the geneticssociety

www.nature.com/hdy

ARTICLE OPEN

Check for updates

Assessing simulation-based supervised machine learning for demographic parameter inference from genomic data

Arnaud Quelin [1,2]✉, Frédéric Austerlitz [1,3] and Flora Jay [2,3]

The ever-increasing availability of high-throughput DNA sequences and the development of numerous computational methods have led to considerable advances in our understanding of the evolutionary and demographic history of populations. Several demographic inference methods have been developed to take advantage of these massive genomic data. Simulation-based approaches, such as approximate Bayesian computation (ABC), have proved particularly efficient for complex demographic models. However, taking full advantage of the comprehensive information contained in massive genomic data remains a challenge for demographic inference methods, which generally rely on partial information from these data. Using advanced computational methods, such as machine learning, is valuable for efficiently integrating more comprehensive information. Here, we showed how simulation-based supervised machine learning methods applied to an extensive range of summary statistics are effective in inferring demographic parameters for connected populations. We compared three machine learning (ML) methods: a neural network, the multilayer perceptron (MLP), and two ensemble methods, random forest (RF) and the gradient boosting system XGBoost (XGB), to infer demographic parameters from genomic data under a standard isolation with migration model and a secondary contact model with varying population sizes. We showed that MLP outperformed the other two methods and that, on the basis of permutation feature importance, its predictions involved a larger combination of summary statistics. Moreover, they outperformed all three tested ABC algorithms. Finally, we demonstrated how a method called SHAP, from the field of explainable artificial intelligence, can be used to shed light on the contribution of summary statistics within the ML models.

Heredity (2025) 134:417–426; https://doi.org/10.1038/s41437-025-00773-x

INTRODUCTION

Deciphering the demographic history of natural populations through the analysis of genetic polymorphism data represents an ongoing challenge in the field of population genetics due to the complexity of these data and their underlying processes. A better understanding of past demographic events can provide, for example, insights on specific questions such as the impact of anthropogenic processes (Pujolar et al. 2017; Dong et al. 2021) or climatic events (Bai et al. 2018; Fedorov et al. 2020) on population dynamics. It can also be helpful in conservation biology (Der Sarkissian et al. 2015; Abascal et al. 2016).

The exploration of genetic diversity also sheds light on the intricate dynamics of human populations (Cavalli-Sforza et al. 1994; Henn et al. 2012; Schraiber and Akey, 2015). Analyzing genomic polymorphisms in the genomes of modern human populations enables us to infer past demographic and evolutionary events on different timescales, including colonization, separation, migration and expansion events, as well as adaptive processes. As such, reconstructing the demographic history of populations is essential to disentangle the effects of demography from those of selection (Akey et al. 2004; Lohmueller 2014; Johri et al. 2023).

Taking full advantage of the information contained in a large number of genomic data remains a methodological and computational challenge. Several kinds of inference methods have been developed to infer past demographic events from these genomic data (Beichman et al. 2018; Marchi et al. 2021). Some of these methods are based on the sequential Markovian coalescent (SMC) model (McVean and Cardin 2005; Marjoram and Wall 2006; Li and Durbin 2011; Sheehan et al. 2013; Schiffels and Durbin 2014; Terhorst et al. 2017). They use the pattern of diversity along the genomes to estimate the inverse coalescent rate through time, which is a proxy for past population sizes under specific assumptions (Chikhi et al. 2018). Other methods use the length of regions that are identical-by-state (IBS) or identical-by-descent (IBD) between haplotypes (Palamara et al. 2012; Harris and Nielsen 2013; Browning and Browning 2015). Finally, some methods rely solely on the site frequency spectrum (Gutenkunst et al. 2009; Liu and Fu 2015). Unlike the other methods, they do not take into account the linkage disequilibrium among SNPs.

Approximate Bayesian computation (ABC) is a simulation-based method used when the likelihood function is unavailable or too expensive to compute. It initially emerged to tackle population genetics issues (Tavaré et al. 1997; Pritchard et al. 1999; Beaumont et al. 2002) and was then used in various domains such as ecology (Jabot and Lohier 2016; Wood and Simon, 2018), epidemiology (Tanaka et al. 2006; McKinley et al. 2018), system biology (Toni

[1]UMR 7206 Eco-Anthropologie (EA), CNRS, Muséum National d'Histoire Naturelle, Université Paris Cité, Paris, France. [2]UMR 9015 - Laboratoire Interdisciplinaire des Sciences du Numérique (LISN), CNRS, INRIA, Université Paris-Saclay, Orsay, France. [3]These authors contributed equally: Frédéric Austerlitz, Flora Jay. Associate editor: Olivier Hardy.
✉email: arnaud.quelin@mnhn.fr

Received: 26 August 2024 Revised: 24 May 2025 Accepted: 24 May 2025
Published online: 6 June 2025

et al. 2008; Liepe et al. 2014) or linguistics (Thouzeau et al. 2017, 2022). This framework appears particularly efficient when dealing with complex demographic models (Beaumont 2010; Bertorelle et al. 2010). It relies on the comparison of summary statistics computed on the genomic data set under investigation and on genomic data sets simulated under a given demographic model, for which the parameters are generally drawn in uninformative prior distributions. In the classical rejection sampling approach, the posterior distributions of the parameters are estimated by keeping only the simulations yielding summary statistics close to those from the real data (Tavaré et al. 1997). More elaborate methods were subsequently designed (Blum and Francois 2010; Estoup et al. 2012).

More recently, simulation-based approaches combined with supervised machine learning algorithms were developed for demographic inference (Schrider and Kern 2018; Collin et al. 2021; Korfmann et al. 2023). Some of these approaches are based on random forests (Pudlo et al. 2016; Raynal et al. 2019; Ghirotto et al. 2021), while others use neural networks (Sheehan and Song 2016; Mondal et al. 2019; Sanchez et al. 2021). As machine learning (ML) methods in demographic inference have gained ground in recent years, the focus has not been on comparing their characteristics and performance.

In this study, we focused on the inference of demographic parameters using coalescent simulations for models in which an ancestral population splits into two at some point in the past, with these populations remaining connected by migration. We implemented three machine-learning methods aiming at inferring demographic parameters from a large number of summary statistics computed on genomic data. These methods were based, respectively, on random forest (Breiman 2001), XGBoost (Chen and Guestrin 2016), and multilayer perceptron (Haykin 1994). To our knowledge, it was the first time that XGB was used in this context.

After comparing their overall performance and dissecting the impact of demographic parameters on the results, we analyzed which summary statistics contributed most and introduced a method to evaluate their influence on predictions.

METHODS

Broadly, our approach consists of (i) producing many simulated datasets under a demographic model, (ii) computing summary statistics on these simulated datasets, and (iii) using these summary statistics as a training dataset for three machine-learning methods.

Demographic models

We first considered a standard model of isolation with migration (IM) (Fig. 1A), where an ancestral population splits into two populations at a given time (*Split_time*). We assumed that, right after this splitting event, a continuous and symmetrical migration rate (*Migration_rate*), defined as the proportion of individuals moving from one population to the other per generation, was maintained between the two populations until the present. These two populations were assumed to have distinct effective population sizes (*N_current_1* and *N_current_2*), themselves distinct from the ancestral effective population size (*N_ancestral*). In this model, effective population sizes were assumed to be constant except for the change at split time.

We then considered a more complex model of secondary contact with varying population sizes (SC) (Fig. 1B). In this model, the effective population sizes of the two populations resulting from the split were no longer assumed to be constant but instead to increase or decrease exponentially, with different growth rates for the two populations. Furthermore, the populations did not undergo gene flow immediately after their separation: migration started at a later stage and lasted until the present time. This allowed us to assess the performances of the ML models on a more complex model.

Simulations

We generated 10,000 simulated data sets using *msprime* (Kelleher et al. 2016; Baumdicker et al. 2021) for each of the two demographic models. For the simulated data sets under the IM model, the five parameters were drawn in uniform prior distributions: Uniform[1;5000] for *Split_time*, corresponding to the number of generations since the split occurred; Uniform [100;10,000] for *N_ancestral*, *N_current_1*, and *N_current_2*; and Uniform[0;0.001] for *Migration_rate* (per generation). While other prior distributions could be considered, we opted for uniform distributions to ensure an even exploration of the parameter space. We note that by setting the lower bound of the *Split_time* prior to 1 generation, predictions can be particularly difficult in such extreme scenarios. For each parameter set, we simulated 20 independent neutral loci for a sample of 10 diploid individuals in each population. Each locus was a genomic sequence of length 2 Mb, with a constant recombination rate of 1.0×10^{-8} per bp per generation and a constant mutation rate of 1.25×10^{-8} per bp per generation, which are standard values for human genomes (Campbell et al. 2012; Kong et al. 2012; Scally and Durbin 2012; Schiffels and Durbin 2014).

For the SC model, the five parameters already present in the IM model were drawn from uniform prior distributions that differed to some extent from the priors of the IM model: Uniform[100;5000] for *Split_time* (in generations); Uniform[1000;5000] for *N_ancestral*, *N_current_1*, and *N_current_2*; and Uniform[0;0.005] for *Migration_rate* (per generation). In addition to the IM model, each population experienced exponential growth with per generation rates *Growth_rate_1* and *Growth_rate_2*, drawn independently in a uniform distribution: Uniform[−0.001;0.002]. Finally, the migration duration parameter was defined as the ratio between the number of generations during which migration occurred and the total number of generations since the split time. It was drawn in a uniform distribution Uniform[0,1], where 0 corresponded to a case with no migration and 1 to a case of continuous migration since the split.

Among the 10,000 simulated data sets, 5000 were used as a training data set (used to train the models), 2500 as a validation data set (used to tune hyperparameters, i.e., the parameters of the models set before the training process), and 2500 as a test data set (used to evaluate the final performance of the trained models).

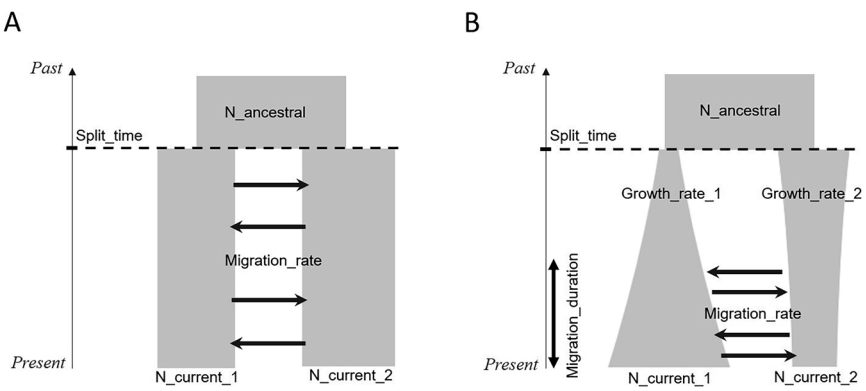

Fig. 1 Demographic models with the varying demographic parameters. A Isolation with migration (IM) model. **B** Secondary contact with varying population sizes (SC) model. Both populations can either grow or decline. *Split_time* is expressed in generations, *Migration_rate* is defined per generation, and *Growth_rate* corresponds to a per-generation exponential rate.

Summary statistics

The summary statistics that we computed on the genomic data belonged to 11 classes: eight classes were computed within populations and three among populations. Note however that the within-population statistics were computed for each population sample separately, and also on the pooled sample combining the two samples. The number of individual summary statistics varied considerably among the different summary statistics classes. Note that during training, these individual features were considered independently from the class to which they belong.

The within-population statistics classes were:

(i) *S*: the proportion of segregating sites.
(ii) *D*: Tajima's *D* statistic (Tajima 1989).
(iii) *PI*: The mean and the standard deviation of the expected heterozygosity across segregating sites (Nei and Li 1979).
(iv) *WinH*: the mean and the standard deviation of the haplotypic heterozygosity computed for non-overlapping windows of size 50 kb.
(v) Site frequency spectrum (*SFS*) statistics: The percentage of single-nucleotide polymorphisms (SNPs) where the derived allele is present at frequency *i*, for all *i* in [1, …, 2*n*], where *n* is the sample size, and the standard deviation of the distances between two adjacent SNPs with the derived allele at frequency *i* (Boitard et al. 2016; Jay et al. 2019).
(vi) *LD*: mean and standard deviation of the squared correlation r^2 of all pairs of SNPs in a given bin of distance. Following Boitard et al. (2016) and Jay et al. (2019), 19 bins of distance with means ranging from 282 bp to 1.4 Mb were considered.
(vii) *IBS*: Deciles of segment length distribution completely identical between *m* haplotypes for $m = 2$, 4, 8, and 16 (Boitard et al. 2016; Jay et al. 2019; inspired by Harris and Nielsen 2013).
(viii) *AFIBS*: For each SNP, the *AFIBS* segment refers to the region surrounding the SNP that remains identical across all haplotypes containing the derived allele at that particular SNP (Theunert et al. 2012). We computed the mean and the standard deviation of the length of *AFIBS* segments for the SNPs of frequency *i* for all *i* in [1, …, 2*n*].

The among-population statistics classes were:

(ix) *Fst*: the fixation index (Wright 1950).
(x) *Dxy*: genetic divergence between the two populations (Nei and Li 1979).
(xi) *JSFS*: Joint site frequency spectrum (Wakeley and Hey 1997): The percentage of SNPs where the derived allele is present at a frequency *i* in population 1 and *j* in population 2.

The statistics of the classes *S*, *D*, *PI*, *Fst*, *Dxy*, and *JSFS* were calculated using the *tskit* package (Kelleher et al. 2016), while those for classes *WinH*, *SFS*, *LD*, *IBS* and *AFIBS* were computed using scripts adapted from Jay et al. (2019) in Python 3.9.7. All summary statistics were computed for each independent locus. We then computed the mean, median, and variance across loci of these values for each simulation, yielding a total of 3024 summary statistics (Table S1).

Parameter inference methods

We explored the effectiveness of three regression methods to infer the demographic parameters: random forest (RF), gradient boosting system XGBoost (XGB), and multilayer perceptron (MLP). All models underwent optimization of their hyperparameters using a validation set. The CPU times used to train and predict these three methods are shown in Fig. S1.

RF (Breiman 2001) is an ensemble method based on the construction of multiple decision trees. This method reduces the uncertainty of the prediction by averaging many predictions from the individual trees. The hyper-optimization procedure led us to set the number of trees between 200 and 300, with depths ranging from 15 to 20 (the exact values depended upon the demographic parameters).

While RF uses independent predictors, XGB (Chen and Guestrin 2016) uses a sequential approach where each new tree is adjusted to correct errors in the existing model. This method assigns higher weights to observations that were poorly predicted in earlier iterations, thus prioritizing their correction in the construction of subsequent decision trees. For this second ensemble learning technique, the tuning of hyperparameters on the validation set led us to set the number of trees to 100, with a depth of seven.

Finally, MLP (Haykin 1994) is a neural network method made up of fully connected layers. We trained a separate network for each demographic parameter, although a multi-output MLP could also be used. Depending on the demographic parameters, a hyper-optimization preliminary study on the validation set led us to set the number of neuron layers to six or seven (Table S2). For all models, summary statistics were standardized before training and the loss function was the mean square error (MSE). Unlike RF or XGB, MLP can lead to predictions outside the prior intervals. In that case, the prediction was set to the closest bound of the prior interval. All MLP models incorporate the ReLU function as an activation function (Agarap 2019). We also performed a preliminary study on the impact on prediction quality of the introduction of an L1 regularization penalty, which led us not to use this option (Fig. S2).

We implemented the RF and MLP methods using the scikit-learn python library (Pedregosa et al. 2011), and the XGBoost library for the XGB method (Chen and Guestrin 2016).

Feature contributions

We used permutation feature importance (PFI) to gauge the contribution of each feature (i.e., each summary statistic) to the performance of a specific model. It consists in randomly shuffling the values of a given feature among replicates and observing the subsequent degradation of the model's performance. By breaking the feature-target relationship, it allows to determine the extent to which the model depends on this particular feature. In this study, we used the Permutation importance function from the sickit-learn library.

Moreover, we used Shapley values to assess the extent to which each summary statistic impacts the model's predictions for each demographic parameter, measuring both the direction and magnitude of their contributions. Shapley values first appeared in a context of cooperative game theory (Shapley 1953) to estimate the contribution of individual players to the outcome of a game. In the ML context, the players are represented by the features and the cooperation is represented by combinations of features. We used the Python SHAP library (Lundberg and Lee 2017), which provides visual representations of the influence of the feature value on the outcome.

Comparison with ABC methods

The three regression methods were compared with several ABC algorithms from the 'abc' package (Csilléry et al. 2012), namely ABC rejection, ABC loclinear and ABC neuralnet, tested on the same simulation set. In addition to a rejection step, the last two methods apply corrections based on a local linear regression and a neural network, respectively. Their efficiency was computed on the same 2500 samples from the test data set used to assess the regression methods. For each of these 2500 samples, we kept the simulations closest to the target among the 7500 simulations not used in the test data set, for tolerance thresholds ranging from 5×10^{-4} to 5×10^{-2} (Table S3).

For the loclinear method, we selected the summary statistics with the highest correlation with the demographic parameters, in order to comply with the limitations of this method on the maximal number of summary statistics. Thus, we retained up to 300 statistics in the case of the 5×10^{-2} tolerance threshold. Similarly, to follow neuralnet specifications, we kept the 300 statistics with the highest correlation with the target parameter while considering a network with three neurons in its hidden layer. For all three algorithms, we then used either the mean or the median of the posterior as a point estimate.

RESULTS

Accuracy of the machine-learning method for parameter estimation under the IM model

After training the three ML methods for each demographic parameter of the standard IM model on the training data set, we evaluated their performances on the test data set. We compared the predictions of these three ML methods using different metrics (Table 1). In the following, *N_current* refers to results for population 1. Since the model is symmetric and both population sizes have the same prior, the results obtained for *N_current_1* and *N_current_2* are similar (see Table S4 for population 2).

Root mean square error (RMSE) and mean absolute error (MAE) are commonly used measures, and the choice between them

Table 1. Prediction errors on the test data set of the three ML methods, for each demographic parameter of the standard IM model.

Demographic parameter	Method	RMSE	MAE *(se)*	NMAE *(se)*
Split_time	RF	7.28E + 02	5.22E + 02 *(1.02E + 01)*	1.04E-01 (*2.03E-03*)
	XGB	6.76E + 02	4.67E + 02 *(1.02E + 01)*	9.34E-02 (*1.95E-03*)
	MLP	**6.35E + 02**	**3.85E + 02** *(9.77)*	**7.71E-02** (*2.01E-03*)
Migration_rate	RF	1.32E-04	8.82E-05 *(1.97E-06)*	8.82E-02 (*1.97E-03*)
	XGB	1.22E-04	8.00E-05 *(1.84E-06)*	8.00E-02 (*1.84E-03*)
	MLP	**1.21E-04**	**7.52E-05** *(1.91E-06)*	**7.52E-02** (*1.91E-03*)
N_ancestral	RF	1.84E + 02	1.29E + 02 *(2.61)*	1.29E-02 (*2.61E-04*)
	XGB	**1.81E + 02**	1.30E + 02 *(2.51)*	1.30E-02 (*2.52E-04*)
	MLP	1.89E + 02	**1.27E + 02** *(2.81)*	**1.27E-02** (*2.81E-04*)
N_current	RF	6.55E + 02	4.27E + 02 *(9.95)*	4.27E-02 (*9.96E-04*)
	XGB	6.09E + 02	3.90E + 02 *(9.36)*	3.91E-02 (*9.36E-04*)
	MLP	**5.98E + 02**	**3.74E + 02** *(9.32)*	**3.75E-02** (*9.33E-04*)

RMSE root mean square error, *MAE* mean absolute error, *NMAE* mean absolute error normalized by the range of the parameter. The lowest errors for each parameter are indicated in bold.

depends on the specificities of the issue to be tackled (Chai and Draxler 2014). The squaring operation in RMSE penalizes large errors more heavily, which is particularly appropriate when avoiding outlier prediction is critical. For three of the four demographic parameters, *Split_time*, *Migration_rate* and *N_current*, MLP performed better than RF and XGB for these two metrics. Conversely, for the *N_ancestral* parameter, the XGB model performed better according to the RMSE metric, while the MLP model performed better according to the MAE metric.

The MAE metric presents the advantage of directly reporting the average magnitude of the errors. *Split_time* was the demographic parameter for which the MLP model outperformed the two other models by the largest margin: the MAE of RF and XGB were 522 and 467 generations respectively, i.e., 36 and 21% higher than the MAE of MLP, which was only of 385 generations. For the migration rate, which ranged between zero and 10^{-3} per generation in the simulations, MLP achieved an average MAE of around 7.5×10^{-5} per generation, compared with 8.0×10^{-5} and 8.8×10^{-5}, respectively, for the XGB and RF models, i.e., a higher MAE by 6 and 17%, respectively. For ancestral and current effective population sizes, the lowest average errors were also achieved by the MLP models. For ancestral effective population size, MLP outperformed only marginally the other methods, with a mean error for MLP of 127 individuals compared with 129 and 130 individuals, respectively, for the RF and XGB models. The difference was more substantial for the current effective population size: for this parameter, the MAE of RF and XGB were higher by 14 and 4%, respectively, than the MAE of 375 of the MLP.

The NMAE represents a normalized version of the mean absolute error (MAE), in which the error for a given parameter is divided by the difference between the minimum and maximum bounds of the prior of this parameter. This allowed us to have a comparison between the estimates of all parameters, whatever their range. *N_ancestral* had the lowest NMAE (0.0127). It was thus the parameter that could be estimated with the highest accuracy, followed by *N_current* (NMAE of 0.0375). *Migration_rate* and *Split_time* parameters, with respective NMAE of 0.0752 and 0.0771, showed the lowest accuracy.

Observing predicted versus observed values allowed us to obtain more detailed information about the predictions for each parameter (Fig. 2; Fig. S3). RF and XGB methods showed both a downward bias for high values of splitting time and migration rate, in contrast to MLP (Fig. 2). We also observed significant upward biased predictions for the RF model for low migration rates, leading to a large overall error for this method on this demographic parameter.

Comparison with ABC methods

The performance of the ABC algorithms depended on both the tolerance thresholds and the choice of using either the median or the mean of the posterior as estimator (Table S3). Whatever the tolerance threshold chosen, the three ML methods all outperformed these ABC algorithms on the same simulation set (Fig. S4). The ABC rejection algorithm always showed lower performances than the ABC loclinear and ABC neuralnet methods. Compared with the three ML methods, the performances of ABC algorithms in predicting current and ancestral effective population sizes were substantially lower. The difference between the three ML methods and the ABC algorithms was less pronounced for the migration rate parameter, but still in favor of the ML methods.

Accuracy of the machine-learning method for parameter estimation under the SC model

We then trained the same hyper-parametrized models on the SC scenarios. In the following, *N_current* and *Growth_rate* refer to results for population 1. As the model is symmetrical and the priors are the same for both populations, the results are similar to those for population 2.

For the parameters already present in the standard IM model, MLP was again generally the best-performing method (Table 2). This was particularly noteworthy for *Split_time*, for which RF and XGB showed both a downward bias for high values of that parameter (Fig. S5). *N_ancestral* remained the most accurately predicted parameter with the lowest NMAE (4.19×10^{-2} for MLP). The model with the lowest RMSE for *Growth_rate* prediction was XGB (6.06×10^{-4}), while MLP showed the lowest MAE (4.51×10^{-4}). Finally, for the migration duration, MLP achieved the best performances in terms of both RMSE and MAE.

In the following sections, we focus on the standard isolation with migration model (IM).

Assessing accuracy discrepancies across demographic scenarios

We focus in this section on MLP, as it showed the lowest MAE. We investigated if there was a heterogeneity in prediction quality according to the demographic scenarios. In order to address this question, we used regression to identify to what extent the different demographic parameters could explain prediction errors

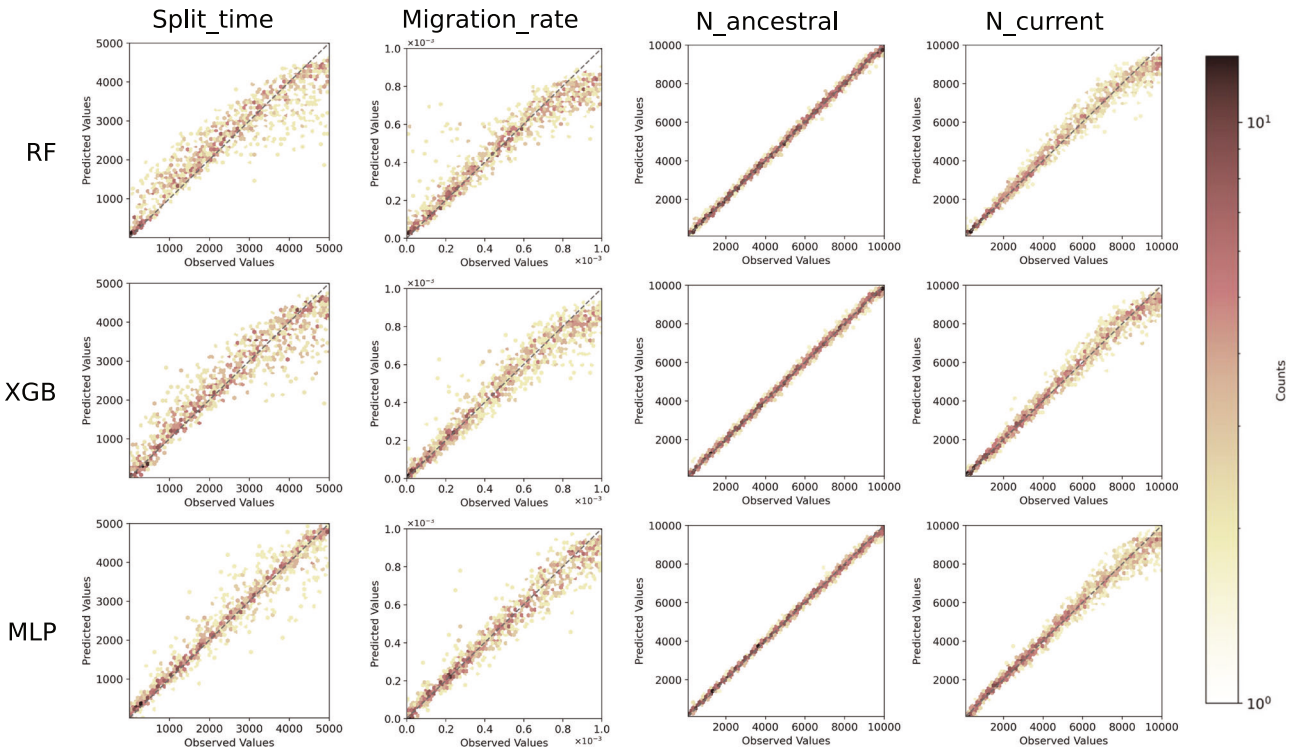

Fig. 2 Predicted values versus observed values for the four isolation-with-migration parameters: *Split_time* (in generations), *Migration_rate* (per generation), *N_ancestral*, and *N_current*, estimated with three different ML methods: RF (Random Forest), XGB (XGBoost), and MLP (Multilayer perceptron).

Table 2. Prediction errors on the test data set of the three ML methods, for each demographic parameter of the secondary contact with varying population sizes model.

Demographic parameter	Method	RMSE	MAE (se)	NMAE (se)
Split_time	RF	8.37E + 02	6.74E + 02 (4.98E + 02)	1.37E-01 (1.02E-01)
	XGB	8.34E + 02	6.45E + 02 (5.28E + 02)	1.32E-01 (1.08E-01)
	MLP	**7.32E + 02**	**5.29E + 02 (5.06E + 02)**	**1.08E-01 (1.03E-01)**
Migration_rate	RF	7.48E-04	5.58E-04 (4.98E-04)	1.12E-01 (9.96E-02)
	XGB	7.33E-04	5.29E-04 (5.08E-04)	1.06E-01 (1.02E-01)
	MLP	**7.12E-04**	**5.08E-04 (4.99E-04)**	**1.02E-01 (9.97E-02)**
N_ancestral	RF	3.29E + 02	1.94E + 02 (2.65E + 02)	4.86E-02 (6.64E-02)
	XGB	3.18E + 02	1.88E + 02 (2.57E + 02)	4.70E-02 (6.43E-02)
	MLP	**3.10E + 02**	**1.68E + 02 (2.61E + 02)**	**4.19E-02 (6.51E-02)**
N_current	RF	4.72E + 02	3.71E + 02 (2.92E + 02)	9.27E-02 (7.30E-02)
	XGB	**4.32E + 02**	3.34E + 02 (2.75E + 02)	8.34E-02 (6.86E-02)
	MLP	4.42E + 02	**3.31E + 02 (2.93E + 02)**	**8.26E-02 (7.33E-02)**
Growth_rate	RF	6.07E-04	4.89E-04 (3.60E-04)	1.63E-01 (1.20E-01)
	XGB	**6.06E-04**	4.71E-04 (3.81E-04)	1.57E-01 (1.27E-01)
	MLP	6.17E-04	**4.51E-04 (4.21E-04)**	**1.50E-01 (1.40E-01)**
Migration_duration	RF	2.19E-01	1.79E-01 (1.26E-01)	1.79E-01 (1.26E-01)
	XGB	2.23E-01	1.78E-01 (1.35E-01)	1.78E-01 (1.35E-01)
	MLP	**2.13E-01**	**1.60E-01 (1.41E-01)**	**1.60E-01 (1.41E-01)**

RMSE root mean square error, *MAE* mean absolute error, *NMAE* mean absolute error normalized by the range of the parameter. The lowest errors for each parameter are indicated in bold.

(Fig. S6). We then plotted these prediction errors as a function of the two most explanatory parameters (Fig. 3).

Regarding the prediction of split time, we observed that the best performance was achieved for smaller ancestral effective population sizes (Fig. 3A). As this effective population size increased, the prediction error increased. This effect was more pronounced for older split times. When predicting the migration rate, MLP performed less efficiently for recent split times. As the

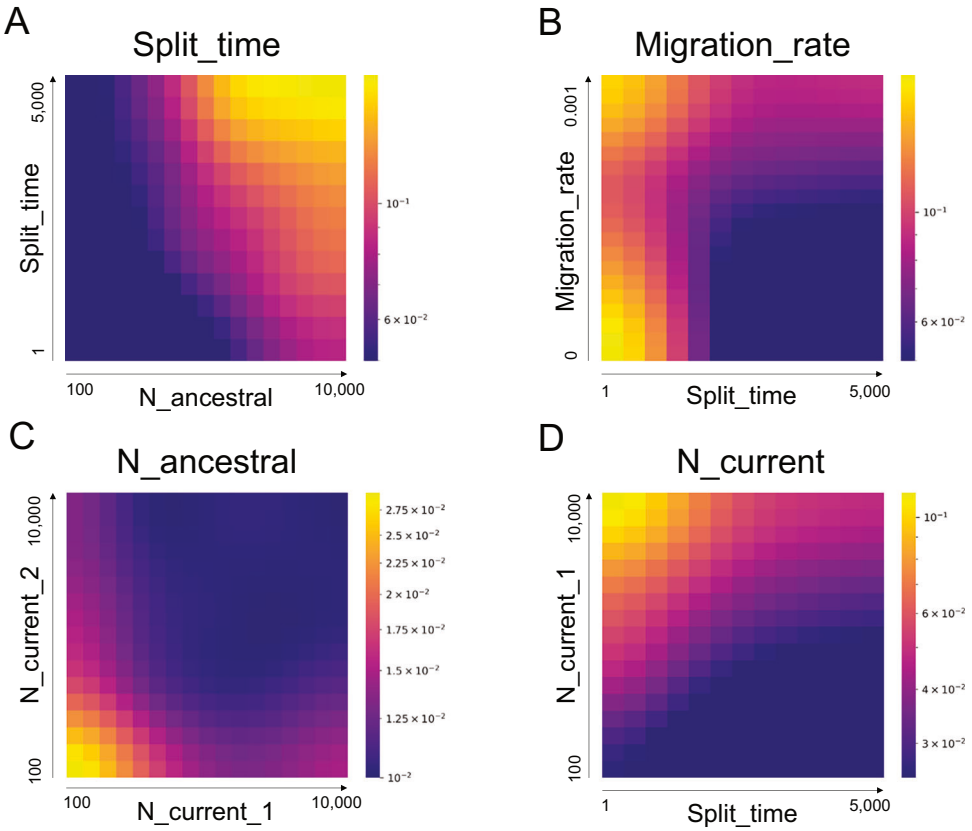

Fig. 3 Standardized error patterns in MLP predictions. Standardized errors $(\hat{y}_i - y_i)/(y_{max} - y_{min})$ where y is the target demographic parameter: *Split_time* (in generations) (**A**), *Migration_rate* (per generation) (**B**), *N_ancestral* (**C**), and *N_current* (**D**), as a function of the two variables most predictive of these errors under the MLP method. Cold colors correspond to areas of low error and therefore better model prediction, while warm colors correspond to areas of higher error.

split time increased, the prediction quality was higher for a lower migration rate (Fig. 3B).

As for the ancestral effective population size, the values taken by the current effective population sizes had the largest impact on the quality of the prediction. The lowest accuracy was observed for the ancestral size when both current populations showed low effective sizes. When one of the two effective population sizes was small, prediction quality remained lower than average, regardless of the effective size of the second population (Fig. 3C).

Split time was the most decisive parameter in the quality of the prediction for current effective population size parameters. As this time increased, the prediction became increasingly accurate, whereas demographic scenarios with recent split time and high current effective population sizes led to the highest errors.

Different patterns of summary statistic usage across ML models

We estimated the importance levels of various summary statistics classes for each ML model and each demographic parameter in order to determine to which extent they differ in that aspect. We computed the importance of each variable, using the PFI method for the three ML methods, which computes the extent to which the quality of a prediction of a given model decreases when a single variable is randomly shuffled.

We observed first that RF and XGB allocated quite similar importance levels to the various classes of summary statistics (Fig. 4), while the MLP method generally showed different behaviors. For *Split_time*, the *SFS* class stood out for the two ensemble methods, with importance levels of 0.71 and 0.66 respectively. The *JSFS* class was second in terms of importance (importance levels of 0.10 and 0.16 for RF and XGB, respectively), but the gap was high with the *SFS* class. Conversely, for MLP, the *JSFS* class was the most important, with an importance of 0.63, ahead of the *SFS* and *AFIBS* classes. The differences among methods were much less pronounced for the *Migration_rate* parameter, where the *AFIBS*, *SFS*, and foremost *JSFS* classes were used predominantly by all methods. Nevertheless, the XGB and RF models also gave significant importance to *Fst*, unlike the MLP model.

The differences were much more striking for the effective population sizes. For the ancestral effective population size, RF and XGB focused almost exclusively on the *IBS* class, with importance levels of 0.99 and 0.96, respectively. Conversely, MLP used the different classes much more evenly, with five classes showing noticeable levels of importance. The discrepancy among methods was a bit lower but still striking for the current effective population sizes. All methods used several classes of summary statistics, but the importance levels were strongly unbalanced toward *AFIBS* for RF and XGB, with values 0.66 and 0.65 respectively, while MLP used again the different classes more evenly.

This similarity in the variables used by XGB and RF was confirmed by observing the individual importance levels of each summary statistics (see Fig. S7). In the case of split time, the variables related to *SFS* were of the utmost importance for the predictions made by these two methods. In a similar manner, these methods assigned the greatest importance to *Fst*, followed by statistics from the *JSFS* class for predicting the migration rate. The most important variables for predicting the ancestral effective population size belonged all to the *IBS* class, while for the current effective population size these variables belonged to different classes: *AFIBS*, *JSFS* and *winHET*. The most important summary statistics chosen by MLP differed strikingly from those chosen by RF and XGB. They also exhibited markedly lower importance

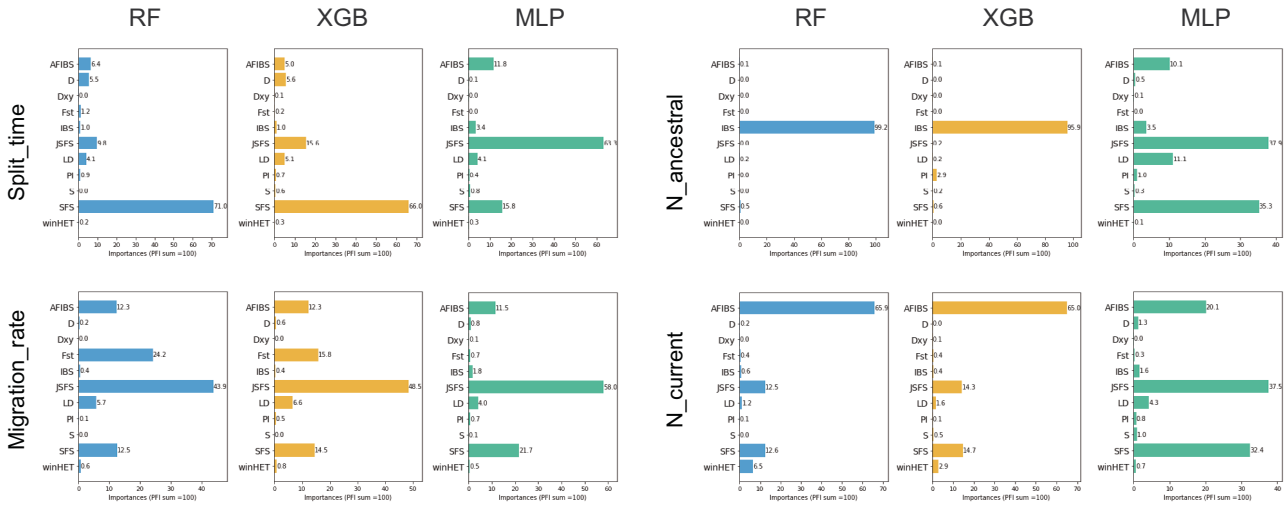

Fig. 4 Importance levels of the different classes of summary statistics as a percentage of the total for each demographic parameter and each method (RF, XGB, MLP).

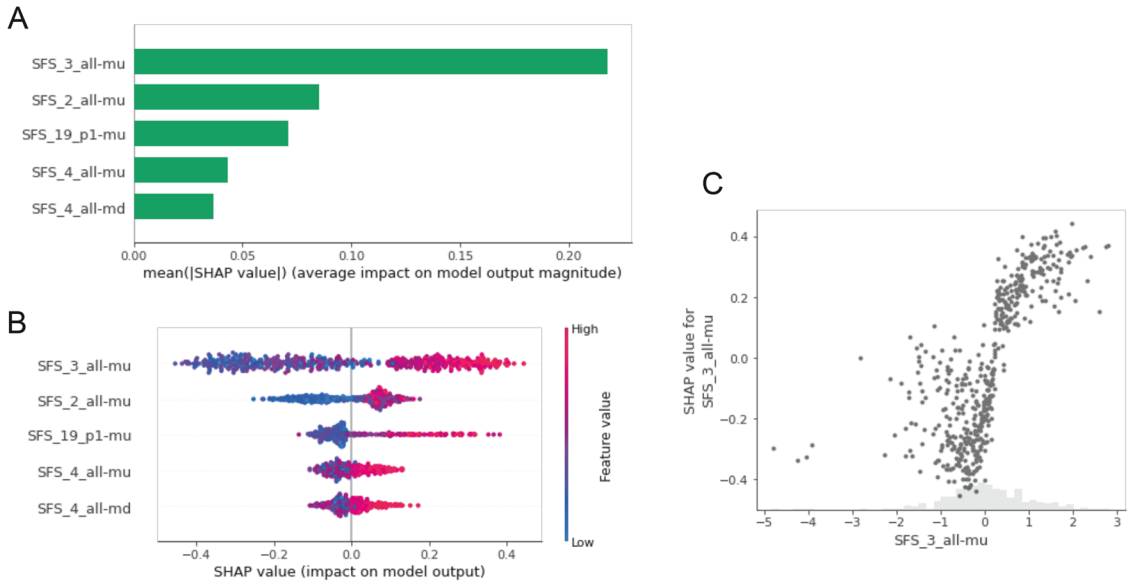

Fig. 5 SHAP-based summary statistics analysis in *Split_time* prediction by XGB model. A Top features with the highest means of absolute SHAP values for the XGB model on the *Split_time* parameter. **B** Feature effects indicated by the Shapley value on the x-axis for the top 5 features with their values represented by color. Each point corresponds to one instance. **C** Shapley values as a function of the normalized *SFS_3_all-mu* (gray points), distribution of the normalized *SFS_3_all-mu* (gray bars).

values. Indeed, the maximum importance attributed to a single feature by MLP was 2.6% for the *Migration_rate* parameter. For this method, importance was clearly spread over a much larger number of variables than for the two other methods (Fig. S8).

Towards improved explainability

The ML methods do not directly provide the impact of each individual statistic on the parameter estimation, due to their inherent complexity, nonlinear relationships, or the fact that they are a combination of a set of predictors. This makes the interpretation of the individual impact of each summary statistics more complex. While some methods such as the PFI method used above are effective in revealing features that are important for prediction, they reach their limit when it comes to quantifying the impact of features while taking into account how they interact with each other. Interpretation techniques may then prove necessary to provide insights into the role of summary statistics in complex models. We therefore used SHAP (Lundberg and Lee 2017), a method based on Shapley values that measures the average marginal contribution of each feature when combined with all other to make a prediction for a given parameter. In addition to identifying the most important summary statistics for each parameter, this method also allows investigating finely how these statistics affect the model outputs, quantifying both the direction and magnitude of their contributions to the prediction of each parameter. We illustrate the contributions of this method in the IM model for the *Split_time* predictions obtained with the XGB model, which combined high accuracy with an ability to bring out statistics with substantial importance (Fig. S7).

Focusing first on the summary statistics with the greatest impact, we observed that the frequency of tripletons in the overall population (*SFS_3_all-mu*) showed the highest mean of absolute SHAP values (Fig. 5A). This statistic was also the one that showed the highest importance (Fig. S7). The distribution of SHAP values for the most important variables allowed us to understand the impact of individual summary statistics on predictions (Fig. 5B).

We observe that *SFS_3_all-mu* had the greatest upward and downward impact on the prediction of *Split_time*. While the highest values of this variable almost always had an upward impact, we see that intermediate values had a substantial impact, particularly downward, when combined with other statistics (Fig. 5C). We also observed that for the variable *SFS_19_p1-mu*, which represents the frequency of SNPs with the ancestral allele found in a single haplotype, higher values have a greater impact on the parameter compared to lower values, which have a more limited impact (Fig. 5B).

DISCUSSION

We have shown here how an approach combining a large number of summary statistics and machine learning methods was particularly effective in predicting demographic parameters in scenarios involving two connected populations. These methods do not require preselecting a subset of summary statistics in order to reduce the number of variables. Both ensemble and neural network methods are indeed suitable for learning from a very large number of statistics, some of which being highly correlated, and some being aggregated indicators that can be drawn from others. We showed that MLP overall outperformed the RF and XGB methods in parameter prediction for the considered IM and SC model. Other demographic and genetic modeling assumptions could be tested in future works using the available scripts, as well as the effect of the number of simulations on accuracy. Other prior distributions could also be tested, such as log-uniform distributions, which are of particular interest when we are interested in inferring the order of magnitude of a parameter. Although achieving lower accuracy, RF and XGB still showed good performances. Depending on the metric and degree of accuracy required, they could still be considered in some cases, as they are less computer-intensive.

The intricate effects of the demographic parameters on the summary statistics undermine the predictive power of these statistics. We therefore emphasize the importance of studying the impact of confounding factors (here, the other demographic parameters) and the heterogeneity of prediction accuracy across the parameter space. Within the isolation with migration framework, we identified which demographic parameters were most decisive in the accuracy of the prediction of our target parameter, and visualized how this accuracy varies as a function of these most explanatory parameters. Taking such an approach when building a new model for a given evolutionary question ensures a precise picture of the inference strengths and weaknesses.

The three machine learning methods showed contrasting patterns in terms of summary statistics usage for predicting demographic parameters in the modeling settings of the IM model. Although XGB showed generally better performances than RF, both methods used quite similar learning schemes. MLP, on the other hand, used strikingly different classes of summary statistics than the two other methods. Furthermore, whereas RF and XGB focused on a limited number of summary statistics to which they gave a high importance, MLP used a much larger number of statistics, each being given a small importance. In other words, the information used by MLP was spread much more evenly over the different statistics. Based on a higher degree of complexity, this method therefore showed its higher capacity to capture intricate patterns and complex relationships from a high number of variables, as for instance those of the *JSFS* class.

Conversely, by pointing at a limited number of important statistics for prediction, tree-based methods reveal few important summary statistics for inference. We therefore note that an error on one of the most important summary statistics is likely to have a strong impact on the estimation of the demographic parameter. However, this has the advantage of allowing us to identify and select the specific classes of summary statistics to which we may wish to restrict ourselves for a particular demographic model, enhancing the interpretability of the results. For this purpose, we can rely on the recently developed interpretability method SHAP (Lundberg and Lee 2017), which was notably used to interpret mutation rate predictions from a neural network based on SFS (Burger et al. 2022) and in a deep learning framework for selection inference (Cecil and Sugden 2023). This method allowed us to assess how the values of a given summary statistic affect the prediction of the demographic parameters under study. Ultimately, while the methods developed here were tested in specific demographic scenarios, with mutation and recombination rates consistent with known values for humans, they could be adapted to other contexts, in which their performances could be tested with the same procedure as we used here. The impact of the sample size and of the number of loci could also be studied either using realistic ranges or matching the exact values of the user's dataset.

Note also that the methods that we implemented here are likely to undergo several developments in the future. Our simulation approach was based indeed on a strictly neutral model, without any selective pressure. However, it has been shown that positive selection (Schrider et al. 2016), as well as background and purifying selection (Ewing and Jensen 2016; Pouyet et al. 2018; Johri et al. 2021), have an impact on demographic inferences. Therefore, while it is difficult for other approaches such as composite-likelihood and SMC-based methods to account for selection explicitly, methods developed in this study could naturally be extended to models that incorporate these selective effects, allowing for a more comprehensive assessment of their impact on demographic inferences.

Similarly, we simulated sequences with even mutation and recombination rates, and without sequencing errors. A relaxation of these assumptions could also be explored (see e.g., Boitard et al. 2016; Jay et al. 2019). We further point out that while all demographic inference methods are sensitive to model assumptions, simulation-based approaches including ABC methods are also sensitive to simulation misspecification. Since the prediction of demographic parameters is restricted to predefined prior bounds, it is essential to ensure that these bounds are set carefully, without being too restrictive. An additional area of focus could be to investigate how these methods perform for model selection, often a preliminary step to parameter estimation.

Finally, even if we showed here the high predictive power of our methods, we suggest to compare their efficiency with that of methods that rely on the direct application of deep learning methods to raw genomic data, in order to infer demographic (Sanchez et al. 2021) or selection events (Flagel et al. 2019; Torada et al. 2019). By learning more complex patterns from the data that may not be captured by summary statistics devised beforehand, these methods may prove particularly effective in case of complex demographic models. However, as they do not benefit from the guidance provided by the summary statistics, it is necessary to compare their performances with those of methods such as the ones that we developed here, as carried out by Sanchez et al. (2021) for the estimation of the effective size of a population over time. In this context, they showed that current neural networks yielded accurate predictions without requiring handcrafted features, but summary statistics were already compelling if carefully optimizing the inference model.

DATA AVAILABILITY

Scripts used to generate simulated data and run the demographic inferences are available on github at: https://github.com/amquelin/SML_demographic_inference.

REFERENCES

Abascal F, Corvelo A, Cruz F, Villanueva-Cañas JL, Vlasova A, Marcet-Houben M et al. (2016) Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biol 17(1):251. https://doi.org/10.1186/s13059-016-1090-1.

Agarap AF (2019) Deep learning using rectified linear units (ReLU) (arXiv:1803.08375). https://doi.org/10.48550/arXiv.1803.08375.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA et al. (2004) Population history and natural selection shape patterns of genetic variation in 132 genes. PLOS Biol 2(10):e286. https://doi.org/10.1371/journal.pbio.0020286.

Bai W-N, Yan P-C, Zhang B-W, Woeste KE, Lin K, Zhang D-Y (2018) Demographically idiosyncratic responses to climate change and rapid Pleistocene diversification of the walnut genus Juglans (Juglandaceae) revealed by whole-genome sequences. New Phytologist 217(4):1726–1736. https://doi.org/10.1111/nph.14917.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G et al. (2021) Efficient ancestry and mutation simulation with msprime 1.0. Genetics 220(3):iyab229. https://doi.org/10.1093/genetics/iyab229.

Beaumont MA (2010) Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst 41(1):379–406. https://doi.org/10.1146/annurev-ecolsys-102209-144621.

Beaumont MA, Zhang W, Balding DJ (2002). Approximate Bayesian computation in population genetics. *Genetics* 162(4), 2025–2035. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1462356/.

Beichman AC, Huerta-Sanchez E, Lohmueller KE (2018) Using genomic data to infer historic population dynamics of nonmodel organisms. Annu Rev Ecol Evol Syst 49(1):433–456. https://doi.org/10.1146/annurev-ecolsys-110617-062431.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: Some cons, many pros. Mol Ecol 19(13):2609–2625. https://doi.org/10.1111/j.1365-294X.2010.04690.x.

Blum MGB, Francois O (2010) Non-linear regression models for approximate Bayesian computation. Stat Comput 20(1):63–73. https://doi.org/10.1007/s11222-009-9116-0.

Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F (2016) Inferring population size history from large samples of genome-wide molecular data—an Approximate Bayesian Computation approach. PLOS Genet 12(3):e1005877. https://doi.org/10.1371/journal.pgen.1005877.

Breiman L (2001) Random forests. Machine Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324.

Browning SR, Browning BL (2015) Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am J Hum Genet 97(3):404–418. https://doi.org/10.1016/j.ajhg.2015.07.012.

Burger KE, Pfaffelhuber P, Baumdicker F (2022) Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown. PLOS Comput Biol 18(8):e1010407. https://doi.org/10.1371/journal.pcbi.1010407.

Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, Vives L, O'Roak BJ et al. (2012) Estimating the human mutation rate using autozygosity in a founder population. Nat Genet 44(11):1277–1281. https://doi.org/10.1038/ng.2418.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press.

Cecil RM, Sugden LA (2023) On convolutional neural networks for selection inference: revealing the effect of preprocessing on model learning and the capacity to discover novel patterns. PLOS Comput Biol 19(11):e1010979. https://doi.org/10.1371/journal.pcbi.1010979.

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.

Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Chikhi L, Rodríguez W, Grusea S, Santos P, Boitard S, Mazet O (2018) The IICR (Inverse Instantaneous Coalescence Rate) as a summary of genomic diversity: insights into demographic inference and model choice. Heredity 120(1):13–24.

Collin F, Durif G, Raynal L, Lombaert E, Gautier M, Vitalis R et al. (2021) Extending approximate Bayesian computation with supervised machine learning to infer demographic history from genetic polymorphisms using DIYABC random forest. Mol Ecol Resour 21(8):2598–2613. https://doi.org/10.1111/1755-0998.13413.

Csilléry K, François O, Blum MGB (2012) abc: an R package for Approximate Bayesian Computation (ABC). Methods Ecol Evol 3(3):475–479. https://doi.org/10.1111/j.2041-210X.2011.00179.x.

Der Sarkissian C, Ermini L, Schubert M, Yang MA, Librado P, Fumagalli M et al. (2015) Evolutionary genomics and conservation of the endangered Przewalski's Horse. Curr Biol 25(19):2577–2583. https://doi.org/10.1016/j.cub.2015.08.032.

Dong F, Kuo H-C, Chen G-L, Wu F, Shan P-F, Wang J et al. (2021) Population genomic, climatic and anthropogenic evidence suggest the role of human forces in endangerment of green peafowl (Pavo muticus). Proc R Soc B: Biol Sci 288(1948):20210073. https://doi.org/10.1098/rspb.2021.0073.

Estoup A, Lombaert E, Marin J-M, Guillemaud T, Pudlo P, Robert CP et al. (2012) Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics. Mol Ecol Resour 12(5):846–855. https://doi.org/10.1111/j.1755-0998.2012.03153.x.

Ewing GB, Jensen JD (2016) The consequences of not accounting for background selection in demographic inference. Mol Ecol 25(1):135–141. https://doi.org/10.1111/mec.13390.

Fedorov VB, Trucchi E, Goropashnaya AV, Waltari E, Whidden SE, Stenseth NChr (2020) Impact of past climate warming on genomic diversity and demographic history of collared lemmings across the Eurasian Arctic. Proc Natl Acad Sci 117(6):3026–3033. https://doi.org/10.1073/pnas.1913596117.

Flagel L, Brandvain Y, Schrider DR (2019) The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol Biol Evol 36(2):220–238. https://doi.org/10.1093/molbev/msy224.

Ghirotto S, Vizzari MT, Tassi F, Barbujani G, Benazzo A (2021) Distinguishing among complex evolutionary models using unphased whole-genome data through random forest approximate Bayesian computation. Mol Ecol Resour 21(8):2614–2628. https://doi.org/10.1111/1755-0998.13263.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLOS Genet 5(10):e1000695. https://doi.org/10.1371/journal.pgen.1000695.

Harris K, Nielsen R (2013) Inferring demographic history from a spectrum of shared haplotype lengths. PLOS Genet 9(6):e1003521. https://doi.org/10.1371/journal.pgen.1003521.

Haykin S (1994) Neural networks: a comprehensive foundation. Prentice Hall PTR.

Henn BM, Cavalli-Sforza LL, Feldman MW (2012) The great human expansion. Proc Natl Acad Sci USA 109(44):17758–17764. https://doi.org/10.1073/pnas.1212380109.

Jabot F, Lohier T (2016) Non-random correlation of species dynamics in tropical tree communities. Oikos 125(12):1733–1742. https://doi.org/10.1111/oik.03103.

Jay F, Boitard S, Austerlitz F (2019) An ABC method for whole-genome sequence data: inferring paleolithic and neolithic human expansions. Mol Biol Evol 36(7):1565–1579. https://doi.org/10.1093/molbev/msz038.

Johri P, Pfeifer SP, Jensen JD (2023) Developing an evolutionary baseline model for humans: jointly inferring purifying selection with population history. Mol Biol Evol 40(5):msad100. https://doi.org/10.1093/molbev/msad100.

Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD (2021) The impact of purifying and background selection on the inference of population history: problems and prospects. Mol Biol Evol 38(7):2986–3003 https://doi.org/10.1093/molbev/msab050.

Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. PLOS Comput Biol 12(5):e1004842. https://doi.org/10.1371/journal.pcbi.1004842.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G et al. (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488(7412), 7412. https://doi.org/10.1038/nature11396.

Korfmann K, Gaggiotti OE, Fumagalli M (2023) Deep learning in population genetics. *Genome Biol Evol* evad008. https://doi.org/10.1093/gbe/evad008.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature* 475 (7357), 7357. https://doi.org/10.1038/nature10231.

Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH (2014) A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc* 9(2), 2. https://doi.org/10.1038/nprot.2014.025.

Liu X, Fu Y-X (2015) Exploring population size changes using SNP frequency spectra. Nat Genet 47(5):555–559. https://doi.org/10.1038/ng.3254.

Lohmueller KE (2014) The impact of population demography and selection on the genetic architecture of complex traits. PLOS Genet 10(5):e1004379. https://doi.org/10.1371/journal.pgen.1004379.

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* 30. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

Marchi N, Schlichta F, Excoffier L (2021) Demographic inference. Curr Biol 31(6):R276–R279. https://doi.org/10.1016/j.cub.2021.01.053.

Marjoram P, Wall JD (2006) Fast 'coalescent' simulation. BMC Genet 7(1):16. https://doi.org/10.1186/1471-2156-7-16.

McKinley TJ, Vernon I, Andrianakis I, McCreesh N, Oakley JE, Nsubuga R et al. (2018) Approximate Bayesian Computation and simulation-based inference for complex stochastic epidemic models. Stat Sci 33(1):4–18. https://doi.org/10.1214/17-STS618.

McVean GAT, Cardin NJ (2005) Approximating the coalescent with recombination. Philos Trans R Soc B Biol Sci 360(1459):1387–1393. https://doi.org/10.1098/rstb.2005.1673.

Mondal, M, Bertranpetit, J, Lao, O (2019) Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nat Commun* 10 (1), 1. https://doi.org/10.1038/s41467-018-08089-7.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci 76(10):5269–5273. https://doi.org/10.1073/pnas.76.10.5269.

Palamara PF, Lencz T, Darvasi A, Pe'er I (2012) Length distributions of identity by descent reveal fine-scale demographic history. Am J Hum Genet 91(5):809–822. https://doi.org/10.1016/j.ajhg.2012.08.030.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel Oet al. (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html.

Pouyet F, Aeschbacher S, Thiéry A, Excoffier L (2018) Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife 7: e36317.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16(12):1791–1798. https://doi.org/10.1093/oxfordjournals.molbev.a026091.

Pudlo P, Marin J-M, Estoup A, Cornuet J-M, Gautier M, Robert CP (2016) Reliable ABC model choice via random forests. Bioinformatics 32(6):859–866. https://doi.org/10.1093/bioinformatics/btv684.

Pujolar JM, Dalén L, Hansen MM, Madsen J (2017) Demographic inference from whole-genome and RAD sequencing data suggests alternating human impacts on goose populations since the last ice age. Mol Ecol 26(22):6270–6283. https://doi.org/10.1111/mec.14374.

Raynal L, Marin J-M, Pudlo P, Ribatet M, Robert CP, Estoup A (2019) ABC random forests for Bayesian parameter inference. Bioinformatics 35(10):1720–1728. https://doi.org/10.1093/bioinformatics/bty867.

Sanchez T, Cury J, Charpiat G, Jay F (2021) Deep learning for population size history inference: design, comparison and combination with approximate Bayesian computation. Mol Ecol Resour 21(8):2645–2660.

Scally, A, & Durbin, R (2012) Revising the human mutation rate: implications for understanding human evolution. *Nat Rev Genet* 13(10), 10. https://doi.org/10.1038/nrg3295.

Schiffels, S, & Durbin, R (2014) Inferring human population size and separation history from multiple genome sequences. *Nat Genet* 46(8), 8. https://doi.org/10.1038/ng.3015.

Schraiber JG, Akey JM (2015) Methods and models for unravelling human evolutionary history. Nat Rev Genet 16(12):727–740. https://doi.org/10.1038/nrg4005.

Schrider DR, Kern AD (2018) Supervised machine learning for population genetics: a new paradigm. Trends Genet 34(4):301–312. https://doi.org/10.1016/j.tig.2017.12.005.

Schrider DR, Shanku AG, Kern AD (2016) Effects of linked selective sweeps on demographic inference and model selection. Genetics 204(3):1207–1223.

Shapley LS (1953) A value for n-person games. In HW Kuhn, AW Tucker (Eds), Contributions to the theory of games (AM-28), Volume II (pp 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Sheehan S, Harris K, Song YS (2013) Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics 194(3):647–662. https://doi.org/10.1534/genetics.112.149096.

Sheehan S, Song YS (2016) Deep learning for population genetic inference. PLOS Comput Biol 12(3):e1004845. https://doi.org/10.1371/journal.pcbi.1004845.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123(3):585–595. https://doi.org/10.1093/genetics/123.3.585.

Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using Approximate Bayesian Computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173(3):1511–1520. https://doi.org/10.1534/genetics.106.055574.

Tavaré, S, Balding, DJ, Griffiths, RC, & Donnelly, P (1997). Inferring coalescence times from DNA sequence data. *Genetics* 145 (2) 505–518. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1207814/.

Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49(2):303–309. https://doi.org/10.1038/ng.3748.

Theunert C, Tang K, Lachmann M, Hu S, Stoneking M (2012) Inferring the history of population size change from genome-wide SNP data. Mol Biol Evol 29(12):3653–3667. https://doi.org/10.1093/molbev/mss175.

Thouzeau V, Affholder A, Mennecier P, Verdu P, Austerlitz F (2022) Inferring linguistic transmission between generations at the scale of individuals. J Lang Evol 7(2):200–212. https://doi.org/10.1093/jole/lzac009.

Thouzeau V, Mennecier P, Verdu P, Austerlitz F (2017) Genetic and linguistic histories in Central Asia inferred using approximate Bayesian computations. Proc R Soc B Biol Sci 284(1861):20170706. https://doi.org/10.1098/rspb.2017.0706.

Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2008) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J R Soc Interface 6(31):187–202. https://doi.org/10.1098/rsif.2008.0172.

Torada L, Lorenzon L, Beddis A, Isildak U, Pattini L, Mathieson S et al. (2019) ImaGene: a convolutional neural network to quantify natural selection from genomic data. BMC Bioinforma 20(Suppl 9):337. https://doi.org/10.1186/s12859-019-2927-x.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. Genetics 145(3):847–855. https://doi.org/10.1093/genetics/145.3.847.

Wood MF, Simon N (2018) ABC in ecological modelling. In Handbook of Approximate Bayesian Computation. Chapman and Hall/CRC.

Wright S (1950) Genetical structure of populations. Nature 166(4215):247–249. https://doi.org/10.1038/166247a0.

ACKNOWLEDGEMENTS

We thank Emilia Huerta-Sanchez and Camille Roux for insightful discussions. We thank the editor and three anonymous reviewers for their comments on the manuscript.

AUTHOR CONTRIBUTIONS

All authors conceived and designed the study. AQ implemented the approach. All authors analyzed the results. AQ wrote the manuscript draft. FA and FJ reviewed and edited the draft. All authors read and approved the final manuscript.

FUNDING

This work was supported by the Sorbonne Center for Artificial Intelligence (AQ) and by the Agence Nationale de la Recherche through grant ANR-20-CE45-0010-01 RoDAPoG (FJ). Open access funding provided by Museum National d´Histoire Naturelle.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41437-025-00773-x.

Correspondence and requests for materials should be addressed to Arnaud Quelin.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.