

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

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

¹Throughout 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 $\mathcal{M}_1 \dots \mathcal{M}_n$, 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 $Y \in 1, \dots, K$ refers to each of k models, and X_1, \dots, X_p refer to p different predictor variables.

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

$\mathbb{P}(Y)$ 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:

$$Y_{pred} = \arg \max_y \mathbb{P}(X_1, \dots, X_p|Y) \quad (2)$$

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 space 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.

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

³There 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 each model to be compared are derived by simulating the dynamics of the model in an agent-based model. 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 has not been seen in the population previously. 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>

⁵The 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 [45–48].

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. Additionally, equifinality should be increased in models where drift occurs due to finite-size effects. Each simulated individual carries 4 different traits at any time, which are treated as separate loci or dimensions. Copying involves no interaction effects between loci in this study. The population is seeded with 10 randomly chosen traits at each Loci as the initial condition. 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 latter figure was chosen for creating the table of simulated samples for classification analysis.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

{edit here}

Since equifinality is a function of the transmission model itself, the variables we employ to measure model outcomes, and the conditions under which those variables are measured, this study examines common variables associated with transmission models over a number of sampling, time averaging, and measurement strategies. Variation in the latter will allow us to potentially separate irreducible and reducible equifinalities. The variables chosen 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). The classifier methods considered here add variables to classification models based upon their performance in separating the data successfully, so it is appropriate to start with all of the variables we might employ, and see which variables actually have relevance to distinguishing between models.

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 recorded the order statistics in addition to the mean value, since it is possible that minima and maxima might 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

⁶The analysis underpinning this decision is available in the Github repository at <https://github.com/mmadsen/experiment-ctmixtures/analysis/verification>.

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

Table 3. Variables measured from each transmission model simulation sample. The middle column records whether the variable is a measurement across traits in a single locus, and then summarized over loci, or whether it applies to trait configurations of all loci. The right column records the variable name used within R statistical models, for examining the relative importance of each variable in classifying observations.

which models 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 [49]. The same variables are then measured as a function of the class counts. 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 a prototype 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.⁷ 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 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.

⁷I confirmed this by inspection of the source code for their simulation model, which was provided by Alex Mesoudi.

2.4.3 Data Collection Treatments

240

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. Measurement 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.

The raw data set for this study thus consists of 900,000 measurements of all 23 variables across the measurement treatments given in Table 4.⁸

241

242

2.4.4 Classifier Selection and Training

243

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 familiar classifiers in archaeological practice are logistic regression and discriminant function analysis, but neither are competitive with contemporary "ensemble" methods which combine many classifier rules into a single prediction. In some cases, combining predictors simply reduces the variance of prediction (e.g., bagging added to traditional classifiers and random forests), while other ensemble classifiers can reduce bias (increasing accuracy) while also decreasing variance (e.g., boosting).

244

245

246

247

248

249

250

251

For our purposes, we simply need to select the best performing classifier method possible. 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 [50, 51]. Thus, for practical applications, it is important to choose a near-optimal method that performs well on the exact type of data to be analyzed. 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 [52], 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, and are I simulated a small initial sample from each of the 4 cultural transmission models investigated here, and evaluated the classification accuracy of random forests against a gradient boosted classifier.⁹ Gradient boosted models outperformed random forests on these simulated data, are comparable in computational costs, and were used for all further results in this paper.

252

253

254

255

256

257

258

259

260

261

262

263

264

265

Gradient boosted classification operates by repeatedly fitting a set of decision trees to the data [53]. Each time classifiers are fit to the data, the error within the training data is calculated, and the next round is fit by finding a function which decreases the total error in fitting the training data. In regression-style problems, this can be visualized as repeatedly refitting the

266

267

268

269

⁸All 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.

⁹The data for this initial comparison are available in the <https://github.com/mmadsen/experiment-ctmixtures> repository under the experiment name "equifinality-2".

residuals until minimum error is achieved [54]. With classification problems, the principle is the same but is less easily visualized. Once a specified number of improvement rounds is reached, each classifier “votes” for the predicted class of each data point, and the majority vote for each data point is recorded as the overall model prediction. In addition to this kind of iterative model averaging, “boosting” is used to improve the fit of the model to difficult-to-predict data points [55–57]. In boosting, the errors from a previous round of fitting are used to weight each data point, with incorrectly predicted data points given higher weight (i.e., “boosted”) and correct predictions downweighted. Thus, in successive fitting rounds, boosting algorithms focus increased attention on the (hopefully dwindling) number of incorrect predictions. The combination of boosting and iterative fitting is powerful, and such methods regularly achieve high accuracy in benchmark studies.

In this study, I employ the standard R package (gbm) for gradient boosted classification, with the tuning parameters chosen by 5 rounds of repeated 10-fold cross-validation [58, 59]. While regression models are evaluated by calculating the mean squared error of their predictions, classification models are usually assessed using the “0-1” loss function, which counts the number of misclassified data points when the true generating class is known. Unfortunately, zero-one loss is a non-differentiable function, which means that calculating its gradient to determine how to improve error is impossible. Instead, for binary classification (distinguishing pairs of transmission models), we use the “bernoulli” option, which uses the binomial deviance function $\log(1 + \exp(-2y\hat{y}))$, where y is the true model for a data point, and \hat{y} is the classifier model’s prediction.

Since evaluating a statistical model on the same data that was used to construct it tends to overestimate its accuracy, all results concerning equifinality of transmission models presented here use metrics derived from applying the best fit model to a hold-out set of data. 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 [59, 60].

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 (an example is given in Table 5).

Predicted	Actual Model:	
	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.

The most basic measure of quality is the *accuracy*, or the ratio of correct predicts to the total number of data points. In the example shown here, the classification is 82.5% accurate. 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” [59], which compares observed accuracy to what one would expect purely from chance, given the marginal totals:

$$\kappa = \frac{O - E}{1 - E} \quad (3)$$

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 [61]. 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.

3 Results

3.1 Unbiased Versus Biased Cultural Transmission

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum dignissim dignissim nunc, et finibus urna aliquam eget. Donec enim dolor, aliquam sed iaculis vitae, vestibulum sed justo. Curabitur fringilla, mauris quis ultrices mattis, neque libero volutpat nisi, vel mollis mi magna a felis. Phasellus a orci ut elit sodales tristique ac placerat nisi. Maecenas orci purus, ullamcorper non neque vel, imperdiet sollicitudin ante. Duis dapibus ante sed gravida imperdiet. Aenean dapibus augue nec vehicula rhoncus. Mauris ac fermentum ante, eget volutpat lectus. Nunc a est auctor, suscipit augue vel, vulputate lectus. Sed eu elit ullamcorper, interdum neque ut, varius nisi. Phasellus leo justo, mattis rutrum leo vitae, consequat auctor diam. Vivamus cursus, ligula et euismod iaculis, odio nulla ullamcorper ex, vitae cursus mi lacus at sem. Aenean dictum odio dolor, sit amet gravida sem scelerisque vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Morbi.

Figure 3. Unbiased versus biased kappa - lorem ipsum

3.2 Unbiased Versus Mixed Conformist/Anticonformist Bias

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum dignissim dignissim nunc, et finibus urna aliquam eget. Donec enim dolor, aliquam sed iaculis vitae, vestibulum sed justo. Curabitur fringilla, mauris quis ultrices mattis, neque libero volutpat nisi, vel mollis mi magna a felis. Phasellus a orci ut elit sodales tristique ac placerat nisi. Maecenas orci purus, ullamcorper non neque vel, imperdiet sollicitudin ante. Duis dapibus ante sed gravida imperdiet. Aenean dapibus augue nec vehicula rhoncus. Mauris ac fermentum ante, eget volutpat lectus. Nunc a est auctor, suscipit augue vel, vulputate lectus. Sed eu elit ullamcorper, interdum neque ut, varius nisi. Phasellus leo justo, mattis rutrum leo vitae, consequat auctor diam. Vivamus cursus, ligula et euismod iaculis, odio nulla ullamcorper ex, vitae cursus mi lacus at sem. Aenean dictum odio dolor, sit amet gravida sem scelerisque vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Morbi.

Figure 4. Unbiased versus balanced biased kappa - lorem ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum dignissim dignissim nunc, et finibus urna aliquam eget. Donec enim dolor, aliquam sed iaculis vitae, vestibulum sed justo. Curabitur fringilla, mauris quis ultrices mattis, neque libero volutpat nisi, vel mollis mi

magna a felis. Phasellus a orci ut elit sodales tristique ac placerat nisi. Maecenas orci purus, ullamcorper non neque vel, imperdiet sollicitudin ante. Duis dapibus ante sed gravida imperdiet. Aenean dapibus augue nec vehicula rhoncus. Mauris ac fermentum ante, eget volutpat lectus. Nunc a est auctor, suscipit augue vel, vulputate lectus. Sed eu elit ullamcorper, interdum neque ut, varius nisi. Phasellus leo justo, mattis rutrum leo vitae, consequat auctor diam. Vivamus cursus, ligula et euismod iaculis, odio nulla ullamcorper ex, vitae cursus mi lacus at sem. Aenean dictum odio dolor, sit amet gravida sem scelerisque vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Morbi.

Data Collection Treatment	% 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 6. Lorem Ipsum

3.3 Conformist Dominated Versus Anticonformist Dominated Populations

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum dignissim dignissim nunc, et finibus urna aliquam eget. Donec enim dolor, aliquam sed iaculis vitae, vestibulum sed justo. Curabitur fringilla, mauris quis ultrices mattis, neque libero volutpat nisi, vel mollis mi magna a felis. Phasellus a orci ut elit sodales tristique ac placerat nisi. Maecenas orci purus, ullamcorper non neque vel, imperdiet sollicitudin ante. Duis dapibus ante sed gravida imperdiet. Aenean dapibus augue nec vehicula rhoncus. Mauris ac fermentum ante, eget volutpat lectus. Nunc a est auctor, suscipit augue vel, vulputate lectus. Sed eu elit ullamcorper, interdum neque ut, varius nisi. Phasellus leo justo, mattis rutrum leo vitae, consequat auctor diam. Vivamus cursus, ligula et euismod iaculis, odio nulla ullamcorper ex, vitae cursus mi lacus at sem. Aenean dictum odio dolor, sit amet gravida sem scelerisque vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Morbi.

Figure 5. Pro versus anticonformist bias kappa - lorem ipsum

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum dignissim dignissim nunc, et finibus urna aliquam eget. Donec enim dolor, aliquam sed iaculis vitae, vestibulum sed justo. Curabitur fringilla, mauris quis ultrices mattis, neque libero volutpat nisi, vel mollis mi

magna a felis. Phasellus a orci ut elit sodales tristique ac placerat nisi. Maecenas orci purus, ullamcorper non neque vel, imperdiet sollicitudin ante. Duis dapibus ante sed gravida imperdiet. Aenean dapibus augue nec vehicula rhoncus. Mauris ac fermentum ante, eget volutpat lectus. Nunc a est auctor, suscipit augue vel, vulputate lectus. Sed eu elit ullamcorper, interdum neque ut, varius nisi. Phasellus leo justo, mattis rutrum leo vitae, consequat auctor diam. Vivamus cursus, ligula et euismod iaculis, odio nulla ullamcorper ex, vitae cursus mi lacus at sem. Aenean dictum odio dolor, sit amet gravida sem scelerisque vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Morbi.

4 Discussion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vestibulum dignissim dignissim nunc, et finibus urna aliquam eget. Donec enim dolor, aliquam sed iaculis vitae, vestibulum sed justo. Curabitur fringilla, mauris quis ultrices mattis, neque libero volutpat nisi, vel mollis mi magna a felis. Phasellus a orci ut elit sodales tristique ac placerat nisi. Maecenas orci purus, ullamcorper non neque vel, imperdiet sollicitudin ante. Duis dapibus ante sed gravida imperdiet. Aenean dapibus augue nec vehicula rhoncus. Mauris ac fermentum ante, eget volutpat lectus. Nunc a est auctor, suscipit augue vel, vulputate lectus. Sed eu elit ullamcorper, interdum neque ut, varius nisi. Phasellus leo justo, mattis rutrum leo vitae, consequat auctor diam. Vivamus cursus, ligula et euismod iaculis, odio nulla ullamcorper ex, vitae cursus mi lacus at sem. Aenean dictum odio dolor, sit amet gravida sem scelerisque vitae. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Morbi.

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. *arXiv.org*. 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 Schleppe 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, Györfi 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. 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>.
46. 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>.
47. 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>.
48. Marin JM, Pudlo P, Robert CP, Ryder RJ. Approximate Bayesian computational methods. *Statistics and Computing*. 2012;22(6):1167–1180.
49. Dunnell RC. *Systematics in prehistory*. New York: Free Press; 1971.
50. Wolpert DH. The supervised learning no-free-lunch theorems. In: *Soft Computing and Industry*. Springer; 2002. p. 25–42.
51. Wolpert DH, Macready WG. No free lunch theorems for optimization. *Evolutionary Computation*, IEEE Transactions on. 1997;1(1):67–82.
52. Breiman L. Random forests. *Machine learning*. 2001;45(1):5–32.
53. 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>.
54. Friedman JH. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*. 2001;p. 1189–1232.

55. Freund Y. Boosting a weak learning algorithm by majority. *Information and computation*. 1995;121(2):256–285.
56. Freund Y, Schapire R, Abe N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*. 1999;14(771-780):1612.
57. Schapire RE, Freund Y. *Boosting: Foundations and algorithms*. MIT Press; 2012.
58. 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>.
59. Kuhn M, Johnson K. *Applied predictive modeling*. Springer; 2013.
60. Kuhn M. Building predictive models in R using the caret package. *Journal of Statistical Software*. 2008;28(5):1–26.
61. Altman DG. *Practical statistics for medical research*. CRC Press; 1991.

FIGURES IN DRAFT - REMOVE THIS SECTION FOR SUBMISSION

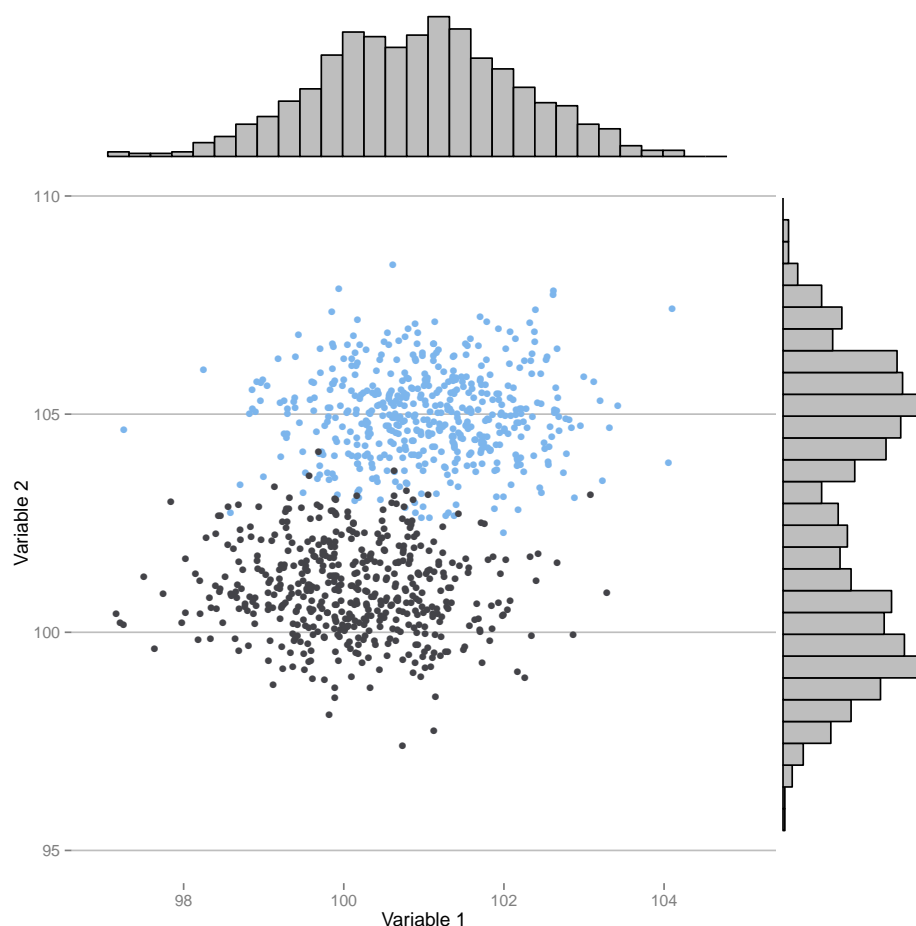


Figure 1. Simple example of the effect of variable choice in distinguishing models. The variable on the X axis displays quite a bit of overlap between models, while the variable on the Y axis distinguishes the models with fairly high accuracy.

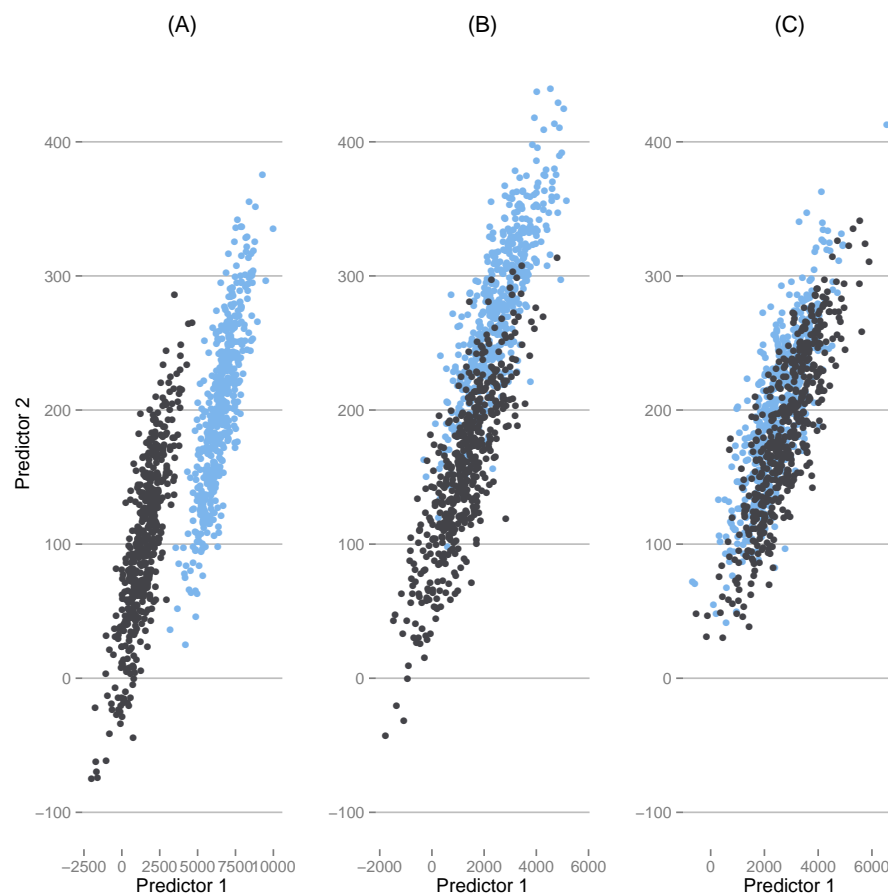


Figure 2. Simple example of model outcomes with different degrees of distinguishability: (A) simulated data point from two fully separate models, (B) two models with a limited overlap region, (C) and two models whose outcomes are highly overlapping.

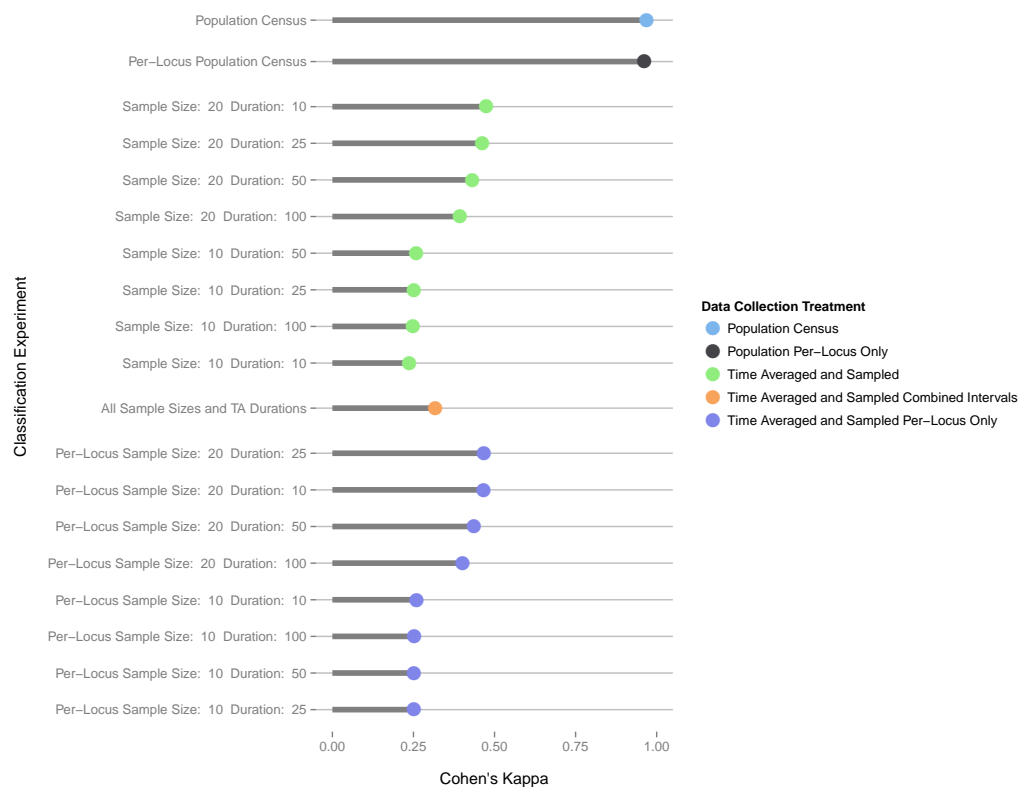


Figure 3. Lorem Ipsum

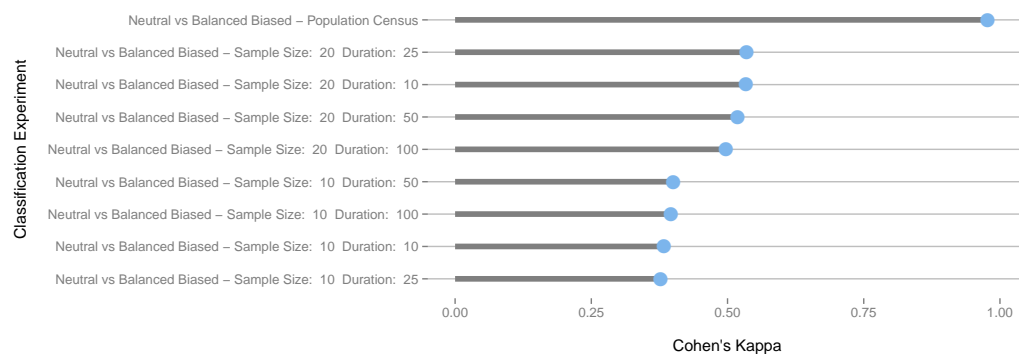


Figure 4. Lorem Ipsum

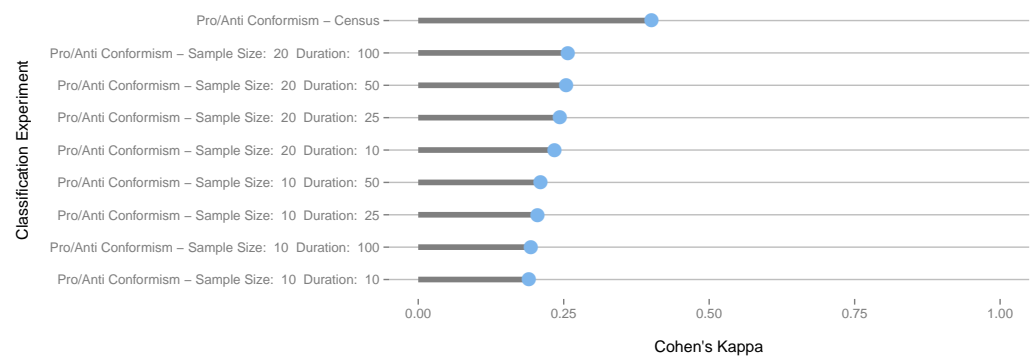


Figure 5. Lorem Ipsum