UTRECHT UNIVERSITY

Department of Information and Computing Science

Applied Data Science Master Thesis

Validation of a Bayesian mixture model for language contact with the use of synthetic language data

First examiner:
Judith Verstegen

Candidate:
David Johannes Keus

Second examiner:
Simon Scheider

In cooperation with:
Joseph Taylor

July 14, 2023

Abstract

Speaker communities typically have some level of interaction and are not completely isolated. When individuals who speak different languages come into contact, it is probable that their respective languages undergo a process of convergence.

Ranacher et al. (2021) have developed a method, *sBayes*, to estimate the relative role of language contact, as opposed to inheritance and universal preference, in creating similarities between languages. The model promises to identify contact areas from empirical data using (Bayesian) inference. However, validation of the approach proves difficult since they use empirical data of real-world language in which, by definition, actual contributions of language contact, inheritance and universal preference are not known.

To further validate the *sBayes* model, a dataset is needed from which we know our expected descriptive contact, inheritance and universal preference values prior to the model run. This dataset can then be compared to the output of *sBayes*.

For this purpose, we created synthetic language datasets using an agent-based model to test the accuracy of *sBayes*. Using these datasets we conducted two experiments, one to validate sBayes ability to detect isolated causal explanations per language feature. The second to test sBayes fit to an artificial language dataset and in determining language areas (clusters) and overall causality counts.

Our results suggest that synthetic language data can successfully be used for validation purposes of the sBayes language model. *sBayes* accuracy on identifying clearly isolated causalities has a combined mean squared error of 0.05 in our simulations. In a simulated real life situation, the model find a similar amount of contact areas. In addition, the overall distribution of feature state causality is the same in our synthetic data when we compare it to a benchmark experiment.

Contents

1. Introduction

Speaker communities typically have some level of interaction and are not completely isolated. When individuals who speak different languages come into contact, it is probable that their respective languages undergo a process of convergence (Bowern and Evans, 2015). In other words, the languages become more similar to each other. In order to communicate with one another, they often need to find a shared language, which can result in situations of bilingualism or multilingualism (Matras, 2009). Exposure to another language, especially if this is widespread within a community and takes place over a long period of time, may lead to horizontal transfer: the incorporation of words or structural features from one language into another (Ranacher et al., 2021).

Ranacher et al. (2021) have developed a method for determining the amount of horizontal transfer between languages. *sBayes* estimates the relative role of language contact, as opposed to inheritance through family and universal preference of one's surroundings, in creating similarities between languages.

Computational models, like *sBayes*, are used to simulate and understand complex systems. However, the accuracy of these models depends on various assumptions, data inputs and parameter values, which can introduce a wide range of uncertainties and errors. To ensure the validity and usefulness of computational models, it is essential to subject them to rigorous validation procedures (Saltelli et al., 2008).

While the model has shown promise in identifying known cases of language contact and change, I believe it needs further validation to determine its accuracy in recognizing new cases of language contact and change. Normally, when we want to assess the content validity of computational models, we compare the outputs with the available empirical data or known

facts to assess their agreement (Saltelli et al., 2008). This proves difficult in the area of language evolution since we simply do not know the facts on the exact feature evolution over time, we can only assume. For this reason, we can not apply any cross-validation or work with train/test models unless we find a way to gather any data we know to be empirically correct.

sBayes was tested on simulated data (951 languages randomly assigned to locations in space) and then on two case studies to reveal language contact in South America and the Balkans (Ranacher et al., 2021). Yet, neither of these experiments can fully validate the approach since the actual contributions of language contact, inheritance and universal preference are unknown. As *sBayes* utilizes an input data set that is formed with assumptions, it is not possible to effectively validate the model using a pre-existing dataset due to all datasets regarding historical languages including partial assumptions on the data points. These contributions are also not known in the simulated data since these are based on existing languages with unknown histories.

This research aims to validate the accuracy of the *sBayes* model by testing its ability to identify areas of language contact and change in several contexts using synthetic datasets of which the contributions are known. For this purpose, we make use of synthetic datasets from which we see the input and output values so that we can compare these with the model outcomes. So, what is the accuracy of *sBayes*?

2. Background

2.1 Contact linguistics

Linguistic contact effects can manifest in various forms and vary in their scope and nature, arising from diverse underlying processes. One of the most easily identifiable effects is the borrowing of linguistic forms and functions from one language to another. This borrowing typically involves the adoption of vocabulary, such as when English borrowed the word "language" from French. However, it can also extend to structural elements, such as affixes or individual sounds, like the borrowing of French suffixes like "-able" (e.g., "readable") into English (Gelman et al., 2013). The areas in which these types of contact effects occur may contain languages similar in their properties. These areas are generally referred to as a linguistic area or Sprachbund (Friedman, 2011).

There are two principled ways to approach the detection and description of a linguistic area: bottom-up and top-down (Muysken et al., 2014). The bottom-up approach starts with recognizing one or more notable attributes in a specific geographic area, indicating potential contact. These attributes can pertain to language or other cultural aspects. This initial observation is used as a prior to further explore the distribution of additional linguistic features within the approximate region. Gradually, a pattern emerges, revealing a generally indistinct area where languages exhibit a shared set of characteristics (Friedman, 2011). The other way to approach linguistic areas, is top-down. Here the approach is to assemble a principled set of features and screen the distribution of the feature values for areal clustering (Friedman, 2011).

Determining linguistic areas is complicated as they result from a number of complex historical processes which are difficult to reconstruct (Ma-

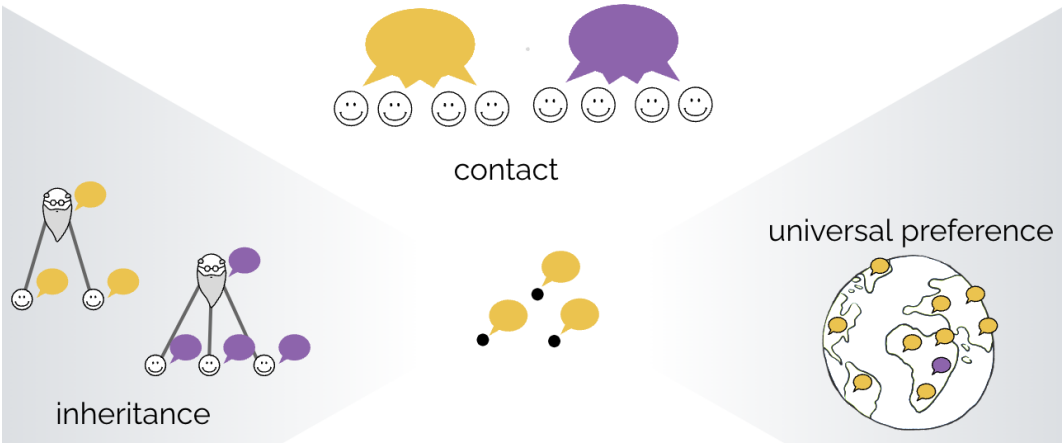

Figure 2.1: Inheritance, universal preference and contact determine horizontal language transfer (Ranacher et al., 2021)

sica, 2001). To detect them, *sBayes* uses a bottom-up approach to determine shared features between geographically close languages. However, assuming these shared features occurred through only contact would ignore any other contributing language processes. In our experiments, we will mainly focus on the confounding factors of inheritance and global preference (fig. 2.1).

Inheritance in language transfer determines language features being transmitted from one generation to the next in an evolutionary process, not unlike the descent with modification that characterizes biological evolution. In language, the modification stems from the variation that each generation adds, mostly for signaling social identities. While this can lead to the split of a language into dialects and eventually into new languages, many properties persist and are inherited faithfully (Croft, 2008).

As for universal preference; the structure of languages is shaped by universal aspects of how they are used for communication and thought, how they are processed in the brain and how they are expressed with our speech and gesture systems. As a result, languages may share a property just because all languages tend to have it (Bickel, 2015).

2.2 Bayes Theorem

Bayes theorem (eq. 2.1) is a fundamental concept in probability theory and statistics that describes how to update our beliefs or probabilities about an event based on new evidence. It mathematically relates the conditional probabilities of two events.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \tag{2.1}$$

$p(A)$ and $p(B)$: $p(A)$ is called the prior probability of event A, while $p(B)$ is the prior probability of event B. These are the probabilities of event A and event B occurring independently of each other.

$p(A|B)$: Posterior probability. This represents the probability of event A occurring given that event B has occurred.

$p(B|A)$: Likelihood. This is the probability of event B occurring given the probability that event A has occurred.

2.3 The sBayes model

sBayes is a Bayesian mixture model that weighs the respective contributions of contact and the confounding effects from inheritance and universal preference in accounting for the similarities between languages in the geographical space.

Ranacher et al.(2021) have adapted Bayesian statistics for locating geographical contact areas (Z) by comparing their feature states. In statistical terms, the task of finding contact areas can be described as clustering, or finding groups of objects whose members share commonalities. However, naive clustering will simply group together languages with similar properties irrespective of the specific processes that have actually made them similar. Instead, *sBayes* infers the relative role of contact, as opposed to the other processes, in creating similarities between languages.

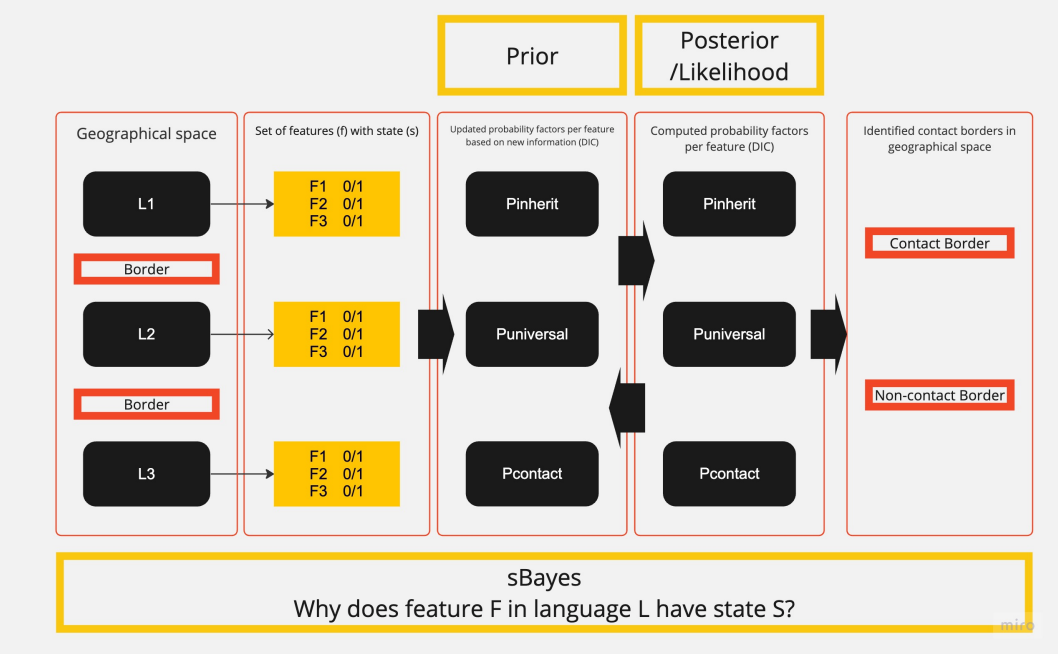

Figure 2.2: Flowchart sBayes model

As can be seen in fig. 2.2, *sBayes* aims to explain why single language (l) has feature (f) with state (s). For example: why a certain vowel sound(f) is present (s) in the Arawakan language Chamicuro (l)? Three effects are proposed and a likelihood function (P) is defined for each:

1. Likelihood for universal preference $(P_{universal})$: the state is universally preferred.
2. Likelihood for inheritance $(P_{inherit})$: the language l belongs to a language family ($\phi(l)$) and the state was inherited from related ancestral languages in the family.
3. Likelihood for contact $(P_{contact})$: the language belongs to area $Z(l)$ and the state was adopted through contact in the area.

The unknown weights $w_{universal}, w_{inherit}$ and $w_{contact}$ quantify the contribution of each of these three effects. For a single language l, which is part of a family $\phi(l)$ and an area $Z(l)$, we define the probability of feature f being in state s as the following mixture likelihood:

The mixture components $P_{universal}$, $P_{inherit}$ and $P_{contact}$ are part of a single categorical distribution parameterized by probability vectors $\alpha f, \beta f, \phi(l)$ and

$\gamma f, Z(l)$. This means that the probability of observing state s in feature f is $\alpha f, s$ if it is the result of universal preference, $\beta f, \phi(l), s$ if it was inherited in family $\phi(l)$ and $\gamma f, Z(l), s$ if it was acquired through contact in area $Z(l)$. Together, these probabilities will always add up to 1 meaning 100%.

This process is illustrated in fig. 2.3, showing universal preference and inheritance acting as confounders to determine contact areas. The figure also introduces the triangular posterior weights density plot which will be used in plotting our results.

Since a Bayesian model requires a start point, prior probabilities are deducted from the entered dataset and the experiment setup parameters if set by the researcher. *sBayes* does not determine the optimal amount of clusters by itself, it does one separate run per n amount of clusters as set by the researcher. After a prior has been established, each run will take an independent posterior sample. In post-processing, users can inspect whether the posterior samples across all runs converge to the same stable distribution (Ranacher et al., 2021).

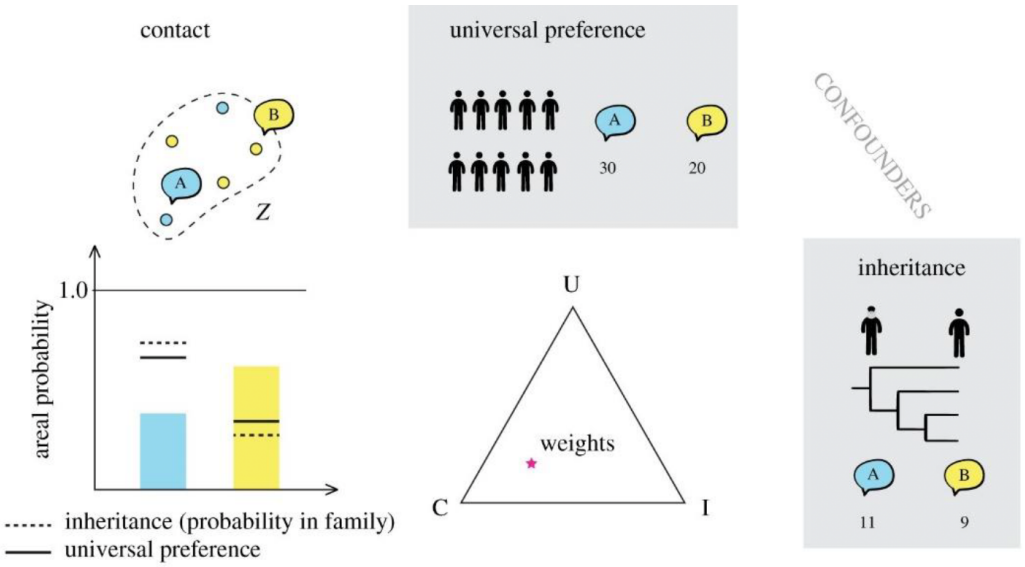

Figure 2.3: Contact, universal preference and inheritance principles illustrated with a posterior weights density plot in the middle (Ranacher et al., 2021)

2.4 Geographical application of sBayes

During each sampling step, Bayesian inference is applied to a moving selection of single language locations l so that the model can find and define contact areas. This moving happens through growing and shrinking the possible area until a possible contact area is defined, not unlike how an amoeba moves through extending (grow operator fig. 2.4) and shrinking (shrink operator fig. 2.5). Areas have a high likelihood of contact if they comprise similar features which cannot be equally well explained by universal preference and inheritance. There are no assumptions about any of the properties of contact areas, such as their shape, size or number, whether they comprise close or distant languages, or cover contiguous or disconnected regions in space. While the assignment of languages to families is fixed, the assignment of languages to areas is inferred from the data. *sBayes* allows for multiple contact areas $Z(fZ1, ..., ZKg)$, each with its own set of areal probability vectors.

sBayes infers the assignment of languages to contact areas from similarities in the language datasets that are poorly explained by the confounding effects of inheritance and universal preference.

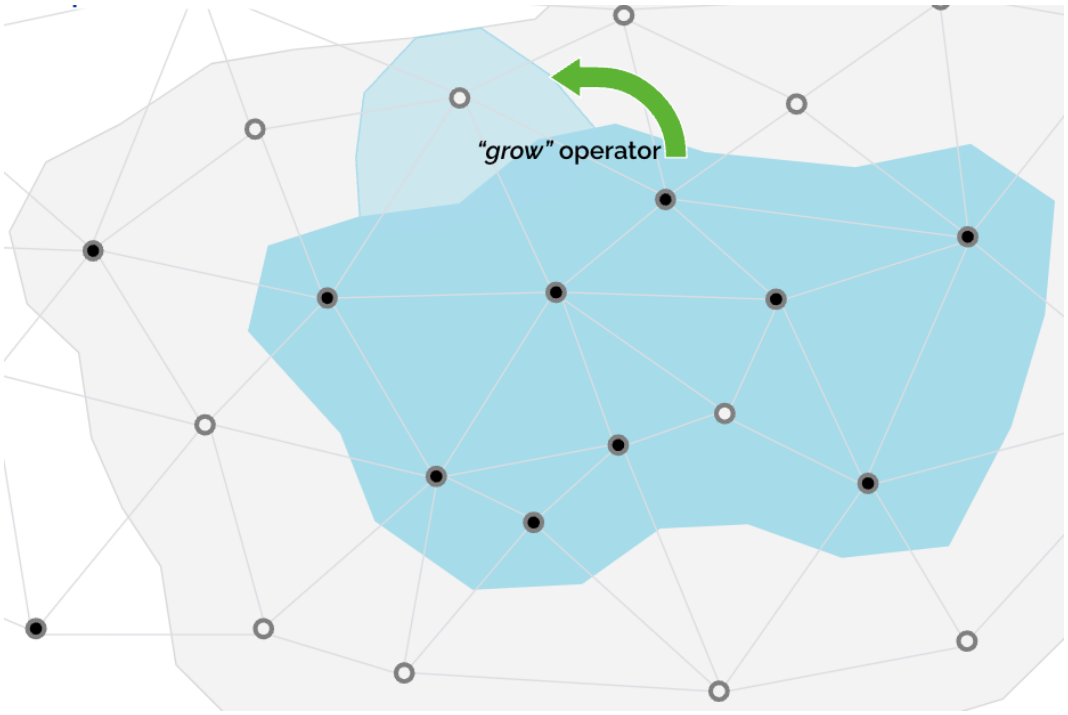

Figure 2.4: Adding a language feature set through geographical point values (Ranacher et al., 2021)

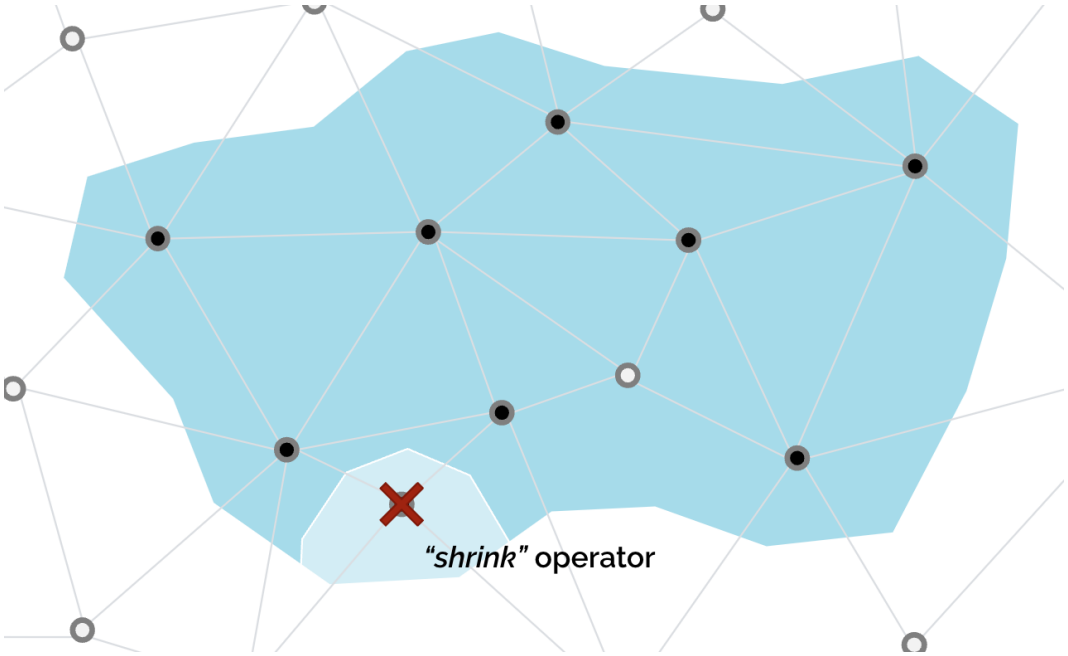

Figure 2.5: Deleting a language feature set through geographical point values (Ranacher et al., 2021)

2.5 sBayes data input

The input of the model is an overview of the known language features and their possible states (fig 2.6). In most cases, a language feature is either present or not present. It is possible to enter any amount of feature states. The chosen features and corresponding states are manually derived from linguistic data in the study area.

The model then requires a description of all known languages in an area in terms of their feature states, each with an identification code and x/y coordinates. If it is known what existing language family the local language stems from, this can also be entered and will give extra weight to the inheritance causality (fig. 2.7).

2.6 sBayes output

The model has the following outputs:

- Statistical overview of the model run

  The log file provides acceptance/rejection statistics for each operator (amount of clusters). A high acceptance rate indicates too frequent and, thus, inefficient sampling, a low acceptance rate indicates too sparse sampling.

At every 2000 iterations:

- Logarithm for prior, posterior and likelihood for the entire dataset
- Logarithm for prior, posterior and likelihood per cluster
- Weights for contact, universal preference and inheritance
- Areal weights per feature state per cluster
- Weights of preference per feature state per language family

Feature	Description	States
F1	Phonemic velar and uvular stops	present, absent
F2	Phonemic /kw/	present, absent
F3	Phonemic glottalized stops/ejectives	present, absent
F4	Phonemic aspirated stops	present, absent
F5	Phonemic retroflex affricates	present, absent
F6	More phonemic affricates than fricatives	present, absent
F7	Phonemic (bi)labial fricative	present, absent
F8	Phonemic voice contrast for fricatives	present, absent
F9	Phonemic palatal nasal	present, absent
F10	Maximally 1 liquid phoneme	present, absent

Figure 2.6: Example of pre-determined spoken language features (Ranacher et al., 2021)

name	id	x	y	family	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Paez	pbb	-1898858.494	3725496.327		N	N	N	Y	N	N	Y	Y	Y	N
Cubeo	cub	-1177837.828	3646741.179	Tucanoan	N	N	N	N	N	N	N	N	N	Y
Kamsá	kbh	-1947342.88	3556364.277		N	N	N	N	Y	N	Y	N	Y	N
Inga	inb	-1941990.01	3540287.708	Quechuan	N	N	N	N	N	N	N	N	Y	N
Koreguaje	coe	-1767807.596	3550662.098	Tucanoan	N	N	N	Y	N	N	N	N	Y	Y
Cha'paalachi	cbi	-2188725.28	3477509.242		N	N	N	N	N	N	Y	N	Y	N
Desano	des	-1092883.205	3570461.932	Tucanoan	N	N	N	N	N	N	N	N	N	Y
Tucano	tuo	-1102073.299	3569861.124	Tucanoan	N	N	N	N	N	N	N	N	N	Y
Imbabura Q	qvi	-2105610.469	3444729.321	Quechuan	N	N	N	N	N	N	Y	Y	Y	N
Siona	snn	-1837407.845	3476434.156	Tucanoan	N	Y	N	N	N	N	N	N	N	Y
Cofan	con	-1955649.321	3458542.23		N	N	N	Y	N	N	Y	Y	Y	Y
Napo Quechua	qvo	-1874842.224	3428875.302	Quechuan	N	N	N	N	N	N	Y	N	Y	N
Tsafiki	cof	-2195939.139	3373951.412		N	N	N	N	N	N	Y	N	N	N

Figure 2.7: Language feature values per language, location and known language family (Ranacher et al., 2021)

3. Methods

3.1 Prediction versus retrodiction

Even though *sBayes* is a probability model, the term prediction is in this context misleading. The model does not actually predict any future outcomes of language, instead, it retraces steps. The term explanation is therefore more fitting and will be used throughout this paper.

When validating explanatory models, validation methods consist of building plausible mechanisms that are able to reproduce simulated behavior similar to real behavior (Nuno et al., 2017). Retrodiction, as opposed to prediction, aims to replicate previously observed elements of the target system. When there is a historical record of factual data from the target system, the justification for retrodictive validity in a predictive model is as follows: If the model consistently and accurately replicates the historical record, it can also be relied upon for future predictions (Gross and Strand, 2000). In order to validate *sBayes* we need to create a dataset of which we have prior knowledge. To achieve this we will work with synthetic language datasets of which we control the setup parameters. When we have generated data with a known history and starting point, we validate the content and construct by running the model on data from which we have evolutionary knowledge in multiple situations. We then compare the output with the input to see whether *sBayes* accurately detects the set parameters.

3.2 Agent based models

For our validation experiments, we will use an agent-based model to create synthetic datasets of which we know the exact values of the model output probabilities, in this case meaning the posterior probability distribution between contact, inheritance and universal preference.

Agent-based models (ABMs) are well-suited for capturing the intricacies of social complexity and language (Civico, 2019). The flexibility of ABMs enables us to explore and depict various facets of the phenomenon being studied. For instance, we can construct societies comprised of diverse agents, organizations, networks, and environments, enabling interactions that reflect the heterogeneous nature of social systems. Through the deliberate selection of specific elements from social reality, ABMs become a valuable tool for representing social dynamics as perceived by researchers and stakeholders alike. This process of delineation allows for a tailored modeling approach that aligns with the nuanced understanding of social phenomena (Nuno et al., 2017).

Agents in our model represent language communities with language features as their attributes. The agents are initialized with some common language features to represent universal preference. When agents interact, features are exchanged with predefined probabilities, leading to contact areas. Over time, new agents are created that inherit features from their 'parents'. Furthermore, agents may migrate. For more information on the agent-based model, please see *Understanding through contact* by Joseph Taylor (2023).

3.3 Experiment 1: Validation by isolation

3.3.1 Creating artificial datasets

First, we need to establish whether the artificial language causal probabilities are identified correctly by *sBayes*. In order to test the causal posterior distribution per feature (inheritance, universal preference and contact) we will generate three datasets in which these factors are simulated separately one by one (see fig. 3.1). We will look at *sBayes* accuracy in identifying the likelihoods of contact, inheritance and universal preferences for language features. These datasets will be used to validate the performance of the model by comparing its posterior distribution outcomes with the now-known cases in language contact over change.

The feature state causality weights will be shown in a Dirichlet distribution for multinomial variables every separate model runs. We will also compute a mean squared error from our known input (100%) per feature per experiment.

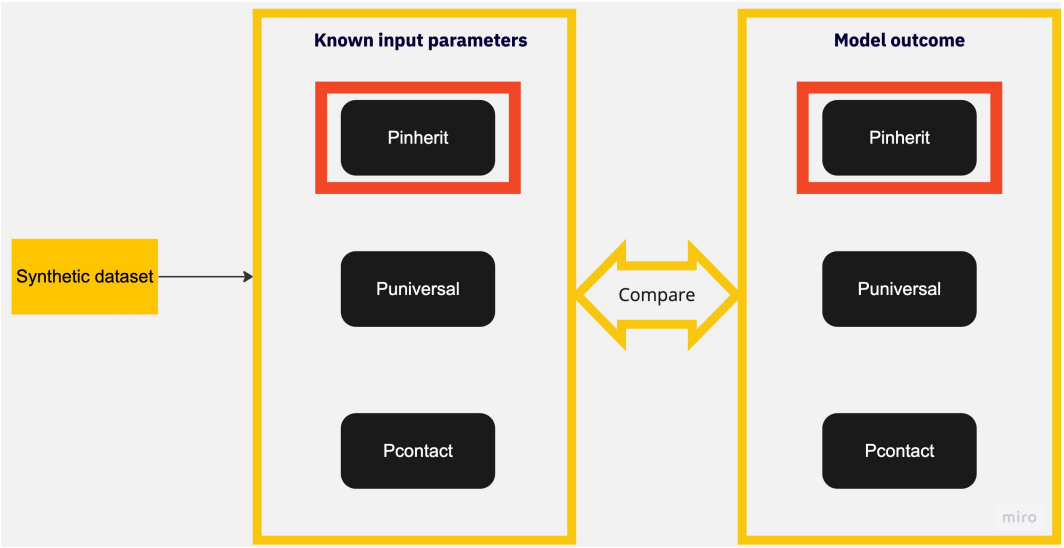

Figure 3.1: Flowchart of isolation detection

3.3.2 Artificial isolation dataset parameters

Table 3.1 shows the possible parameters for our agent-based model and their corresponding values in our 3 isolation datasets. For comparison purposes, each dataset was created with 50 starting agents, which automatically starts 50 language families since every starting agent brings their own language branch.

In our inheritance dataset, inheritance happens with a small number of starting agents before adding children per agent. We achieve this by having a low contact rate between children (new languages) once they are born and a very high rate of feature inheritance (the number of feature states taken from a parent). In our universal preference dataset, all agents have exactly the same feature states per feature. Lowest possible amount of children and every language belongs to its own language family. In our contact dataset we set a high interaction rate, low birth likelihood and low inheritance rate.

	Inheritance	Universal Preference	Contact
Number of agents (start)	50	50	50
Number of agents (end)	535	54	51
Number of language families	50	50	50
Number of features	20	20	10
Max number of states	3	3	3
Grid size	20	20	20
Shortcut pct	0.4	0.4	0.4
Interaction likelihood	0.01	0.01	1
Birth likelihood	0.5	0.01	0.01
Inheritance rate lower	0.9	0.51	0.51
Inheritance rate higher	0.99	0.6	0.6
Child relocation likelihood	0.2	0.2	0.2

Table 3.1: Test caption

3.3.3 sBayes setup

Setting up an experiment in sBayes requires parameters. This will give the model information about how we want to analyse at our language data. It also gives the model information on computing the prior likelihoods. We feed these setting into the model through a configuration file separated into three sections:

1. The setting for our Markov Chain Monte Carlo (MCMC). This is a technique used in statistics and computational science for generating samples from complex probability distributions. The purpose of MCMC sampling is to approximate the distribution of a target variable or set of variables that may be difficult to sample directly (Congdon, 2005). Its purpose in *sBayes* is to generate a series of observations that approximate a given multivariate probability distribution. These settings will be the same for every model run in this paper in order to be able to compare our outcomes (fig. 3.2).
2. Model settings. The amount of clusters we want sBayes to look for where each cluster is a separate model run. We will run the model for clusters 1 through 5. With more area clusters sBayes will find it easier to explain variance in the data. in the data. Universal preference is always modeled as a confounder, while the effect of inheritance can be turned on or off by the analyst (fig. 3.3).
3. Prior settings. Both universal preference and inheritance priors can be uniform or adjusted to empirical data about the language (Gelman et al., 2013. Preference in an area (contact) is unknown as a prior, the contact areas and their defining features will be the output of *sBayes* (fig. 3.4).

The following parameters are used for every experiment in this paper.

Markov Chain Monte Carlo:	
steps	4000000
samples	2000
runs	10
grow to adjacent	0.85
Operators	
cluster steps (growing, shrinking, swapping)	50
weight steps (changing weights)	10
cluster effect (changing probabilities in clusters)	20
confounding effects (probabilities in confounders)	10
source	10
initial objects per cluster	10
Warmup	
warmup steps	10000
warmup chains	10
sample from prior	false

Table 3.2: sBayes setting for our Markov Chain Monte Carlo (MCMC)

Model:	
clusters	[1,2,3,4,5]
sample source	true
confounders	
universal	"<ALL>"
family:	All known families here

Table 3.3: sBayes setting for our model

Priors:	
Clusters	min. size of 3 and a max. size of 100
Geo-prior	cost based
Probability function	exponential
Aggregation policy	mean
Scale	200.0
Cost-matrix inferred from geo-locations:	
Prior on cluster size	uniform area
Prior on weights	uniform
Prior on cluster effect	uniform
Prior on confounding effect universal:	
Dirichlet prior for confounder universal	<ALL>
Prior on confounding effect family:	
Uniform prior for confounder family	All known families here

Table 3.4: sBayes setting for our priors

3.4 Experiment 2: Benchmark run versus synthetic language data

3.4.1 South America experiment parameters

Since our last experiment portrayed theoretical and extreme situations, it should be relatively easy for *sBayes* to pick up on pre-set parameters. In real life however, all three causes will appear together. We want to know whether *sBayes* is able to identify similar contact patterns in a (synthetic) real-life situation.

By doing a benchmark run with *sBayes*, we are able to determine the values of a 'regular' verified model run. For this, we will use the provided South America experiment setup for testing purposes of the model. This dataset is put together from empirical data collected by Ranacher et al. The researchers chose the area of Western South America because of its extreme

genealogical diversity and major split between two cultural/linguistic areas (Andean and Amazonian). They were expecting strong signals to point towards linguistic contact (Ranacher et al., 2021).

After running our benchmark experiment, we will generate a dataset similar to the South American dataset using our agent-based model. We will use the settings shown in figure 3.5. Figure 3.6 shows the shape of the South America language dataset compared to our generated artificial language dataset. We chose a higher number of language families to combat bias created by unknown historical links in language heritage (the missing links that are likely present but unknown to us).

We want sBayes to interpret our synthetic data in a similar way as the official South America experiment. Since we can not reproduce the actual language history, we will not be able to compute any comparisons in posterior weights. We can however compute a deviance information criterion (DIC), which is used to measure model performance considering model fit and complexity. The most suitable K clusters is where the DIC levels off, so that adding more areas does not improve the fit of the model. We run *sBayes* iteratively increasing the number of areas K and evaluate the DIC for each run. We expect the DICs to level out around the same amount of areas.

Number of agents	10
Number of features	36
Max number of states	5
Grid size	10
Shortcut pct	0.4
Interaction likelihood	0.4
Birth likelihood	0.5
Inheritance rate lower	0.8
Inheritance rate higher	0.004
Child relocation likelihood	0.2

Table 3.5: Agent-based model settings for creating our synthetic dataset

	Benchmark experiment	**Synthetic experiment**
Number of features	36	36
Number of feature states	min 2 max 5	min 2 max 5
Number of languages	100	131
Number of language families	6	10

Table 3.6: Details on South America experiment

3.4.2 sBayes parameters compared

Model:		
	Benchmark experiment	Synthetic experiment
clusters	[1,2,3,4,5]	[1,2,3,4,5]
sample source	true	true
confounders		
universal	"<ALL>"	"<ALL>"
family:	Tucanoan	family 0
	Panoan	family 1
	Tacanan	family 2
	Arawak	family 3
	Quechuan	family 4
	Tupian	family 5
		family 6
		family 7
		family 8
		family 9

Table 3.7: Model settings for both of our datasets

Priors:		
	Benchmark experiment	Synthetic experiment
Cluster size	min. 3 max. 100	min. 3 max. 100
Geo-prior	cost based	cost based
Probability function	exponential	exponential
Aggregation policy	mean	mean
Scale	200.0	100.0
Cost-matrix inferred from geo-locations:		
Prior on cluster size	uniform area	uniform area
Prior on weights	uniform	uniform
Prior on cluster effect	uniform	uniform
Prior on confounding effect universal:		
Dirichlet prior for confounder universal	<ALL>	<ALL>
Prior on confounding effect family:		
Uniform prior	Tucanoan.	family 0
Uniform prior	Panoan.	family 1
Uniform prior	Tacanan.	family 2
Uniform prior	Arawak.	family 3
Uniform prior	Quechuan.	family 4
Uniform prior	Tupian.	family 5
Uniform prior		family 6.
Uniform prior		family 7.
Uniform prior		family 8.
Uniform prior		family 9.

Table 3.8: Prior settings for both datasets

4. Results

4.1 Experiment 1: Isolation validation

4.1.1 Inheritance isolation

Our results are shown in two visualizations, the posterior weights density distributions per feature (fig. 4.1) and the error chart per feature (fig. 4.2) where 0 is perfect and 1 is the worst. Our weight distributions show the spread of the likelihood outcomes per feature for universal preference (U), inheritance (I) and contact (C). The spread of the likelihood is shown in green, ranging from light to dark showing the computed weights by density. These can be observed more clearly in later figures 4.3 and 4.5. The purple dot represents the mean point of the spread between these three likelihoods.

The isolation of inheritance as a confounding factor is clearly picked up by *sBayes*. All features overall causalities are almost exclusively inherited. Our errors from 1 (100%) have a mean squared error of 0.0004 and a barely visible spread in our weights visualization.

In our weights table (fig. 4.1) we can observe the means of all density plots to be in absolute corner. However, F1 and F3 show a slight diversion. Our error table (fig. 4.2) supports these slight outliers.

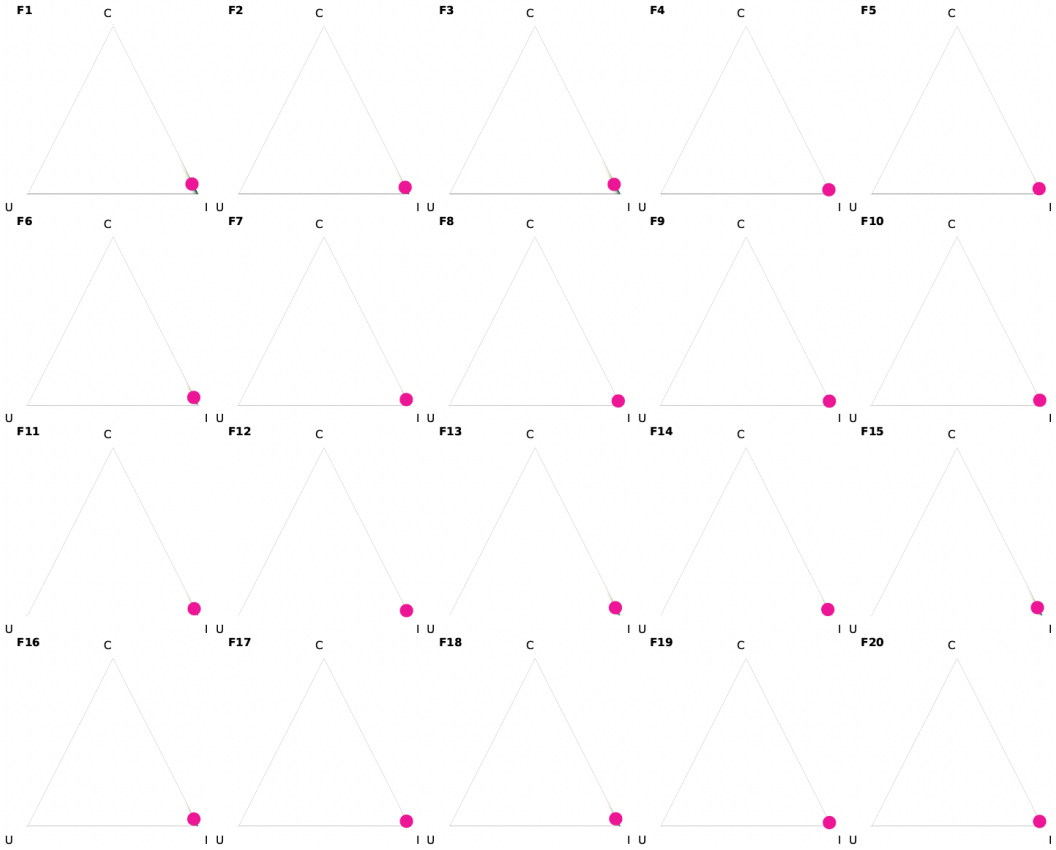

Figure 4.1: Posterior weight plots per feature for for the experiment in which only inheritance takes place

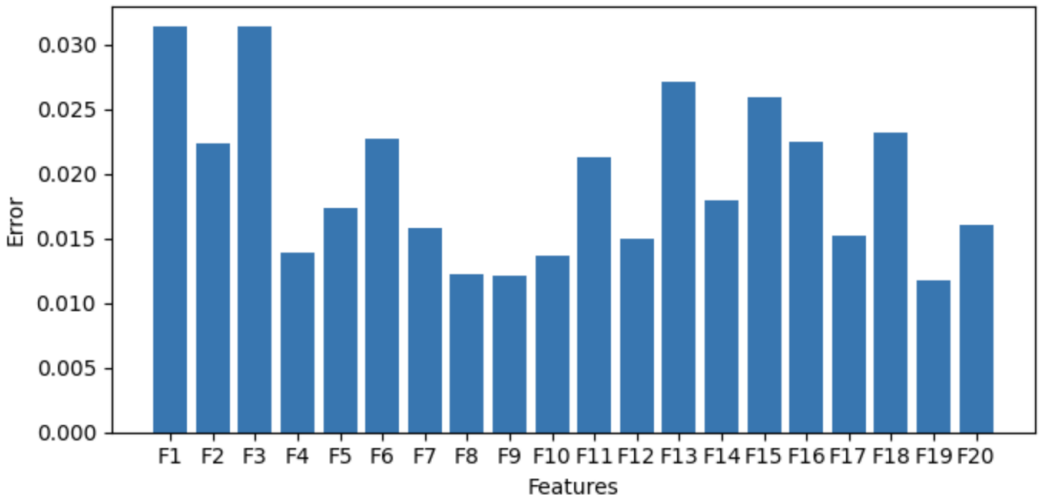

Figure 4.2: Error per feature for the experiment in which only inheritance takes place

4.1.2 Universal preference isolation

Figure 4.3 shows the posterior density weight plots per feature for our universal preference dataset. For universal preference to be detected, the majority of our languages needs to share a similar feature state. When a majority is reached, other languages will convince other languages to assume the same feature state. Looking at fig. 4.4 we see that the errors of our estimated probabilities are low for the majority features, with a mean squared error of 0.01 showing that our likelihoods are accurate.

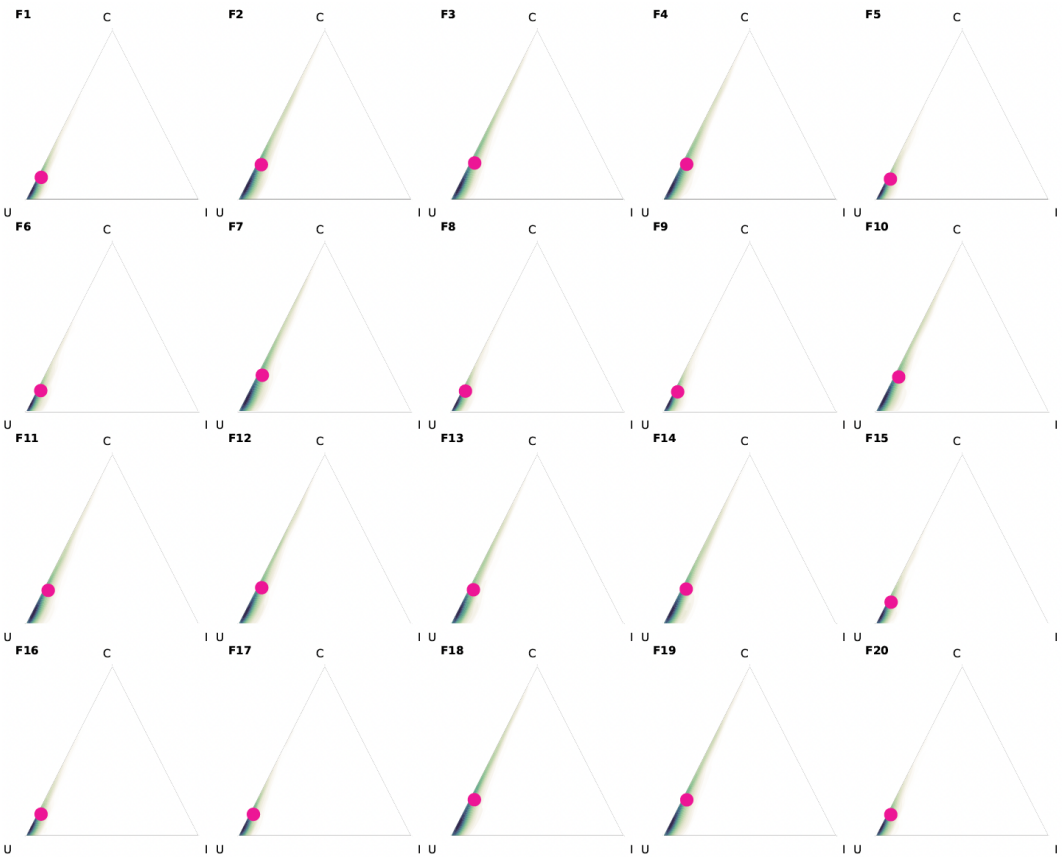

Figure 4.3: Posterior weight plots per feature for the experiment in which only universal preference takes place

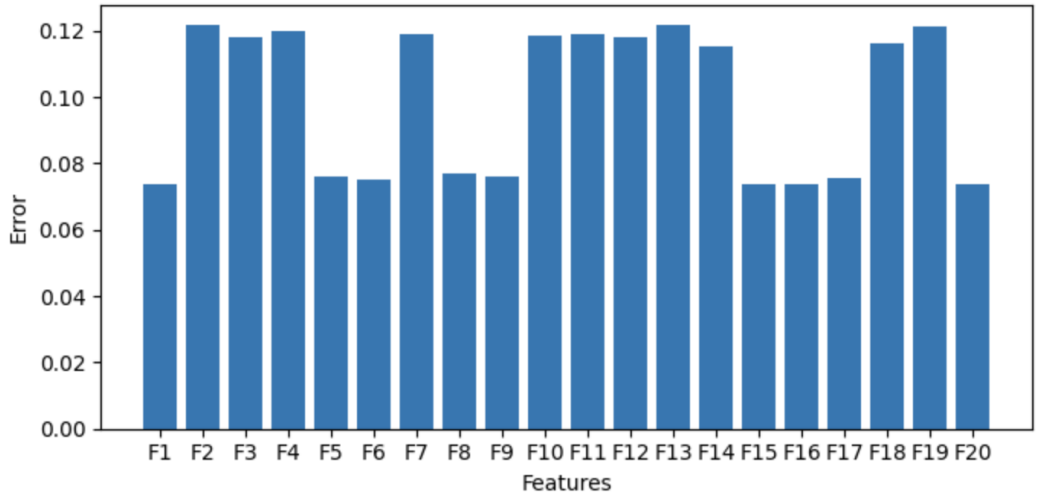

Figure 4.4: Error per feature for the experiment in which only universal preference takes place

4.1.3 Contact isolation

Our contact-focused dataset provides the most varying results. The spread of our weight distributions is visibly higher than in our other isolation experiments (fig. 4.5). Our error graph confirms this (fig. 4.6). While all of our results are still pointing towards contact, some of our features move towards universal preference (F4) and slightly towards inheritance (F3, F7). The mean squared error of our combined features is 0.13.

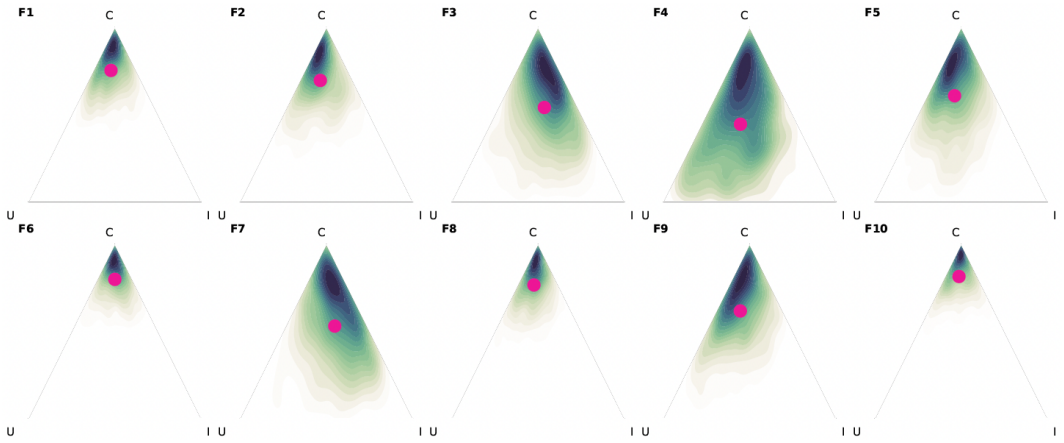

Figure 4.5: Posterior weight plot per feature for experiment in which only contact takes place

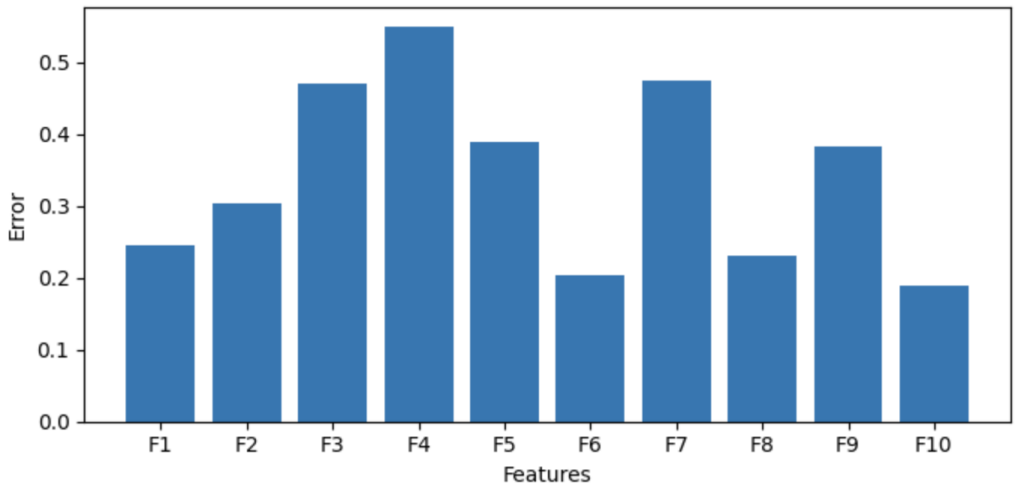

Figure 4.6: Error per feature for the experiment in which only contact takes place

4.2 Experiment 2: Benchmark run versus synthetic language data

We are testing sBayes fit to an artificial language dataset in determining language areas (clusters). In fig. 4.8 and fig. 4.7 we see the deviance information criterion (DIC) for different contact clusters. We observe both models flattening out around $K = 3$. The DIC however, has much higher values and variation in the synthetic dataset compared to our real life situation. This can be cause by the larger dataset.

Weight plots for the South America dataset (fig. 4.9) and our synthetic dataset (fig. 4.10) show that the distribution of feature causalities is comparable between the two. Our causality outcomes are summarized in table 4.1. Here we observe that inheritance is more present in our synthetic dataset. Universal preference however, is mostly absent. Overall though, the distribution between causes is the same. Inheritance occurs most, contact less and universal preference least.

	South America experiment	**Synthetic language set**
Number of I features	25	30
Number of U features	4	2
Number of C features	7	4

Table 4.1: Feature causality count comparison

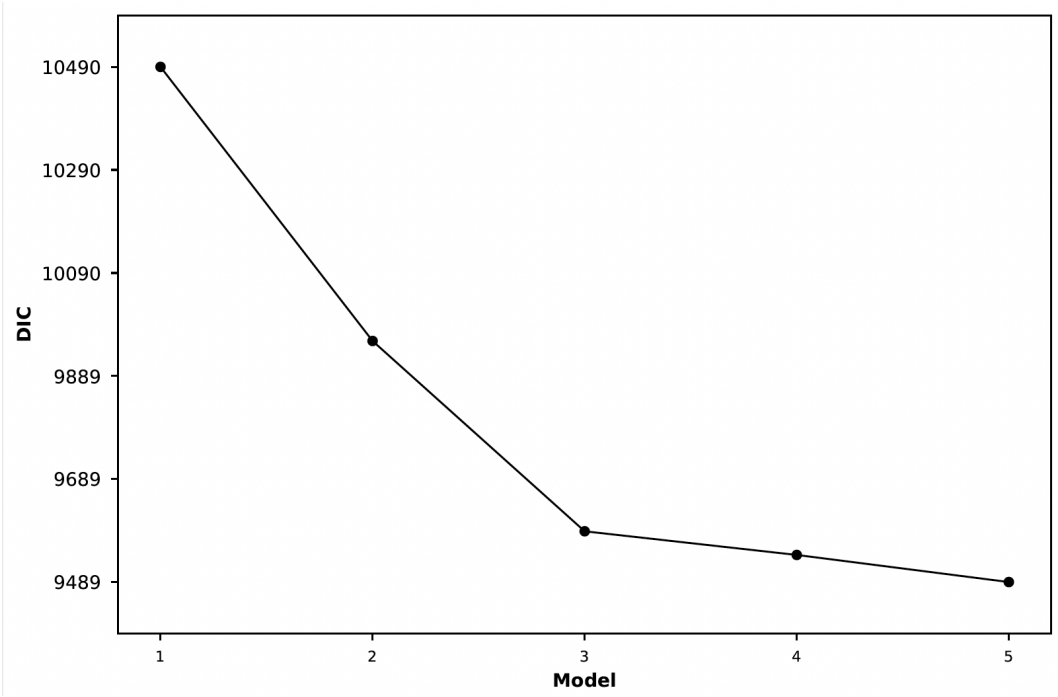

Figure 4.7: Deviance information criterion for the benchmark dataset

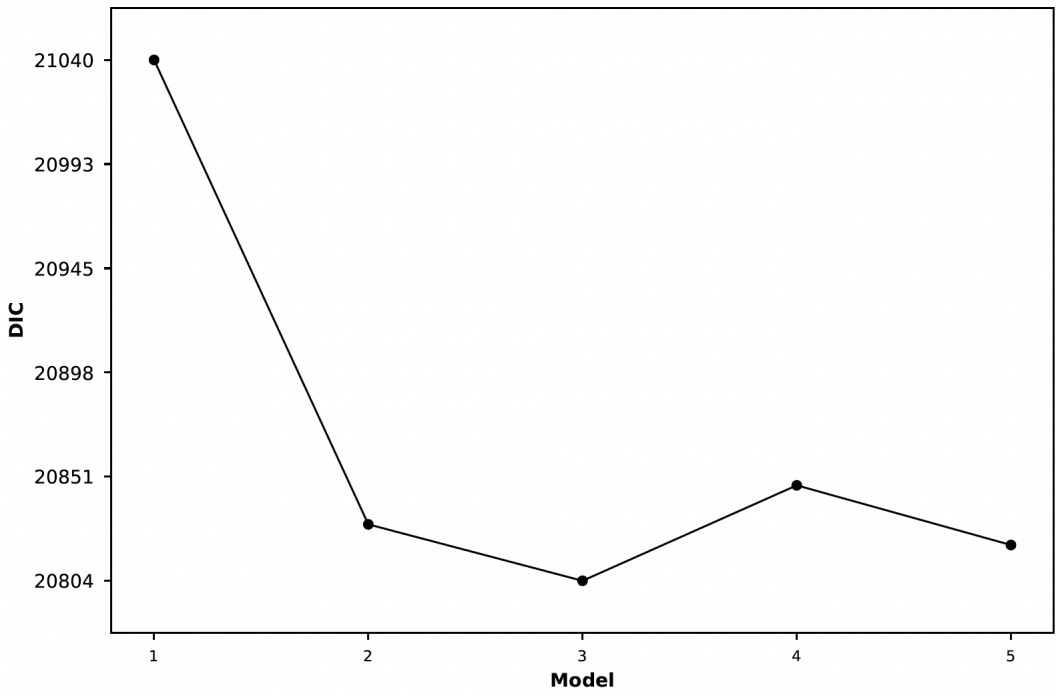

Figure 4.8: Deviance information criterion for the synthetic dataset

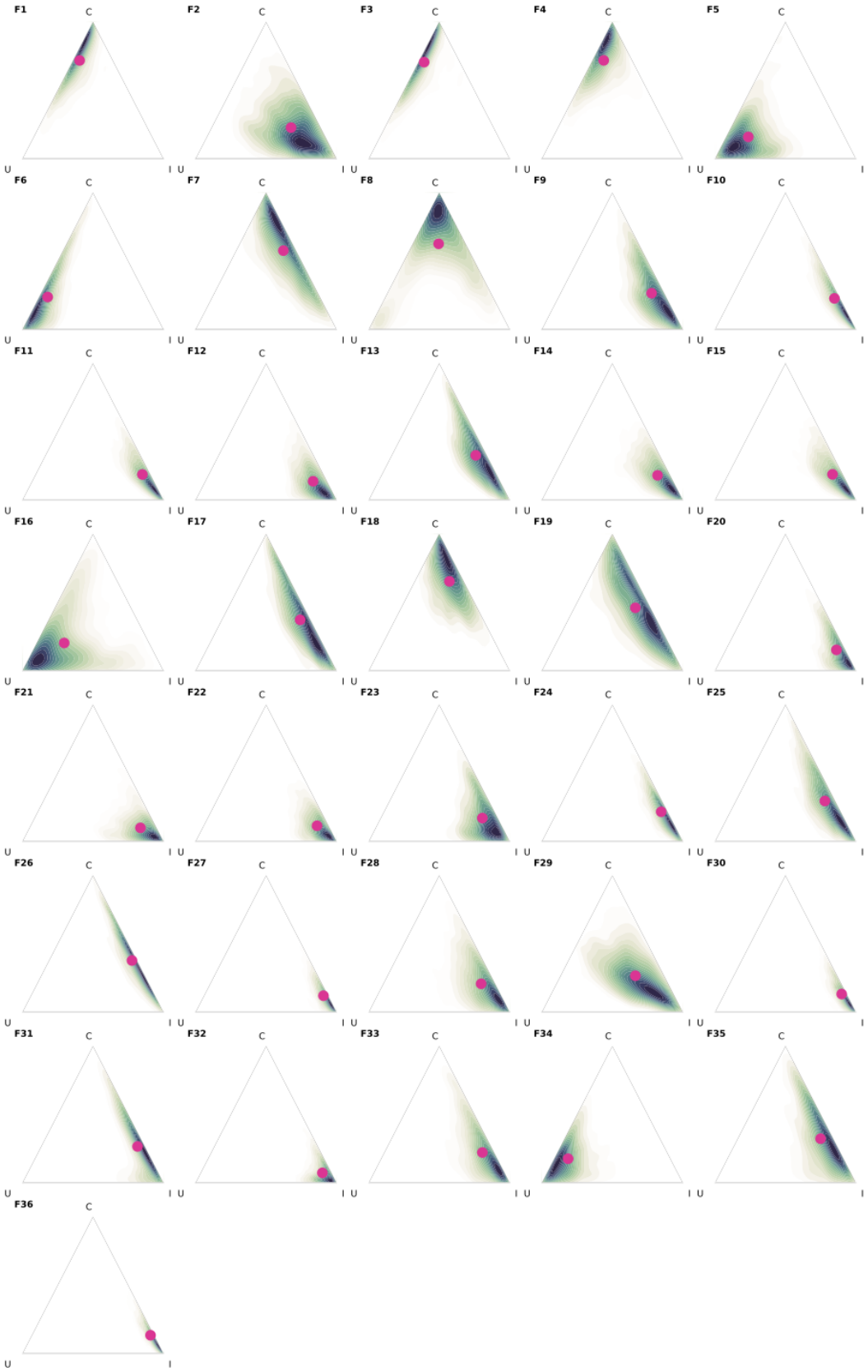

Figure 4.9: Posterior density weight plots for benchmark data

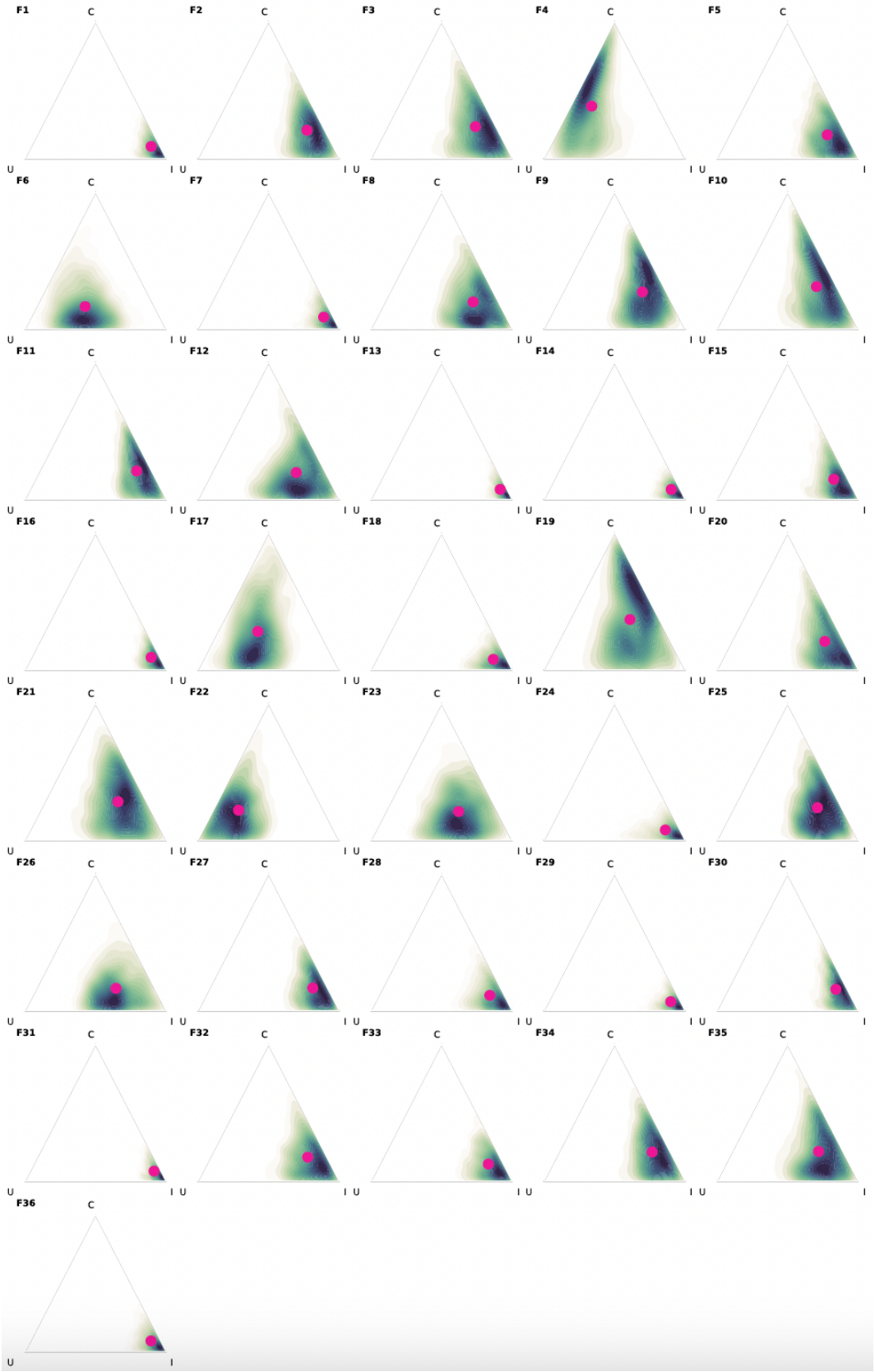

Figure 4.10: Posterior density weight plots for synthetic language data

5. Discussion

5.1 Implications

We presented a validation method for sBayes, a Bayesian mixture model for language contact while accounting for confouders. Since no empirical data can be easily collected on historical language evolution, we proposed the use of synthetic language data to determine the accuracy of sBayes. For this purpose we conducted two experiments, one to validate sBayes ability to detect isolated causal explanations per language feature. The second to test sBayes fit to an artificial language dataset and in determining language areas (clusters). Our results suggest that synthetic language data can successfully be used for validation purposes of the sBayes language model. sBayes accuracy on identifying clearly isolated causalities has a combined mean squared error of 0.05 in our simulations.

In addition to explaining language contact, sBayes can now be further utilized in the exploration of various other fields. When different groups interact, there are numerous aspects beyond language that can be examined through sBayes, as long as they can be represented as features. Culture is one such dimension: whenever individuals come into contact, they often exchange not only language but also artifacts, cultural customs, ideas, rituals, mythology, and more.

Further research can be done by generating more datasets for specific purposes and testing their outcomes in sBayes. More research can also be done on parameter validation for the agent-based model that was used for generating our datasets.

5.2 Context

Analysis leads us to the conclusion that there is still a need for a systematic quantitative approach to identify contact areas, one that considers both the process leading to contact effects and the influence of other factors that can confound the results. A preliminary attempt to address this research gap was presented in (Daumé, 2009), where a non-parametric Bayesian model was used to reconstruct language areas, capturing both areal and phylogenetic effects, but without distinguishing universal preference and inheritance. A similar concept was explored in (Towner et al., 2012), where an autologistic model, along with family and neighbor graphs, was employed to examine the impact of inheritance and areality on cultural macroevolution in North America. Although this model does not directly infer areas, it assumes that spatial influence occurs within a fixed radius of 175 km. Later, this approach was extended to infer hidden areas based on language data (Murawaki, 2020).

An alternative approach is proposed in (Michael et al., 2014), where a set of languages is assigned to a potential contact area, referred to as a "core," based on prior knowledge. Subsequently, a naive Bayes classifier is used to determine whether other languages belong to the core or a control set, consisting of languages unlikely to have been in contact with the core. The same authors also suggested a relaxed admixture model to identify language contact (Michael and Chang, 2014). This mixture model focuses on detecting borrowings between pairs of languages but does not account for the possibility of larger contact areas.

Ranacher et al. (2021) developed a method for determining the amount of horizontal transfer between languages. *sBayes* estimates the relative role of language contact, as opposed to inheritance through family and universal preference of one's surroundings, in creating similarities between languages. *sBayes* draws inspiration from these approaches. However, it differs in that it explicitly infers the assignment of languages to contact areas based on the available data. The shape and size of the areas are not predetermined but can vary. Additionally, our model incorporates a geographical

prior to enforce spatial coherence and account for the influence of geography. Furthermore, the model effectively addresses the confounding factors of inheritance and universal preference, ensuring that only signals of language contact are detected.

sBayes was tested on simulated data and on two case studies to reveal language contact in South America and the Balkans (Ranacher et al., 2021). Yet, neither of these experiments fully validated the approach since the actual contributions of language contact, inheritance and universal preference are unknown.

Building upon previous work, the validation of the sBayes model provides a valuable contribution to the existing knowledge by offering a comprehensive and flexible tool for analyzing language contact and other domains influenced by social complexities.

5.3 Limitations

This research assumes the validity of the agent-based model used to create our validation data sets. It has not been fully tested and validated.

For clarification, sBayes uses a greatly simplified projection of language and language feature development. The model also greatly simplifies geographical contact and does not take into account the social status of language speakers, speaking goals, trade networks, or geographical landmarks like rivers or mountain ranges.

The model can not plot if any of the three confounding factors turn out to be an actual 0. We therefore had the adjust the datasets to not output any zeroes but as close to 0 as possible. This only caused problems with our isolation experiments and would be highly unlikely to occur in real life data.

The reliability of our synthetic data is impacted by 'islanding' or isolated clustering of the agents (fig. 5.1) in dataset generation. Since the languages do not move, they can only interact when they are close to each other. If the grid size is set too high, agents will be randomly generated and spawn

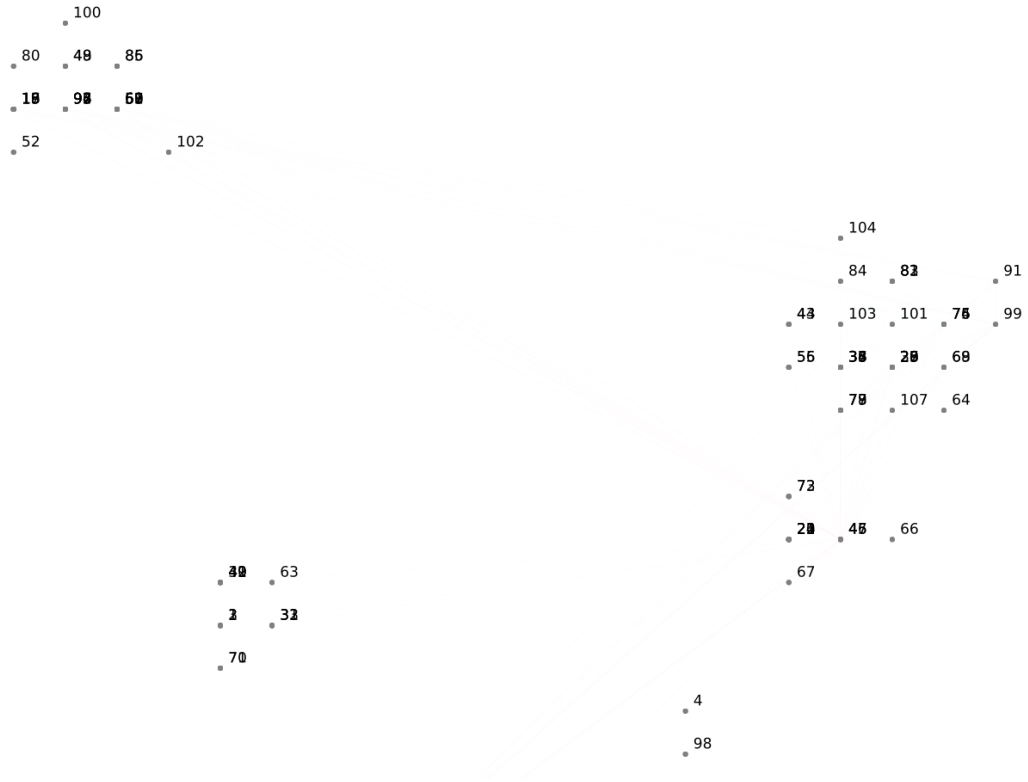

Figure 5.1: "Islanding" or isolated clustering of agents without ever contacting neighboring families

children around them over time. They may never come into contact with other language families in different locations on the grid. This can cause our results to have a relatively low representation of contact and an overinflated presence of inheritance.

6. Conclusion

sBayes recognized our experimental artificial language data in every scenario. By looking at the weight distributions per feature in 3 hypothetical situations, we determined that *sBayes* correctly identified the causalities of inheritance, universal preference and contact. We see that the average causality per feature clearly points towards our simulated feature state heritage with a combined mean squared error of 0.05.

Our second experiment tested sBayes' fit to an artificial language dataset in determining language areas (clusters). We performed a benchmark experiment with an empirically correct dataset and reproduced these dataset characteristics with an agent based model. We then calculated the deviance information criterion (DIC) for different contact clusters and observed both models flat out around K = 3. The overall distribution of feature state causality is the same in our synthetic data when we compare it to a benchmark experiment,

The accuracy of *sBayes* in our situations has proven to be high. The model performs well in recognizing historical language data and reproducing real life language situation.

Bibliography

Bowern, C., & Evans, B. (2015). The routledge handbook of historical linguistics. *Routledge*.

Matras, Y. (2009). Language contact. https://doi.org/10.1017

Ranacher, P., Neureiter, N., van Gijn, R., & Sonnenhauser, B. (2021). Contact-tracing in cultural evolution: A bayesian mixture model to detect geographic areas of language contact. *Journal of the Royal Society Interface*. https://doi.org/10.1098/20201031

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., & Tarantola, S. (2008). Global sensitivity analysis: The primer. *John Wiley and Sons*.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). Bayesian data analysis (3rd ed.) *Chapman and Hall/CRC*.

Friedman, V. A. (2011). The balkan languages and balkan linguistics.

Muysken, P., Hammarstrom, H., Birchall, J., Gijn, R. V., Krasnoukhova, O., & Muller, N. (2014). Linguistic areas: Bottom-up or top-down? https://doi.org/10.1017

Masica, C. (2001). The definition and significance of linguistic areas: Methods, pitfalls, and possibilities. *Tokyo Symposium on South Asian languages: contact, convergence, and typology*.

Croft, W. (2008). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*. https://doi.org/10.1146

Bickel, B. (2015). Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. *The Oxford handbook of linguistic analysis*.

Nuno, D., Nuno, F., & Rosa, A. C. (2017). Verifying and validating simulations. *Simulating social complexity*.

Gross, D., & Strand, R. (2000). Can agent-based models assist decisions on large-scale practical problems: A philosophical analysis. *Complexity*.

Civico, M. (2019). The dynamics of language minorities: Evidence from an agent-based model of language contact.

Congdon, P. (2005). Bayesian model assessment via parameter estimation and model comparison. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Daumé, H. (2009). Non-parametric bayesian areal linguistics.

Towner, M. C., Mark N. Grote, J. V., & Mulder, M. B. (2012). Cultural macroevolution on neighbor graphs.

Murawaki, Y. (2020). Latent geographical factors for analyzing the evolution of dialects in contact.

Michael, L., Chang, W., & Stark, L. (2014). Exploring phonological areality in the circum-andean region using a naive bayes classifier.

Michael, L., & Chang, W. (2014). A relaxed admixture model of language contact.