carson (2012)	
Santo	3200, 3000	Vanuatu + Temotu	Santo	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Tamambo	3200, 3000	Vanuatu + Temotu	Tamambo	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Tangoa	3200, 3000	Vanuatu + Temotu	Santo	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Tanna	3200, 3000	Vanuatu + Temotu	Tanna	2	Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Taveuni	3130, 2870	Fiji	Vanua Levu	2	Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Te Anu	3185, 2639	Vanuatu + Temotu	Te Anu	2	Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)

Tegua	3185, 2639	Vanuatu + Temotu	Loh-Toga	2		Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)
Tikopia	3000	Vanuatu + Temotu	Tikopia	2		Tikopia	Carson (2012)	
Tongoa	3200, 3000	Vanuatu + Temotu	Tongoa	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Tutuba	3200, 3000	Vanuatu + Temotu	Tutuba	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Ureparapara	3200, 3000	Vanuatu + Temotu	Ureparapara	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Utupua	3185, 2639	Vanuatu + Temotu	Utupua	2		Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)
Vanikoro	3185, 2639	Vanuatu + Temotu	Vanikoro	2		Reef Islands and Santa Cruz Islands	Green (1991), Green et al (2008) and Sheppard et al (2015)	Rieth & Cochrane (2018)
Vanua Lava	3200, 3000	Vanuatu + Temotu	Vanua Lava	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Vanua Levu	3130, 2870	Fiji	Vanua Levu	2		Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)

Vao	3200, 3000	Vanuatu + Temotu	Vao	2		Vanuatu	Bedford et al (2006)	Rieth & Cochrane (2018)
Vatulele	3130, 2870	Fiji	Viti Levu + Yasawa	2		Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Viti Levu	3130, 2870	Fiji	Viti Levu + Yasawa	2		Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Yasawa (greater)	3130, 2870	Fiji	Viti Levu + Yasawa	2		Fiji	Hope et al (2009), Denham et al (2012) and Nunn and Petchey (2013)	Rieth & Cochrane (2018)
Sāmoa	2800, 2400	Sāmoa	Sāmoa	3		Samoa	Cochrane et al (2013), Kurch and Hunt (1993) and Petchey (2001)	Rieth & Cochrane (2018)
Tonga	2846, 2750	Tonga	Tonga	3		Tonga	Burley and Connaughton (2007) and Burley et al (1999, 2001, 2012)	Rieth & Cochrane (2018)
Chuuk	2300, 1750	Chuuk	Chuuk	4		Chuuk	Shutler (1984)	Rieth & Cochrane (2018)

East Futuna	2300, 2200	East Futuna	East Futuna	4		Futuna	Kirch (1981)	Rieth & Cochrane (2018)
Mortlock	2300, 1750	Chuuk	Mortlock (greater)	4	Yes	Chuuk	Shutler (1984)	Rieth & Cochrane (2018)
Namonuito	2300, 1750	Chuuk	Namonuito	4	Yes	Chuuk	Shutler (1984)	Rieth & Cochrane (2018)
Nguluwan	2400, 2100	Nguluwan	Nguluwan	4	Yes	Nguluwan	Napolitano et al. (2019)	
Niuafoou	2300, 2200	Niuafoou + Niuatoputapu	Niuafoou	4	Yes	Niuafoou	Kirch (1981)	Rieth & Cochrane (2018)
Niuatoputapu	2300, 2200	Niuafoou + Niuatoputapu	Niuatoputapu	4	Yes	Niuatoputapu	Kirch (1981)	Rieth & Cochrane (2018)
Oroluk	2300, 1750	Oroluk, Chuuk	Mortlock (greater)	4		Chuuk	Shutler (1984)	Rieth & Cochrane (2018)
Paafang	2300, 1750	Chuuk	Paafang	4	Yes	Chuuk	Shutler (1984)	Rieth & Cochrane (2018)
Pollap	2300, 1750	Chuuk	Pollap	4	Yes	Chuuk	Shutler (1984)	Rieth & Cochrane (2018)
Puluwat	2300, 1750	Chuuk	Puluwat	4	Yes	Chuuk	Shutler (1984)	Rieth & Cochrane (2018)
Uvea (Wallis)	2300, 2200	Uvea (Wallis)	Uvea (Wallis)	4	Yes	Futuna	Kirch (1981)	Rieth & Cochrane (2018)
Yap	2400, 2100	Ulithi + Yap	Yap	4		Yap	Napolitano et al. (2019)	
Banaba	2150, 1750	Banaba	Tungaru + Tuvalu	5	Yes	Tungaru	DiPazza (1999)	Rieth & Cochrane (2018)

Kosrae	2100, 1750	Kosrae	Kosrae	5			Kosrae	Athens (1995)	Rieth & Cochrane (2018)
Nauru	2150, 1750	Nauru	Nauru	5		Yes	Tungaru	DiPazza (1999)	Rieth & Cochrane (2018)
Niue	2000, 1600	Niue	Niue	5			Niue	Walter and Anderson (2002)	Rieth & Cochrane (2018)
Ratak	2000, 1600	Ratak + Rālik	Ratak + Rālik	5			Marshall Islands	Beardsley (1994), Riley (1987), weisler (1999, 2001) and Weisler et al (2012)	Rieth & Cochrane (2018)
Rālik	2000, 1600	Ratak + Rālik	Ratak + Rālik	5			Marshall Islands	Beardsley (1994), Riley (1987), weisler (1999, 2001) and Weisler et al (2012)	Rieth & Cochrane (2018)
Sorol	2000, 1700	Sorol	Ulithi (greater)	5		Yes	Ulithi	Intoh and Leach (1985) and Takayama (1982)	
Tungaru	2150, 1750	Tungaru	Tungaru + Tuvalu	5			Tungaru	DiPazza (1999)	Rieth & Cochrane (2018)
Ulithi	2000, 1700	Ulithi + Yap	Ulithi (greater)	5			Ulithi	Intoh and Leach (1985) and Takayama (1982)	Rieth & Cochrane (2018)

Ānewetak	2000, 1600	Ānewetak	Ratak + Rālik	5	Yes	Marshall Islands	Beardsley (1994), Riley (1987), weisler (1999, 2001) and Weisler et al (2012)	Rieth & Cochrane (2018)
Ngatik	1900, 1700	Ngatik	Pohnpei	6	Yes	Pohnpei	Athens (1990) and Galipaud (2000)	Rieth & Cochrane (2018)
Pohnpei	1900, 1700	Pohnpei	Pohnpei	6		Pohnpei	Athens (1990) and Galipaud (2000)	Rieth & Cochrane (2018)
Satawal	1900, 1700	Woleai	Satawal	6	Yes	Pohnpei	Napolitano et al. (2019)	
Satawal (Woleai speaking)	1900, 1700	Woleai	Woleai	6	Yes	Pohnpei	Napolitano et al. (2019)	
Wolei	1900, 1700	Woleai	Woleai	6	Yes	Pohnpei	Napolitano et al. (2019)	
Mwoakilloa	1700	Pohnpei	Mwoakilloa	7	Yes	Pohnpei	Poteate et al (2016)	Levin et al. (2019)
Pingelap	1700, 1550	Pohnpei	Pingelap	7	Yes	Pohnpei	Levin et al. (2019)	
Rotuma	1300, 1050	Rotuma	Rotuma	8		Rotuma	Ladefoged et al (1998)	Rieth & Cochrane (2018)
Emae	1000	Vanuatu + Temotu	Emae	9	Yes	Emae	Kirch (2012) and Carson (2012)	
Hawaii	950, 850	Hawaii	Hawaii	9		Hawaii	Athens et al (2014)	Rieth & Cochrane (2018)
Hereheretue	1100, 770	Hereheretue	Tuamotu	9		Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Kapingamarangi	1000, 700	Kapingamarangi	Kapingamarangi	9		Kapingamarangi	Carson (2012)	

Mangaia	960, 780	Mangaia	Avaiki Nui (south)	9	Southern Cook Islands	Walter and Reilly 2010	Walworth (2015)
Mangareva	920, 660	Mangareva	Mangareva	9	Mangareva	Anderson et al (2003), Green and Weisler (2002) and Kirch et al (2010)	Rieth & Cochrane (2018)
Marutea	1100, 770	Marutea	Tuamotu	9	Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Morane	1100, 770	Morane	Tuamotu	9	Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Napuka	1100, 770	Napuka	Tuamotu	9	Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Ngā Pū Toru	900, 680	Ngā Pū Toru	Avaiki Nui (south)	9	Southern Cook Islands	Allen and Morrison (2013), Allen and Wallace (2007), Kirch et al (1995) and Parkes (1997)	Rieth & Cochrane (2018)
Nia Matai	960, 800	Raro Matai + Nia Matai	Tahiti	9	Society Islands	Anderson and Sinoto (2002), Kahn (2012), Wilmshurst et al (2011) and Parkes (1997)	Rieth & Cochrane (2018)

Nukutaveke	1100, 770	Nukutaveke 61, Nukutaveke 59	Tuamotu	9		Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Puka-puka	1100, 770	Tuamotu + Puka-puka	Tuamotu	9		Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Rapa Iti	850, 750	Rapa Iti	Rapa Iti	9		Austral islands (Rapa)	Bollt (2008), Kennett et al (2012) and Prebble and Anderson (2012)	Rieth & Cochrane (2018)
Raro Matai	960, 800	Raro Matai + Nia Matai	Tahiti	9		Society Islands	Anderson and Sinoto (2002), Kahn (2012), Wilmshurst et al (2011) and Parkes (1997)	Rieth & Cochrane (2018)
Rarotonga	900, 680	Rarotonga	Avaiki Nui (south)	9		Southern Cook Islands	Allen and Morrison (2013), Allen and Wallace (2007), Kirch et al (1995) and Parkes (1997)	Rieth & Cochrane (2018)
Reao	1100, 770	Tatakoto + Reao	Tuamotu	9		Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Tatakoto	1100, 770	Tatakoto + Reao	Tuamotu	9		Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)

Te Fenua 'Enata	950, 750	Te Henua 'Enana + Te Fenua 'Enata	Te Fenua 'Enata	9			Marquesas	Allen (2004) and Allen and McAllister (2010, 2013)	Rieth & Cochrane (2018)
Te Henua 'Enana	950, 750	Te Henua 'Enana + Te Fenua 'Enata	Te Henua 'Enana	9			Marquesas	Allen (2004) and Allen and McAllister (2010, 2013)	Rieth & Cochrane (2018)
Tuamotu	1100, 770	Tuamotu + Puka-puka	Tuamotu	9			Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Tureia	1100, 770	Tureia	Tuamotu	9			Tuamotu	Chazine (1985) and Hatanaka et al (1978)	Rieth & Cochrane (2018)
Tuvalu	1070, 770	Tuvalu 31, Niu + Tuvalu, Tuvalu 32	Tungaru + Tuvalu	9			Tuvalu	Dickinson et al (1990)	Rieth & Cochrane (2018)
North Island (NZ)	750, 670	Aotearoa	Aotearoa	10			New Zealand, Auckland Islands	Higham et al (1999), McGlone and Wilmshurst (1999) and Wilmshurst et al (2008)	Rieth & Cochrane (2018)
Pukapuka	750, 550	Pukapuka	Pukapuka	10		Yes	Tokelau	Petchey et al (2010)	Rieth & Cochrane (2018)
Raivavae	670, 550	Raivavae	Raivavae	10		Yes, Rurutu	Austral islands (Rurutu)	Bollt (2008), Kennett et al (2012) and Prebble and Anderson (2012)	Rieth & Cochrane (2018)

Rapa Nui	750, 700	Rapa Nui	Rapa Nui	10		Rapa Nui	Hunt and Lipo (2006) and Mann et al (2008)	Rieth & Cochrane (2018)
Rimatara	670, 550	Rimatara	Rimatara	10	Yes, Rurutu	Austral islands (Rurutu)	Bollt (2008), Kennett et al (2012) and Prebble and Anderson (2012)	Rieth & Cochrane (2018)
Rurutu	670, 550	Rurutu	Rurutu	10		Austral islands (Rurutu)	Bollt (2008), Kennett et al (2012) and Prebble and Anderson (2012)	Rieth & Cochrane (2018)
Sonsorol	600, 450	Sonsorol	Sonsorol	10		Sonsorol	Intoh (2008)	
South Island (NZ)	750, 670	Aotearoa	Aotearoa	10		New Zealand, Auckland Islands	Higham et al (1999), McGlone and Wilmshurst (1999) and Wilmshurst et al (2008)	Rieth & Cochrane (2018)
Tokelau	750, 550	Tokelau	Tokelau	10		Tokelau	Petchey et al (2010)	Rieth & Cochrane (2018)
Tupuai	670, 550	Tupuai	Tupuai	10	Yes, Rurutu	Austral islands (Rurutu)	Bollt (2008), Kennett et al (2012) and Prebble and Anderson (2012)	Rieth & Cochrane (2018)
Luangiua	500	Luangiua + Nukumanu	Luangiua	11	Yes	Luangiua	Kirch (2012) and Carson (2012)	

Mapia	500, 400	Mapia	Mapia	11	Yes	Mapia	Intoh & Ono (2007)	
Māngarongaro	550, 300	Māngarongaro	Māngarongaro	11		Northern Cook Islands	Chikamori (1998) and Chikamori and Yoshida (1988)	Rieth & Cochrane (2018)
Nukumanu	500	Luangiua + Nukumanu	Nukumanu	11	Yes	Nukumanu	Kirch (2012) and Carson (2012)	
Nukuoro	500	Nukuoro	Nukuoro	11		Nukuoro	Kirch (2012) and Carson (2012)	
Nukuria	500	Nukuria	Nukuria	11	Yes	Nukuria	Kirch (2012) and Carson (2012)	
Ouvea (West Uvean)	500	Kanaky	Ouvea	11	Yes	None	Kirch (2012) and Carson (2012)	
Rakahanga-Manihiki	550, 300	Rakahanga-Manihiki	Rakahanga-Manihiki	11		Northern Cook Islands	Chikamori (1998) and Chikamori and Yoshida (1988)	Rieth & Cochrane (2018)
Rēkohou	450, 400	Rēkohou	Rēkohou	11		Aotearoa	McFadgen (1994)	Rieth & Cochrane (2018)
Sikaiana	500	Sikaiana	Sikaiana	11	Yes	Sikaiana	Kirch (2012) and Carson (2012)	
Takuu	500	Takuu	Takuu	11	Yes	Takuu	Kirch (2012) and Carson (2012)	
Tobi	500, 400	Tobi	Tobi	11		Tobi	Intoh & Ono (2007)	

Northern Marianas	185	Laguas yan gåni	Laguas yan gåni	12		Mariana Islands (Carolinian settlement)	Fritz (1911), Spoehr (1954), Bowers (1953) and Quackenbush (1968)	Ellis (2012)

Table 2: Table of settlement time depth.

Input data tables

All measurements, except the response variable, were scaled and centred to make the coefficients easier to interpret and compare. Shorelines were also log-10-transformed (before scaling).

All code and data will also be available on OSF/GitHub/Zenodo.

Anonymised OSF project for double-blind peer review: `https://osf.io/amnh9/?view_only=403d57d715254539a3a99c798b4`

Island group (shared language)	Lang count	Shoreline	environ PC1	environ PC2	environ PC3	Political complexity (EA033)	Time depth
Malakula	33	1.010	-0.612	-0.836	-0.662	-1.110	0.864
Kanaky (New Caledonia main island)	29	2.275	-1.420	-0.254	-1.210	-1.110	0.864
Santo	24	1.038	-0.340	-0.853	-0.347	-1.110	0.864
Ambrym	6	-0.706	0.372	1.623	-2.086	-1.110	0.864
Efate	5	0.559	-0.675	-0.429	-1.082	0.182	0.864
Pentecost	5	-0.173	-0.218	-0.356	-0.601	-1.110	0.864
Tanna	5	-0.308	-1.063	-0.592	-0.950	-1.110	0.864
Viti Levu + Yasawa	4	1.238	-0.597	-0.572	-0.337	1.474	0.864
Erromango	3	-0.141	-0.775	-0.101	-1.303	0.182	0.864
Laguas yan gåni	3	0.703	-0.303	-1.148	-1.140	-1.110	1.146
Maewo	3	-0.194	-0.123	-0.383	-0.448	-1.110	0.864
Vanua Lava	3	-0.272	0.160	-0.438	0.068	-1.110	0.864
Ambae	2	-0.804	0.557	1.558	-1.813	-1.110	0.864
Duff + Reef Islands	2	-0.636	0.906	-0.532	1.537	-1.110	0.864
Tungaru + Tuvalu	2	1.363	0.691	-0.930	-0.013	0.182	0.018
Ureparapara	2	-0.774	-0.050	-1.023	0.392	-1.110	0.864
Aneityum	1	-0.500	-0.970	0.121	-1.272	0.182	0.864
Anuta	1	-2.042	1.360	1.510	-0.124	-1.110	0.864
Aore	1	-0.578	-0.274	-0.384	-0.787	-1.110	0.864
Aotearoa	1	3.025	-2.534	2.598	2.541	0.182	-1.393
Avaiki Nui (south)	1	0.210	-0.708	0.082	0.110	0.182	-1.111
Chuuk	1	0.800	0.831	-1.221	1.477	0.182	0.300
East Futuna	1	-0.408	0.272	-0.306	0.597	0.182	0.300
Hawaii	1	1.331	-0.745	0.375	0.746	2.766	-1.111
Hiu	1	-0.709	0.045	-1.031	0.588	-1.110	0.864
Kadavu	1	0.324	-0.740	-0.163	-0.472	1.474	0.864
Kapingamarangi	1	-1.305	1.535	1.011	-0.955	-1.110	-1.111
Kosrae	1	-0.269	1.040	-1.206	1.262	1.474	0.018
Lau	1	1.435	-0.633	-0.523	-0.057	1.474	0.864
Lifou	1	-0.010	-1.095	0.145	-1.496	0.182	0.864
Loh-Toga	1	-0.070	0.191	-0.617	0.322	-1.110	0.864
Luangiua	1	-0.797	1.763	1.102	-0.133	-1.110	-1.675

Mangareva	1	0.123	-0.675	0.783	1.372	0.182	-1.111
Mo-ava-mo-iki	1	-0.106	0.237	-1.095	0.876	0.182	0.864
Mota	1	-1.077	-0.045	-1.002	0.451	-1.110	0.864
Mota Lava	1	-0.694	0.168	-0.442	0.075	-1.110	0.864
Māngarongaro	1	-0.785	1.351	1.137	-0.988	0.182	-1.675
Nengone	1	-0.165	-1.172	0.290	-1.364	1.474	0.864
Niue	1	-0.523	-0.771	-0.410	0.075	0.182	0.018
Nukuoro	1	-1.283	1.732	0.870	-0.718	-1.110	-1.675
Paama	1	-0.654	-0.603	-0.874	-0.523	-1.110	0.864
Palau	1	0.534	0.818	-1.211	1.437	0.182	0.864
Pohnpei	1	0.509	0.932	-1.175	1.414	1.474	-0.264
Pukapuka	1	-0.834	0.780	-0.384	0.952	0.182	-1.393
Raivavae	1	-0.360	-0.714	1.501	0.072	0.182	-1.393
Rakahanga-Manihiki	1	-0.952	1.396	1.320	-0.483	0.182	-1.675
Rapa Nui	1	-0.498	-1.583	0.218	-0.397	0.182	-1.393
Ratak + Rālik	1	1.259	0.604	-1.046	0.334	1.474	0.018
Rimatara	1	-1.082	-0.787	0.588	0.125	0.182	-1.393
Rotuma	1	-0.779	0.848	0.316	0.722	1.474	-0.829
Rurutu	1	-0.416	-0.946	0.479	0.429	0.182	-1.393
Rēkohou	1	0.707	-2.574	2.634	2.163	0.182	-1.675
Sāmoa	1	1.179	0.150	-0.788	0.972	1.474	0.582
Tahiti	1	1.346	-0.179	-0.409	0.555	1.474	-1.111
Te Henua ‘Enana	1	0.823	-0.201	-1.951	-2.525	-1.110	-1.111
Tikopia	1	-1.595	1.233	1.512	-0.402	0.182	0.864
Tokelau	1	-0.508	1.755	1.333	0.378	0.182	-1.393
Tonga	1	1.477	-0.871	-0.331	-0.074	1.474	0.582
Tuamotu	1	2.367	-0.210	-0.450	0.723	0.182	-1.111
Tupuai	1	-0.674	-0.897	0.828	0.398	0.182	-1.393
Ulithi (greater)	1	-0.003	0.710	-0.636	0.973	0.182	0.018
Uvea (Wallis)	1	-0.382	0.466	-0.358	0.919	0.182	0.300
Woleai	1	-0.828	1.691	1.176	-0.123	0.182	-0.264
Yap	1	-0.741	1.507	1.351	-0.168	1.474	0.300

Table 3: Table of input values to model, shared language island groups.

Island group (overnight-sailing)	Lang count	Shoreline	environ PC1	environ PC2	Political complexity (EA033)	Time depth
Vanuatu + Temotu	130	1.293	-0.338	0.998	-1.860	1.781
New Caledonia (incl loyalties)	34	2.129	-1.348	1.066	1.447	1.781
Fiji	8	1.756	-0.644	0.923	1.447	1.781
Chuuk	6	0.783	0.650	0.707	-0.207	1.109
Laguas yan gåni	3	0.634	-0.368	1.859	-1.860	2.117
Pohnpei	3	0.451	0.649	0.777	1.447	0.438
Luangiua + Nukumanu	2	-0.671	1.413	-0.680	-1.860	-1.241
Niu + Tuvalu	2	-0.221	1.502	-1.125	-0.207	-0.570
Te Henua ‘Enana + Te Fenua ’Enata	2	0.869	-0.223	2.945	-1.860	-0.570
Ulithi + Yap	2	-0.004	0.439	0.538	-0.207	1.109
Woleai	2	-0.802	1.379	-0.732	-0.207	0.438
Aotearoa	1	2.840	-2.490	-2.493	-0.207	-0.905
Banaba	1	-1.522	1.033	-0.235	-0.207	0.773
Bellona + Rennelle	1	-0.134	0.015	0.981	-0.207	1.781
East Futuna	1	-0.421	0.067	0.384	-0.207	1.109
Hawaii	1	1.231	-0.828	-0.174	3.100	-0.570
Hereheretue	1	-0.518	-0.485	-0.191	-0.207	-0.570
Kapingamarangi	1	-1.273	1.268	-0.317	-1.860	-0.570
Kosrae	1	-0.289	0.714	0.928	1.447	0.773
Mangaia	1	-0.807	-1.086	0.253	-0.207	-0.570
Mangareva	1	0.084	-0.805	-0.816	-0.207	-0.570
Marutea	1	0.076	-0.508	-0.887	-0.207	-0.570
Morane	1	-1.294	-0.069	-2.198	-0.207	-0.570
Māngarongaro	1	-0.779	1.113	-0.345	-0.207	-1.241
Napuka	1	-0.402	-0.055	0.668	-0.207	-0.570
Ngatik	1	-1.469	1.506	-0.551	1.447	0.438
Ngā Pū Toru	1	0.089	-0.607	0.371	-0.207	-0.570
Niue	1	-0.530	-0.843	0.727	-0.207	0.773
Nukuoro	1	-1.252	1.432	-0.244	-1.860	-1.241
Nukutaveke 59	1	0.315	-0.609	-0.690	-0.207	-0.570
Nukutaveke 61	1	0.315	-0.609	-0.690	-0.207	-0.570
Palau	1	0.474	0.513	0.878	-0.207	1.781

Pukapuka	1	-0.826	0.510	0.323	-0.207	-0.905
Raivavae	1	-0.376	-0.781	-0.977	-0.207	-0.905
Rakahanga-Manihiki	1	-0.938	1.132	-0.706	-0.207	-1.241
Rapa Nui	1	-0.506	-1.529	0.529	-0.207	-0.905
Raro Matai + Nia Matai	1	1.246	-0.330	0.530	1.447	-0.570
Rarotonga	1	-0.436	-0.925	0.147	-0.207	-0.570
Ratak + Rālik	1	1.162	0.364	1.178	1.447	0.773
Rimatara	1	-1.061	-0.853	-0.180	-0.207	-0.905
Rotuma	1	-0.773	0.580	-0.252	1.447	-0.234
Rurutu	1	-0.429	-1.010	-0.186	-0.207	-0.905
Rēkohou	1	0.638	-2.495	-2.312	-0.207	-1.241
Sorol	1	-1.436	1.326	-0.843	-0.207	0.773
Sāmoa	1	1.087	-0.058	0.695	1.447	1.445
Tatakoto + Reao	1	0.259	-0.386	0.320	-0.207	-0.570
Tokelau	1	-0.516	1.411	-1.063	-0.207	-0.905
Tonga	1	1.370	-0.927	0.708	1.447	1.445
Tuamotu + Puka-puka	1	2.211	-0.340	0.721	-0.207	-0.570
Tungaru	1	1.251	0.210	1.531	-0.207	0.773
Tupuai	1	-0.673	-0.962	-0.487	-0.207	-0.905
Tureia	1	-0.488	-0.654	-0.650	-0.207	-0.570
Tuvalu 31	1	-0.221	1.502	-1.125	-0.207	-0.570
Tuvalu 32	1	-0.221	1.502	-1.125	-0.207	-0.570
Uvea (Wallis)	1	-0.397	0.227	0.305	-0.207	1.109
Ānewetak	1	-0.882	0.711	0.287	1.447	0.773

Table 4: Table of input values to model, overnight-sailing island groups.

Directed Acyclic Graph

Directed Acyclic Graphs are a helpful tool in causal reasoning (Pearl (1995) and McElreath (2020)). Fig. 19 shows a full DAG of all variables discussed in the literature that are related to this topic. Nodes with dashed outlines indicate that the variable cannot be instantiated in a measurable predictor for modelling, most often due to a lack of data. Variables that only relate to the response variable through measurable intermediate nodes (e.g. isolation) are excluded.

Model settings

The Bayesian Regression model used is implemented in the R-package BRMS (Bürkner, 2017). There is no clear evidence for an informative prior for the predictors (a.k.a. fixed effects), therefore a prior was not specified which means the default is used, i.e. a flat prior.

The list below specifies further settings of the model functions.

- brms::brm()
  - iter = 30000
  - warmup = 1000
  - chains = 4
  - cores = 4
  - seed = 10
  - family = poisson
  - backend = cmdstanr
- brms::posterior_predict()
  - ndraws = 10000
  - cores = 4

All code and data will also be available on OSF/GitHub/Zenodo.

Anonymised OSF project for double-blind peer review: `https://osf.io/amnh9/?view_only=403d57d715254539a3a99c798b4`

BRMS model tables

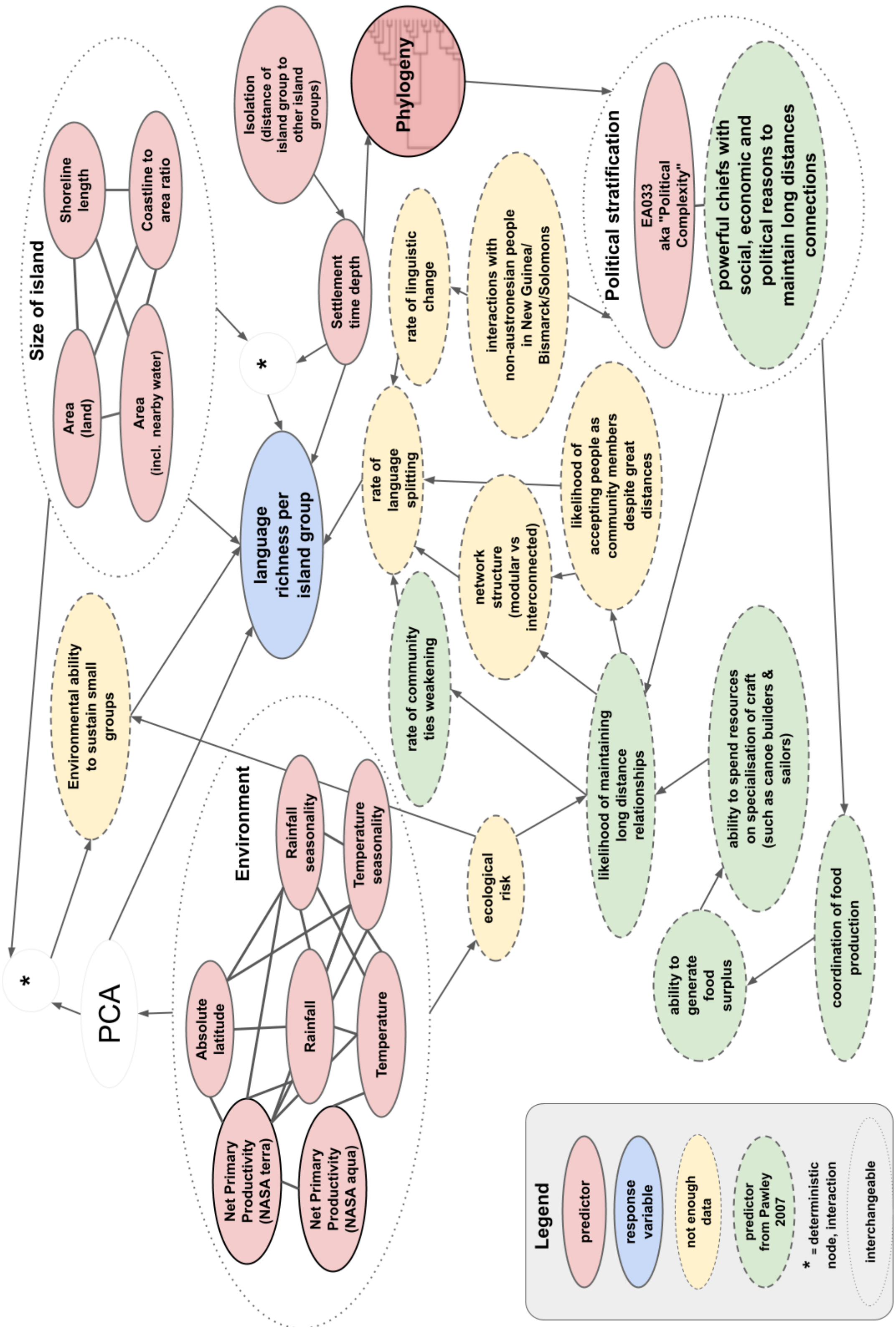

Fig. 19. Directed Acyclic Graph of the variables in this study. Blue = variable to be predicted (response). Asterisk-nodes represent interactions.

term	Estimate	Est.Error	l-95% CI	u-95% CI	Does 95% interval straddle zero?	Bulk ESS	Tail ESS
EA033	-0.854	0.119	-1.093	-0.626	no	84425.289	81421.505
environ_PC1	-0.389	0.197	-0.776	-0.004	no	53786.906	72777.062
environ_PC1:Shoreline	0.145	0.146	-0.130	0.441	yes	51663.978	71060.623
environ_PC2	0.196	0.198	-0.194	0.586	yes	53715.694	71858.101
environ_PC3	0.109	0.135	-0.158	0.371	yes	75057.067	79671.438
Intercept	0.248	0.170	-0.093	0.574	yes	58590.121	72978.752
Time depth	0.468	0.135	0.214	0.743	no	86204.009	74655.200
Shoreline	0.632	0.174	0.296	0.980	no	53563.638	70289.754
Shoreline:environ_PC2	-0.353	0.182	-0.697	0.019	yes	56356.997	64884.923
Shoreline:environ_PC3	0.207	0.138	-0.070	0.473	yes	63281.946	74922.550
Shoreline:Time depth	0.488	0.131	0.241	0.754	no	88157.074	78784.599

Table 5. Table of BRMS model outcomes, shared-language island groups (all observations included).

term	Estimate	Est.Error	l-95% CI	u-95% CI	Does 95% interval straddle zero?	Bulk ESS	Tail ESS
EA033	-0.917	0.063	-1.042	-0.797	no	82802.809	81343.825
environ_PC1	0.053	0.137	-0.215	0.324	yes	89077.323	80524.829
environ_PC1:Shoreline	0.145	0.107	-0.061	0.357	yes	69870.197	77242.507
environ_PC2	-0.710	0.158	-1.017	-0.398	no	79125.306	81539.120
Intercept	-0.231	0.157	-0.546	0.066	yes	82655.407	80735.962
Time depth	0.415	0.163	0.093	0.735	no	66168.261	74639.742
Shoreline	0.996	0.163	0.673	1.318	no	73783.374	79021.266
Shoreline:environ_PC2	0.089	0.134	-0.180	0.343	yes	62464.650	75072.193
Shoreline:Time depth	0.814	0.136	0.554	1.087	no	60210.832	71135.019

Table 6. Table of BRMS model outcomes, overnight-distance island groups (all observations included).

Supplementary figures

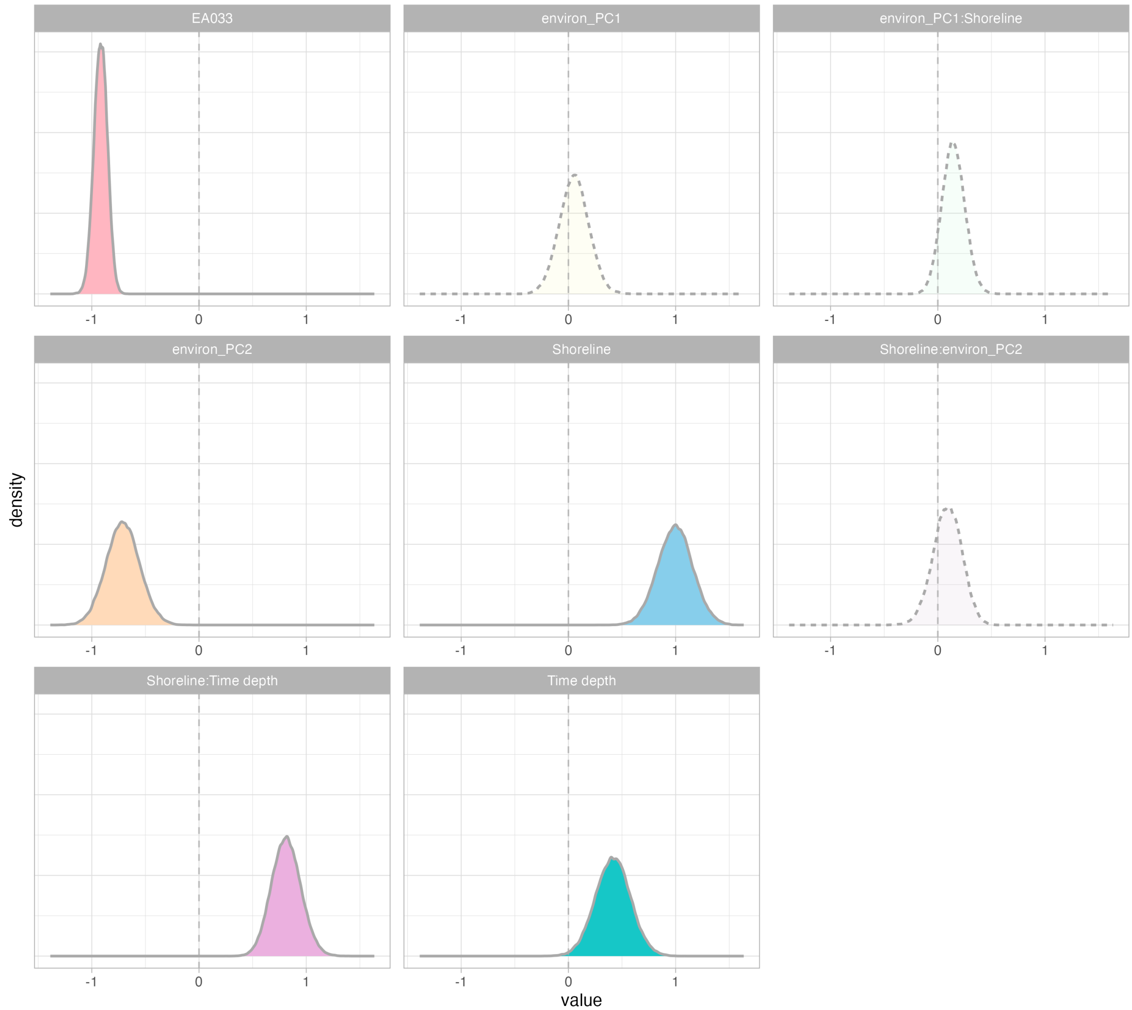

Fig. 20: Coefficient distribution of BRMS model, using overnight sailing distances island groups with all observations. Distributions, where the 95% confidence interval straddles zero, are marked as more transparent and with dashed outlines.

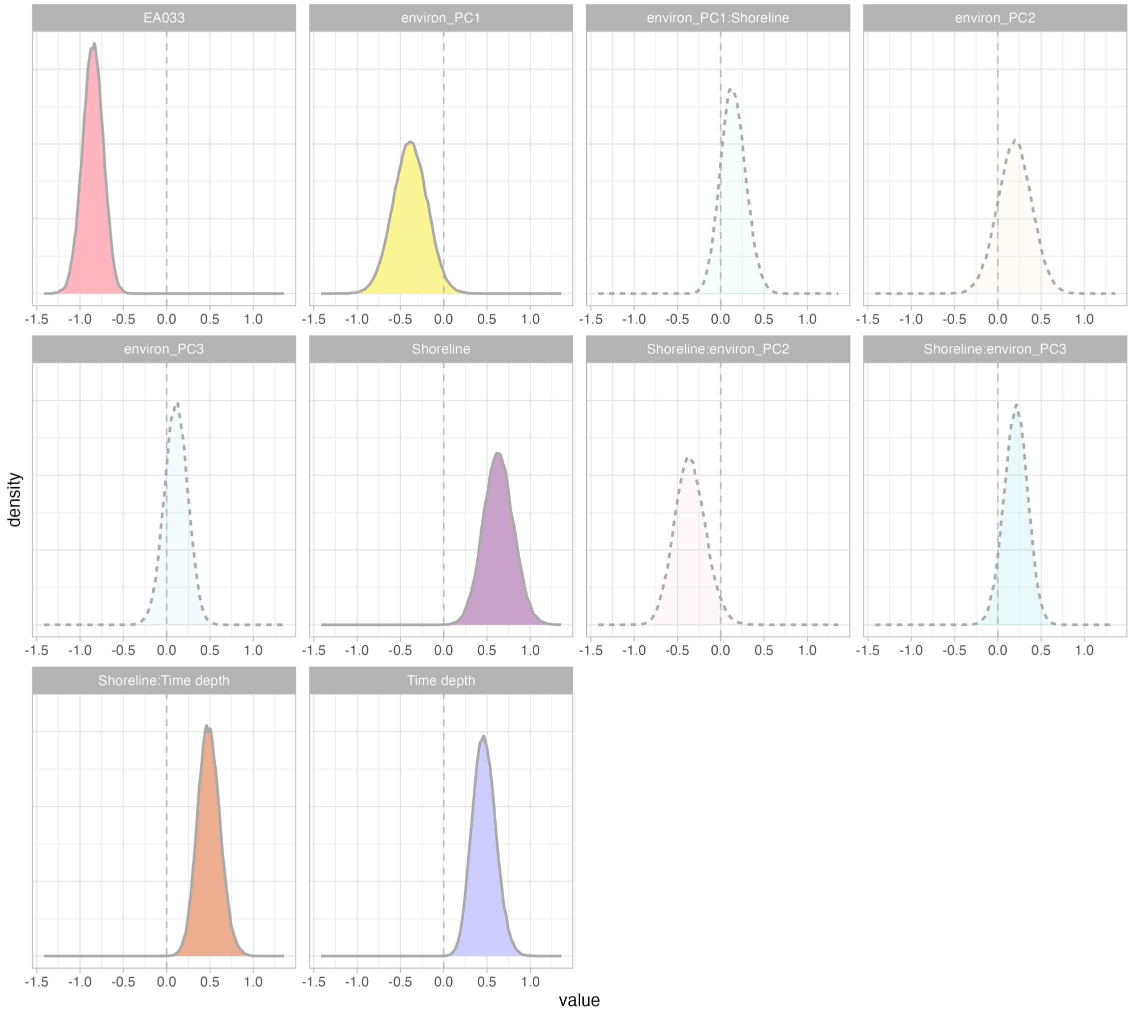

Fig. 21: Coefficient distribution of BRMS model, using shared-language island groups with all observations. Distributions, where the 95% confidence interval straddles zero, are marked as more transparent and with dashed outlines.

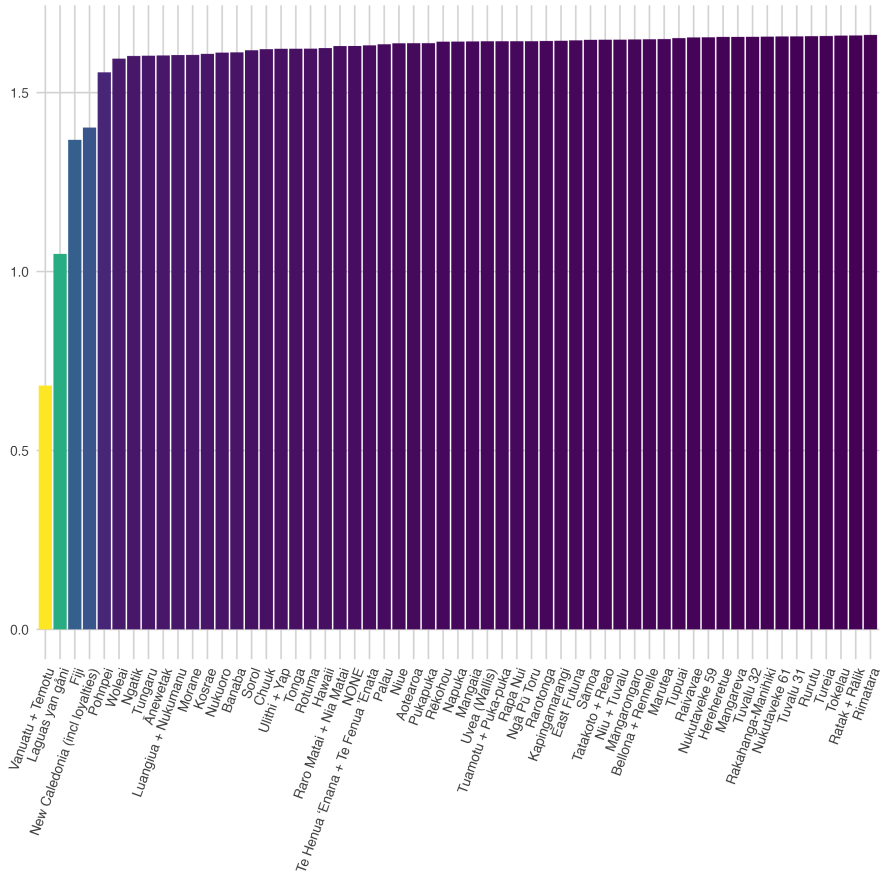

Fig. 22: Mean absolute difference of predicted and observed number of languages per island group (overnight sailing distances group). Every column represents a new run of the model with a particular group dropped out from the observations. When Vanuatu + Temotu is dropped out, the model predicts numbers that are more similar to the observed values than when that island group is included in the model.

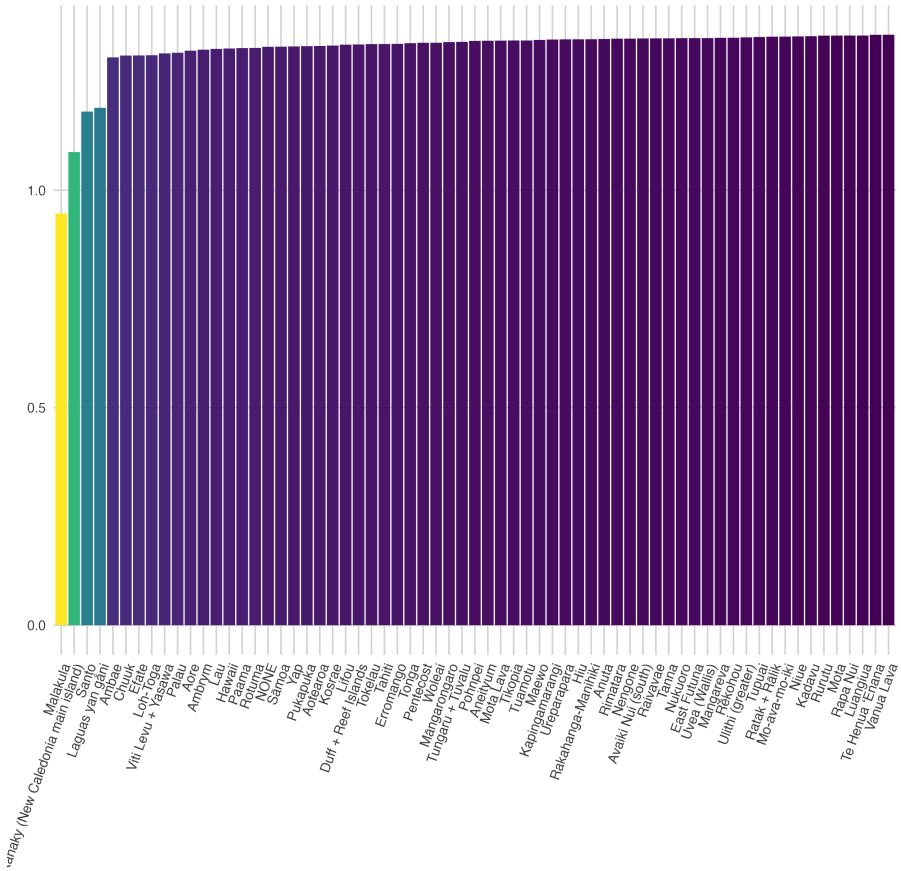

Fig. 23: Mean absolute difference of predicted and observed number of languages per island group (shared language grouping). Every column represents a new run of the model with a particular group dropped out from the observations. When Malakula is dropped out, the model predicts numbers that are more similar to the observed values than when that island group is included in the model.

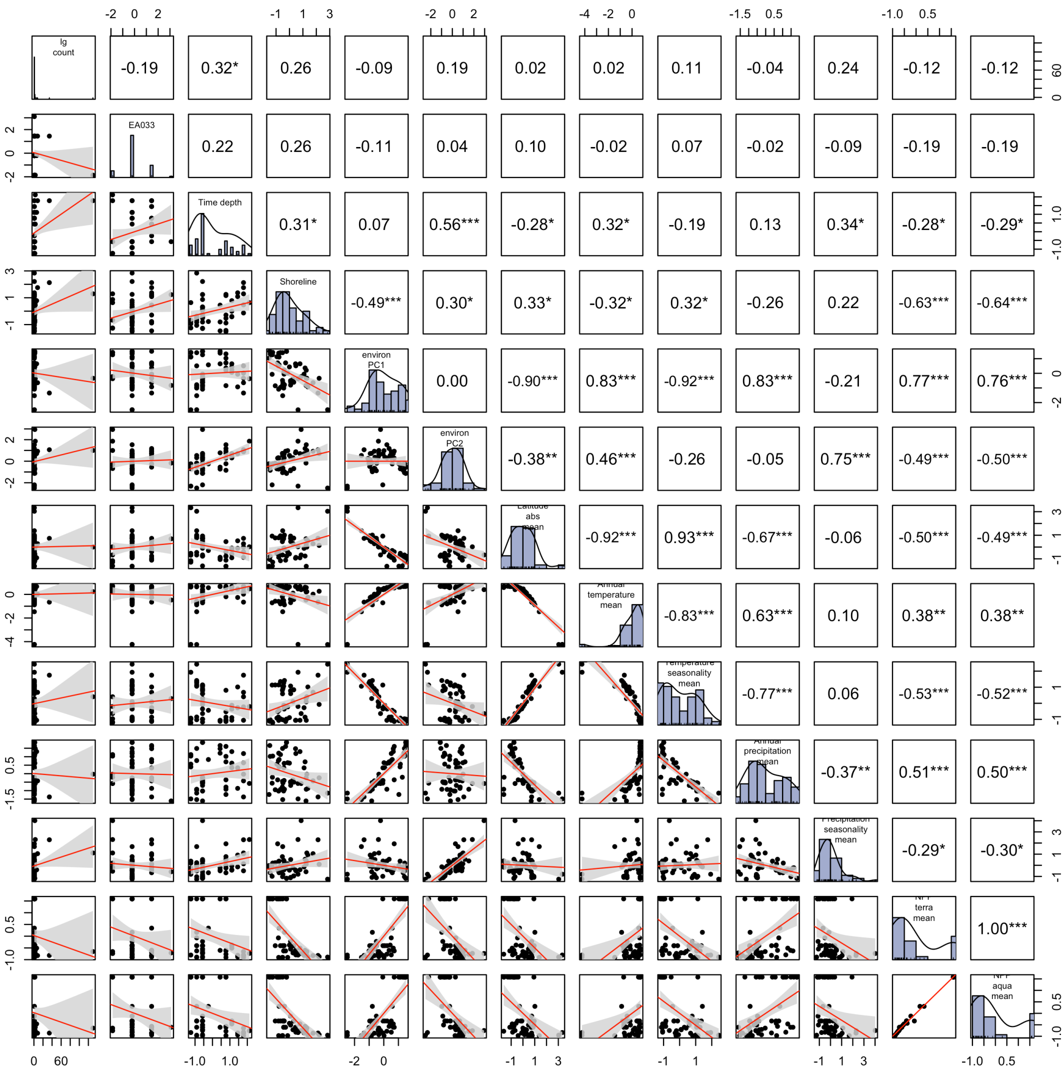

Fig. 24: Scatterplot matrix of all variables available, over overnight-sailing-groups.

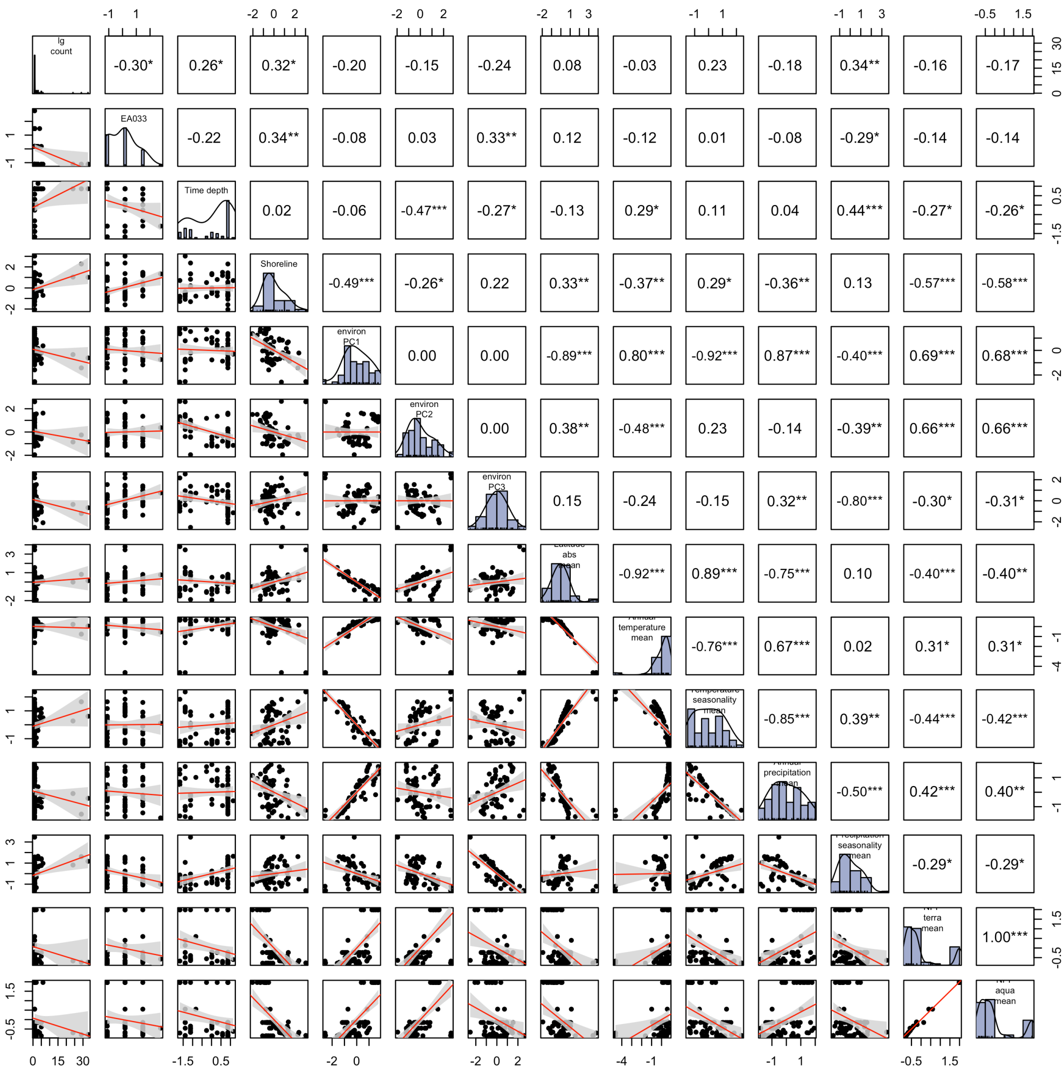

Fig. 25: Scatterplot matrix of all variables available, over shared-language-groups.

R-packages used

All the analysis for this research project was done in the free and open source programming language R, using a multitude of packages. All code and data for this project are available in the locations listed in appendix A. The scripts have been written so that any user of R can execute them. Please see the bibliography for information on package versions. Below are citations for all used packages. The citations also include which version of the package was used.

R Core Team (2023), Ottolinger (2019), Bürkner (2023), Gabry & Češnovar (2022), Ihaka et al. (2023), Dowle & Srinivasan (2023), Wickham et al. (2022), Wickham et al. (2023b), Nychka et al. (2022), Wickham (2023), Rudis et al. (2017), Wickham et al. (2023a), Kassambara (2023), Slowikowski (2023), Arnold (2021), Ooms (2023), Xie (2023), Sarkar (2023), Edwards (2022), Ripley (2023), Tierney et al. (2023), Cooper (2022), Raiche & Magis (2022), Revelle (2023), Wickham & Henry (2023), Ammar (2019), Hijmans (2023), Schutten et al. (2023), Wickham et al. (2023c), Wickham (2020), Guo et al. (2023), Wickham & Seidel (2022), Pebesma & Bivand (2023), Furrer et al. (2022), Wickham (2022), Müller & Wickham (2023), Wickham et al. (2023d), Garnier (2023), Dahl et al. (2019), Bürkner (2017), Bürkner (2018), Bürkner (2021), Zeileis et al. (2020), Zeileis et al. (2009), Stauffer et al. (2009), Wickham (2016), Ooms (2014), Xie (2015), Xie (2014), Sarkar (2008), Venables & Ripley (2002), Tierney & Cook (2023), Wickham (2007), Pebesma & Bivand (2005), Bivand et al. (2013), Furrer & Sain (2010), Gerber & Furrer (2015), Gerber et al. (2017) and Garnier et al. (2023) .

Competing interests

No competing interest is declared.