RESEARCH ARTICLE

Machine learning techniques for continuous genetic assignment of geographic origin of forest trees

Bernd Degen[1*], **Yulai Yanbaev**[2], **Niels A. Müller**[1]

1 Thünen Institute of Forest Genetics, Grosshansdorf, Germany, **2** Bashkir State Agrarian University, Ufa, Russia

* bernd.degen@thuenen.de

Abstract

Origin tracking is important to ensure use of the right seed source and trade with legally harvested timber. Additionally, it can help to reconstruct human-caused historical long-distance seed transfer and to spot mislabelling in forest field trials. So far, genetic assignment approaches were mostly discrete, assigning test samples to predefined groups. The main limitation of this approach is the justification of these discrete groups when genetic variation across the landscape is actually continuous. Here, we compare the accuracy of five continuous assignment methods. Specifically, we test a nearest neighbour method (NN), direct gaussian process regression (GPR-D) using the radial basis kernel function, grid based gaussian process regression (GPR-G) applying the Matérn kernel function, genomic prediction (GP) and deep learning (DL), using two genome-wide single nucleotide polymorphism (SNP) datasets of trees from across Europe. The first dataset comprises 30,000 SNPs from 865 European beech (*Fagus sylvatica*) trees, the second dataset consists of 381 SNPs from 1,883 pedunculate oak (*Quercus robur*) trees. The accuracy, as measured by the geographic distance between true and predicted locations, was highest for the GPR-G and DL methods with the beech dataset with a median distance of only 55 km and 76 km, respectively. For the oak data GPR-G and DL also performed best with median distances of 263 km and 278 km, respectively. The relative error (distance/max distance among tree pairs) was below 8% for 90% of all samples for the best method for both datasets. We detected 35 individuals and 10 groups as outliers in the beech data and 27 individuals and 18 groups in the oak data. These outliers may be caused by mislabelling or historical human-caused long distance seed transfer. We discuss the differences in performance of the approaches and highlight future applications and potential for further improvements.

OPEN ACCESS

Citation: Degen B, Yanbaev Y, Müller NA (2025) Machine learning techniques for continuous genetic assignment of geographic origin of forest trees. PLoS One 20(6): e0324994. https://doi.org/10.1371/journal.pone.0324994

Editor: Maher Maalouf, Khalifa University, UNITED ARAB EMIRATES

Received: December 6, 2024

Accepted: May 5, 2025

Published: June 6, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0324994

Data availability statement: The genotype data, including geographic coordinates of the oak samples, have been deposited in the Dryad Digital Repository and are accessible at the following DOI: https://doi.org/10.5061/dryad.8931zcrrd. All beech sequencing data are publicly available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under the accession number PRJNA1005581: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1005581.

Funding: The work on the oak dataset was supported via a grant from the Waldklimafonds WKF-22WC4111 01 (German Federal Ministry of Food and Agriculture & German Federal German Ministry for the Environment, Nature Conservation and Nuclear Safety), grant No 23RL-1F017 from the Higher Education and Science Committee of the Republic of Armenia in support of the contribution of Yulay Yanbaev. The research on the beech data was funded by a core grant from the Thünen Institute and another grant from the Waldklimafonds WKF-2219WK60A4.

Competing interests: NO authors have competing interests Enter: The authors have declared that no competing interests exist.

Introduction

Most tree species have a broad natural distribution range. In many cases, they are genetically adapted to the different local environmental conditions [1] although the environmental variables and phenotypic traits underlying this process still remain mostly elusive. Local adaptation has been experimentally demonstrated by common garden experiments [2,3]. It can make the selection of appropriate seed sources essential for the success of forest plantations [4] and thus calls for efficient methods to control the declared origin of seed material. For this purpose, genetic methods can play a vital role [5,6].

Several studies have shown historical long-distance transfer of seed material for afforestation. Jansen and Geburek [7] analysed the transnational use of Larch seeds for afforestation in different European countries. The intensive historical transfer of seed material of Norway spruce and Scots pine in Scandinavian countries was reviewed by Myking et al. [8]. Mal-adaptation of the planted trees can be the result of such long-distance seed transfer, especially when the environmental conditions of the seed origin and the later place of use differ strongly [9]. Thus, it would be useful to have a tool that can reconstruct the origin of planted trees that originate from a distant seed source [10]. Genetic control of origin would also be valuable for field trials, as it could help to identify mislabelled plants or errors in the execution of the planting scheme. Moreover, genetic assignment of origin has been applied for timber tracking as a tool to fight illegal logging [11–13]. Finally, ecological genetic studies could benefit from genetic assignment of geographic origins providing information about distances and directions of natural gene flow. Reconstructing the origin of successful pollen and seeds is feasible on smaller geographic scales using paternity analysis and genetic screening with highly variable gene markers e.g., microsatellites [14,15]. However, the precise origin can only be reconstructed if all potential parents in an area have been studied. This is creating an exponentially growing workload with increasing spatial scale. For geographically isolated populations paternity analysis can give minimum distance estimates [16], but still cannot determine the exact origin.

Spatial-genetic assignment methods can be subdivided into discrete and continuous approaches. The distinction aligns closely with the fundamental machine learning terms—classification and regression, respectively. The discrete genetic assignment approaches split the reference samples into groups and try a falsification of the origin of the test samples (named “classification” in machine learning). The assignment is done based on allele frequencies and the likelihood of the genotype of the test individual to occur in the different groups of reference samples [17]. Bayesian approaches have been shown to be more powerful for this assignment [17,18]. Also, a nearest neighbour approach has been proposed as this method is less sensitive to datasets with reference samples of different taxonomical status [19]. Other approaches use clustering methods such as STRUCTURE-like [10] or parameter-free methods like random forests [13]. In the best case, all but one group can be excluded and the geographic area of this group can thus be considered to be the true origin of the test sample.

The continuous spatial-genetic assignment methods used so far predict the expected location of an individual based on the modelling of allele frequencies in a two or three-dimensional space taken from reference samples (called "regression" in machine learning). This approach has been applied to trees [13] but also endangered animals such as elephants [20]. Continuous assignment of origin has also been done with the programs SPASIBA [21] and SPA [22]. SPASIBA stands for Spatial Bayesian Interference and is a further development of the SCAT program [23]. It uses the allele counts of training samples to model the two-dimensional distribution of the allele frequencies. Other than the SCAT software, SPASIBA does not require Monte Carlo simulations but applies Nested Laplace Approximation for the optimisation of the underlying functions. The spatial ancestry analysis (SPA) also models allele frequencies as continuous functions in geographic space but using logistic functions optimized by Newton's based methods.

The advances in next generation DNA sequencing provide much larger genetic datasets that are now also available for studies on geographic origin. In addition, machine learning approaches are progressively used to analyse these data. In forestry, both developments have been merged for genomic predictions (GP) of phenotypic traits important for tree breeding [24,25]. Also, Gaussian process regression (GPR) is a modern machine learning approach that has recently been applied to predict the origin of timber using stable isotopes and trace elements [26], and to reconstruct ancient migration routes of humans based on whole genome data [27]. Maldonado et al. [28] used different machine learning models to assign the genetic sub-population in a Eucalyptus progeny-provenance trial based on foliar spectral information as features and SNP-data to define the sub-populations (output). Recently, deep learning (DL) has also been applied to assign the origin of horse breeds using data of a large SNP array [29].

In the present paper we test and compare the power of a nearest neighbour approach (NN), genomic prediction (GP), gaussian process regression (GPR) and deep learning (DL) using SNP data of georeferenced tree samples from across Europe to predict the geographic origin of trees (continuous spatial-genetic assignment). Further, we identify outliers with large differences between true and predicted geographic origin, which may be due to long-distance seed transfer or mislabelling within trial sites.

Materials and methods

Genetic data

Beech. We used a subsample of 30.000 polymorphic SNPs from the whole-genome data of 865 *Fagus sylvatica* trees [30]. The SNPs were randomly sampled over all chromosomes. The trees represent 6 to 10 individuals from 99 different beech provenances distributed across the natural distribution range (Fig 1).

Oaks. For the oaks, 381 polymorphic SNPs (359 nuclear, 17 chloroplast and 5 mitochondrial SNPs) were taken from 1,883 trees from 188 locations with 6 to 20 individuals per location [32] sampled over a large part of the natural range (Fig 2). The chloroplast and mitochondrial SNPs were maternally inherited in oaks (gene flow only via seeds). In contrast the nuclear SNPs have a bi-parental inheritance and are more broadly dispersed via pollen and seed. The nuclear SNPs were developed by high coverage ddRAD-sequencing [33] and the plastid makers were developed by low coverage whole-genome sequencing. The location of these markers in the chloroplast and mitochondria was confirmed by mapping against the reference plastid sequences [34]. We only took samples that were located inside the natural range [31]. All of the trees had an admixture coefficient for *Quercus robur* of more than 0.8. The admixture was estimated with the program STRUCTURE based on the nuclear SNPs [35].

Statistical analysis

The statistical analyses were done with R [36] and custom scripts written in Python (version 3.12).

Data structure and pre-processing

The data matrix X represents the genetic information used for the geographic assignment. It consists of rows corresponding to individuals and columns corresponding to single nucleotide polymorphisms (SNPs). Each element in the matrix

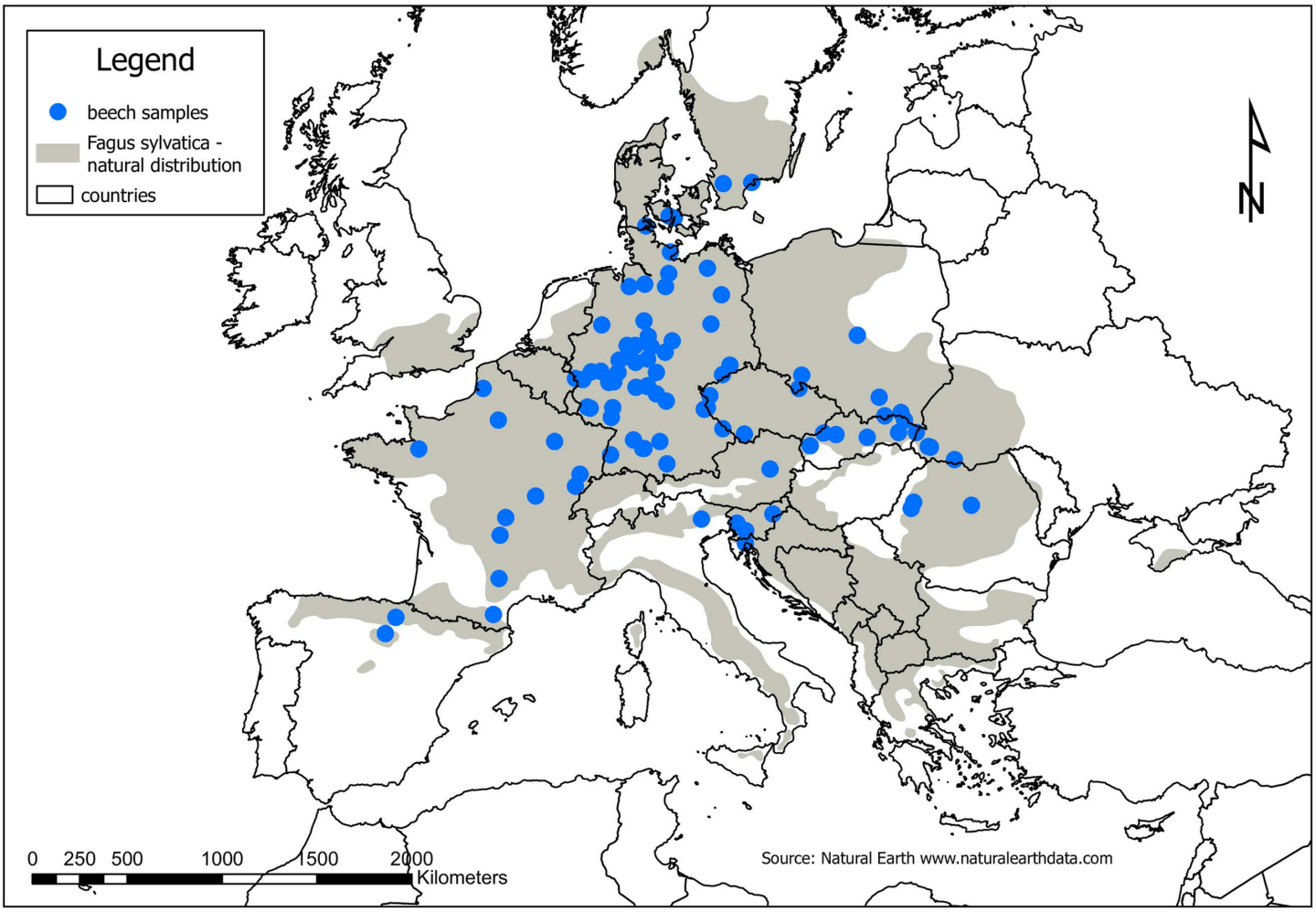

Fig 1. Distribution of sampled trees and natural range of *Fagus sylvatica* [31], shapefile of country borders from www.naturalearthdata.com.

https://doi.org/10.1371/journal.pone.0324994.g001

encodes the genotype of an individual at a given SNP locus. Rows (*n*): Each row represents a single tree individual ($i = 1, 2,\ldots,n$), where *n* n is the total number of sampled individuals. Columns (*p*): Each column corresponds to a specific SNP marker ($j = 1,2,\ldots,p$), where *p* is the total number of SNPs included in the analysis.

The selected individuals and the selected SNPs had less than 5% missing data. The genotypes at the SNPs were transformed to 1 (homozygote reference allele), 0.5 (heterozygote) and 0 (homozygote for alternative allele) for the nearest neighbour approach (NN) and for the gaussian process regression grid approach (GPR-G). They were transformed to 1 (homozygote reference allele), 0 (heterozygote) and −1 (homozygote for alternative allele) for the other three machine learning algorithms. For the nearest neighbour approach (NN) and the genomic prediction (GP), missing values were coded as "NA" and for the two types of gaussian process regression (GPR-D, GPR-G) and deep learning (DL) missing values were imputed using average allele values of all other individuals for that particular SNP. The imputation method should only have a very limited influence on the results because of the low amount of missing data. For the GPR-D, GPR-G and DL only SNPs were used that had a statistically significant correlation either with longitude or latitude. After applying a p-value threshold with Bonferroni correction, 4,335 out of the 30,000 SNPs were selected for GPR and DL analysis of the beech dataset, and 223 (203 nuclear SNPs, 16 chloroplast and

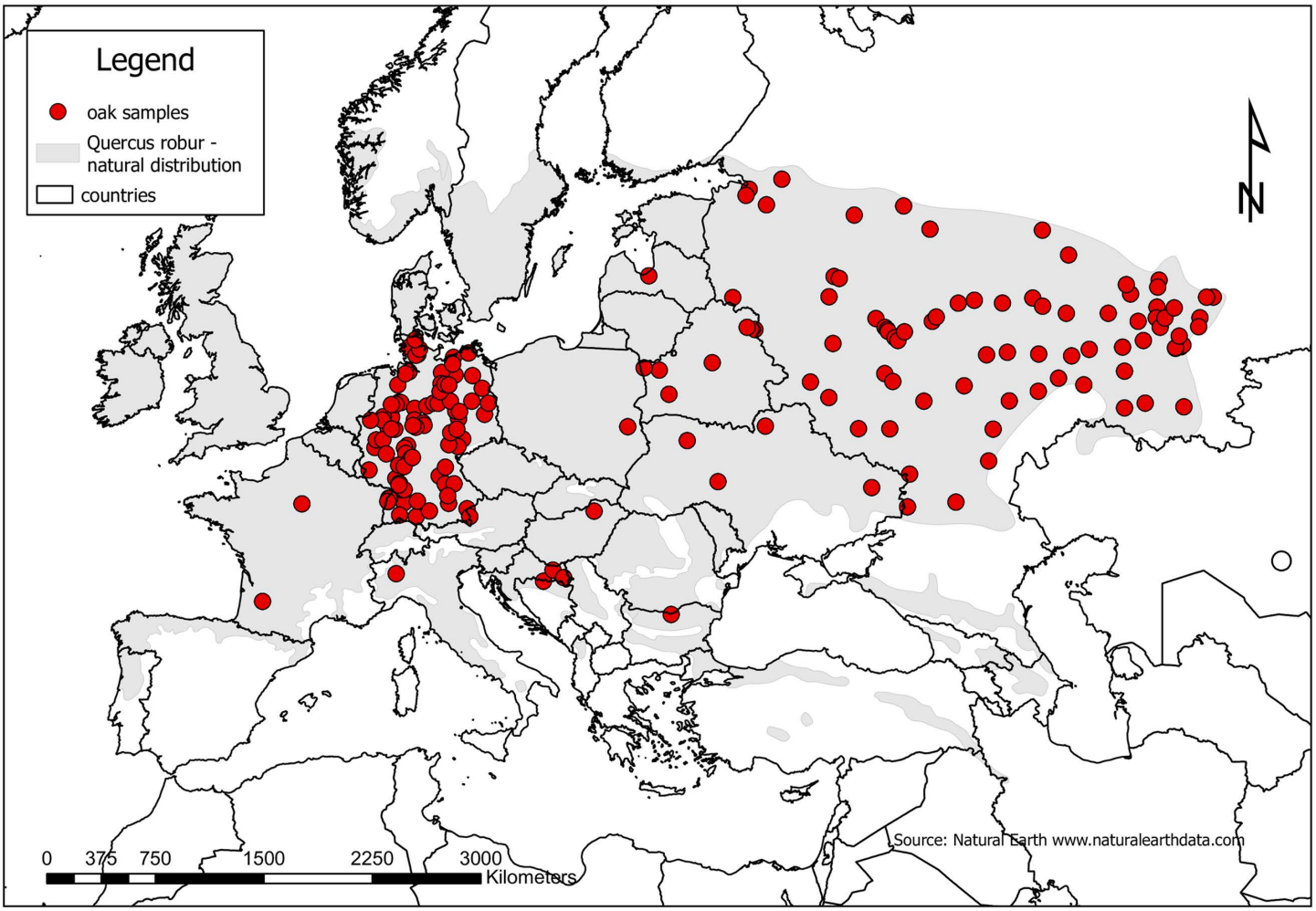

Fig 2. Distribution of sampled pedunculate oak trees and natural distribution range of *Quercus robur* [31], shapefile of country borders from www.naturalearthdata.com.

https://doi.org/10.1371/journal.pone.0324994.g002

4 mitochondrial SNPs) out of 381 SNPs for the oak dataset. The plastid SNPs of the oaks were treated in the data processing as homozygotes.

Nearest neighbour approach (NN)

Following Degen et al. [19] we computed the genetic distance (mean absolute difference of allele frequencies) between all pairs of trees and used this to generate a continuous assignment of the mean latitude and longitude of the *k* genetically most similar individuals as the predictor. We tested different k values (5, 10, 15, 20) but got the highest accuracy in the cross-validation for k = 5 (see below), which was then applied for both datasets.

Genomic prediction (GP)

We utilized the genomic Best Linear Unbiased Prediction (gBLUP) algorithm to compute the predicted latitude and longitude for each tree. For that we applied the "kin.blup" function of the "rrBLUP" R package [37]. An essential part of gBLUP is the construction of a kinship matrix. This matrix quantifies the genetic similarity between pairs of individuals based on

their allele profiles at the SNPs and thus the kinship matrix provided estimates of co-variation between individuals for the predicted target values (latitude, longitude).

Gaussian Process Regression- direct approach (GPR-D)

We implemented GPR using the GPyTorch library in Python. The GPR model was formulated with a constant mean function to represent the mean of the Gaussian process and the Kernel Function. A Radial Basis Function (RBF) kernel was chosen for its ability to model smooth functions. The kernel had the form: $k\ (x_i,\ x_j) = \sigma^2 \exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right)$, with $k(x_i, x_j)$ representing the covariance function or kernel function. It measured the similarity between two input points x_i and x_j, in our case the allele composition of individuals across loci (genotypes). The kernel function determines the shape of the functions that the Gaussian process can model. In that σ^2 is the variance parameter. It scales the overall magnitude of the kernel. The exponential function $\exp\left(-\frac{\|x_i - x_j\|^2}{2l^2}\right)$ ensures that the kernel function produces values between 0 and σ^2. The term inside the exponential function controls how rapidly the covariance decreases with distance between points. The term *l* is the lengthscale parameter. Larger values of *l* imply a smoother function, where points farther apart in the input space are still considered simi*l*ar.

An optimization of hyperparameters was done using a random search approach. A total of 10,000 random configurations of hyperparameters were evaluated. Each configuration included:

- lengthscale_lat: Lengthscale for the latitude predictions.
- noise_lat: Noise level for the latitude predictions.
- lengthscale_lon: Lengthscale for the longitude predictions.
- noise_lon: Noise level for the longitude predictions.

The random values were sampled uniformly within the following ranges:

- Lengthscales: [0.001, 400]
- Noise levels: [1e-4, 1]

For each hyperparameter-configuration, a cross-validation was performed. For this the data was randomly split, with 90% used for training and 10% for testing in each iteration. This was repeated five times. The mean haversine distance between the predicted and actual locations was used as the evaluation metric. The GPR was used in this direct approach to train two models using the entire allele profiles of the training individuals: one to predict latitude and one to predict longitude.

Gaussian Process Regression- grid approach (GPR-G)

In the grid-based Gaussian Process Regression (GPR-G), we again used the GPyTorch library in Python but with the Matérn kernel function and a parameter $\nu = 2.5$ that controlled the smoothness of the function. The study area was divided into a fixed grid with each cells measuring 20 km x 20 km for the beech data and 50 km by 50 km for the oak data. The grid spanned a latitude and longitude range from the minimum to the maximum value observed among our individuals. The allele frequencies for each SNP were pre-aggregated for all training individuals located within each grid cell. A different GPR model was trained to predict allele frequencies for each SNP across the grid cells. This means for beech 4,335 and for the oaks 223 different GPR models were created. Hyperparameters were not explicitly tuned. Instead, the model parameters were optimized over 10 epochs by minimizing the marginal log likelihood of the training data with the Adam optimizer. In each grid cell we used the predicted allele frequencies at all SNPs to compute the likelihood to obtain the

genotype of the test individual. The underlying assumption was that genotypes occur as expected in a Hardy-Weinberg Equilibrium (HWE). The grid cell with the highest probability was selected as predicted location (maximum likelihood approach).

Deep learning (DL)

We employed a feedforward deep neural network implemented in the R package "H2O" [38] to predict latitude and longitude from genetic data. The H2O cluster offers optimized parallel processing, enhancing computational speed. The model architecture consisted of five hidden layers with the following neuron configuration: 100, 100, 100, 50, and 50 neurons, respectively. We used the "RectifierWithDropout" activation function with dropout ratios of 0.1 for the input layer and varied dropout ratios of 0.3, 0.3, 0.3, 0.2, and 0.2 for the hidden layers [39]. L1 and L2 regularization (1e-6) were applied to mitigate overfitting. Dropout is a technique where a proportion of neurons are randomly set to zero during training, which helps in preventing overfitting and improving generalization by reducing dependency on specific neurons. The model was trained for up to 500 epochs with early stopping based on Root Mean Square Error (RMSE) and a tolerance of 0.001. RMSE measures the average magnitude of errors between predicted and actual values. The training process halts if RMSE does not improve by at least 0.001 for ten consecutive rounds. To ensure robust predictions, an ensemble of five models was used. The outputs of these models were aggregated to compute the mean and variance of the predictions (ensemble learning).

Cross validation – leave-one-out approach

We assessed the accuracy of the NN, GP, GPR-D, GPR-G and DL analyses using a leave-one-out cross-validation approach. This was done by a loop over all individuals treating each individual in the dataset as the test data while using the remaining individuals as training data. For this, the GP analysis set the true values for latitude and longitude of the individual under test temporarily to "NA" (to simulate unseen data). In the NN, GPR-D, GRG-G and DL analysis, all individuals except the one being tested (test data) were used as training data. The predicted latitude and longitude values obtained from these analyses were then compared with the true values, i.e., the passport data. We calculated the spatial distance between the predicted and actual locations with the Haversine formula [40]. The formula computes the shortest distance over the earth's surface between two geographical points. As another measure of accuracy, we computed the Pearson's correlation coefficient between given and predicted latitude as well as longitude values.

Identification of outliers

In each of the five methods we computed the standard deviation or variance for the predicted location. The standard deviation of the model predictions is directly computed by the Gaussian Process Regression (GPR) models. The function for the genomic prediction (GP) also provides an error estimation called "prediction error variance" (PEV). The "kin.blup" function used the mixed model equations and the genomic relationship matrix to compute the PEV for each prediction. Same as for GPR these variance values were used to computed the upper limit of the 95% distance interval. For the nearest neighbour approach (NN) we computed mean and standard deviation of the pairwise distances between the five identified nearest neighbours, and for the deep learning (DP) we computed the pairwise distances between the predicted locations of five repetitions. Based on these standard deviations we calculated the 95% upper limit for the distance radius. We searched for outliers among the individuals but also for outliers of entire groups. For the group approach the means of all predicted latitude and longitude values as well as the means of the 95% upper limits were computed. The distribution of distances was used for an interquartile range (IQR) methodology to identify outliers in the predicted locations. This approach computes as range the difference between the 25th and 75th percentile of the quotient distribution (IQR). Our application used the 75th percentile + 1.5 times the IQR (1.5 x IQR) as threshold for the group outliers and the 75th percentile + three times the IQR (3 x IQR) as threshold for the individual outliers. The IQR approach does not assume normal

distributed values and is robust to extreme values. In both approaches only those outliers were kept for which the distance between given and predicted location was larger than the 95% upper limit for the distance radius.

Results

Pairwise geographic distances

The pairwise geographic distances between the beech samples (Fig 3a) had a mean of 621 km (standard deviation: 395 km). The oak samples covered a much larger area with a mean pairwise geographic distance between trees of 1452 km and a standard deviation of 1030 km (Fig 3b). These distributions are important to assess the accuracy of the used assignment methods.

Accuracy of predicted locations

For the beech dataset, the grid based Gaussian process regression (GPR-G) and deep learning (DL) gave the highest accuracies with median distances of 55 km and 76 km, followed by the nearest neighbour (NN) method, direct Gaussian process regression (GPR-D) and genomic prediction (GP) (Table 1). Notably, the standard deviations of NN and GPR-G were larger than those of the other three methods. Based on the Pearson's correlation between given and predicted latitude and longitude values, DL and GPR-G performed best with mean values of 0.97 and 0.95 (Table 2). The ranking of methods is also visible by the curves of the relative assignment errors, calculated as the distance between true and predicted location divided by the maximum distance between two data points (Fig 4). Here, the good performance of DL and

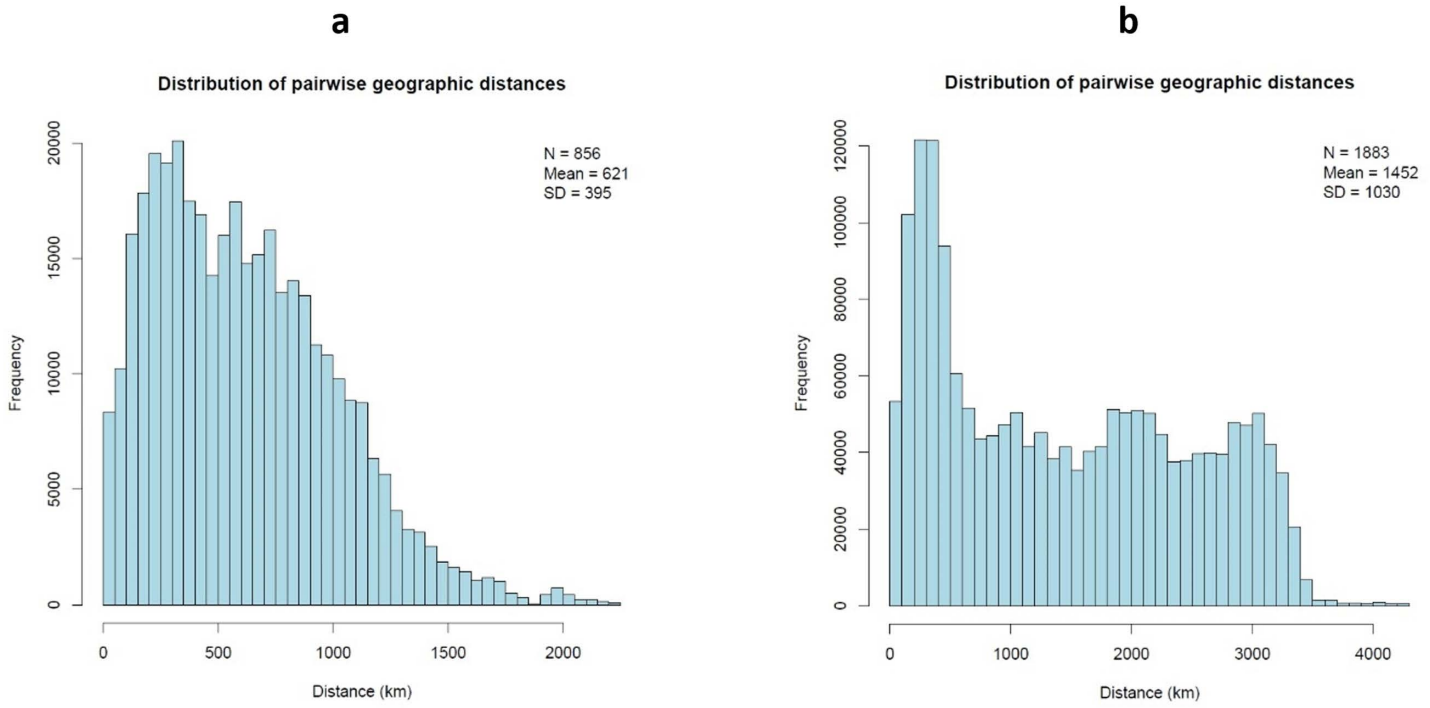

Fig 3. Histogram of pairwise distances between samples. (a, b) The y-axis shows the frequency and the x-axis the geographic distance in km of all pairwise comparisons between samples in beech (a) and oak (b). The number of samples (N), their mean pairwise distance and the standard deviation (SD) of that distance are given in the top right corner.

https://doi.org/10.1371/journal.pone.0324994.g003

Table 1. Results of the leave-one-out cross validation for the two datasets "Fagus" and "Quercus" on the accuracy of the predicted locations measured as distance between true and predicted locations.

Dataset	Method	Mean distance (km)	Median distance (km)	SD Distance (km)	N outliers individuals	N outliers groups
Fagus	DL	107	76	100	12	4
Fagus	NN	135	97	130	4	0
Fagus	GPR-D	154	141	97	5	6
Fagus	GPR-G	98	55	122	29	5
Fagus	GP	147	127	96	2	0
Quercus	DL	312	278	210	7	4
Quercus	NN	365	281	300	8	3
Quercus	GPR-D	356	295	243	3	8
Quercus	GPR-G	363	263	343	10	10
Quercus	GP	366	312	238	6	10

https://doi.org/10.1371/journal.pone.0324994.t001

Table 2. Pearson's correlation coefficient between given and predicted latitude and longitude values for the five methods.

	Latitude	Longitude	Mean
beech			
DL	0.964	0.978	0.971
NN	0.915	0.943	0.929
GPR-D	0.891	0.958	0.924
GPR-G	0.937	0.963	0.950
GP	0.917	0.959	0.938
oaks			
DL	0.874	0.983	0.928
NN	0.670	0.945	0.808
GPR-D	0.614	0.960	0.787
GPR-G	0.615	0.945	0.780
GP	0.656	0.957	0.807

https://doi.org/10.1371/journal.pone.0324994.t002

GPR-G is also visible with the lowest relative error distribution. The correlations between the distances of most methods were only moderate with values between 0.35 and 0.61 (Table 3). Higher correlations were identified between GP and GPR-D (0.77) and between DL and GP (0.72).

Also, in the oak dataset with fewer SNPs, GPR-G and DL performed best, with a median distance of 263 km and 278 km (Table 1). However, the other methods showed similar performance with median distances of 281 km for NN, 295 km for GPR and 312 km for GP (Table 1). The standard deviation of distances was smaller for DL but relatively high for GPR-G compared to the other methods. In the oak dataset the mean Pearson's correlations between given and predicted locations were generally weaker compared to the beech dataset, especially the correlations for latitude (Table 2). Based on these criteria, DL with a mean correlation of nearly 0.93 performed better than the other four methods. There was a high Pearson's correlation coefficient of 0.92 between the distances of GPR-D and GP (Table 4). All other correlations were equal or smaller compared to the ones from the beech data. The relative errors (Fig 4b) showed the good performance of DL but an ambivalent performance for GPR-G with very low errors for about 50% of the cases but higher errors compared to DL for the other 50% of samples.

It should be noted that the best methods, that is DL and GPR-G, for the beech dataset both had sample errors below 7.9% for 90% of the samples and for the oak dataset this was DL with an error of less than 8.5% for 90% of the samples.

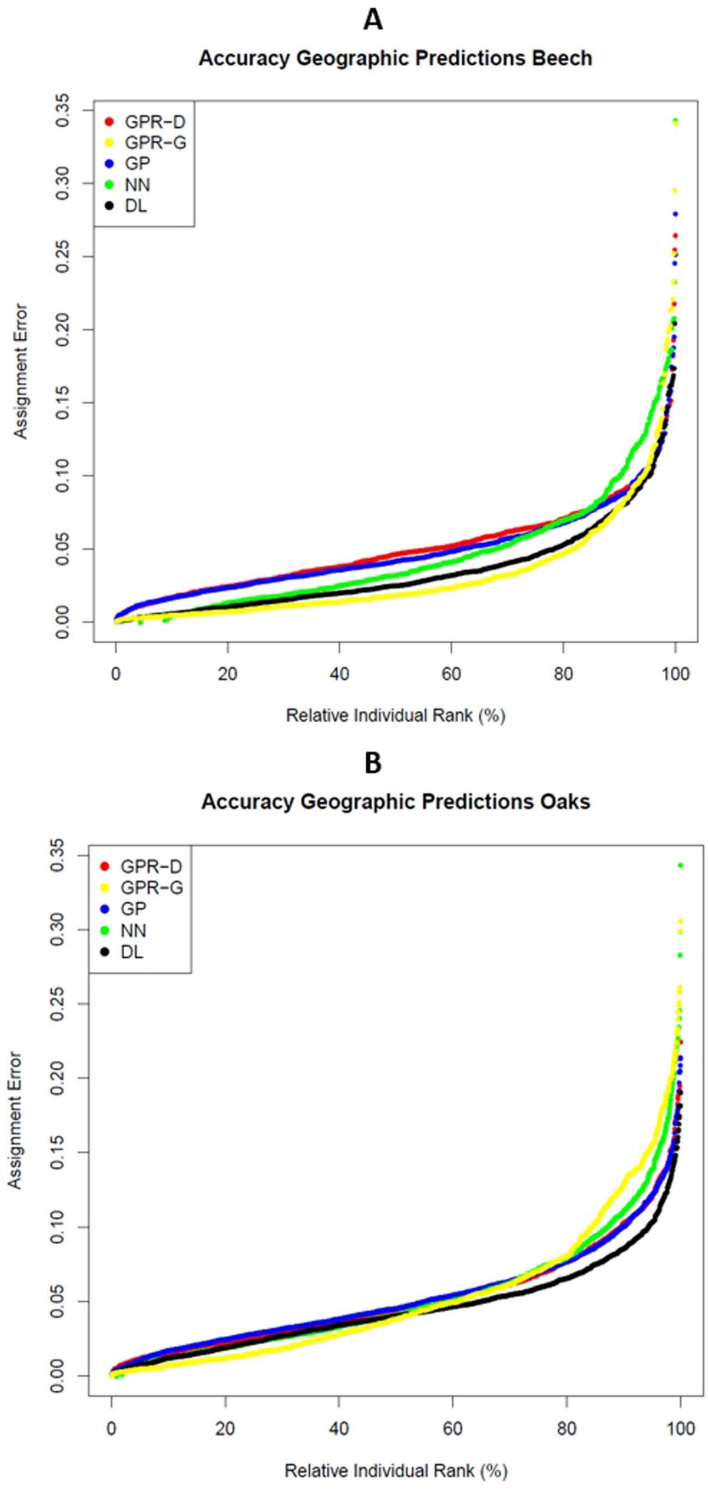

Fig 4. Assignment errors estimated using a dataset of beech (A) and oak (B). The errors are expressed as fraction of the distances between true and predicted location from the maximum distance among individuals. For each approach the individuals are ordered from left to right according to the error.

https://doi.org/10.1371/journal.pone.0324994.g004

Table 3. Pearson's correlation coefficient between distances between given and predicted locations for the five methods for the beech dataset.

	DL	NN	GPR-D	GPR-G
NN	0.350			
GPR-D	0.615	0.465		
GPR-G	0.425	0.540	0.504	
GP	0.721	0.533	0.776	0.562

https://doi.org/10.1371/journal.pone.0324994.t003

Table 4. Pearson's correlation coefficient between distances between given and predicted locations for the five methods for the oak dataset.

	DL	NN	GPR-D	GPR-G
NN	0.336			
GPR-D	0.467	0.546		
GPR-G	0.275	0.432	0.478	
GP	0.437	0.532	0.918	0.447

https://doi.org/10.1371/journal.pone.0324994.t004

Outliers with large distance between given and predicted geographic origin

For the beech dataset a total of ten group-outliers were identified out of 99 groups (between 0 and 6 per method, Table 1). The methods GP and NN did not identify outlier groups. The same provenances 32, 111 and 9 were identified by the three methods DL, GPR-G and GPR-D (Sup 1 Fig). In each case the provenance 32 with a given location in North-East Germany had group predictions much more in the South-West of Germany, the provenance 111 with given location in the West of Czechia got predicted positions much more in the East (Slovakia or Northern Hungary). The provenance 9 in North-West France had predicted locations more in the south-eastern part of France. The other seven group-outliers were obtained only with one method (Sup 1 Fig). A total of 35 individuals (2–29 per method) were identified as outliers (Sup 2 Fig). One individual, B9 of provenance 51 was identified as outlier by all five methods and likely represents a planting error. The given location in West Germany was in all cases projected a few hundred kilometres more to the East. The individual H65 of provenance 32 was identified by four methods as outlier, two individuals were identified by three methods (E33 from provenance 32 and F75 from provenance 101). Individual F75 had a given location in South Germany but was predicted to be located much more in the North of Germany. Six individuals were identified as outlier by two methods and the remaining 28 just by one method. In most cases the individual outliers belonged to groups that were also entirely identified as an outlier (Sup 1 Fig, Sup 2 Fig).

For the oak data the different methods identified between three and ten outlier-groups, summing up to a total of 18 different groups out of 188 groups assigned as outlier (Table 1, Sup 3 Fig). The location 213 in Latvia was identified by all five approaches as an outlier-group with a predicted location much more in the South-East in West Russia. Four methods spotted location 210 (Ukraine) to be from a region more North-East in West Russia. Three outlier-groups were identified by three methods (437 Russia, 5 France, 219 Croatia). Here, samples from the French location were predicted to have an origin in West Germany, the Croatian samples were supposed to be from a region much more in the North-East and the samples from location 437 in Russia are assigned to a region much more in the West of Russia. All other group-outliers were identified by only two methods (4) or a single method (9). On the individual level between three and ten outliers were identified per method, summing up to a total of 27 individuals (Table 1, Sup 4 Fig). Most of the individual outliers were from locations that have also identified as a group outlier (Latvia 213, Russia 437, Croatia 219, Ukraine 210, France 5) but also a few that were not spotted on the level of groups were identified. Examples for this are samples from location 233, 236 and 238 (Germany) that were predicted to be from a much more eastern origin.

Discussion

Performance of different continuous assignment methods

The two datasets for beech and oak differed largely in the spatial scale of the covered area. The oak samples are distributed across an area from France to Siberia which is more than two times larger than the region covered by the beech samples. Thus, it is useful to generate a relative measure for the error to compare the accuracy among methods, datasets and different studies. We followed the suggested approach of Guillot et al. [21] and computed the relative error as the distance between passported and predicted distances in relation to the maximum distance between two samples in the data set.

The grid-based gaussian process regression (GPR-G) and deep learning (DL) performed best for the beech data with a median difference between passport and predicted location of 55 and 76 km (error = 1.8 and 2.4%) and for the oak data with 263 and 278 km (error = 3.8 and 4%). But DL performed better than GPR-G if we consider also the correlation between given and predicted latitude and longitude values. The differences in mean correlation between the two methods were small for the beech data set (0.97 versus 0.95) but larger for the oak data set (0.93 versus 0.78). The nearest neighbour approach (NN) was on the third rank in both cases based on the median but had also a large proportion of individuals with large distances. The other two approaches genomic prediction (GP) and direct gaussian progress regression (GPR-D) performed slightly worse, although differences of relative errors between methods were minor especially for the oak data.

It may not be surprising that DL was the best method for both data sets. Deep learning is known to be highly suitable for large-scale studies where complex pattern of genetic markers need to be captured [41,42]. DL models can automatically learn from the patterns associated with different geographic origins. Not much user input in terms of parameter tuning is needed for the implementation of DL in the computing environment H2O [38]. The disadvantages of DL were the higher requirements for computing resources and the difficulty to interpret the “black-box” neural network [43]. We tested if we can increase the precision of the predicted origin by further enlarging the complexity of the neural network but there was no significant improvement neither by increasing the number of layers nor by using more neurons per layer.

The performance of the two variations of the gaussian process regression was quite different. The grid-based approach (GPR-G) with aggregation of allele frequencies of all individuals in the same grid cell, application of the Matérn kernel function, modelled allele frequencies and a maximum-likelihood approach for test genotype origin was clearly better than the direct prediction (GPR-D) of latitude and longitude based on the complete genotypes. The grid-wise approach is similar to the approaches implemented in the SCAT and SPACIBA programs discussed below. Our direct approach without the intermediate step of allele frequency modelling has not been used so far. The weaker performance of GPR-D can be explained by the less suitable radial kernel function compared to the Matérn kernel function. The Matérn kernel function is known to be good for modelling spatially autocorrelated data often found in ecology research [26].

The strongest correlation between methods was observed in both datasets for GP and GPR-D which can be explained by the fact that they are both picking up the same pattern and using a kernel functions for the predictions (although the kernel function it linear for GP).

So far, genomic predictions or genomic selection have been mostly used in breeding programs to estimate breeding values of individuals [44]. This was first done for animal breeding, followed by crop breeding and since a few years also tree breeding [24]. There have been many different algorithms proposed to compute breeding values or to predict phenotypes [45]. Among them genomic best linear unbiased prediction (gBLUP) is widely used. For the predictions with gBLUP the co-variation between kinship-similarities and similarities of phenotypes is used. In breeding programmes, the success of this approach is linked to the complex genetic architecture of the predicted traits that are highly polygenic with small effects of each causal allele that is well captured with a kinship-matrix. In our application “the complex genetic architecture” of the traits “longitude” and “latitude” is generated by different processes and the involved genetic variants causing spatial genetic structures at different spatial scales. Among them are demographic processes affecting the whole genome,

especially the recolonisation from the last glacial refugia [46] in combination with limited pollen and seed dispersal creating an isolation by distance pattern [47,48] but also other processes that only impact parts of the genome such as local adaptation [30,49,50] or spatially differential introgression [51,52].

The nearest neighbour approach (NN) is a non-parametric method used for classification and regression. Among the used methods the calculation of NN is the fastest one because it is just an averaging of latitude and longitude values of the k nearest neighbours. The k nearest neighbours have been identified based on genetic similarity to the tested individual. The method has been successfully used for discrete geographic assignment of different tropical tree species genotyped with different gene markers (SNP and microsatellites): Sapelli [19], trees of the genus *Dipteryx* [53], and *Hymenaea* [54]. In these cases, NN performed better compared to Bayesian methods because it could pick up complex pattern in the reference data caused by a mixture of genotypes from related taxa or a mixture of individuals from different postglacial recolonisation routes. A clear disadvantage is that NN cannot predict longitude and latitude values outside the range of the reference data and it is proposed to be sensitive for non-representative sampling of reference data.

An open question is to which extend the selection of the species has an impact on the accuracy of the machine learning based prediction of genetic origin. Generally, the accuracy is linked to the extent of genetic differentiation among populations and more explicitly to the spatial autocorrelation of the genetic data used as input features. Recently, Milesi et al. [55] calculated in a comparative study with large sets of SNPs of seven European tree species for *Fagus sylvatica* and *Quercus petraea* (a related species to *Quercus robur*) intermediate measures of genetic differentiation (F_{ST} = 0.05, 0.04) whereas other species such as *Pinus pinaster* and *Populus nigra* had much stronger values of genetic differentiation (0.13, 0.16) and the species *Pinus sylvestris* and *Betula pendula* only very small F_{ST}-values (0.01, 0.03).

Also, the sample design has an impact on the precision of the geographic predictions. We assume a systematic sampling scheme with sampled locations matching grid points distributed over the whole species range would be ideal. Each sampled location should be represented by three to ten individuals. The distances among neighboured grid points are limited by the available budget for sampling and genotyping.

Other studies on continuous assignment

One of the first continuous assignment of geographic locations based on genetic data has been done by Wasser et al. [23] with the SCAT program. They used genetic data of 16 microsatellites from 399 elephant samples collected all over the natural distribution range in Africa. In their spatial smoothing approach, they modelled the allele frequencies in a two-dimensional space. The distribution of allele frequencies followed independent Gaussian processes and were allowed to vary in a spatially correlated way. They used Monte Carlo simulations for the optimisation of the underlying functions. Their median distance between predicted and true location of the elephants was 499 km. That corresponds to a relative error of 7.9%. Guillot et al. [21] introduced the program SPASIBA, which stands for Spatial Bayesian Interference and is a further development of the SCAT program. It also uses the allele counts of training samples to model the two-dimensional distribution of the allele frequencies. Other than the SCAT software it does not require computing intense Monte Carlo simulations but applies nested Laplace approximation for the optimisation of the underlying functions. Two examples of SPASIBA predictions of individual locations were given by Guillot et al. [21]. First, they predicted the bird origin of the Florida scrub jay based on 41 SNPs. The median distance was 26.4 km corresponding to a relative error of 9.1%. Second, they predicted the origin of *Arabidopsis thaliana* in Europe using genetic data of 1007 individuals at 1000 SNP loci. The distance at the 75% percentile was 93 km corresponding to a relative error of less than 2%.

The SCAT program was also used by Puckett and Eggert [56] to predict the geographic origin of American black bear (*Ursus americanus*) based on 1000 SNPs and 15 microsatellite loci. The accuracy ranged between 192−902 km (error 3.4%−16.4%) depending on the gene maker set used. Further, the SPASIBA program was used by Finch et al. [13] to predict the geographic origin of Spanish Cedar (*Cedrela odorata*) in Latin-America. They used 140 SNPs for 386 individuals and came up with a median distance of 189 km between true and predicted location (error = 4%).

Another modelling approach named SPA for continuous assignment has been proposed by Yang et al. [22]. The spatial ancestry analysis (SPA) also models allele frequencies as continuous functions in geographic space but using logistic functions which were optimized by Newton's based methods. Unfortunately, both programs SPASIBA and SPA were not updated anymore and we could not apply them to our oak and beech data. Nevertheless, the supplementary material of Guillot et al. [21] contains results on comparative application of SPASIBA and SPA for a data set of *Arabidopsis thaliana* from Horton et al. [57]. These data are available and have genotypes at 1000 SNPs, sampled from a total of 215 k variants, for 1107 individual distributed over Eurasia. In a comparison we used this data with our five methods. We compared the accuracy of the predicted geographic locations of our five methods (GPR-D, GPR-G, GP, NN, DL) with results obtained by the programs for continuous assignment SPASIBA and SPA (supplementary 5). The four machine learning approaches (GP, GPR-D, GPR-G, DL) provided errors between the results of SPA and SPASIBA. The median error of NN and GPR-G was even better than that of SPASIBA.

Mortier et al. [26] used Gaussian process regression and the Matérn kernel function to predict the origin of timber in East-Europe based on data for stable isotopes and trace elements. In frame of a four-fold cross-validation they observed average distances between true and predicted location of 179 to 228 km depending of the tree species.

Speciality of our approaches

There is an important difference between four of our methods and the above-mentioned examples. For NN, GPR-D, GP and DL we used a direct *prior* probability approach to predict locations and the others applied *posterior* probability approach. The four methods focused on predicting geographic coordinates directly from genetic data without relying on intermediate steps such as allele frequency modelling or computing *posterior* probabilities across different grid cells. But our GPR-G method matches with the above cited other approaches used in SPASIBA, SCAT and SPA.

Another direct approach has been done by Battey et al. [58] using the deep learning network named "Locator" for geo-referenced whole genome data of *Plasmodium parasistes*, *Anopheles mosquitoes* and global humans to predict geographic origin. They tested architectures of different complexity for the predictions and finally used a network with ten layers and 256 nodes each for runs with GPUs. Our network was less complex (5 hidden layers, three with 100 and two with 50 neurons). Their error rates were below 1%. They trained their network with 90% of the samples running repetitions along windows in the genomes of 500kbp to 10Mbp. The variation between the windows served as estimator of uncertainty. We used all genetic data but repeated the network training five times to gain an estimator of uncertainty.

Sampling design in future studies

The effectiveness of continuous genetic assignment methods depends heavily on the structure of the reference dataset. For future studies with the objective to capture the spatial genetic structure, a critical choice has to be made between individual-based sampling and population-based sampling. In the individual based approach, individuals are sampled across the landscape in a spatially uniform manner to ensure a broad geographic coverage. This is particularly useful when genetic variation follows a continuous gradient (i.e., isolation by distance) and there are no discrete populations [48,59]. Individual-based sampling improves the precision of spatial assignment models, especially for machine learning approaches like Gaussian Process Regression (GPR-G) and Deep Learning (DL), which benefit from fine-scale spatial data [58,60]. In contrast, population-based sampling focuses on well-defined populations. Here, multiple individuals are collected per site but with fewer total locations. This approach is useful for species with strong population structure due to historical refugia, geographic barriers, or limited gene flow [61]. This sampling approach is particularly suited for discrete assignment methods, such as Bayesian clustering or STRUCTURE-like models. However, when genetic variation is not strongly clustered but varies continuously over space, population-based sampling can introduce bias, leading to an overestimation of differentiation between populations and an underestimation of within-population variability [59,62].

Practical implications of the identified outliers

The identified outliers in the study have practical implications for forest management and genetic research. First, these outliers highlight potential cases of long-distance seed transfer or mislabelling. These issues are critical to ensure the accuracy of provenance trials and the selection of future seed harvesting stands. For example, detecting outliers where the predicted and actual locations of tree origins differ substantially may indicate historical human interventions in seed distribution (translocations or unintentional admixture). This information is elementary for future seed sourcing strategies and may help to correct past errors in seed origin documentation. Special importance is given for tree species and regions with known human-mediated seed transfer such as Larch seeds used for afforestation in different European countries [7], seed material of Norway spruce and Scots pine in Scandinavian countries [8] or for introduced exotic tree species with a huge natural distribution range on the continent of origin such as red oak (*Quercus rubra*) or Douglas fir (*Pseudotsuga menziesii*) in Europe [10].

Genetic timber tracking as law enforcement tool to fight illegal logging [11] can also benefit from our new approaches of continuous prediction of tree origin. For timber harvested in natural forests or forests generated with local seed sources, differences between predicted and passported origin (outliers) can indicate misdeclarations and uncover fraud. In addition to former categorial classification approaches the continuous assignment provides useful information on true origin and thus potential locations of possible crime.

Moreover, the ability to identify outliers combined with information of growth and viability of the outliers can serve as a powerful tool for monitoring the adaptation and survival of tree species in changing climates. Trees that are significantly out of place might exhibit maladaptation, which can be detrimental to their survival. Therefore, identifying and understanding the reasons behind these outliers can lead to more effective strategies for selecting of future seed sources that are better suited to current and future environmental conditions.

Furthermore, our findings underscore the importance of using advanced continuous genetic assignment methods, such as deep learning, genomic prediction and Gaussian process regression, which have proven effective in identifying geographical origins and outliers. As the study shows, applying these methods can improve the accuracy of geographic origin predictions by avoiding definition of groups, making them invaluable for forest genetic management, conservation efforts, and law enforcement related to seed trade, timber tracking and illegal logging.

Supporting information

S1 Fig. Maps showing the group outliers identified in the beech dataset.
(DOCX)

S2 Fig. Maps showing the individual outliers identified in the beech dataset.
(DOCX)

S3 Fig. Maps showing the group outliers identified in the oak dataset.
(DOCX)

S4 Fig. Maps showing the individual outliers identified in the oak dataset.
(DOCX)

S5 "Supplementary 5.docx". Comparison of prediction accuracy with previously reported methods.
(DOCX)

S6 "Supplementary 6.zip". Python programs, r-scripts and input data for the five continuous geographic assignment methods.
(ZIP)

Acknowledgments

We thank two anonymous reviewers for their helpful comments on a former version of the manuscript and we are thankful to Malte Mader for his support in using the High-Performance Computing Cluster of the Thünen-Institute.

Author contributions

Conceptualization: Bernd Degen, Niels A. Müller.

Data curation: Bernd Degen, Niels A. Müller, Yulai Yanbaev.

Formal analysis: Bernd Degen.

Funding acquisition: Bernd Degen, Niels A. Müller, Yulai Yanbaev.

Investigation: Bernd Degen, Niels A. Müller, Yulai Yanbaev.

Methodology: Niels A. Müller, Yulai Yanbaev.

Software: Bernd Degen.

Writing – original draft: Bernd Degen, Niels A. Müller.

Writing – review & editing: Yulai Yanbaev.

References

1. Alberto FJ, Aitken SN, Alía R, González-Martínez SC, Hänninen H, Kremer A, et al. Potential for evolutionary responses to climate change - evidence from tree populations. Glob Chang Biol. 2013;19(6):1645–61. https://doi.org/10.1111/gcb.12181 PMID: 23505261
2. Park A, Rodgers JL. Provenance trials in the service of forestry assisted migration: A review of North American field trials and experiments. For Ecol Manage. 2023;537:16.
3. Leites L, Garzon MB. Forest tree species adaptation to climate across biomes: Building on the legacy of ecological genetics to anticipate responses to climate change. Glob Change Biol. 2023;20.
4. Girard Q, Ducousso A, de Gramont CB, Louvet JM, Reynet P, Musch B. Provenance variation and seed sourcing for sessile oak (*Quercus petraea* (Matt.) Liebl.) in France. Annals of Forest Science. 2022;79(1).
5. Kembryte R, Danusevicius D, Buchovska J, Baliuckas V, Kavaliauskas D, Fussi B. DNA-based tracking of historical introductions of forest trees: the case of European beech (*Fagus sylvatica* L.) in Lithuania. Eur J For Res. 2021;140(2):435–49.
6. Blanc-Jolivet C, Liesebach M. Tracing the origin and species identity of *Quercus robur* and Quercus petraea in Europe: a review. Silvae Genet. 2015;64(4):182–93.
7. Jansen S, Geburek T. Historic translocations of European larch (*Larix decidua* Mill.) genetic resources across Europe - A review from the 17th until the mid-20th century. For Ecol Manage. 2016;379:114–23.
8. Myking T, Rusanen M, Steffenrem A, Kjær ED, Jansson G. Historic transfer of forest reproductive material in the Nordic region: drivers, scale and implications. Forestry. 2016;89(4):325–37.
9. Ying CC, Yanchuk AD. The development of British Columbia's tree seed transfer guidelines: Purpose, concept, methodology, and implementation. For Ecol Manage. 2006;227(1–2):1–13.
10. Merceron NR, Leroy T, Chancerel E, Romero-Severson J, Borkowski DS, Ducousso A, et al. Back to America: tracking the origin of European introduced populations of *Quercus rubra* L. Genome. 2017;60(9):778–90. https://doi.org/10.1139/gen-2016-0187 PMID: 28750176
11. Dormontt EE, Boner M, Braun B, Breulmann G, Degen B, Espinoza E, et al. Forensic timber identification: It's time to integrate disciplines to combat illegal logging. Biol Conserv. 2015;191:790–8.
12. Degen B, Ward SE, Lemes MR, Navarro C, Cavers S, Sebbenn AM. Verifying the geographic origin of mahogany (*Swietenia macrophylla* King) with DNA-fingerprints. Forensic Sci Int Genet. 2013;7(1):55–62. https://doi.org/10.1016/j.fsigen.2012.06.003 PMID: 22770645
13. Finch KN, Cronn RC, Richter MCA, Blanc-Jolivet C, Guerrero MCC, Beltrán LD, et al. Predicting the geographic origin of Spanish Cedar (*Cedrela odorata* L.) based on DNA variation. Conserv Genet. 2020;21(4):625–39.
14. Chybicki IJ, Oleksa A. Seed and pollen gene dispersal in *Taxus baccata*, a dioecious conifer in the face of strong population fragmentation. Ann Bot. 2018;122(3):409–21. https://doi.org/10.1093/aob/mcy081 PMID: 29873697
15. Gerber S, Chadœuf J, Gugerli F, Lascoux M, Buiteveld J, Cottrell J, et al. High rates of gene flow by pollen and seed in oak populations across Europe. PLoS One. 2014;9(1):e85130. https://doi.org/10.1371/journal.pone.0085130 PMID: 24454802
16. Buschbom J, Yanbaev Y, Degen B. Efficient long-distance gene flow into an isolated relict oak stand. J Hered. 2011;102(4):464–72. https://doi.org/10.1093/jhered/esr023 PMID: 21525180

17. Ogden R, Linacre A. Wildlife forensic science: A review of genetic geographic origin assignment. Forensic Sci Int Genet. 2015;18:152–9. https://doi.org/10.1016/j.fsigen.2015.02.008 PMID: 25795277
18. Rannala B, Mountain JL. Detecting immigration by using multilocus genotypes. Proc Natl Acad Sci U S A. 1997;94(17):9197–201. https://doi.org/10.1073/pnas.94.17.9197 PMID: 9256459
19. Degen B, Blanc-Jolivet C, Stierand K, Gillet E. A nearest neighbour approach by genetic distance to the assignment of individual trees to geographic origin. Forensic Sci Int Genet. 2017;27:132–41. https://doi.org/10.1016/j.fsigen.2016.12.011 PMID: 28073087
20. Wasser SK, Brown L, Mailand C, Mondol S, Clark W, Laurie C, et al. Conservation. Genetic assignment of large seizures of elephant ivory reveals Africa's major poaching hotspots. Science. 2015;349(6243):84–7. https://doi.org/10.1126/science.aaa2457 PMID: 26089357
21. Guillot G, Jónsson H, Hinge A, Manchih N, Orlando L. Accurate continuous geographic assignment from low- to high-density SNP data. Bioinformatics. 2016;32(7):1106–8. https://doi.org/10.1093/bioinformatics/btv703 PMID: 26615214
22. Yang W-Y, Novembre J, Eskin E, Halperin E. A model-based approach for analysis of spatial structure in genetic data. Nat Genet. 2012;44(6):725–31. https://doi.org/10.1038/ng.2285 PMID: 22610118
23. Wasser SK, Shedlock AM, Comstock K, Ostrander EA, Mutayoba B, Stephens M. Assigning African elephant DNA to geographic region of origin: applications to the ivory trade. Proc Natl Acad Sci U S A. 2004;101(41):14847–52. https://doi.org/10.1073/pnas.0403170101 PMID: 15459317
24. Grattapaglia D. Twelve years into genomic selection in forest trees: climbing the slope of enlightenment of marker assisted tree breeding. Forests. 2022;13(10):25.
25. Degen B, Mueller NA. A simulation study comparing advanced marker-assisted selection with genomic selection in tree breeding programs. G3-Genes Genomes Genet. 2023;13(10):9.
26. Mortier T, Truszkowski J, Norman M, Boner M, Buliga B, Chater C, et al. A framework for tracing timber following the Ukraine invasion. Nature Plants. 2024:1–12.
27. Schmid C, Schiffels S. Estimating human mobility in Holocene Western Eurasia with large-scale ancient genomic data. Proc Natl Acad Sci U S A. 2023;120(9):e2218375120. https://doi.org/10.1073/pnas.2218375120 PMID: 36821583
28. Maldonado C, Mora-Poblete F, Echeverria C, Baettig R, Torres-Díaz C, Contreras-Soto RI. A neural network-based spectral approach for the assignment of individual trees to genetically differentiated subpopulations. Remote Sens. 2022;14(12):15.
29. Manzoori S, Farahani AHK, Moradi MH, Kazemi-Bonchenari M. Detecting SNP markers discriminating horse breeds by deep learning. Sci Rep. 2023;13(1).
30. Lazic D, Geßner C, Liepe KJ, Lesur-Kupin I, Mader M, Blanc-Jolivet C, et al. Genomic variation of European beech reveals signals of local adaptation despite high levels of phenotypic plasticity. Nat Commun. 2024;15(1):8553. https://doi.org/10.1038/s41467-024-52933-y PMID: 39362898
31. Caudullo G, Welk E, San-Miguel-Ayanz J. Chorological maps for the main European woody species. Data Brief. 2017;12:662–6. https://doi.org/10.1016/j.dib.2017.05.007 PMID: 28560272
32. Degen B, Yanbaev Y, Mader M, Ianbaev R, Bakhtina S, Schroeder H. Impact of gene flow and introgression on the range wide genetic structure of *Quercus robur* (L.) in Europe. Forests. 2021;12(10):17.
33. Degen B, Blanc-Jolivet C, Bakhtina S, Ianbaev R, Yanbaev Y, Mader M. Applying targeted genotyping by sequencing with a new set of nuclear and plastid SNP and indel loci for *Quercus robur* and *Quercus petraea*. Conserv Genet Resour. 2021;13(3):345–7.
34. Schroeder H, Cronn R, Yanbaev Y, Jennings T, Mader M, Degen B, et al. Development of Molecular Markers for Determining Continental Origin of Wood from White Oaks (*Quercus* L. sect. *Quercus*). PLoS One. 2016;11(6):e0158221. https://doi.org/10.1371/journal.pone.0158221 PMID: 27352242
35. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59. https://doi.org/10.1093/genetics/155.2.945 PMID: 10835412
36. R Core Team R. R: A language and environment for statistical computing. 2024.
37. Endelman JB. Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome. 2011;4(3):250–5.
38. LeDell E, Poirier S, editors. H2o automl: Scalable automatic machine learning. Proceedings of the AutoML Workshop at ICML. ICML San Diego, CA, USA; 2020.
39. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research. 2014;15(1):1929–58.
40. Vincenty T. Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey Review. 1975;23(176):88–93.
41. Alharbi WS, Rashid M. A review of deep learning applications in human genomics using next-generation sequencing data. Hum Genomics. 2022;16(1):26. https://doi.org/10.1186/s40246-022-00396-x PMID: 35879805
42. Montesinos-López OA, Montesinos-López A, Pérez-Rodríguez P, Barrón-López JA, Martini JWR, Fajardo-Flores SB, et al. A review of deep learning applications for genomic selection. BMC Genomics. 2021;22(1):19. https://doi.org/10.1186/s12864-020-07319-x PMID: 33407114
43. Lourenço VM, Ogutu JO, Rodrigues RAP, Posekany A, Piepho H-P. Genomic prediction using machine learning: a comparison of the performance of regularized regression, ensemble, instance-based and deep learning methods on synthetic and empirical data. BMC Genomics. 2024;25(1):152. https://doi.org/10.1186/s12864-023-09933-x PMID: 38326768
44. Meuwissen T, Hayes B, Goddard M. Genomic selection: A paradigm shift in animal breeding. Anim Front. 2016;6(1):6–14.

45. Wang J, Zhang Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. Genomics Proteomics Bioinformatics. 2021;19(4):629–40. https://doi.org/10.1016/j.gpb.2021.08.005 PMID: 34492338
46. Petit RJ, Brewer S, Bordács S, Burg K, Cheddadi R, Coart E, et al. Identification of refugia and post-glacial colonisation routes of European white oaks based on chloroplast DNA and fossil pollen evidence. For Ecol Manage. 2002;156(1–3):49–74.
47. de Lafontaine G, Ducousso A, Lefèvre S, Magnanou E, Petit RJ. Stronger spatial genetic structure in recolonized areas than in refugia in the European beech. Mol Ecol. 2013;22(17):4397–412. https://doi.org/10.1111/mec.12403 PMID: 23980761
48. Sexton JP, Hangartner SB, Hoffmann AA. Genetic isolation by environment or distance: which pattern of gene flow is most common?. Evolution. 2014;68(1):1–15. https://doi.org/10.1111/evo.12258 PMID: 24111567
49. Modica A, Lalagüe H, Muratorio S, Scotti I. Rolling down that mountain: microgeographical adaptive divergence during a fast population expansion along a steep environmental gradient in European beech. Heredity. 2024.
50. Zhou BF, Shi Y, Chen XY, Yuan S, Liang YY, Wang BS. Linked selection, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence in *Quercus dentata*. Journal of Systematics and Evolution. 2022;60(6):1344–57.
51. Leroy T, Louvet J-M, Lalanne C, Le Provost G, Labadie K, Aury J-M, et al. Adaptive introgression as a driver of local adaptation to climate in European white oaks. New Phytol. 2020;226(4):1171–82. https://doi.org/10.1111/nph.16095 PMID: 31394003
52. Degen B, Blanc-Jolivet C, Mader M, Yanbaeva V, Yanbaev Y, Dodd R. Introgression as an important driver of geographic genetic differentiation within European white oaks. Forests. 2023;14(12).
53. Honorio Coronado EN, Blanc-Jolivet C, Mader M, García-Dávila CR, Aldana Gomero D, Del Castillo Torres D, et al. SNP Markers as a Successful Molecular Tool for Assessing Species Identity and Geographic Origin of Trees in the Economically Important South American Legume Genus *Dipteryx*. J Hered. 2020;111(4):346–56. https://doi.org/10.1093/jhered/esaa011 PMID: 32402074
54. Chaves CL, Degen B, Pakull B, Mader M, Honorio E, Ruas P, et al. Assessing the Ability of Chloroplast and Nuclear DNA Gene Markers to Verify the Geographic Origin of Jatoba (*Hymenaea courbaril* L.) Timber. J Hered. 2018;109(5):543–52. https://doi.org/10.1093/jhered/esy017 PMID: 29668954
55. Milesi P, Kastally C, Dauphin B, Cervantes S, Bagnoli F, Budde KB, et al. Resilience of genetic diversity in forest trees over the Quaternary. Nat Commun. 2024;15(1):8538. https://doi.org/10.1038/s41467-024-52612-y PMID: 39402024
56. Puckett EE, Eggert LS. Comparison of SNP and microsatellite genotyping panels for spatial assignment of individuals to natal range: A case study using the American black bear (*Ursus americanus*). Biol Conserv. 2016;193:86–93.
57. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet. 2012;44(2):212–6. https://doi.org/10.1038/ng.1042 PMID: 22231484
58. Battey CJ, Ralph PL, Kern AD. Predicting geographic location from genetic variation with deep neural networks. Elife. 2020;9:e54507. https://doi.org/10.7554/eLife.54507 PMID: 32511092
59. Schwartz MK, McKelvey KS. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conserv Genet. 2009;10(2):441–52.
60. Sylvester EVA, Bentzen P, Bradbury IR, Clément M, Pearce J, Horne J, et al. Applications of random forest feature selection for fine-scale genetic population assignment. Evol Appl. 2017;11(2):153–65. https://doi.org/10.1111/eva.12524 PMID: 29387152
61. Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics. Trends Ecol Evol. 2003;18(4):189–97.
62. Lotterhos KE, Whitlock MC. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol Ecol. 2015;24(5):1031–46. https://doi.org/10.1111/mec.13100 PMID: 25648189