Can We Identify Transmission Bias in the Archaeological Record? An Investigation Using Boosted Classifier Models

Mark E. Madsen[1,*]
1 Department of Anthropology, Box 353100, University of Washington, Seattle, WA 98195-3100, USA

*** mark@madsenlab.org**

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Curabitur eget porta erat. Morbi consectetur est vel gravida pretium. Suspendisse ut dui eu ante cursus gravida non sed sem. Nullam sapien tellus, commodo id velit id, eleifend volutpat quam. Phasellus mauris velit, dapibus finibus elementum vel, pulvinar non tellus. Nunc pellentesque pretium diam, quis maximus dolor faucibus id. Nunc convallis sodales ante, ut ullamcorper est egestas vitae. Nam sit amet enim ultrices, ultrices elit pulvinar, volutpat risus.

1 Introduction

A major use of cultural transmission models in archaeology is inference regarding the mode of transmission operative within past populations. Identifying cognitive biases is central, for example, to several hypotheses for the origin of cumulative cultural transmission and complex culture [1–4]. In more recent archaeological contexts, the identification of frequency-biased social learning has been used to support inferences concerning sociopolitical structure in past societies [5]. Simulation and mathematical studies have yielded many insights into the empirical patterns we can expect from different transmission models [6–10], although much of this knowledge is derived from very simplified population models. In particular, theoretical analyses of transmission models have ignored until recently the effect of data collection methods and coarse-grained observations on the patterns we should expect in archaeological data. As a result, we simply do not know whether the mode of transmission can be reliably inferred from samples of the archaeological record, if it is possible at some time scales but not others, or how we might tailor data collection strategies to maximize the accuracy of such inferences.

We do know that the coarse graining of observable variables that occurs given time averaging reduces our ability to distinguish between unbiased and biased transmission models [11–13], that nonstationary population sizes reduce our ability to infer transmission modes [14], and that diachronic statistics and nonequilibrium models are better than synchronic measures and equilibrium models [15, 16]. The effect of these factors is such that when deposits are highly time averaged, equifinality occurs, with different models yielding the same empirical distributions despite describing different underlying processes [17]. Equifinality between theoretical models is a serious concern whenever we study complex systems, and has been discussed in geomorphology, hydrology, climatology, and within archaeology itself [18–25]. If models which represent different modes of cultural transmission cannot be distinguished when we include aggregation, heterogeneity, or sampling in our models, then there may be questions concerning past cultural transmission that we cannot answer. As a result, there may be classes of models which are useful for contemporary or historical research, but not for the coarse grained scales of observation that archaeologists often confront.

Existing theoretical studies have almost exclusively focused upon distinguishing models based on the ability of a single statistic or variable to distinguish the distribution of outcomes from different social learning modes. Scores from the power law exponent in a log-log plot of trait frequencies have received the most attention along with more recent application of neutrality tests [6, 8, 10, 26, 27]. More recently, Kandler's work has demonstrated that diachronic measures such as trait survival time, or the length of time the most common trait stays ranked the most common, can be robust predictors of different classes of transmission models [15, 16]. But there is little reason to suspect that single statistics will be adequate in most cases to cleanly separate and identify different transmission models, given the strong convergence in distribution that characterize diffusion processes. Instead, we should expect that statistical models employing multiple predictors would be the best discrimination tools, if any exist for a set of transmission models. In this paper, I employ a robust machine learning classifier algorithm and multiple ways of measuring trait richness, diversity, and survival times to test whether equifinalities exist between various combinations of unbiased and biased transmission rules when measurements come from realistic data collection scenarios.[1]

The results indicate that while neutral and biased transmission models can be distinguished very accurately given measurements from entire populations taken when no temporal aggregation occurs, the introduction of sampling and the interaction between sampling and time averaging markedly degrades our ability to distinguish these transmission rules. Furthermore, the degradation is not symmetric. With sampled, time averaged data, we are extremely likely to conclude that samples represent biased transmission, even when this is not the case. Other mixtures of conformist and anti-conformist transmission rules are even less distinguishable given time averaging and limited samples. As a result, I conclude that it may be difficult or impossible to infer the details of cognitively biased transmission rules from frequency data alone, when we lack data from an entire population and when only coarse grained, aggregated data are available.

2 Analysis

2.1 Reducible and Irreducible Equifinality Among Transmission Models

Equifinality among cultural transmission models can arise from several sources. First, equifinality may occur because of our measurement and analysis procedures. There is growing evidence, for example, that assemblage duration affects our ability to distinguish biased from neutral transmission across a variety of statistical predictors [11–13]. Equifinality among transmission models is thus possibly reducible by collecting finer-grained samples during fieldwork, if deposits are well stratified. However, in situations where the depositional environment actively creates temporal aggregation (e.g., in the plowzone, or in deflated aeolian contexts), there may be little that an investigator can do to improve the temporal resolution of data collection. And when we employ published data sets, obviously we cannot easily subdivide the data into assemblages finer than the original investigation supported. When studying living populations, of course, equifinalities may be addressed by converting a purely observational study to a controlled experiment in some cases [28–30], but of course this is not an option in archaeological contexts.

Second, equifinality is partially determined by the predictors or variables we use in trying to separate the behavior of models. Fig. 1 shows an artificial example with two distributions, measured on two variables. The marginal distribution of each variable demonstrates how models might be distinguishable given one variable (Y axis) but not another (X axis). In the published literature on transmission modes, single variables are usually examined, but we gain huge power

[1] Throughout this paper, I used "classification" in the statistical and machine-learning sense of a statistical model whose dependent variable is a binary or discrete value, such that the model predicts which value a data point takes from a labelled set. Archaeologists will be used to using the term in the sense of systematics and taxonomy, which is not the intent here.

Figure 1. Simple example of the effect of variable choice in distinguishing models. The variable on the X axis displays quite a bit of overlap between models, while the variable on the Y axis distinguishes the models with fairly high accuracy.

by considering statistical models with multiple variables.

Not all equifinalities may be reducible. The statistical distributions generated by diffusion processes can be highly convergent among related models, and almost all cultural transmission models are, at base, diffusion processes. This type of equifinality is **irreducible**, and is not solved by changing how we perform the analysis or by changing data collection. Irreducibly equifinal models form an **equivalence class** of models that we cannot distinguish given our data. Instead, all we can say is that our data could have been generated by any of the models in the equivalence class. If the equivalence classes of equifinal models are coarse enough (at worst, if they form a single group), then we cannot meet our original inferential goals at all.

In some cases, irreducible equifinalities can become reducible given advances in measurement technologies that open up new sets of predictor variables. After his seminal works of the 1970's on drift and the infinite-alleles neutral model, Warren Ewens stopped working on neutrality tests because tests using allele count data lack statistical power. Ewens moved to studying the population genetics of human diseases instead [31], recognizing that further progress would require sequence data unavailable at the time. This judgment proved accurate: a new suite of neutrality tests did arise starting the late 1980's and 1990's when sequence data became widely available [32, 33].

2.2 Equifinality As Classification Error

Since our evolutionary models of cultural transmission are stochastic, and generate a variety of outcomes for the same parameter values, I take a statistical approach to examining equifinality of transmission models in archaeological data. Transmission modes are separable and thus identifiable in archaeological data if the distribution of model outcomes are non-overlapping, when measured in a space created by a set of predictor variables. With stochastic models like the ones currently used by archaeologists, the most efficient method of studying the outcome distribution is to simulate values from the model, and examine our ability to correctly predict which model generated each data point, given a function of the predictor variables.

This general approach can be visualized as in Fig. 2. Here, three pairs of probability models are represented by 500 measurements each of two continuous predictors variables (e.g., a diversity index). In the left panel, the pair of models do not overlap in their outcomes. Given a data point, we can assign it to Model 1 or Model 2 with virtually no error, and thus we would consider models 1 and 2 to be distinct and not equifinal at all. The situation in the middle and right panels of Figure 1 is different. There is some overlap in the middle panel, and very strong overlap in the right panel. In the right hand panel, in fact, there is enough overlap that on average, our ability to assign a randomly chosen data point to the correct model is no better than chance. Intuitively, we would say that there is some equifinality in the middle panel, and that the two models were strongly equifinal in the right hand panel.

Figure 2. Simple example of model outcomes with different degrees of distinguishability: (A) simulated data point from two fully separate models, (B) two models with a limited overlap region, (C) and two models whose outcomes are highly overlapping.

We can formalize the analysis of overlap between models as a problem of "classification" or "pattern recognition" in the sense of statistical or machine learning [34]. Given a set of models $\mathscr{M}_1 \ldots \mathscr{M}_n$, we can measure equifinality as the minimum possible error achievable in correctly assigning simulated data points to the models which generated them, given measurement of a set of predictor variables. In general, the classification problem asks which model has the highest

probability for a given data point, given the conditional density of the data and models. This sounds exactly like Bayes' theorem, and in fact we can write the classification problem as follows, where $Y \in 1,\ldots,K$ refers to each of k models, and $X_1,\ldots,X_p$ refer to p different predictor variables.

$$\mathbb{P}(Y|X_1,\ldots,X_p) = \frac{\mathbb{P}(Y_i)\mathbb{P}(X_1,\ldots,X_p|Y)}{\mathbb{P}(X_1,\ldots,X_p)} \quad (1)$$

$\mathbb{P}(Y)$ plays the role of the prior distribution, and is the prevalence of each model in the population. This is a constant in situations where we are simulating values from each model to test for equifinality. The data points in a classification problem are given, and thus the denominator is a constant. The most probable class for a given data point is just the mode of the likelihood function, which is given by:

$$Y_{pred} = \arg\max_{y} \mathbb{P}(X_1,\ldots,X_p|Y) \quad (2)$$

This is the *Bayes classifier* for a controlled simulation experiment, and its error rate in separating data points by model is called the *Bayes error*. This is the lowest possible error in separating the models given the data [34–36]. The Bayes error is zero when we can correctly identify each data point as to its model of origin (as in the left panel of Fig. 2, and rises as two models overlap in the measurement space. With sufficient overlap, the Bayes error could approach 0.5, which represents a prediction rule which is no better than chance.[2]

Unfortunately, we can almost never directly calculate the Bayes error rate for a prediction or classification rule, because we rarely have an expression for the likelihood function of our transmission models in the spaced formed by the predictor variables. Bayes error can be directly calculated, in fact, only for a small number of cases, such as Gaussian distributions with a shared covariance matrix.[3] Despite the fact that we can rarely calculate the Bayes error rate, it is useful as an operational definition for equifinality, since it measures our uncertainty about model choice given a set of measurable variables. In practice, we approximate the Bayes error by employing algorithms which are known to have near-optimal performance in classification problems. In particular, boosting, bagging, and ensemble approaches that combine many classifier rules are attractive since each achieves some of the best generalization error in prediction tests [34], and thus come closest to estimating the Bayes rate [41].

2.3 Study Design

To examine whether a classifier approach with multiple predictor variables, allows us to identify modes of cultural transmission in archaeological data, this study simulates four different transmission models, and compares our ability to correctly identify data points in a series of three pairwise comparisons between models. The transmission models include:

1. Unbiased or neutral cultural transmission
2. Mixture of equal numbers of conformists and anti-conformists.
3. A mixture dominated by conformists, but with 30% anti-conformists.
4. A mixture dominated by anti-conformists, but with 30% anti-conformists.

[2]Predictors can achieve even worse error levels, performing more poorly than coin-flipping, but in the current study we will not encounter such rates.

[3]There is a large literature, especially in pattern recognition and language classification, on approximating upper bounds for the Bayes error of a classifier, because it is highly useful to know when you cannot improve a recognition system or classifier any further [37–39]. Most such upper bounds are based upon parametric models, and use estimates of a distance metric between the classes being distinguished (typically, the Mahalanobis or Bhattacharyya distance) [35]. Such bounds are difficult to justify in situations where we have complex social learning models, whose probability density functions in the space of measured variables are typically unknown and are unlikely to be Gaussian. Nonparametric bounds are possible, using nearest-neighbor methods [40], but in most cases the values obtained are not very tight and the performance of boosting and bagged classifiers easily surpasses such methods.

Unlike previous theoretical studies of cultural transmission models in anthropology, I study mixtures of transmission modes since real populations are typically heterogeneous. We might also expect that certain mixtures of biased transmission might be statistically difficult to distinguish. For example, an equal mixture of conformist and anti-conformists might cancel out each other's biases, and thus be difficult to separate from a population of unbiased individuals who copy one another randomly. This study focuses, therefore, on three pairwise comparisons drawing upon the four models just described:

Comparison	Model #1	Model #2
Neutral vs. Biased	Unbiased transmission	All 3 biased models
Neutral vs. Balanced Bias	Unbiased transmission	Equal number of pro/anti conformists
Pro/Anti Conformism	Conformist dominated	Anti-conformist dominated

Table 1. Model comparisons tested in this study for equifinality.

In addition to comparing different mixtures of transmission modes, we want to determine the extent to which equifinality (if it occurs) is affected by the conditions of data collection. Thus, for each set of simulations from a model, a standard set of predictor variables is measured for a variety of levels of sampling from the simulated population, and temporal aggregation of copying events. The extent to which equifinality occurs for some data collection regimes, but not others, is an indication that the equifinality is potentially reducible by augmenting sample size or obtaining finer-grained, less time averaged samples. To the extent that equifinality exists regardless of data collection strategy, the study might identify irreducible equifinalities between specific model pairs.

The general process followed throughout the study is:

- Simulate a large number of samples from each cultural transmission model, given a range of parameters.
- Measure a set of archaeologically relevant variables (e.g., richness, diversity) on each sample
- Perform each variable measurement across different data collection regimes (e.g., duration of accumulation, sample size)
- Train a predictive classifier model for each data collection regime, to predict the model of origin given the measured variables
- Assess the classifier error rate using additional samples simulated from each transmission model

2.4 Methods

2.4.1 Simulated Samples of Cultural Transmission Models

The outcomes of all four transmission models are drived by simulating the dynamics of the model in an agent-based framework that allows each agent to be assigned a different transmission rule. All simulations employ the Moran dynamics, where one individual engages in a copying event at each elemental step [42–44]. Innovations are modeled using the "infinite alleles" approximation, where every innovation is new to the population [45]. Simulations were performed using the CTMixtures software package, available as open source software.[4] The parameters for all simulation runs are given in Table 2. Where there is a range given (e.g., innovation rate), the parameter is treated as a prior distribution and each simulation run is assigned a uniform random value from the range. This ensures good coverage of the parameter space given 25,000 replicates for each of the 4 models.[5]

[4] `https://github.com/mmadsen/ctmixtures`

[5] The use of a good prior distribution for parameter ranges also results in simulation data that are usable for later data fitting by approximate Bayesian inference [46–49].

Parameter	Value or Interval
Innovation rate (in θ scaled units)	$[0.1, 5.0]$
Probability of conformism	$[0.05, 0.25]$
Probability of anti-conformism	$[0.05, 0.25]$
Sample fractions	0.1 and 0.2
Time averaging intervals (units of 100 individuals)	10, 20, 50, 100
Population size	100
Number of trait dimensions (loci)	4
Initial traits per dimension	10

Table 2. Parameters for simulation runs across the four models studied. Intervals are treated as prior distributions, and each simulation run is assigned values derived from a uniform random sample on the interval indicated. Lists of values are all applied to every simulation run (e.g., there is both a 10% and a 20% sample from each simulation run. Single values are applied to every simulation run, and represent a point prior.)

Simulated populations are 100 individuals in size, because most archaeological studies of cultural transmission have focused upon situations where population sizes are assumed to be small. Each simulated individual carries 4 different traits at any time, which are treated as separate loci or dimensions. Trait frequencies are tracked on a per-locus basis, and combinations of loci are tracked in order to simulate archaeological "types" or classes which include multiple dimensions of variation.

Regardless of transmission model, social learning involves no interaction effects between loci in this study. The population is seeded with 10 randomly chosen traits at each locus as a starting configuration. The evolution of each simulated population proceeds for 4 million elemental steps, which is equivalent to about 40,000 copying events on average per individual. This value was chosen by performing simulations at 1 million time step intervals and verifying that the distribution of a key statistic (the number of traits per Loci) had stabilized. This occurred in most cases between 2 and 3 million steps, and in all cases between 3 and 4 million, so the last value was chosen.[6] At the end of 4 million simulation steps, a suite of variables are measured from each of the 25,000 replicates and stored for analysis.

2.4.2 Variable Selection

Since most previous work on identifying transmission mode from archaeological data employ single diagnostic variables, and begin to display equifinality under realistic data collection conditions, it is reasonable to examine whether using multiple variables will yield more discriminatory power in the same contexts. By representing the outcomes of transmission models in a higher dimensional space, it should be easier to find a decision boundary ("separating hyperplane") that correctly predicts the model which generated each data point, if such a boundary exists.

The predictor variables chosen in this study focus upon measures of richness and diversity, trait survival over time [15], and the Slatkin neutrality test [26, 27]. Each has been employed in the archaeological literature on identifying cultural transmission modes, or is a variant on such measures (e.g., IQV is a normalized version of Shannon entropy), and crucially, all are measurable in standard archaeological contexts using type frequency data. This additionally makes most of the variables applicable to the re-analysis of already published data, which is an important usage scenario in archaeological research.

For the locus-centric variables, each statistic was applied to each locus separately, and the mean, minimum, and maximum of the values obtained for each locus were recorded. I collect order statistics in addition to the mean value, since it is possible that minima and maxima might

[6] The analysis underpinning this decision is availble in the Github repository at `https://github.com/mmadsen/experiment-ctmixtures/analysis/verification`.

Variable	Model Variable
Cross-Tabulated Class Richness (Class)	num_trait_configurations
Slatkin Exact (Class)	configuration_slatkin
Shannon Entropy (Class)	config_entropy
IQV Diversity (Class)	config_iqv
Neiman T_f (Class)	config_neiman_tf
Slatkin Exact (Max for Locus)	slatkin_locus_max
Slatkin Exact (Min for Locus)	slatkin_locus_min
Slatkin Exact (Mean for Locus)	slatkin_locus_mean
Shannon Entropy of Trait Frequencies (Min)	entropy_locus_max
Shannon Entropy of Trait Frequencies (Max)	entropy_locus_min
Shannon Entropy of Trait Frequencies (Mean)	entropy_locus_mean
IQV Diversity Index (Min)	iqv_locus_max
IQV Diversity Index (Max)	iqv_locus_min
IQV Diversity Index (Mean)	iqv_locus_mean
Trait Richness (Min)	richness_locus_max
Trait Richness (Max)	richness_locus_min
Trait Richness (Mean)	richness_locus_mean
Kandler-Shennan Trait Survival (Min)	kandler_locus_max
Kandler-Shennan Trait Survival (Max)	kandler_locus_min
Kandler-Shennan Trait Survival (Mean)	kandler_locus_mean
Neiman T_f (Min)	neiman_tf_locus_max
Neiman T_f (Max)	neiman_tf_locus_min
Neiman T_f (Mean)	neiman_tf_locus_mean

Table 3. Variables measured from each transmission model simulation sample. The parenthetical expression records whether the variable was calculated for cross-tabulations of all 4 loci (Class) or represent the order statistics from individual loci (Min/Mean/Max). The right column records the variable name used within R statistical models, for examining the relative importance of each variable in classifying observations.

be a better discriminator between models than averages. In addition to the variables calculated upon each of the 4 loci, the traits at each locus were combined into a cross-tabulation of "classes" which simulates the process of archaeological classification. Each class represents a different combination of traits from the 4 loci, and very roughly simulates observing cultural variation through the lens of a standard paradigmatic classification [50]. The same variables are then measured as a function of the class counts.[7] This allows us to understand whether transmission models are better distinguished on a per-locus (dimension) basis or by operating on more complex classes that combine several traits together. The full list of measured variables is given in Table 3.

As a final note on variable selection, in an exploratory analysis for this project, I tried to include the power law exponent from a log-log transformation of trait frequency, given the important work by Bentley [8] and Mesoudi and Lycett [10]. It is not clear, however, that previous uses of this variable have been comparable to measurements we can make on archaeological assemblages. As an example, Mesoudi and Lycett [10] use the cumulative number of adoptions of each trait over the entire timespan of the simulation as the "frequency" used to calculate power law exponents.[8] Given the measurement strategies described in Table 4, the number of traits present at any given time is often small, and their prevalence in a small population makes it difficult to fit a power law to the data. Despite its importance in

[7] The sole exception is the Kandler-Shennan survival time, which is not measured here for the cross-tabulated classes. Understanding the quantitative behavior of this measure for multidimensional classes of traits is an important open research question, however.

[8] I confirmed this by inspection of the source code for their simulation model, which was provided by Alex Mesoudi.

archaeological discussions of neutral versus biased transmission, I have omitted power law exponents from the published analysis, pending investigation of the proper method for calculating them in situations with small N and small numbers of trait categories.

2.4.3 Data Collection Treatments

At the end of each simulation run, after the model has reached a quasi-stable equilibrium (measured as stability in per-locus trait richness), a series of samples are taken from the evolving population. These samples are taken in ways that correspond to various real-world data collection strategies. First, a census of the entire population is taken. This functions as a baseline for the "most complete" information we can use to identify transmission modes, and there are also conditions during observational studies or in laboratory experiments where census is possible. In archaeological studies, anything approximating a census is usually impossible, although Jonathan Scholnick's study of New England gravestones and their makers may approximate this quality of data collection [51]. Second, the simulated population is sampled, at the 10% and 20% levels. Sampled data is ubiquitous in archaeological research, and although the issues involved in mapping artifact samples to their meaning for the underlying population of social learners is complex and unresolved, it is useful to determine whether the overall sample fraction has a measurable effect upon model equifinality.

Archaeological data are rarely synchronic or "point in time" samples of the results of human activity, and are typically aggregated over an appreciable duration of time through both data recovery conventions and formation processes [11–13, 52, 53]. Thus, the sampled data employed in this study is also temporally aggregated over a number of time steps, and the aggregate trait counts and then used to determine the frequencies of cultural traits over the entire interval. The population census has no temporal aggregation, and thus does represent a synchronic census.

Time averaging is implemented according to the schematic in Fig. 3. At the end of the simulation run, sampling begins at a time index calculated to allow time averaged samples to be taken twice, with a gap of 50 "generations" to allow the calculation of the Kandler-Shennan trait survival statistic (although unlike their original study, the values at the start and end times are inherently time averaged in this study, which would be the base in any real archaeological context) [15].[9]

Figure 3. Schematic of how sampling is implemented in this study. Time runs from the start of the simulation run at the top, to the end at the bottom. The interval of time over which we calculate the Kandler-Shennan trait survival is given as a simulation parameter, and represents the gap in the middle of the diagram. Before and after that gap are windows of successive duration, representing aggregation over 10, 25, 50, and 100 "generations" of the simulation.

The data collection strategies employed in this study are given in Table 4. Applied to all 23 variables, the study yielded approximately 900,000 samples from the four transmission models.[10] This raw data was then formed into the three pairwise comparisons shown in Table 1 for equifinality analysis with a classifier model.

2.4.4 Classifier Selection and Training

Classifier algorithms are supervising learning models from statistics and machine learning that predict a categorical response from a mixture of discrete or continuous variables [34]. The most

[9] The effect of time averaging on the start and end values used to calculate the Kandler-Shennan trait survival is not directly studied in this paper, but is a necessary component of using their method to study archaeological assemblages, I believe.

[10] All data and analyses for this study are available as part of a Github repository, although large data files are kept on Amazon S3 for long-term storage. See `https://github.com/mmadsen/experiment-ctmixtures` for details. The published analysis described here is the "equifinality-4" data set.

Sampling Strategy	Time Averaging Duration
Population Census	0
10% Sample	10
10% Sample	25
10% Sample	50
10% Sample	100
20% Sample	10
20% Sample	25
20% Sample	50
20% Sample	100

Table 4. Data collection strategies, applied to every simulation run. Time averaging duration is given in units of "generations," which are units of 100 time steps (given the population size). 100 generations thus represents 10,000 elemental time steps in the Moran simulation dynamics.

familiar classifiers in archaeological practice are logistic regression and discriminant function analysis, but neither is competitive with contemporary "ensemble" methods which combine many classifier rules into a single prediction. In such models, combining predictors can both reduce the variance of prediction (e.g., bagging added to traditional classifiers and random forests), and reduce bias. Some classifiers, like boosted trees, can do both.

Since the Bayes error rate of comparing two complex transmission models is not something we can calculate or even estimate, we must approximate it using the best performing classifier model available. A very general result in statistical decision theory (called, appropriately, the "No Free Lunch" theorems) guarantee that there is no single prediction model that can achieve the best result with every data set and problem [54, 55]. Thus, I took a compromise approach, selecting several algorithms that are known to have excellent performance across a range of data sets, and then performing a pilot study using the four transmission models previously described. A recent study compared 179 classifier algorithms on 121 different data sets (representing the entire UC Irvine Machine Learning Database), and found that random forests [56], support vector machines, and gradient boosted classifiers performed the best [34]. Additionally, some ensemble methods (random forests and gradient boosted classifiers) provide information on variable importance as an integral part of the algorithm. Since understanding which of our 23 variables are useful for separating transmission models is an important aspect of this study, I evaluated random forests against gradient boosted classification trees using small simulated samples from each transmission model.[11] Gradient boosted models outperformed random forests on these simulated data, are comparable in computational costs, and are used for all further results in this paper.

Gradient boosted classification operates by repeatedly fitting a set of decision trees to the data [34, 57]. In each round, decision trees are fit to the training data, and individual data points scored as errors or successful predictions. Subsequent trees are fitted by modifying the trees in the direction that minimizes the residual error. This is equivalent to finding the gradient of the loss function in the space of possible classifier functions, hence the name of the method. The impact of each gradient step is smoothed by including a "shrinkage" factor. Finally, the gradient steps are "boosted" to weight data points by the success in prediction, such that data points that are frequently misclassified become targeted by the algorithm until they can be correctly predicted [58–60]. After a specified number of iterations, the class or label membership of each data point is obtained by having each gradient step classifier tree "vote" for class membership, and the final answer is the majority vote. This class of models can also be visualized as repeated refitting of residuals until error is minimized [61]. This combination of boosting and iterative function search is very powerful, and gradient boosted models regularly achieve top accuracy in

[11] The data for this initial comparison are available in the `https://github.com/mmadsen/experiment-ctmixtures` repository under the experiment name "equifinality-2".

benchmark studies.

In this study, I employ the R package (**gbm**) for gradient boosted classification [62], with the binomial deviance $\log(1+\exp(-2y\hat{y}))$ as our loss function, where y is the true model for a data point, and $\hat{y}$ is the classifier model's prediction. Binomial deviance approximates the "zero-one" loss function with one which is differentiable, which is needed for a gradient descent method. The tuning parameters for this study (number of boosting iterations, depth of classification trees) were selected using 5 rounds of repeated 10-fold cross-validation on the training data [63, 64].

The full data set is split into two chunks. 80% of the data are used to train the classifier model, and 20% are held back to provide an unbiased evaluation of classifier performance. For each comparison of models reported here, the training data are thus fitted 50 times across different values of the tuning parameters (number of boosting iterations, and depth of decision trees), and the best performing parameters chosen from the repeated cross-validation sets. The final model is then constructed using the entire training set and the optimal parameter values. All classifier tuning, final model fitting, and test error evaluation was performed using Max Kuhn's superb **caret** package for R [64, 65].

	Actual Model:	
Predicted	Model 1	Model 2
Model 1	**9000**	2500
Model 2	1000	**7500**

Table 5. Example confusion matrix. Columns correspond to the actual model for data points, rows correspond to predictions from a classification model. Bold numbers on the diagonal correspond to correct predictions, the off diagonal elements correspond to classification errors.

2.4.5 Classification Error and Equifinality Assessment

The basic data for assessing the quality of a classifier model is the *confusion matrix*, which compares classification successes and errors for a data set. A hypothetical example is given in Table 5. The most basic measure of classification quality is the *accuracy*, or the ratio of correct predictions to the total number of data points. In the confusion matrix, this is the ratio of the sum of diagonal elements to the sum of off-diagonal elements. In the example given in Table 5, the classifier is 82.5% accurate. We often also use the misclassification rate, which is simply $1-\text{accuracy}$.

When the classes being predicted are not balanced, and especially if there are a small number of one class compared to another, a better statistic is Cohen's "kappa" [64], which compares observed accuracy to what one would expect purely from chance, given the marginal totals:

$$\kappa = \frac{O-E}{1-E} \tag{3}$$

where O is the observed accuracy, and E is the expected accuracy due to chance given the ratio of classes in the marginal totals of the confusion matrix. Kappa ranges from -1 to $+1$, with 0 indicating no agreement between predictions and the real class memberships. High values indicate good agreement, while values below 0.5 and especially less than 0.2 indicate very poor predictive ability [66]. In the present context, a classifier comparison (for example, biased versus neutral models with no sampling or time averaging) that yield a high kappa value are strong evidence that no equifinality exists between the two situations, since the classifier is highly accurate. Low kappa values are evidence that despite strong statistical methods and many variables to choose from, we cannot distinguish between models, and thus the models may be equifinal.

In studies where one outcome or class represents the presence of something (e.g., a positive test for a disease marker) and the other the absence, we may look at the individual cells of the confusion matrix rather than the bulk accuracy. The "false positive rate" (FPR), for example, is

the number of cases which are not members of the "positive" class, but which the classifier falsely identifies as such (in the example shown here, if Model 1 is the positive class, the cell in the upper right corner of the matrix is the FPR. A number of other statistics build from FPR and the false negative rate to handle asymmetric experiments. In the present study, we are interested simply in the misclassification rate, or bulk accuracy, of predicting the correct model. Throughout these results, I use the misclassification rate and Cohen's kappa values exclusively.

3 Results

In the next three sections, I review the results of applying the gradient boosted classifier to the three pairwise comparisons described in Table 1.

3.1 Unbiased Versus Biased Cultural Transmission

In the first comparison, all data points generated by unbiased (neutral) cultural transmission form one class, and the data points generated by each of the 3 biased models. As a reminder, the latter are:

1. Mixture of equal numbers of conformists and anti-conformists.
2. A mixture dominated by conformists, but with 30% anti-conformists.
3. A mixture dominated by anti-conformists, but with 30% anti-conformists.

This comparison examines the question of whether multiple predictor variables give us the power to discriminate between unbiased and any mixture of biased transmission, across different data collection regimes. For this comparison, classifier models were also developed for all of the predictor variables, and just for the per-locus variables, to determine the effect of using multidimensional classes that mimic archaeological classification, as opposed to simply examining single dimensions of variation (which has been the most common practice in archaeological studies to date).

The results are summarized in Fig. 4 as Cohen's kappa values across the different predictor variable sets and data collection regimes. It is immediately apparent that multiple variables in our classifier model gives us great power in distinguishing biased from unbiased transmission, in the case where we have population census data which is not subject to temporal aggregation. The use of multidimensional classes *in addition* to per-locus variables offers a tiny increase in accuracy, but these two comparisons are otherwise equivalent and display no equifinality given 97% accuracy in predicting biased versus unbiased transmission in the hold-out test set.

Figure 4. Cohen's kappa for correctly predicting whether simulated data points originate from unbiased copying or any of 3 other biased transmission models. High values of kappa correspond to high accuracy in correctly distinguishing between transmission models, while values well below 0.5 indicate great difficult and low classifier accuracy. Each line in the dotchart represents a different data collection treatment, and overall the results indicate that significant equifinality exists except when time averaging is absent and a population census (or near equivalent) is available.

Accuracy rapidly declines, however, when data points are derived from samples of the evolving population and where time averaging is present. As one might expect, larger samples ofer more accurate predictions than smaller samples. Within the larger, 20% sample, when cross-tabulated class and per-locus predictors are included, accuracy is highest with the smallest amount of time averaging (10 generations), and decreases as time averaging increases. When we remove cross-tabulated class predictors, and simply look at per-locus variables, this clean pattern is not apparent, and accuracy is not a function of time averaging duration. Furthermore, for the

smaller 10% sample with all variables included, accuracy is not a function of time averaging duration. In these cases, Cohen's kappa values are close to 0.25, indicative of a very poor classification model whose output bears little relation to the underlying transmission models. Finally, pooling all sample sizes and time averaging durations simply yields the average performance of the sampled and aggregated models, as one might expect.

Importance	Predictor Variable
100.00	Cross-Tabulated Class Richness
50.71	Slatkin Exact for Classes
29.50	Shannon Entropy (Mean for Locus)
23.13	Shannon Entropy for Classes
19.66	IQV Diversity (Mean for Locus)
11.58	Kandler-Shennan Trait Survival (Mean for Locus)

Table 6. Relative importance of predictor variables for population census data, in the comparison between unbiased transmission and all biased models. The most important variable is (by convention) scaled to 100, and the values indicate the ratio of variable importance to the variable which is most effective at classifying data points. Only values greater than 10 are shown. The remainder of the predictor variables are 1/100th as effective as class richness or less.

Gradient boosting algorithms allow measurement of how much each predictor variable contributes the classification model. The importance of a variable is assessed over the iterations of tree construction by estimating the relative improvement in training set misclassification error from adding the variable to the model. The importance values are usually scaled such that the most important variable has a score of 100, and variables with smaller importance values are less important to classificatory power. Table 6 gives the relative importance of predictor variables for the comparison between biased and unbiased models using population census data, and we can see that most of the classification power comes from the richness of cross-tabulated classes, about half as much from the Slatkin Exact test for cross-tabulated classes, and then an entropy measure of diversity among traits and classes. This is followed by a normalized version of the Shannon entropy, and finally by the Kandler-Shennan survival time, averaged across the 4 loci.

3.2 Unbiased Versus Balanced Conformist/Anticonformist Bias

The second comparison pairs unbiased transmission with a model the simulated population is composed of an equal number of conformists and anti-conformists. The probabilities of biased copying events are simulation parameters and are chosen uniformly from the prior distribution given in Table 2. This comparison examines the question of whether mixtures of biases can cancel each other out and appear to be unbiased, previously raised by Mesoudi and Lycett [10] and others. I believe this to be a likely scenario, and the likelihood that such mixtures would be indistinguishable from unbiased or neutral transmission when sampled, time averaged, or observed at larger regional scales was the original impetus for this study. The comparison was performed in the same manner as the first, except that the minor differences between using all predictors and only per-locus variables in the first comparison led to dropping separate comparisons given the computational cost of doing so. In this and the third comparison, all results refer to the full suite of 23 variables, across the same set of data collection regimes.

Figure 5. Cohen's kappa for correctly predicting whether simulated data points originate from unbiased copying or a balanced mixture of pro- and anti-conformist individuals. Each line in the dotchart represents a different data collection treatment, and overall the results indicate that significant equifinality exists except when time averaging is absent and a population census (or near equivalent) is available.

Fig. 5 displays the results of comparing "balanced biases" against unbiased transmission. We

can see again that excellent separation is achieved with population census data, but with all of the sampled and time averaged data collection strategies, there is considerable difficulty in correctly predicting the model from which a data point originated. With larger sample sizes, of course, there is less equifinality than with the smaller 10% but in both cases Cohen's kappa is 0.5 or less, indicating substantial equifinality though not complete overlap in the model outcomes.

Importance	full_variable
100.00	Cross-Tabulated Class Richness
85.49	Slatkin Exact for Classes
47.12	Shannon Entropy (Mean for Locus)
24.12	Kandler-Shennan Trait Survival (Mean for Locus)
18.88	IQV Diversity (Mean for Locus)
10.85	Shannon Entropy for Classes

Table 7. Relative importance of predictor variables for population census data, in the comparison between unbiased transmission and a balanced mixture of pro- and anti-conformists. The most important variable is (by convention) scaled to 100, and the values indicate the ratio of variable importance to the variable which is most effective at classifying data points. Only values greater than 10 are shown. The remainder of the predictor variables are 1/100th as effective as class richness or less.

The same predictor variables are responsible for almost all of the classification power, but there are subtle differences. Slatkin's "exact" test has more relative importance for this comparison than in differentiating between unbiased and all biases, and the entropy measures of diversity also have higher importance. This suggests that subtle differences in the evenness of classes and individual loci are very important in determining whether a population is truly engaged in unbiased transmission, or whether transmission biases are simply "cancelling out" at the macroscopic scale. Unfortunately, it appears that working with small samples of the population in the presence of time averaging strongly compromises our ability to differentiate those scenarios.

3.3 Conformist Dominated Versus Anticonformist Dominated Populations

The final comparison pairs two simulated populations, one of which is dominated by 70% conformists, with 30% anti-conformists, and the opposite with 70% anti-conformists and 30% conformists. In previous efforts to model the statistic signatures of conformism and anti-conformism, several authors have argued that there are clear patterns which separate these two modes of transmission (especially see Mesoudi and Lycett [10]). My own view is that these modes of transmission are much harder to detect in heterogeneous populations. This comparison is meant to test a simplified version of this conjecture. In this analysis, the number of conformists and anti-conformists is fixed by each of the models to the ratios given above, but the probability of a biased copying event is set for each simulation run to randomly chosen values drawn from the prior distribution given in Table 2.

Figure 6. Cohen's kappa for correctly predicting whether simulated data points originate from a conformist-dominated mixed population versus a mixed population dominated by anti-conformists. Each line in the dotchart represents a different data collection treatment, and overall the results indicate that strong equifinality exists regardless of the data collection treatment.

Fig. 6 displays the result of this comparison across data collection treatments. None of the results indicate an ability to cleanly separate these two models. Population census data and the

absence of time averaging certainly help, but the accuracy of classification is dismal in all cases. Strong equifinality exists between these models, as one might expect given their similarity. It is possible that with even stronger propensities to engage in conformity or its opposite, that we may be able to detect it in a heterogeneous population, but at the levels probed here, the models are indistinguishable, even in the high-dimensional space created by all 23 predictor variables.

4 Discussion

The classifier models used in this study provide a sensitive probe into the issue of equifinality between models of cultural transmission modes. Using both accuracy measures and measures of variable importance, this work highlights the variables we need to use in order to reliably distinguish between modes of transmission, given particular data collection conditions, in population models that are more realistic than those previously used.

This study seems to substantiate previous claims that time averaged data make the task of identifying the mode of transmission difficult. I propose to extend those claims by noting that small sample sizes dramatically worsen our ability to separate models, and to note that distinguishing *among* detailed models of transmission bias given frequency data alone (on individual dimensions/loci and multidimensional classes) appears to be impossible without new predictor variables. A better understanding how we can best calculate and use the power law exponent for trait or class diversity may help. To date I believe there are inconsistent ways in which the statistic has been applied to simulated data, some of which seem incompatible with the measurements we can make on archaeological assemblages.

But simple equifinality is not the whole story. It is not simply the case that sampling and time averaging render our predictions of transmission mode random with respect to the set of models tested. There is substantial bias (in the statistical sense) in the classifier models for sampled and time averaged data collection regimes. We can see this by looking at individual confusion matrices from predictions made on the hold-out test data.

Table 8. Two confusion matrices arising from the first model comparison, between unbiased and all biased models.

(a) Population Census Data

	biased	neutral
biased	14898	132
neutral	102	4868

(b) Sample Size: 20 Duration: 50

	biased	neutral
biased	13926	2724
neutral	1074	2276

The left hand side of Table 8 shows the confusion matrix for the population census data collection treatment, while the right hand panel represents predicts for a sample size of 20%, time averaged over 50 generations. The top row of the table shows data points for which the model predicted an origin in a biased model, with the columns representing the "real" origin of the data points. In the left panel, the population census data only identified 132 data points as biased, when they really arose from an unbiased model. However, in the right panel, we see a very different pattern. Of the 5000 data points arising from an unbiased model, *more* of the points were identified as coming from biased transmission models than as unbiased.

By looking at the ratio of the right column in each confusion matrix, across all data collection treatments, we can see the magnitude of this "preference" for predicting data points as coming from biased transmission models (Table 9). Immediately apparent is that sampling and time averaging have a dramatic effect on predictions of transmission bias, making it extremely likely that we will find sampled and time averaged data samples to fit our models of conformist and anti-conformist bias.

I believe that this asymmetry in discriminatory ability means that archaeologists must be extremely careful in identifying transmission bias from archaeological samples. Only under very

Data Collection Treatment	% of Unbiased Data Misclassified
Population Census	2.6
Per-Locus Population Census	3.4
Sample Size: 20 Duration: 10	49.1
Sample Size: 20 Duration: 25	49.5
Per-Locus Sample Size: 20 Duration: 25	49.7
Per-Locus Sample Size: 20 Duration: 10	50.1
Per-Locus Sample Size: 20 Duration: 50	54.4
Sample Size: 20 Duration: 50	54.5
Sample Size: 20 Duration: 100	57.6
Per-Locus Sample Size: 20 Duration: 100	58.1
All Sample Sizes and TA Durations	68.0
Per-Locus Sample Size: 10 Duration: 25	73.5
Sample Size: 10 Duration: 25	73.6
Sample Size: 10 Duration: 50	73.7
Per-Locus Sample Size: 10 Duration: 10	73.8
Per-Locus Sample Size: 10 Duration: 100	74.0
Per-Locus Sample Size: 10 Duration: 50	74.1
Sample Size: 10 Duration: 100	74.4
Sample Size: 10 Duration: 10	74.7

Table 9. Percentage of data points from the unbiased transmission model that are falsely identified as arising from a biased model.

rare preservation and sedimentary conditions, or in historical contexts, will we find data collection regimes that approximate the population census treatment studied here. Whenever we deal with small samples and data that come from aggregated deposits, we would do well to exhibit healthy skepticism about our ability to detect transmission bias, or indeed to say much about social learning modes. In the terms developed in this study, we face some irreducible equifinalities (as between conformism and anti-conformism), and many equifinalities that are reducible given more precise and complete data collection. Unfortunately, some of these potentially reducible equifinalities may be irreducible in practical terms.

This conclusion, however, is relative to the exact details of transmission models, predictor variables, and data collection regimes. As we develop better models and additional predictor variables that are measurable from archaeological data, we may be able to achieve better resolution of social learning processes. The classifier approach demonstrated here is capable of identifying whether we have sucdessfully reduced equifinalities, or whether social learning modes remain out of reach for most archaeological contexts, and whether our efforts at applying cultural transmission modeling to the archaeological record would be better served by focusing upon different questions tailored to the data we possess.

Acknowledgments

Cras egestas velit mauris, eu mollis turpis pellentesque sit amet. Interdum et malesuada fames ac ante ipsum primis in faucibus. Nam id pretium nisi. Sed ac quam id nisi malesuada congue. Sed interdum aliquet augue, at pellentesque quam rhoncus vitae.

References

1. Boyd R, Richerson PJ. Culture and the Evolutionary Process. Chicago: University of Chicago Press; 1985.

2. Cavalli-Sforza LL, Feldman MW. Cultural Transmission and Evolution: A Quantitative Approach. Princeton: Princeton University Press; 1981.

3. Henrich J, Boyd R. The Evolution of Conformist Transmission and the Emergence of Between-Group Differences. Evolution and Human Behavior. 1998 Jul;19(4):215–241. Available from: http://linkinghub.elsevier.com/retrieve/pii/S109051389800018X.

4. Wakano JY, Aoki K. Do social learning and conformist bias coevolve? Henrich and Boyd revisited. Theoretical Population Biology. 2007 Dec;72(4):504–512. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0040580907000433.

5. Kohler TA, VanBuskirk S, Ruscavage-Barz S. Vessels and villages: evidence for conformist transmission in early village aggregations on the Pajarito Plateau, New Mexico. Journal of Anthropological Archaeology. 2004;23(1):100–118.

6. Bentley RA, Shennan SJ. Cultural transmission and stochastic network growth. American Antiquity. 2003;68(3):459–485.

7. Bentley RA, Lipo CP, Herzog HA, Hahn MW. Regular rates of popular culture change reflect random copying. Evolution and Human Behavior. 2007;28(3):151–158.

8. Bentley RA, Hahn MW, Shennan SJ. Random drift and culture change. Proceedings of the Royal Society of London Series B: Biological Sciences. 2004;271(1547):1443–1450.

9. Evans TS, Giometto A. Turnover Rate of Popularity Charts in Neutral Models. arXivorg. 2011;http://arxiv.org/abs/1105.4044.

10. Mesoudi A, Lycett SJ. Random Copying, Frequency-dependent Copying and Culture Change. Evolution and Human Behavior. 2009;30:41–48.

11. Madsen ME. Unbiased Cultural Transmission in Time-Averaged Archaeological Assemblages. ArXiv e-prints. 2012;1204.2043. Available from: http://arxiv.org/abs/1204.2043.

12. Porčić M. Exploring the Effects of Assemblage Accumulation on Diversity and Innovation Rate Estimates in Neutral, Conformist, and Anti-Conformist Models of Cultural Transmission. Journal of Archaeological Method and Theory. 2014;p. 1–22. Available from: http://dx.doi.org/10.1007/s10816-014-9217-8.

13. Premo LS. Cultural Transmission and Diversity in Time-Averaged Assemblages. Current Anthropology. 2014;55(1):105–114.

14. Rorabaugh AN. Impacts of drift and population bottlenecks on the cultural transmission of a neutral continuous trait: an agent based model. Journal of Archaeological Science. 2014 Sep;49:255–264. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0305440314001940.

15. Kandler A, Shennan S. A non-equilibrium neutral model for analysing cultural change. Journal of theoretical biology. 2013;330:18–25.

16. Wilder B, Kandler A. Inference of cultural transmission modes based on incomplete information; 2015. Forthcoming.

17. von Bertalanffy L. Problems of organic growth. Nature. 1949;163(4135):156–158.

18. Aronica G, Hankin B, Beven K. Uncertainty and equifinality in calibrating distributed roughness coefficients in a flood propagation model with limited data. Advances in Water Resources. 1998 Oct;22(4):349–365. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0309170898000177.

19. Beven K. A manifesto for the equifinality thesis. Journal of Hydrology. 2006 Mar;320(1-2):18–36. Available from: http://linkinghub.elsevier.com/retrieve/pii/S002216940500332X.

20. Bonham SG, Haywood AM, Lunt DJ, Collins M, Salzmann U. El Niño-Southern Oscillation, Pliocene climate and equifinality. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2009 Jan;367(1886):127–156. Available from: http://rsta.royalsocietypublishing.org.offcampus.lib.washington.edu/content/367/1886/127.full.

21. Cicchetti D, Rogosch FA. Equifinality and multifinality in developmental psychopathology. Development and Psychopathology. 1996 Sep;8(04):597–600. Available from: http://journals.cambridge.org.offcampus.lib.washington.edu/action/displayAbstract?aid=4495040.

22. Culling WEH. Equifinality: Modern Approaches to Dynamical Systems and Their Potential for Geographical Thought. Transactions of the Institute of British Geographers. 1987;12(1):57. Available from: http://www.jstor.org/stable/622577?origin=crossref.

23. Marean CW, Spencer LM, Blumenschine RJ, Capaldo SD. Captive hyaena bone choice and destruction, the Schlepp effect and olduvai archaeofaunas. Journal of Archaeological Science. 1992 Jan;19(1):101–121. Available from: http://linkinghub.elsevier.com/retrieve/pii/030544039290009R.

24. Rogers AR. On Equifinality in Faunal Analysis. American Antiquity. 2000 Oct;65(4):709–723.

25. Savenije HHG. Equifinality, a blessing in disguise? Hydrological Processes. 2001 Oct;15(14):2835–2838. Available from: http://doi.wiley.com/10.1002/hyp.494.

26. Slatkin M. An exact test for neutrality based on the Ewens sampling distribution. Genetical Research. 1994;64(01):71–74.

27. Slatkin M. A correction to the exact test based on the Ewens sampling distribution. Genetical research. 1996;68(03):259–260.

28. Kempe M, Mesoudi A. Experimental and theoretical models of human cultural evolution. Wiley Interdisciplinary Reviews: Cognitive Science. 2014;5(3):317–326.

29. Mesoudi A. Experimental Studies of Modern Human Social and Individual Learning in an Archaeological Context: People Behave Adaptively, But Within Limits. In: Dynamics of Learning in Neanderthals and Modern Humans Volume 2. Springer; 2014. p. 65–76.

30. Schillinger K, Mesoudi A, Lycett SJ. Copying Error and the Cultural Evolution of Additive vs. Reductive Material Traditions: an Experimental Assessment. American Antiquity. 2014;79(1):128–143.

31. Plutinski A. Interview with Warren Ewens; 2004. Available from: http://authors.library.caltech.edu/5456/1/hrst.mit.edu/hrs/evolution/public/ewens.html [cited 12/14/2014].

32. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133(3):693–709.

33. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123(3):585–595.

34. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. The elements of statistical learning. vol. 2. Springer; 2009.

35. Devijver PA, Kittler J. Pattern recognition: A statistical approach. vol. 761. Prentice-Hall London; 1982.

36. Fukunaga K. Introduction to statistical pattern recognition. Academic press; 1990.

37. Antos A, Devroye L, Gyorfi L. Lower bounds for Bayes error estimation. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1999 Jul;21(7):643–645. Available from: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=777375`.

38. Dobbin KK. A method for constructing a confidence bound for the actual error rate of a prediction rule in high dimensions. Biostatistics (Oxford, England). 2009 Apr;10(2):282–296. Available from: `http://biostatistics.oxfordjournals.org/cgi/doi/10.1093/biostatistics/kxn035`.

39. McLachlan GJ. Confidence Intervals for the Conditional Probability of Misallocation in Discriminant Analysis. Biometrics. 1975 Mar;31(1):161. Available from: `http://www.jstor.org/stable/2529717?origin=crossref`.

40. Loizou G, Maybank SJ. The Nearest Neighbor and the Bayes Error Rates. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 1987 Mar;PAMI-9(2):254–262. Available from: `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4767899`.

41. Tumer K, Ghosh J. Bayes error rate estimation using classifier ensembles. International Journal of Smart Engineering System Design. 2003;5(2):95–109.

42. Moran PAP. The statistical processes of evolutionary theory. Clarendon Press; Oxford University Press.; 1962.

43. Moran P. Random processes in genetics. In: Mathematical Proceedings of the Cambridge Philosophical Society. vol. 54. Cambridge Univ Press; 1958. p. 60–71.

44. Aoki K, Lehmann L, Feldman MW. Rates of cultural change and patterns of cultural accumulation in stochastic models of social transmission. Theoretical population biology. 2011;79(4):192–202.

45. Ewens WJ. Mathematical Population Genetics, Volume 1: Theoretical Introduction. 2nd ed. New York, Springer; 2004.

46. Beaumont MA. Approximate Bayesian computation in evolution and ecology. Annual Review of Ecology. 2010;Available from: `http://www.annualreviews.org/doi/abs/10.1146/annurev-ecolsys-102209-144621`.

47. Crema ER, Edinborough K, Kerig T, Shennan SJ. An Approximate Bayesian Computation approach for inferring patterns of cultural evolutionary change. Journal of Archaeological Science. 2014 Jul;Available from: `http://linkinghub.elsevier.com/retrieve/pii/S0305440314002593`.

48. Csilléry K, Blum MGB, Gaggiotti OE, François O. Approximate Bayesian Computation (ABC) in practice. Trends in Ecology & Evolution. 2010 Jul;25(7):410–418. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0169534710000662.

49. Marin JM, Pudlo P, Robert CP, Ryder RJ. Approximate Bayesian computational methods. Statistics and Computing. 2012;22(6):1167–1180.

50. Dunnell RC. Systematics in prehistory. New York: Free Press; 1971.

51. Scholnick JB. The spatial and temporal diffusion of stylistic innovations in material culture. Advances in Complex Systems. 2012;15(01n02).

52. Grayson DK, Delpech F. Changing diet breadth in the early Upper Paleolithic of southwestern France. Journal of Archaeological Science. 1998;25:1119–1129.

53. Lyman RL. The influence of time averaging and space averaging on the application of foraging theory in zooarchaeology. Journal of Archaeological Science. 2003;30(5):595–610.

54. Wolpert DH. The supervised learning no-free-lunch theorems. In: Soft Computing and Industry. Springer; 2002. p. 25–42.

55. Wolpert DH, Macready WG. No free lunch theorems for optimization. Evolutionary Computation, IEEE Transactions on. 1997;1(1):67–82.

56. Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

57. Alexey Natekin AK. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics. 2013;7:21. Available from: http://journal.frontiersin.org/Journal/10.3389/fnbot.2013.00021/full.

58. Freund Y. Boosting a weak learning algorithm by majority. Information and computation. 1995;121(2):256–285.

59. Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence. 1999;14(771-780):1612.

60. Schapire RE, Freund Y. Boosting: Foundations and algorithms. MIT Press; 2012.

61. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001;p. 1189–1232.

62. Ridgeway G. The state of boosting. Computing Science and Statistics. 1999;p. 172–181.

63. Kim JH. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics & Data Analysis. 2009 Sep;53(11):3735–3745. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0167947309001601.

64. Kuhn M, Johnson K. Applied predictive modeling. Springer; 2013.

65. Kuhn M. Building predictive models in R using the caret package. Journal of Statistical Software. 2008;28(5):1–26.

66. Altman DG. Practical statistics for medical research. CRC Press; 1991.

FIGURES IN DRAFT - REMOVE THIS SECTION FOR SUBMISSION

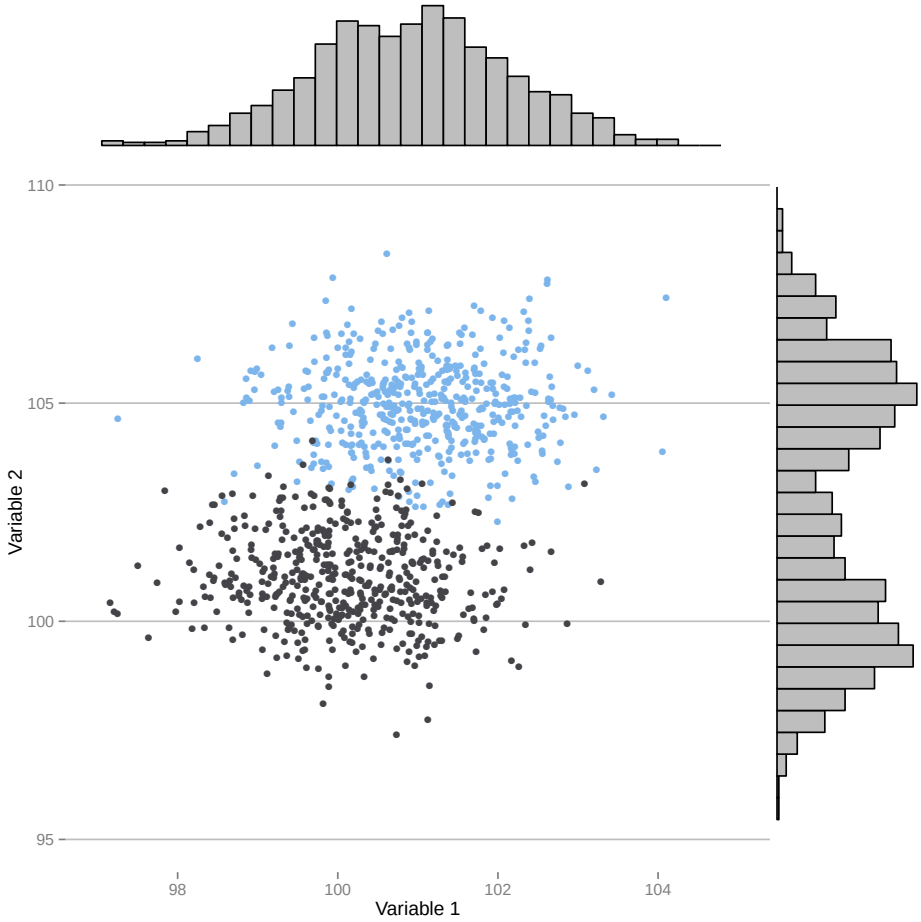

Figure 1. Simple example of the effect of variable choice in distinguishing models. The variable on the X axis displays quite a bit of overlap between models, while the variable on the Y axis distinguishes the models with fairly high accuracy.

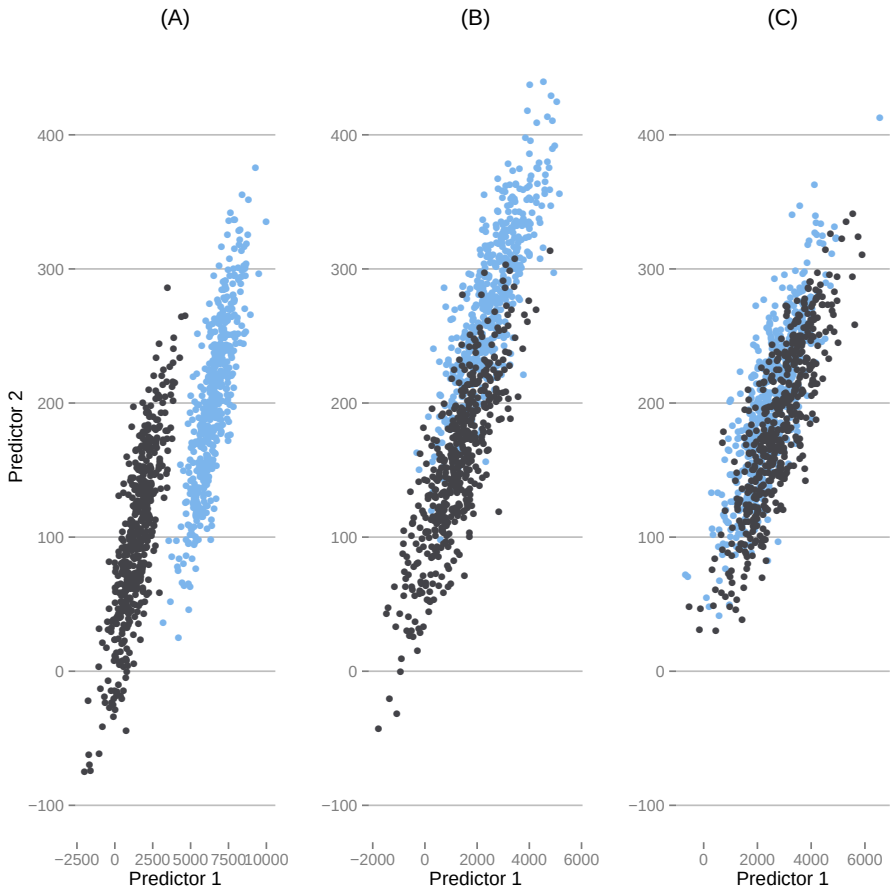

Figure 2. Simple example of model outcomes with different degrees of distinguishability: (A) simulated data point from two fully separate models, (B) two models with a limited overlap region, (C) and two models whose outcomes are highly overlapping.

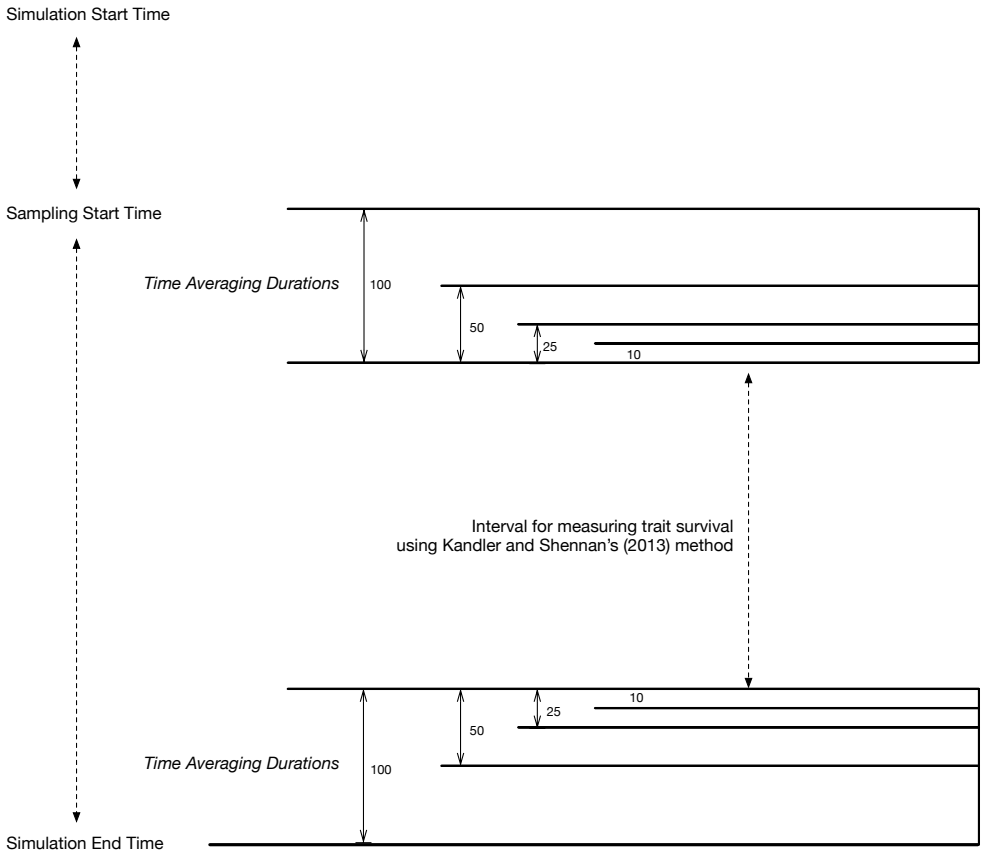

Figure 3. Schematic of how sampling is implemented in this study. Time runs from the start of the simulation run at the top, to the end at the bottom. The interval of time over which we calculate the Kandler-Shennan trait survival is given as a simulation parameter, and represents the gap in the middle of the diagram. Before and after that gap are windows of successive duration, representing aggregation over 10, 25, 50, and 100 "generations" of the simulation.

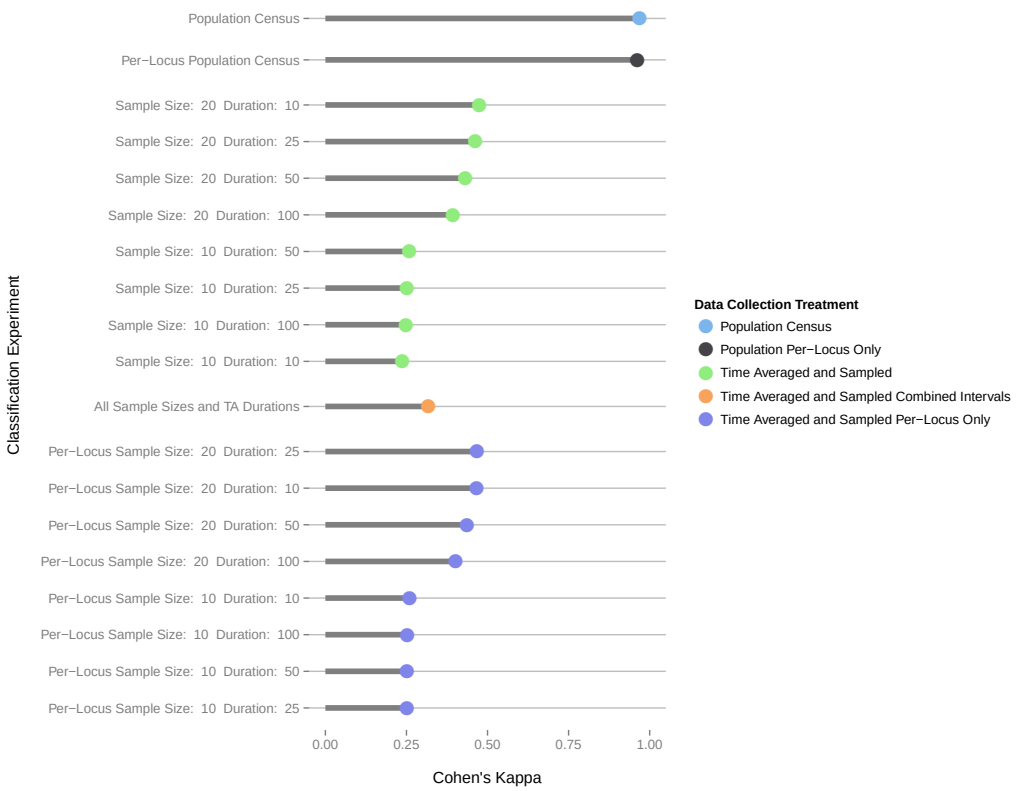

Figure 4. Cohen's kappa for correctly predicting whether simulated data points originate from unbiased copying or any of 3 other biased transmission models. High values of kappa correspond to high accuracy in correctly distinguishing between transmission models, while values well below 0.5 indicate great difficult and low classifier accuracy. Each line in the dotchart represents a different data collection treatment, and overall the results indicate that significant equifinality exists except when time averaging is absent and a population census (or near equivalent) is available.

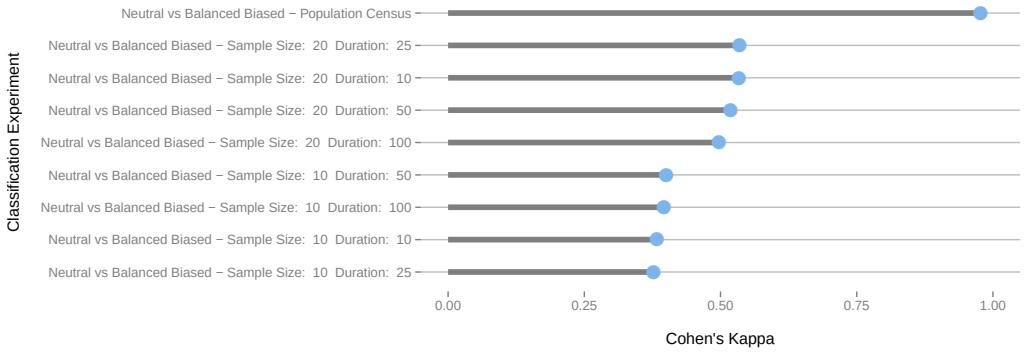

Figure 5. Cohen's kappa for correctly predicting whether simulated data points originate from unbiased copying or a balanced mixture of pro- and anti-conformist individuals. Each line in the dotchart represents a different data collection treatment, and overall the results indicate that significant equifinality exists except when time averaging is absent and a population census (or near equivalent) is available.

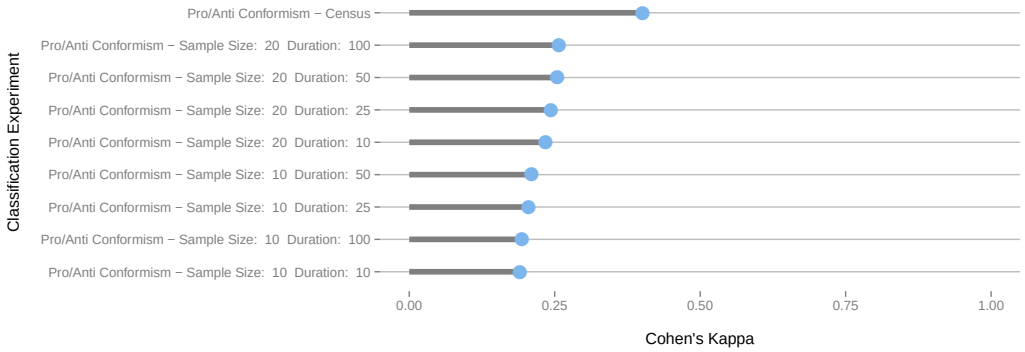

Figure 6. Cohen's kappa for correctly predicting whether simulated data points originate from a conformist-dominated mixed population versus a mixed population dominated by anti-conformists. Each line in the dotchart represents a different data collection treatment, and overall the results indicate that strong equifinality exists regardless of the data collection treatment.