Molecular Ecology Resources
Article In Press
Acceptation date : 2023
https://doi.org/10.1111/1755-0998.13778
https://archimer.ifremer.fr/doc/00822/93394/

Estimating resistance surfaces using gradient forest and allelic frequencies

Vanhove Mathieu [1,*], Launey Sophie [1]

[1] DECOD (Ecosystem Dynamics and Sustainability), INRAE, Institut Agro, IFREMER, Rennes, France

* Corresponding author : Mathieu Vanhove, email address : mathieu.vanhove@gmail.com

Abstract :

Understanding landscape connectivity has become a global priority for mitigating the impact of landscape fragmentation on biodiversity. Connectivity methods that use link-based methods traditionally rely on relating pairwise genetic distance between individuals or demes to their landscape distance (e.g., geographic distance, cost distance). In this study, we present an alternative to conventional statistical approaches to refine cost surfaces by adapting the gradient forest approach to produce a resistance surface. Used in community ecology, gradient forest is an extension of random forest, and has been implemented in genomic studies to model species genetic offset under future climatic scenarios. By design, this adapted method, resGF, has the ability to handle multiple environmental predicators and is not subjected to traditional assumptions of linear models such as independence, normality and linearity. Using genetic simulations, resistance Gradient Forest (resGF) performance was compared to other published methods (maximum likelihood population effects model, random forest-based least-cost transect analysis and species distribution model). In univariate scenarios, resGF was able to distinguish the true surface contributing to genetic diversity among competing surfaces better than the compared methods. In multivariate scenarios, the gradient forest approach performed similarly to the other random forest-based approach using least-cost transect analysis but outperformed MLPE-based methods. Additionally, two worked examples are provided using two previously published datasets. This machine learning algorithm has the potential to improve our understanding of landscape connectivity and inform long-term biodiversity conservation strategies.

Keywords : functional connectivity, gradient forest, isolation by resistance, landscape genetics, resistance surface, machine learning

1 Introduction

Assessing connectivity has become a global conservation priority to respond to habitat fragmentation. The increasing anthropogenic pressures threaten natural populations and the need to maintain connectivity among the remaining habitats appears central for biodiversity (McClure *et al.* 2016). At the species level, landscape constraints affect dispersal, causing range shifts, fitness reduction and pushing species to the brink of extinction (Steffen *et al.* 2015). These deleterious effects to wildlife could only be mitigated if these challenges can be efficiently assessed.

Introduced in 2003, this field of landscape genetics has strengthened our understanding on species ability to move through heterogeneous landscapes (Manel *et al.* 2003). The term “functional connectivity” has been defined as “the degree to which the landscape facilitates or impedes movement among resource patches” (Taylor *et al.*, 1993). Measuring connectivity by estimating the degree of gene flow between individuals represents a method of interest to tackle habitat fragmentation, especially in areas where direct observation is complicated or impossible. Gene flow influences evolutionary trajectories of populations due to the modification of allelic frequencies, and “genetic connectivity” refers to the degree to which gene flow affects evolutionary processes within natural subpopulations. Genetic connectivity depends on several criteria including sufficient gene flow in order to avoid harmful effects of local inbreeding (inbreeding connectivity), to maintain equal allele frequencies (drift connectivity) or to allow potential advantageous alleles to spread across species range (adaptive connectivity) (Lowe & Allendorf 2010).

In the absence of direct observation of species movement, resistance estimations are used to fill the gap and provide a quantitative estimation of how environmental space is affecting dispersal. Resistance describes the willingness of individuals to move through their environment and the physiological costs associated to movement (Zeller *et al.* 2012).

Resistance surfaces are driven by a combination of movement and successful reproduction (Pflüger & Balkenhol 2014). High resistance values reveal restricted movement or barriers impeding gene flow, whereas low values translate into an ease of movement. In an 'isolation-by-distance' framework, the degree of genetic isolation is a sole function of Euclidean distances between populations (or individuals), while in 'isolation-by-resistance' (IBR) scenarios (McRae 2006), landscape variables (e.g. temperature gradient, roads, or topography) are used where each variable is characterized by a maximum resistance (R_{MAX}) and a functional form relating the environmental/anthropogenic predicator to resistance values (Shirk *et al.* 2018).

To date, no consensus has been established on the parametrization of resistance surfaces. Initially, resistance design relied on expert opinions (Stevens *et al.* 2006) but expert-based methods presented arbitrary costs with no consensus about resistance values (Spear & Storfer 2010). To avoid subjectivity in assigning resistance values, other approaches have been deployed like the use of telemetry (Driezen *et al.* 2007; Zeller *et al.* 2017) or habitat suitability models (Wang *et al.* 2008). The latter can provide different information than genetic connectivity models. However, a major drawback of habitat suitability models is their inability to move beyond the species level and assess the intraspecific variation due to local adaptation (Jay *et al.* 2012). Therefore, several studies have suggested that resistance parametrization should include genetic data to estimate gene flow, as the most basic level of biological diversity remains genetic variation (Khimoun *et al.* 2017; Peterman 2018).

Exhaustive search approaches based on genetic distance data have been developed to optimize resistance values (Wang *et al.* 2009). Grid search approaches were first introduced, where a constrained parameter space can be explored, limiting the number of models being assessed (Graves et al., 2013). Another stepping-stone in the development of unbiased resistance surface was the use a machine learning (i.e., genetic) algorithm to maximise pairwise genetic distances and effective resistance distances correlations as implemented by Peterman et al., (2014). Their subsequent R packages *ResistanceGA* (Peterman 2018) and *Radish* (Peterman & Pope 2021) can evaluate simultaneously multiple surfaces and a wide range of cost values. The algorithm relies on pairwise genetic distances (e.g., F_{ST} or D_a) and performs data transformation using monomolecular and Ricker functions. However, the optimization procedure requires high computer power to explore parameter space, which limits the choice of predicator variables to be optimized at the same time. Corridor or transect-based approaches have also been implemented where corridors are presumed to be the favoured paths between locations across the landscape. In this least-cost transect analysis (LCTA), landscape features are calculated along straight lines (or using a least cost path approach) between locations, with buffers of various widths (Emaresi *et al.* 2011; Van Strien *et al.* 2012). A movement model is then applied to the obtained resistance surfaces to calculate effective distances between sample points. Different modelling frameworks can be implemented to obtain these measures of landscape connectivity including least-cost path (LCP) modelling hypothesis where the organism is hypothesised to follow an ideal path (Adriaensen *et al.* 2003) or circuit-based modelling approaches which include all possible paths into the final measure of resistance (McRae 2006). Fletcher *et. al* (2019) introduced the spatial absorbing Markov chain (SAMC) framework to interpret matrix resistance and discern between movement behaviour and mortality. For additional methods to infer resistance-based connectivity, please refer to the review by Dutta *et al.* (2022).

An interesting alternative to conventional statistical approaches to parameterize resistance-surfaces (and its inverse: connectivity) is the use of machine learning techniques (Etherington 2016). These methods have been developed to predict models for complex and non-linear data which by design have the potential to avoid violations of assumptions of independence, linearity and normality (Balkenhol *et al.* 2009). Machine learning techniques have rarely been used in landscape genetics but have been recently gaining momentum (Hether & Hoffman 2012; Sylvester *et al.* 2018). In 2010, Murphy *et al.* used a random forest approach to assess the ecological processes limiting *Bufo boreas* connectivity in Yellowstone National Park. Recently, Pless *et al.*, (2021) implemented the first application of LCTA since Van Strien *et al.* (2012). The method combined a random forest framework with a LCP iterative optimization process to map the genetic connectivity of *Aedes aegypti* in North America and was subsequently used to understand the connectivity of tsetse flies (*Glossina pallidies*) in Kenya (Bishop *et al.* 2021). Convolutional neural networks (CNNs) have also been introduced to the field using image data to predict spatial patterns of genetic variation (Kittlein *et al.* 2022).

Another promising ensemble learning method is gradient forest (GF), an extension of random forest that uses regression trees to fit a model of associations between individual responses variables (e.g. single-nucleotide polymorphisms (SNP)) to predicator variables (e.g. climate variables). Initially developed in community ecology to model species turnover (Ellis *et al.* 2012), this regression tree-based method has been successfully used to map turnover of allele frequencies along environmental gradients (Fitzpatrick & Keller 2015). The GF approach is becoming increasingly implemented in landscape genomics and conservations studies (Gugger *et al.* 2018; Martins *et al.* 2018; Ingvarsson & Bernhardsson 2020; Vanhove *et al.* 2021). Once trained, GF models are used to predict continuous distribution of allelic turnover across the species range to estimate species maladaptation (i.e. genetic offset) when projected in time according to future climate scenarios (reviewed in Capblancq *et al.* 2020; Rellstab *et al.* 2021). The present study aims to investigate the potential of the gradient forest approach to generate landscape resistance surfaces (*res*GF) as an alternative to traditional link-based linear models which are subjected to normality, independence and linearity (Balkenhol *et al.* 2009).

Moving along environmental gradients, few changes in allelic frequencies might be observed while in some places large changes might occur. In the GF approach, partition of the data is distributed on either side of a split produced in the random forest. Split values are obtained from the ensemble of regression models and explain changes in allelic frequencies across each environmental predicator. The amount of variations, known as the 'raw split importance' (Fitzpatrick & Keller 2015), is cumulatively added in a step-wise manner and a compositional frequency curve, $F_p(x)$ is computed. Similarly, as building a staircase, the importance values are cumulatively summed and each step strength is proportional to the importance split at that location. The predictive performance for each SNP is quantified using the out-of-bag proportion (R^2) which provides a cross-validated estimate of the generalization error (Ellis *et al.* 2012). These goodness-of-fit R^2 estimates allow to assess the relative contribution of each environmental variables in explaining changes in allelic frequencies. Lastly, the accuracy importance of predicators is determined as the decrease in performance when a predictor is randomly permuted.

In our proposed method, these large steps along the gradient observed on the compositional frequency curve $F_p(x)$ are considered as evidence of resistance impeding gene flow whereas flatter regions represent regions where gene flow is facilitated. Therefore, *resGF* approach uses the derivative $f_p(x)$, known as the compositional turnover rate of $F_p(x)$, as a specialised transformation function to convert raster values into resistance values (Fig. 1). This transformation function acts similarly as the monomolecular or Ricker functions in *ResistanceGA* but appears specific to the examined predicator. Additionally, gradients which are strongly associated with biological variations will have larger steps and will reach a greater overall importance than other gradients. Therefore, each environmental predicator can be weighted by its R^2-weighted importance to build a multilayer resistance surface. In the present study, we evaluated ability of the gradient forest approach to generate resistance surfaces. Regression models rely on genetic distance (i.e. F_{ST} or D_{PS}) whereas GF methods use individual genotypes which makes this method a good candidate to potentially generate resistance surface. Through landscape and genetic simulations, we first assess the ability to recover the *true* resistance among alternative resistance surfaces. The ability of *resGF* to handle multiple environmental predicators is then assessed and its performance is compared to other published methods using previously published datasets.

2 Materials and Methods

A simulation approach was undertaken to assess the performance of gradient forest in generating a landscape surface. This framework aimed to compare *resGF* to methods commonly used in landscape genetics: Least-cost Transect Analysis, ResistanceGA and Species Distribution Model.

2.1 Methods to estimate functional connectivity

2.1.1 Least-cost Transect Analysis (LCTA)

1) The least-cost Transect Analysis (LCTA) method was first introduced by Van Strien *et al.* (2012). The method was first applied by Pless *et al.* (2021) to generate resistance surface of *Aedes aegypti* in North America. 2) LCTA has the potential to handle multiple variables. Briefly, the mean across straight lines between population pairs is calculated for each environmental predicator and a measure of genetic distance (1 - D_{PS}) is used as a response variable in a random forest (RF) model. The genetic distance for each pixel is predicted resulting in a resistance surface. In each iteration, the mean least cost path values are calculated through the previous iteration's connectivity surface (i.e. the inverse of the resulting resistance surface). 3) To account for spatial autocorrelation, a point kernel density

surface is implemented as described in the original method (Pless *et al.* 2021). The latter is then used to weight the RF bootstrapping to allow lower-density point to be more frequently sampled (Shen *et al.* 2020). 4) The method is assessed by selecting the surface from the optimal iteration (i.e. the iteration which displayed the lowest root-mean-square error). A leave-one-out cross-validation (loocv) can also be implemented to improve the robustness of the RSME values.

2.1.2 Optimisation of surface resistance with ResistanceGA

1) Peterman *et al.* (2014) were the first to maximise pairwise genetic distance and effective resistance distances using a genetic algorithm combined with linear mixed-effects models. 2) The *ResistanceGA* v4.1 package (Peterman 2018) optimizes landscape resistance surface using a genetic algorithm (R package *GA* v3.2; Scrucca, 2013). The genetic algorithm uses a unique combination of parameters to transform raster layer into a resistance surface seeking to maximize the relationship between pairwise landscape resistances (or least-cost distances) and pairwise genetic distances. Pairwise least-cost distances were calculated with the *costDistance* function as performed by Flores-Manzanero et al., (2019) using the R package *gdistance* v1.3 (Van Etten 2017). All these processes were performed while allowing an exploration of resistance values up to 2,500 and using an eight-neighbour connection scheme. 3) However, the method does not account for spatial autocorrelation. 4) To assess the resulting surface, an objective function was selected during optimization (i.e. Akaike information criteria (AIC)). It was determined from linear mixed-effects models using the maximum-likelihood population effects (MLPE) parameterization. The model fit with pairwise genetic distance (1 – D_{PS}) as dependent variable and commute distance between individual pairs as predicator variable. The maximum-likelihood population effects (MPLE) parameterization (Clarke *et al.* 2002; Van Strien *et al.* 2012) implemented in the R package *LME4* v1.1 (Bates *et al.* 2014) accounted for the non-independence among the pairwise genetic and ecological distances. Additionally,

MLPE models have been found to performed better than regression models in landscape genetic model selection (Shirk *et al.* 2018). The support of the optimized resistance surfaces was assessed using the AICc (Akaike information criteria corrected for small sample; Akaike, 1974).

2.1.3 Species Distribution Model

1) Habitat suitability model have been first used to assess landscape connectivity by Wang *et al.* (2008). 2) Resistance surfaces were derived from Species distribution model (SDM). These modelling approach do not include a genetic component and are based on environmental characteristics of sample locations. The resultant habitat suitability map can be incorporated into least-cost path analyses. We included resistance surface generated using the package *rmaxent* v0.8.5 (Baumgartner *et al.* 2017). Habitat suitability values were generated for each pixel across the landscape, normalise between 0 and 1 and converted to resistance values (1 – suitability value following Spear *et al.* (2010)). 3) However, spatial autocorrelation remains a largely unresolved problem in species distribution modelling. 4) The method was assessed using area under a receiver operating characteristic (ROC) curve (AUC) and withholding 50% of the data. An iteration was included in the final analysis when AUC > 0.70.

2.1.4 Resistance Gradient Forest - *resGF*

1) The Gradient forest (GF) approach was initially developed in community ecology to model species turnover (Ellis *et al.* 2012) before successfully used to map turnover of allele frequencies along environmental gradients (Fitzpatrick & Keller 2015). The present study aims to test GF ability to generate resistance surfaces.

2) GF models compositional turnover in allele frequencies using monotonic non-linear functions along environmental gradients. These turnover functions transform environmental variables into a common biological scale of compositional turnover allowing the conversion of multidimensional environmental space to multidimensional genetic space while considering the weight of the selected variables which best describe genetic variation (Ellis *et al.* 2012; Capblancq *et al.* 2020). The *gradientForest* v0.1 package first produces a random forest model for each of the input SNP using the R package *extendedForest* v1.6.1 (Liaw & Wiener 2002). In the random forest models, regression or classification tree models describe the relationship between an individual SNP and environmental variables (Breiman 2001).

Each random forest is composed of an ensemble of regression trees which recursively partition the data. At each split, the partitioning is performed to obtain the smallest total impurity, defined as the sum of squared deviations about the group mean. The partition is repeated until a minimum number of sites is attained and the last partition becomes the terminal node. Each node in the decision tree is characterised by its split importance which is the reduction in impurity of the node created by the split. This split hold information regarding the sensitivity of a set SNP along a gradient as it measures the degree of variation explained by the partitioning. If a certain SNP is absent under a specific threshold, the split is likely to uncover this specific threshold.

The predictive power of random forest models is assessed using i) the goodness-of-fit R_f^2 for each SNP f which is the proportion of out-of-bag variance explained, ii) the accuracy importance I_{fp} for a predictor p within the forest as well as iii) the raw importance I_{fpts} for a predictor p at a split value s in a particular tree t (Ellis *et al.* 2012). Random forest uses out-of-bag (OOB) estimates as cross-validated estimate of error by comparing the expected variance of the OOB samples to the variance of the observations. The proportion of the variance explained $R^2{}_f$ can be used as a measure of the information of a SNP provided by a particular predicator. The goodness-of-fit $R^2 f$ is partitioned among the predicator in proportion to their accuracy importance (I_{fp}) yielding R^2_{fp}, the predictive accuracy of the SNP f for the predictor p. The cumulative turnover function, $F_p(x)$ is compiled along each environmental gradient using an aggregate of the tree splits values from the random forest models for all species' models, $F_{fp}(x)$, which display a positive fit ($R^2_f > 0$) (see Ellis et al., 2012 for details).

3) To account for spatial autocorrelation, Ellis *et al.,* (2012) introduced the scaled density. In the original paper, the derivative $f_p(x)$ of $F_p(x)$ is defined as the compositional turnover rate at a predicator value x where $f_p(x)$ would be equal to the expected value of observed importance density, $I(x)$, if sampling was uniformed. To account for non-uniformity of sampling the density of the observed splits distributed across the gradient are scaled over the observed range. A combined importance density $I_{ap}(s)$ is computed for each predicator value x and $f_p(x)$ can be estimated as $\hat{f}_{ap}(x) = I_{ap}(s)/d_{ap}(x)$ where $d_{ap}(x)$ if the scaled density of the predicator values over the observed range Δ_a, normalized to satisfy $\int d_{ap}(x)dx = \Delta_a$. The scaled density is computed using Gaussian kernel with bandwidth given by Silverman's rule-of-thumb and "whitened" as descried in Ellis et al. (2012).

4) In the classical use of the gradient forest approach, model selection is implicit in the fitting process. The shape of these turnover functions describes the rate of compositional change along environmental gradients. Species with higher predictive random forest models (high R^2_f values) contributing more than those with low predictive power. Steep parts of the turnover function indicate rapid turnover of allelic frequencies (or species assemblage), whereas flatter regions of the curve describe more homogenous parts (Pitcher et al., 2012). The *resGF* package uses the derivative $f_p(x)$ of the compositional turnover function $F_p(x)$ as a transformation function to convert raster values into resistance values (Fig. 1). The monotonic turnover function for each predicator ranges from 0 to R^2_{fp} and the rationale behind

resGF is that steep parts of the turnover function represent barriers to gene flow whereas flatter regions facilitate connectivity. When using multiple environmental layers, each predictor variable is weighted by its accuracy importance to generate a multilayer resistance surface. Each $F_p(x)$ function allows for the transformation from arbitrary scales to common biological units of compositional turnover. Then, these multidimensional environment spaces can be weighted and transformed into a multidimensional biological space (Ellis *et al.* 2012; Fitzpatrick & Keller 2015).

2.2 Simulations and method comparisons

To evaluate the robustness of each model and optimization procedure given different combination of samples, multiple tests were performed. Landscape genetics data were simulated using 1) an univariate framework to validate *resGF* ability to generate a resistance surface and was compared to other published methods (*ResistanceGA*, LCTA and SDM). The ability of each method to identify a *true* resistance surface among a competing set of surfaces. The SDM method serves as a non-genetic-based method to contrast the results. Correlations between resulting surfaces using *resistanceGA*, *resGF*, LCTA and SDM were recorded at each iteration. 2) A multilayer resistance surface approach investigated the ability of each of the three genetic methods to correctly weight the contribution of each landscape. 3) Finally, *resGF* was compared with the LCTA optimization method and Estimated Effective Migration Surface (EEMS) (Petkova *et al.* 2016) using on a previously published datasets.

2.2.1 Univariate scenarios

In our univariate scenario, we aimed to represent populations which were naturally distributed along a gradient (Script: sim_single_cont_surface.R). For each simulation, we used the same neutral landscape surface (100x100 cells) as described in Peterman & Pope, (2021). We generated a spatially correlated Gaussian random field ("gaus") with an autocorrelation range of seven and a magnitude of variation of 25; a fractional Brownian motion (“fbm”) with a fractal dimension of 0.5 using *NLMR* v1.0 R package (Sciaini *et al.* 2018) as well a composite surface (“composite_surface”) combining the two surfaces using the *Combine_Surfaces* function in *ResistanceGA* which allowed to obtain a landscape correlated to the other two surfaces. The resulting rasters cover a gradient and resemble the temperature and precipitation variables commonly used in habitat modelling (Fig. 2 and Fig. S1). In this univariate scenario, we aimed at maximising the effects of isolation by environment and minimizing the effects due to distance to investigate the power of *resGF* and *ResistanceGA* to identify the true surface. The different surfaces served alternatively as our *true* surface (only surface influencing movement). An artificial gradient of 200 SNPs for 10,000 individuals was generated using the R package *coenocliner* v0.2 (Simpson 2016) with the *coenocline* function and the Gaussian response model and Bernoulli countModel options. The resulting count matrix was converted into a *genind* object and subsampled to retain 50 individuals and 200 biallelic SNPs. The 50 final individuals were mapped to the closest value on the *true* raster (Fig. 2). Mapping the genotypes to their closest value onto the true raster allowed to obtain individuals which where spatially structured according to the gradient generated with *coenocliner*. Our gene flow estimate (calculated as 1 - pairwise proportion of shared alleles (D_{PS})) were calculated in *adegenet* v2.1.3 (Jombart 2008). Simulated genetic diversity estimates were summarised on Fig.1 and Fig. 2 using Principal Component Analysis (PCA) computed in ade4 v1.7-18 (Dray & Dufour 2007). Although the species movement was not explicitly modelled, this framework allowed to obtain individuals structured according to our simulated environmental variable which allowed to generate resistance surfaces under the tested methods. To incorporate spatial components, principal coordinates of neighbour matrices (PCNMs) were computed in *vegan* v2.5-7 R package and the first half of the positive PCNMs were retained for the gradient forest models as previously suggested (Manel *et al.*

2010). This set of orthonormal variables are computed calculated through eigenvalue decomposition of a spatial weighting matrix (x–y-coordinates) (Dray *et al.* 2006). An example of the residuals is provided on Fig. S7.

Five hundred trees were run per iterations using default parameters and including a correlation threshold of 0.5 (Strobl *et al.* 2008). The resulting GF object was passed into the *resGF* function to obtain our final resistance surface (https://github.com/MVan35/resGF).

For each iteration of 100 iterations performed, three resistance surfaces ("gaus", "fbm" and "composite_surface") were generated and the best model according to the different performance metrics were recorded. The true surface was alternatively the "gaus" variable for 100 iterations, the "fbm" variable and finally the composite surface. The correlation between the *true* resistance surfaces obtained from the different models (*ResistanceGA, LCTA*, *resGF and SDM*) were calculated for each iteration as well as the correlation between landscape distance and our gene flow estimate.

To evaluate the robustness of the assessed methods given the different combination of samples, we examined the effects of the resistance surfaces by refitting a maximum likelihood population effects model (MLPE; Clarke et al., (2002)), implemented in the R package *ecodist* v2.0.5 (Goslee & Urban 2007). Euclidian distances were computed into a distance matrix and effective resistance distance were obtained using a movement model (*least-cost path*) implemented in the R package *gdistance* v1.3 (Van Etten 2017). All MLPE models were fit using the *mlpe_rga* function in *ResistanceGA* (Peterman 2018). Each scenario was assessed using a model selection approach with different performance metrics: AIC, BIC, conditional and marginal R^2 were computed in the R package *performance* v0.7 (Lüdecke *et al.* 2020), AICc was calculated in *AICcmodavg* v2.3 (Mazerolle & Mazerolle 2017). The accuracy importance of the GF model was also recorded to capture the ability of the model to identify the true model.

To investigate the ability of the different methods to generate resistance surfaces from categorical variables, we repeated the same simulation with the rasters described above, but the surfaces were transformed into categorical surfaces following the procedure described by Savary *et al.,* (2021) (script: sim_single_cat_surface.R). Five categories were obtained for each surface (proportion: category 1 - 30%; category 2 – 30%; category 3 – 15%; category 4 – 15% and category 5 – 20%).

Finally, the resulting habitat suitability map obtained for each iteration was integrated in an individual-based model (details provided in *Supplementary Information* (SI)). These simulations using the R package *RangeShiftR* v1.0.4 (Malchow *et al.* 2021), an individual-based model which allows to incorporate ecological and evolutionary processes as well as population dynamics (demography, emigration, transfer and mortality). This modelling framework extended the simulation performed in *coenocliner* where an artificial gradient of 200 SNP over the landscape was generated. The individual-based model yielded allelic frequencies simulated across 400 years and allowed for populations to evolve in a landscape habitat map (generated under the SDM model under the univariate scenario) using a correlated random walk. From the resulting allelic frequencies outputted by the individual-based model at year 400, resistance surfaces were obtained using the tested methods using 500 individuals across 50 populations randomly selected. Correlations between the initial SDM input and the resulting resistance surfaces and were recorded (See Supplementary Information).

2.2.2 Multivariate scenarios

Simulations were performed using either one, three, five or ten landscapes generated using the *NLMR* package (Script: sim_multi_surface.R – 20 iterations). As described above an artificial gradient was created for each raster resulting in 200 SNPs and 50 individuals. To investigate the ability of each method to account for individual surfaces in a multivariate scenario, a number of individuals were selected from each individual surface and pooled together into a final *genind* object using the *repool* function in adegenet. In our three variables scenario, raster 1 contributed to 70% of the allelic frequencies, raster 2 contributed 30% with no contribution from raster 3. Similarly, in the five variables scenario, only raster 1 (50%), raster2 (30%) and raster 3 (20%) contributed to the final surface.

In the ten variables approach, rasters which had an impact on the final generated resistance surface varied (Fig. S9). In the first scenario, the expected contribution of each raster was as followed: raster 1 (35%), raster 2 (25%), raster 3 (15%), raster 4 (10%), raster 5 (10%) and raster 6 (5%). In the second scenario, only the first four rasters contributed to gene flow: raster 1 (40%), raster 2 (30%), raster 3 (20%), raster 4 (10%), and in the third scenario, the four first rasters contributed: raster 1 (25%), raster 2 (25%), raster 3 (25%), raster 4 (25%).

To produce a multiple resistance surface using *ResistanceGA*, a composite surface was generated using the *Combine_Surfaces* function. However, the relative contribution of each raster cannot be accounted for without adequate parametrization. Therefore, to compare the result obtained from multisurface scenarios, we compare *resGF* to the least-cost transect analysis (LCTA) implemented as described above.

2.2.3 Comparison using published datasets

To compare the two methods, a previously published dataset of *Andropadus virens*, an African tropical bird, was obtained on Dryad (Zhen *et al.* 2017). In this study, 15 populations (182 individuals) were sequenced using RADseq resulting in 47,482 SNPs using a minimum allelic frequency of 0.02. Twenty-three environmental variables were used to estimate the resistance surface of *Andropadus virens* (See Supplementary Information).

Additionally, the microsatellites dataset used in the original publication of the LCTA method (Pless *et al.* 2021) was also investigated (Fig. 5 and Fig. S4). Microsatellite datasets were included as this type of markers remains widely used and the ability of the gradient forest approach has not been tested using a limited set of markers. For this dataset, iterative random forest approach was performed without using the leave-one-out cross-validation.

To compare the resulting resistance surfaces, Estimating Effective Migration Surfaces (EEMS) were generated (Petkova *et al.* 2016). EEMS was developed to visualise non-homogeneous gene-flow on geographic map. This method explores patterns of isolation by distance (IBD) and uses effective migration to investigate the relationship between genetic distance and geography. Regions with low effective migration are associated with reduced gene-flow over time, whereas regions with relatively high effective migration can be interpreted as evidence of elevated gene-flow.

3 Results

3.1 Univariate simulations

For the continuous variable, the “fbm” and “gaus” rasters showed no sign of correlation (Pearson’s ρ = -0.047, SD = 0.259) whereas the “gaus” and “composite surface” were strongly correlated (ρ = 0.648, SD = 0.136). The average correlation between Euclidean distance and genetic distance (1 – D_{PS}) was 0.443 indicating that simulations of patterns of isolation by distance was successful. For the different genetic-based models which were investigated (*resGF, ResistanceGA* and *LCTA*), the correlation between genetic distance and the effective resistance distance was on average 0.486 for *resGF*, 0.477 for *ResistanceGA*, 0.412 *LCTA*. One iteration is provided as an example in Fig. S1 with the resulting R^2 weighted importance for the GF method (Fig. S8). The SDM approach achieved an average pattern of isolation by resistance of 0.415 (average AUC = 0.734).

The MLPE regressions on our independent variables using the *resGF* approach were accurate (67.0% with AIC) while *ResistanceGA* identified the true model in 47.0% of cases and 26.0% for LCTA. *ResGF* and *ResistanceGA* were generally successful in identifying the true surface. Their performance metrics yielded similar inferences (Table 1) whereas the LCTA approach was not able to robustly detect the true surface under the MLPE evaluation approach. In the gradient forest models, the accuracy importance was able to consistently identify the *true* surface (100% of iterations). When using the "gaus" raster as *true* surface, the average contribution was 42.1% with an average of 86 SNPs (42.8% of all SNPs) with a positive R^2 signalling that model correctly identified the contribution of landscape predicator as a driver of genetic differentiation. The contribution of the two other variables in their respective single layer gradient forest model was 4.1% for the "fbm" surface and 23.0% for the "composite surface". The resulting resistance surfaces for the two models appeared reasonably correlated (ρ = 0.299, Table S8). The *resGF* approach achieved on average the highest correlation with the other methods tested reaching ρ = 0.669 with the LCTA approach and ρ = 0.319 with *ResistanceGA* (Table 2). Resistance surfaces obtained using the GF-based method appeared most strongly correlated with the ones produced by the non-genetic-based method (SDM) with a Pearson correlation of 0.570. Fig. 2 depicts the resulting resistance/cost surfaces for two iterations obtained for each of the methods investigated.

When running our univariate simulations on a local computer, the *resGF* approach took on average 1.63 min, *LCTA* 13.43 min and 1h55 for *ResistanceGA*. The computation time for *resGF* increased as the number of individuals and number of loci used in the analysis (Table S2). Using different subsets of a large SNPs dataset (10,000 SNPs and 1,000 individuals) and comparing the resulting surfaces from each subset dataset yielded correlated results (average

Pearson's correlation > 0.497 using 50 individuals and > 0.748 using 100 individuals; Table S3).

Regarding simulations for categorical variables, the accuracy importance of the gradient forest appeared as the most reliable metric. The gradient forest approach identified the *true* surface 85% of the time (Fig. S2). When fitting a MLPE model, the different metrics performed unevenly (Table S1). AIC values for LCTA approach were the most consistent at finding the *true* surface although the RSME value for LCTA identified the correct surface in 35% of cases. For conditional and marginal R^2, *resGF* was the most supported method.

Additionally, in the individual-based simulations (presented in SI), the average surface correlations across the different iterations resulted in *resGF* being the genetic-based method which appeared the most strongly correlated with the initial surface (ρ = 0.233), followed by *resistanceGA* (ρ = 0.192) and LCTA (ρ = -0.040).

3.2 Multivariate simulations

The multivariate approach aimed to examine the ability of gradient forest to correctly account for the contribution of each surface. For the scenario using three surfaces, *resGF* approached the expected contribution of each raster (Table 3). For raster 1, the average contribution of *resGF* was 61% (expected value = 70%), 29% for raster 2 (expected value = 30%) and 10% for raster 3 (expected value = 0%). The LCTA approach achieved 48%, 31% and 21% for the 3-surfaces scenario. Since the multisurface layers for *ResistanceGA* is based on a composite surface, the relative contributions for this method were different than the expected values.

The correlation of the resistance surfaces across the different methods was the highest between *resGF* and LCTA (ρ = 0.246). Similarly, the two RF-based approaches were able to identify the relative contribution of each raster using five variables (Table S4). For the 5-

surfaces scenario, *resGF* and LCTA achieved respectively 43% and 31% for raster 1 (expected 50%), 24% and 21% for raster 2 (expected 30%) and 19% and 22% for raster 3 (expected 20%). However, the correlation achieved between the two methods was equal to 0.109 in the 5-surfaces scenario. In the 10-surfaces approach, the correlation between the two RF methods reached 0.106. In each of the three different scenarios (Table S4-7), the contribution of single surface to the overall model for each method (*resGF* and LCTA) was similar.

3.3 Comparison using published datasets

A total of 47,482 SNPs (MAF > 0.2) were included in the analyses of the *Andropadus virens* dataset. The migration map estimated using Estimating Effective Migration Surface (EEMS) (where blue indicates relative high migration and red indicates lower migration) revealed elevated historical gene flow in the northern and eastern parts of the distribution with an area of reduced gene flow western Cameroon (Fig. 3A). The gradient forest approach resulted in 4,469 SNPs with a $R^2 > 0$ which summed to around 10% of the dataset. Regarding the iterative random forest optimization, the present result was performed using the leave-one-out cross-validation procedure. The third iteration obtained the lowest root-mean-square error (RSME = 0.0091; with a nodesize = 2 and Mtry = 8) with a Pearson correlation coefficient of 0.528 between predicted and observed genetic distance (Fig. S3). The two methods produced a Pearson's correlation between the two resulting surfaces of 0.474 (Fig. 3).

The contribution of the different surfaces between the two approaches differed. In our gradient forest analysis, spatial predicators contributed the most to the model (Fig. 4) and the combined R^2 for the remaining environmental predicators was 41.0 %. In the GF analysis, precipitation variables contributed strongly to the model. Among environmental variables, precipitation in the warmest quarter was the variable contributed the most (3.05%) followed by precipitation of the wettest month (2.83%) and temperature seasonality (2.38%). However, in the iterative random forest optimization approach, only precipitations in the warmest quarter appeared among the top ten variables. Maximum and Minimum temperatures followed by NDVI and altitude were the predicators contributing more strongly to the model. The two methods were further tested by removing the Bioko island population and comparing the resulting surfaces in order to investigate the robustness of the different approaches (Fig. S5). In this process, the leave-one-out cross-validation procedure for LCTA was compared to the full method. The *resGF* method appeared the most consistent with a correlation of 0.895 with and without the Bioko island population while the correlation among the LCTA methods varied from 0.051 to 0.880.

When investigating the distribution of *Ae. aegypti* in the United States using the microsatellites dataset published by Pless *et al.* (2021), the migration surface revealed the highest migration in Florida and some parts of central Mexico. Reduced migration rates were inferred in the center east of the distribution, specifically in Louisiana. The two connectivity surfaces generated using *resGF* and LCTA had a correlation of 0.461 (Fig. 5). The two methods identified similar areas as regions of high connectivity: in Florida, in central United States and in Western United States (Fig. 5). Variable contribution varied between the two methods, where barren land cover was ranked eighth most important predicator for the LCTA method but its contribution as well as the contribution of snow or flood in the GF model was null. Both human density and urban land cover contributed to the GF model but urban land cover was the least contributing variable in LCTA while human density was fourth (Fig. S4).

4 Discussion

To understand animal movement, it is crucial to develop effective landscape-level conservation approaches (Zeller *et al.* 2012). Resistance surfaces represent the interaction between gene flow and landscape variables and therefore appear as adequate candidates for conservation prioritization analyses (Spear & Storfer 2010; Hanson *et al.* 2019). The *resGF* method is derived from the gradient forest approach (Ellis *et al.* 2012) which models allelic changes over the landscape. In the present study, we tested this approach with other methods designed to generate map of gene flow. Our *resGF* method was able to generate resistance surfaces and to assemble multisurface resistance layers using the R^2-weighted importance as a weighting factor.

4.1 Univariate and multivariate Simulations

In our univariate simulation studies, the average correlation between geographic and genetic distances was 0.443 indicating that the simulation framework was able to generate patterns of isolation-by-distance. Our simulation framework relied on generating individuals with SNPs associated with a gradient using *coenocliner* package and these individuals were subsequently mapped onto a raster. On average, 42.8% of the 200 simulated SNPs were found to have a positive R^2 in gradient forest analyses indicating that this simulation approach succeeded in modelling SNP responding to environmental gradients. We first aimed to recover the *true* surface from competing raster layers, replicating the framework described by Peterman & Pope (2021). This testing approach relied on obtaining pairwise effective distance across individuals using a least-cost path algorithm and to perform model selection on the resulting distance matrices by fitting a MLPE model. The different metrics used in model selection did not vary, but we referred to AIC as suggested by Row *et al.* (2017). The investigated methods were able to achieved an overall robust correlation between cost and genetic distance (*resGF* = 0.477; *ResistanceGA* = 0.486, LCTA = 0.412 and SDM = 0.411 - Fig. S1-2).

The ability of *ResistanceGA* to recover patterns of isolation-by-distance is due to its genetic algorithm which by design aims to maximise the pairwise relationship between genetic and cost distances. On the other hand, *resGF* algorithm maps changed in allelic frequencies over the landscape and the GF algorithm was also able to correctly capture the overall patterns of isolation-by-distance when present. The *resGF* method performed slightly better than *ResistanceGA* in recovering the *true* resistance surface from competing surfaces (Table 1 & S1). In our continuous scenario, we examined the accuracy importance of the different surfaces for the GF analyses and *resGF* always correctly identified the *true* surface. In the categorical scenario, the studied methods performed unevenly (Table S1), but the accuracy importance of the gradient forest appeared as the most consistent metric. However, it is worth noting that the different methods performed better using continuous variables.

Among the different methods tested, *resGF* and the LCTA approach achieved the highest correlation between their resulting resistance surfaces reaching 0.699 in the univariate scenario. *ResGF* also displayed the strongest correlation among genetic-based methods with the resistance surface generated using species distribution model (SDM) (Table 2). *ResistanceGA* was moderately correlated with either *resGF*, LCTA or SDM methods reaching 0.299, 0.319 and 0.337 respectively. Although we observed strong correlation between resistance surfaces generated with *resGF* and LCTA, the latter was not able to distinguish the *true* resistance surface among competing surfaces when refitting a MLPE model (Table 1). However, the LCTA method performed well when using RMSE as a metric where LCTA approach identified the true surface >60% of the time (Table 1) for continuous variable. These findings ascertain that *resGF* performed at least as well as *ResistanceGA* under the univariate scenario to generate resistance surface with continuous and categorical variables. Moreover, in our individual-based model (provided in SI), *resGF* was the genetic-based method which was the most strongly correlated surface with initial resistance map (1-habitat map).

The resolution provided by the gradient forest offered a refined complexity over the landscape. As opposed to regression models which uses summary statistics in the form of F_{ST} or D_{PS}, *resGF* takes advantage of the potential of machine learning applications, it uses individual genotypes and environmental predicators experienced by these individuals rather than population estimates to refine a cost surface.

Peterman & Pope (2021) raised concerned about fitting regression models with effective distance calculated independently from several single resistance surfaces. In this study, the two RF-based approaches tested were successfully able to incorporate multiple environmental predicators to generate a resistance surface (Table 3). In *ResistanceGA*, multisurface optimization was performed using a composite surface which does not take into account the weighted importance of the different environment predicators and therefore could not calculate the relative contribution of the different landscape variables. Link-based linear models refer to methods relating pairwise genetic distance to their landscape distance and these approaches are deeply subjected to multicollinearity among predictors (Cayuela *et al.* 2018). RF-based approaches appear free from these assumptions and by design, can accommodate correlated variables (Ellis *et al.* 2012).

In our multivariate simulations, *resGF* performed slightly better than LCTA in its ability to match the expected contribution of each individual raster value (Table 3, Table S4-7). The two methods maintained a moderate average correlation of 0.246 for the three-surfaces scenario but the correlation dropped to 0.109 and 0.106 in the five and ten-surfaces scenarios. The lack of observed correlation between the two methods under these scenarios might reflect the unrealistic aspect of the cumulative effect of the generated landscapes. A different simulation framework outputting allelic frequencies where populations are allowed to evolve according to multiple abiotic factors could further improve the assessment of these methods. The two RF-based methods are designed to incorporate multiple environmental predicators and have been previously used successfully (Fitzpatrick & Keller 2015; Bishop *et al.* 2021; Pless *et al.* 2021).

The runtime analysis (Table S2) indicates that *resGF* is faster than *ResistanceGA* when the number of individuals is below 500 with 1,000 loci investigated. However, using different subsets of a large dataset, the resulting resistance surface appeared strongly correlated (Table S3) These findings indicate that the method would be appropriate for RAD-seq or Genotyping-in-Thousands by sequencing (GT-seq) datasets. *ResGF* could be implemented using larger dataset using cluster computing resources or using a subset of SNPs selected using a landscape genomic approach. Regarding the number of features to be included in an analysis, Breiman (2001) suggested to limit the number of variables to $\log_2(N + 1)$ features in order to limit the correlation among trees and reduce the impact of the features on the generalization of error.

4.2 Comparison using published datasets

In the two real-dataset examples, the two approaches converged. *ResGF* and LCTA performance converged with an average correlation of 0.474 for the *Andropadus virens* dataset (Zhen *et al.* 2017) (Fig. 3) and 0.461 for the *Aedes aegypti* microsatellite dataset (Pless *et al.* 2021) (Fig. 5). Interestingly, the migration surface calculated using EEMS offered a different picture of gene flow. This approach under the stepping-stone model approximate all possible migration histories using resistance distance to adjust the migration rate in the population grid before interpolating across the entire habitat (Petkova *et al.* 2016). In the two datasets, EEMS appeared significantly different than the resistance surfaces. Migration surface are designed to highlight areas where genetic similarities decay faster (i.e. low effective migration) or where the relationship between genetic similarities and geographic distance remain constant. In the *Andropadus virens* dataset, high effective migrations were observed in the North of the studied area, whereas RF-based methods identified this region as high resistance to gene flow (i.e. low connectivity). In the West, low migration rates were inferred, a pattern which was observed in the *resGF* resistance surface. In the microsatellite dataset, Florida represented an area of high connectivity for *Ae. aegypti* as well as a place with high effective migrations. Other pockets of high connectivity were observed using both *resGF* and the LCTA approaches, which were not inferred by EEMS excepted for the stretch of high effective migration rates inferred in Texas. The combination of effective migration surface with resistance-based methods can be complementary when designing conservation areas or developing control strategy for infectious diseases propagated by dispersing animals. Resistance surface seemed to offer a refined precision, whereas EEMS has the ability to reveal broad scale patterns of gene flow.

To investigate the robustness of RF-based methods, one population was removed from *A. virens* dataset (Bioko island population). This also allowed to obtain a dataset which was not fragmented by the sea as non-fully connected populations could potentially be problematic for some of the tested methods. In the resulting resistance surfaces, some variation was observed across the different implementations of the LCTA methods (full model and loocv) and when using 14 and 15 populations with correlation among cost surfaces ranging from 0.051 to 0.880 (Fig. S6). The LCTA optimization method, like most landscape genetic approach, rely on sampling unit underlying some *a priori* decision about demes delineation (Manel *et al.* 2003). In this analysis, the genetic distance measure used was proportion of shared alleles (D_{PS}) as this metric responds faster to landscape change (Savary *et al.* 2021b). F-statistics are predominantly used to reflect gene flow and assume equal effective population sizes and demographic equilibrium (Prunier *et al.* 2017). By modelling the change in allelic frequencies over the landscape, gradient forest approach does not rely on demes delineation or F-statistics. In the random forest model, each allele is used and contributes to establish species connectivity over the landscape. In the LCTA optimization approach, mean values between population pairs calculated through each environmental raster are used to predict the resistance surface using a RF model. The optimization method therefore models the potential distance through each raster between populations, whereas *resGF* uses changes in allelic frequencies to obtain transformation functions. These variations might explain the differences in variable contributions where *resGF* performed slightly better in matching the expected contribution of environmental predicator in the multivariate scenario (Table 3). LCTA relies on least-cost path algorithm where the algorithm assumes that dispersing individuals possess a perfect knowledge of the entire landscape (Adriaensen *et al.* 2003). *ResGF* is an ensemble learning approach which relies on specialised transformation functions where each function is suited to the data.

Machine learning algorithms are designed to develop predictive models using complex and non-linear data (Olden *et al.* 2008). In this study, we examined the potential of random forest approaches to generate cost-surface. The *resGF* approach provides an alternative to link-based linear models as it does not violate traditional assumptions of linear model, independence, normality and linearity (Balkenhol *et al.* 2009). The study has shown that gradient forest approach could be applied to wide range of genetic datasets including whole-genome, RAD-seq, microsatellite or even Genotyping-in-Thousands by sequencing (GT-seq) datasets. This method appeared more precise that the LCTA optimization although the two approaches converged when using real datasets. These methods can be use in complement of effective migration surfaces when implementing conservation or disease elimination strategies. In a world subjected to increasing landscape fragmentation, we hope that this gradient forest approach and other machine learning techniques will prove useful to understand landscape connectivity.

5 Acknowledgements

This research was funded by the European Regional Development Fund through the Interreg Channel VA Programme (project: 'Salmonid Management Around the Channel'SAMARCH).

6 Conflict of Interest statement

No conflict of interest to declare

7 References

Adriaensen F, Chardon JP, De Blust G, *et al.* (2003) The application of "least-cost"modelling as a functional landscape model. *Landscape and urban planning*, **64**, 233–247.

Akaike H (1974) A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**, 716–723.

Balkenhol N, Waits LP, Dezzani RJ (2009) Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography*, **32**, 818–830.

Bates D, Mächler M, Bolker B, Walker S (2014) Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Baumgartner J, Wilson P, Esperon-Rodriguez M (2017) rmaxent: Tools for working with Maxent in R. *R package version 0.4*, **9**.

Bishop A, Amatulli G, Hyseni C, *et al.* (2021) A machine learning approach to integrating genetic and ecological data in tsetse flies (Glossina pallidipes) for spatially explicit vector control planning. *Evolutionary Applications*.

Breiman L (2001) Random forests. *Machine learning*, **45**, 5–32.

Capblancq T, Fitzpatrick MC, Bay RA, Exposito-Alonso M, Keller SR (2020) Genomic prediction of (mal) adaptation across current and future climatic landscapes. *Annual Review of Ecology, Evolution, and Systematics*, **51**, 245–269.

Cayuela H, Rougemont Q, Prunier JG, *et al.* (2018) Demographic and genetic approaches to study dispersal in wild animal populations: A methodological review. *Molecular Ecology*, **27**, 3976–4010.

Clarke RT, Rothery P, Raybould AF (2002) Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics*, **7**, 361.

Dray S, Dufour A-B (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software*, **22**, 1–20.

Dray S, Legendre P, Peres-Neto PR (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *ecological modelling*, **196**, 483–493.

Driezen K, Adriaensen F, Rondinini C, Doncaster CP, Matthysen E (2007) Evaluating least-cost model predictions with empirical dispersal data: a case-study using radiotracking data of hedgehogs (Erinaceus europaeus). *Ecological modelling*, **209**, 314–322.

Dutta T, Sharma S, Meyer NF V, Larroque J, Balkenhol N (2022) An overview of computational tools for preparing, constructing and using resistance surfaces in connectivity research. *Landscape Ecology*, 1–30.

Ellis N, Smith SJ, Pitcher CR (2012) Gradient forests: calculating importance gradients on physical predictors. *Ecology*, **93**, 156–168.

Emaresi G, Pellet J, Dubey S, Hirzel AH, Fumagalli L (2011) Landscape genetics of the Alpine newt (Mesotriton alpestris) inferred from a strip-based approach. *Conservation Genetics*, **12**, 41–50.

Etherington TR (2016) Least-cost modelling and landscape ecology: concepts, applications, and opportunities. *Current Landscape Ecology Reports*, **1**, 40–53.

Van Etten J (2017) R package gdistance: distances and routes on geographical grids.

Fitzpatrick MC, Keller SR (2015) Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecology letters*, **18**, 1–16.

Fletcher Jr RJ, Sefair JA, Wang C, *et al.* (2019) Towards a unified framework for connectivity that disentangles movement and mortality in space and time. *Ecology letters*, **22**, 1680–1689.

Flores-Manzanero A, Luna-Bárcenas MA, Dyer RJ, Vázquez-Domínguez E (2019) Functional connectivity and home range inferred at a microgeographic landscape genetics scale in a desert-dwelling rodent. *Ecology and evolution*, **9**, 437–453.

Goslee SC, Urban DL (2007) The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, **22**, 1–19.

Graves TA, Beier P, Royle JA (2013) Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular ecology*, **22**, 3888–3903.

Gugger PF, Liang CT, Sork VL, Hodgskiss P, Wright JW (2018) Applying landscape genomic tools to forest management and restoration of Hawaiian koa (Acacia koa) in a changing environment. *Evolutionary applications*, **11**, 231–242.

Hanson JO, Fuller RA, Rhodes JR (2019) Conventional methods for enhancing connectivity in conservation planning do not always maintain gene flow. *Journal of Applied Ecology*, **56**, 913–922.

Hether TD, Hoffman EA (2012) Machine learning identifies specific habitats associated with genetic connectivity in Hyla squirella. *Journal of evolutionary biology*, **25**, 1039–1052.

Ingvarsson PK, Bernhardsson C (2020) Genome-wide signatures of environmental adaptation in European aspen (Populus tremula) under current and future climate conditions. *Evolutionary Applications*, **13**, 132–142.

Jay F, Manel S, Alvarez N, *et al.* (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular ecology*, **21**, 2354–2368.

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*, **24**, 1403–1405.

Khimoun A, Peterman W, Eraud C, *et al.* (2017) Landscape genetic analyses reveal fine-scale effects of forest fragmentation in an insular tropical bird. *Molecular ecology*, **26**, 4906–4919.

Kittlein MJ, Mora MS, Mapelli FJ, Austrich A, Gaggiotti OE (2022) Deep learning and satellite imagery predict genetic diversity and differentiation. *Methods in Ecology and Evolution*, **13**, 711–721.

Liaw A, Wiener M (2002) Classification and regression by randomForest. *R news*, **2**, 18–22.

Lowe WH, Allendorf FW (2010) What can genetics tell us about population connectivity? *Molecular ecology*, **19**, 3038–3051.

Lüdecke D, Makowski D, Waggoner P, Patil I (2020) Performance: assessment of regression models performance. *R package version 0.4*, **5**.

Malchow A, Bocedi G, Palmer SCF, Travis JMJ, Zurell D (2021) RangeShiftR: an R package for individual-based simulation of spatial eco-evolutionary dynamics and species' responses to environmental changes. *Ecography*, **44**, 1443–1452.

Manel S, Poncet BN, Legendre P, Gugerli F, Holderegger R (2010) Common factors drive adaptive genetic variation at different spatial scales in Arabis alpina. *Molecular ecology*, **19**, 3824–3835.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in ecology & evolution*, **18**, 189–197.

Martins K, Gugger PF, Llanderal-Mendoza J, *et al.* (2018) Landscape genomics provides evidence of climate-associated genetic variation in Mexican populations of Quercus rugosa. *Evolutionary Applications*, **11**, 1842–1858.

Mazerolle MJ, Mazerolle MMJ (2017) Package "AICcmodavg." *R package*.

McClure ML, Hansen AJ, Inman RM (2016) Connecting models to movements: testing connectivity model predictions against empirical migration and dispersal data. *Landscape Ecology*, **31**, 1419–1432.

McRae BH (2006) Isolation by resistance. *Evolution*, **60**, 1551–1561.

Murphy MA, Evans JS, Storfer A (2010) Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology*, **91**, 252–261.

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, **83**, 171–193.

Peterman WE (2018) ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods in Ecology and Evolution*, **9**, 1638–1647.

Peterman WE, Connette GM, Semlitsch RD, Eggert LS (2014) Ecological resistance surfaces predict fine-scale genetic differentiation in a terrestrial woodland salamander. *Molecular Ecology*, **23**, 2402–2413.

Peterman WE, Pope NS (2021) The use and misuse of regression models in landscape genetic analyses.

Petkova D, Novembre J, Stephens M (2016) Visualizing spatial population structure with estimated effective migration surfaces. *Nature genetics*, **48**, 94–100.

Pflüger FJ, Balkenhol N (2014) A plea for simultaneously considering matrix quality and local environmental conditions when analysing landscape impacts on effective dispersal.

Pitcher C, Lawton P, Ellis N, *et al.* (2012) Exploring the role of environmental variables in shaping patterns of seabed biodiversity composition in regional-scale ecosystems. *Journal of Applied Ecology*, **49**, 670–679.

Pless E, Saarman NP, Powell JR, Caccone A, Amatulli G (2021) A machine-learning approach to map landscape connectivity in Aedes aegypti with genetic and environmental data. *Proceedings of the National Academy of Sciences*, **118**.

Prunier JG, Colyn M, Legendre X, Flamand M (2017) Regression commonality analyses on hierarchical genetic distances. *Ecography*, **40**, 1412–1425.

Rellstab C, Dauphin B, Exposito-Alonso M (2021) Prospects and limitations of genomic offset in conservation management. *Evolutionary Applications*.

Row JR, Knick ST, Oyler-McCance SJ, Lougheed SC, Fedy BC (2017) Developing approaches for linear mixed modeling in landscape genetics through landscape-directed dispersal simulations. *Ecology and Evolution*, **7**, 3751–3761.

Savary P, Foltête J, Moal H, Vuidel G, Garnier S (2021a) Analysing landscape effects on dispersal networks and gene flow with genetic graphs. *Molecular Ecology Resources*.

Savary P, Foltête J, Moal H, Vuidel G, Garnier S (2021b) graph4lg: A package for constructing and analysing graphs for landscape genetics in R. *Methods in Ecology and Evolution*, **12**, 539–547.

Sciaini M, Fritsch M, Scherer C, Simpkins CE (2018) NLMR and landscapetools: An integrated environment for simulating and modifying neutral landscape models in R. *Methods in Ecology and Evolution*, **9**, 2240–2248.

Scrucca L (2013) GA: a package for genetic algorithms in R. *Journal of Statistical Software*, **53**, 1–37.

Shen LQ, Amatulli G, Sethi T, Raymond P, Domisch S (2020) Estimating nitrogen and phosphorus concentrations in streams and rivers, within a machine learning framework. *Scientific data*, **7**, 161.

Shirk AJ, Landguth EL, Cushman SA (2018) A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Molecular Ecology Resources*, **18**, 55–67.

Simpson GL (2016) coenocliner: a coenocline simulation package for R.

Spear SF, Balkenhol N, FORTIN M, McRae BH, Scribner KIM (2010) Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Molecular ecology*, **19**, 3576–3591.

Spear SF, Storfer A (2010) Anthropogenic and natural disturbance lead to differing patterns of gene flow in the Rocky Mountain tailed frog, Ascaphus montanus. *Biological Conservation*, **143**, 778–786.

Steffen W, Broadgate W, Deutsch L, Gaffney O, Ludwig C (2015) The trajectory of the Anthropocene: the great acceleration. *The Anthropocene Review*, **2**, 81–98.

Stevens VM, Verkenne C, Vandewoestijne S, Wesselingh RA, Baguette M (2006) Gene flow and functional connectivity in the natterjack toad. *Molecular Ecology*, **15**, 2333–2344.

Van Strien MJ, Keller D, Holderegger R (2012) A new analytical approach to landscape genetic modelling: Least-cost transect analysis and linear mixed models. *Molecular ecology*, **21**, 4010–4023.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. *BMC bioinformatics*, **9**, 1–11.

Sylvester EVA, Beiko RG, Bentzen P, *et al.* (2018) Environmental extremes drive population structure at the northern range limit of Atlantic salmon in North America. *Molecular ecology*, **27**, 4026–4040.

Taylor PD, Fahrig L, Henein K, Merriam G (1993) Connectivity is a vital element of landscape structure. *Oikos*, 571–573.

Vanhove M, Pina-Martins F, Coelho AC, *et al.* (2021) Using gradient Forest to predict climate response and adaptation in Cork oak. *Journal of Evolutionary Biology*.

Wang IJ, Savage WK, Bradley Shaffer H (2009) Landscape genetics and least-cost path analysis reveal unexpected dispersal routes in the California tiger salamander (Ambystoma californiense). *Molecular ecology*, **18**, 1365–1374.

Wang Y-H, Yang K-C, Bridgman CL, Lin L-K (2008) Habitat suitability modelling to correlate gene flow with landscape connectivity. *Landscape ecology*, **23**, 989–1000.

Zeller KA, McGarigal K, Cushman SA, *et al.* (2017) Sensitivity of resource selection and connectivity models to landscape definition. *Landscape ecology*, **32**, 835–855.

Zeller KA, McGarigal K, Whiteley AR (2012) Estimating landscape resistance to movement: a review. *Landscape ecology*, **27**, 777–797.

Zhen Y, Harrigan RJ, Ruegg KC, *et al.* (2017) Genomic divergence across ecological gradients in the Central African rainforest songbird (A ndropadus virens). *Molecular ecology*, **26**, 4966–4977.

8 Data Accessibility

The *resGF* function is available on GitHub (https://github.com/MVan35/resGF)as well as the simulation scripts for single continuous and categorical variable and multi-variable simulations. The dataset of *Andropadus virens,* an African tropical bird, was obtained on Dryad - https://datadryad.org/stash/dataset/doi:10.5061/dryad.8n8t0 (Zhen et al., 2017) and the microsatellite dataset for the LCTA (Pless *et al.* 2021) was obtained from GitHub - https://github.com/evlynpless/MOSQLAND/tree/master/ModelingConnectivity.

Data citation:

- Zhen Y, Harrigan RJ, Ruegg KC, *et al.* 2017, Data from: Genomic divergence across ecological gradients in the Central African rainforest songbird (Andropadus virens) - https://datadryad.org/stash/dataset/doi:10.5061/dryad.8n8t0
- Pless E, Saarman NP, Powell JR, Caccone A, Amatulli G 2021, ModelingConnectivity - https://github.com/evlynpless/MOSQLAND/tree/master/ModelingConnectivity

9 Authors' contributions

MV designed the study and performed the simulation as well as writing the manuscript under SL supervision.

10 Figures and tables

Tables

Table 1 – Simulation summary for the continuous univariate scenario varying the *true* surface using either the "gaus", "fbm" or "composite" surface as the true surface influencing movement over 100 iterations. Evaluation metrics for each surface are provided, they were obtained either by refitting a maximum likelihood population effects model (MLPE) or using method specific metrics.

True surface	**Method tested**	**Evaluation metrics by surfaces by refitting a MLPE**					**Method specific metrics**
		AIC	**BIC**	**R2c**	**R2m**	**AICc**	
gaus	resGF	67%	67%	42%	61%	67%	Accuracy importance: 100%

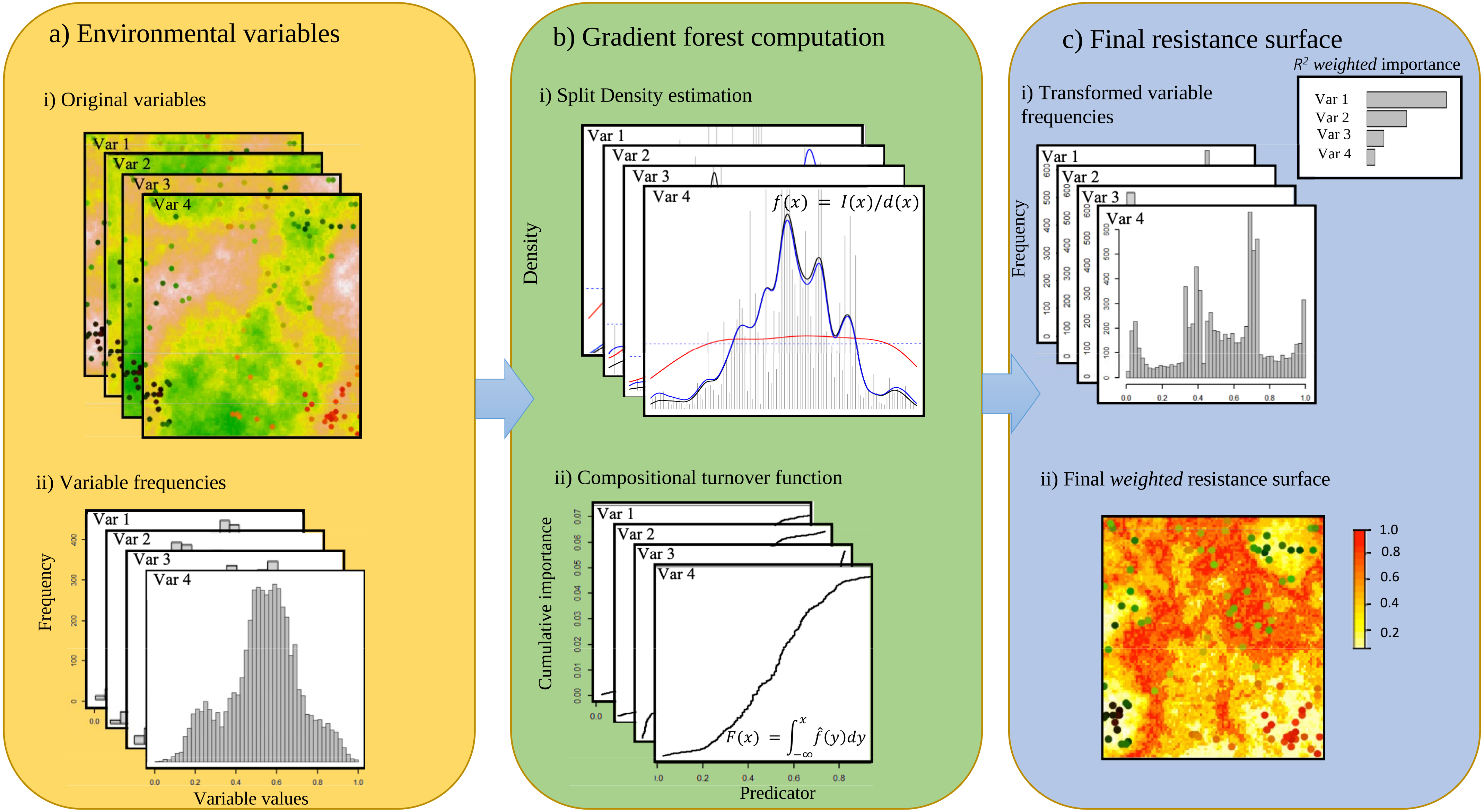
	ResistanceGA	47%	47%	55%	50%	47%	-
	LCTA	26%	26%	35%	37%	35%	RSME: 63%
fbm	resGF	61%	58%	55%	67%	59%	Accuracy importance: 100%
	ResistanceGA	43%	49%	55%	43%	43%	-
	LCTA	43%	37%	20%	26%	43%	RSME: 63%
Composite	resGF	63%	59%	61%	61%	64%	Accuracy importance: 100%
	ResistanceGA	44%	47%	50%	46%	41%	-
	LCTA	46%	42%	31%	29%	45%	RSME: 59%

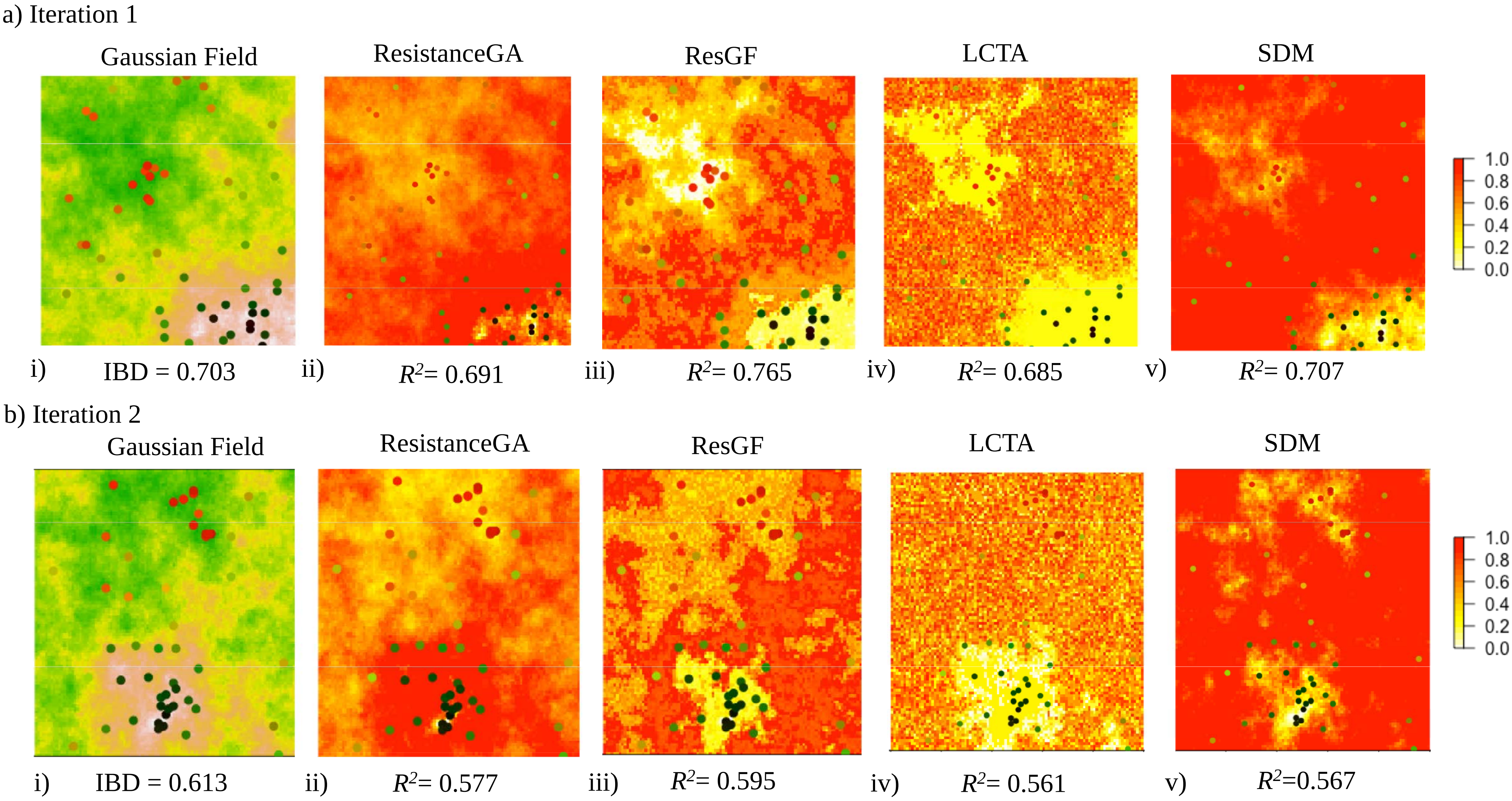
Table 2 – Correlation between the final resistance surfaces generated using the different methods

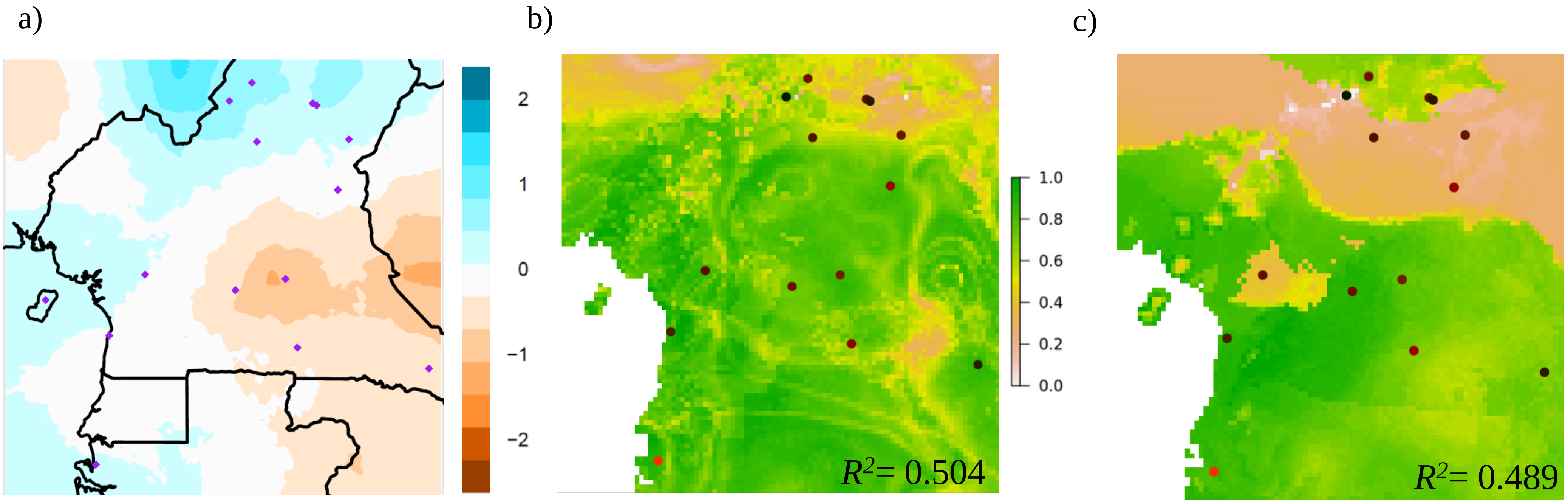
	resGF	**ResistanceGA**	**LCTA**	**SDM**
resGF	1.000	0.442	0.669	0.570
ResistanceGA	0.442	1.000	0.319	0.337
LCTA	0.669	0.319	1.000	0.030
SDM	0.570	0.337	0.030	1.000

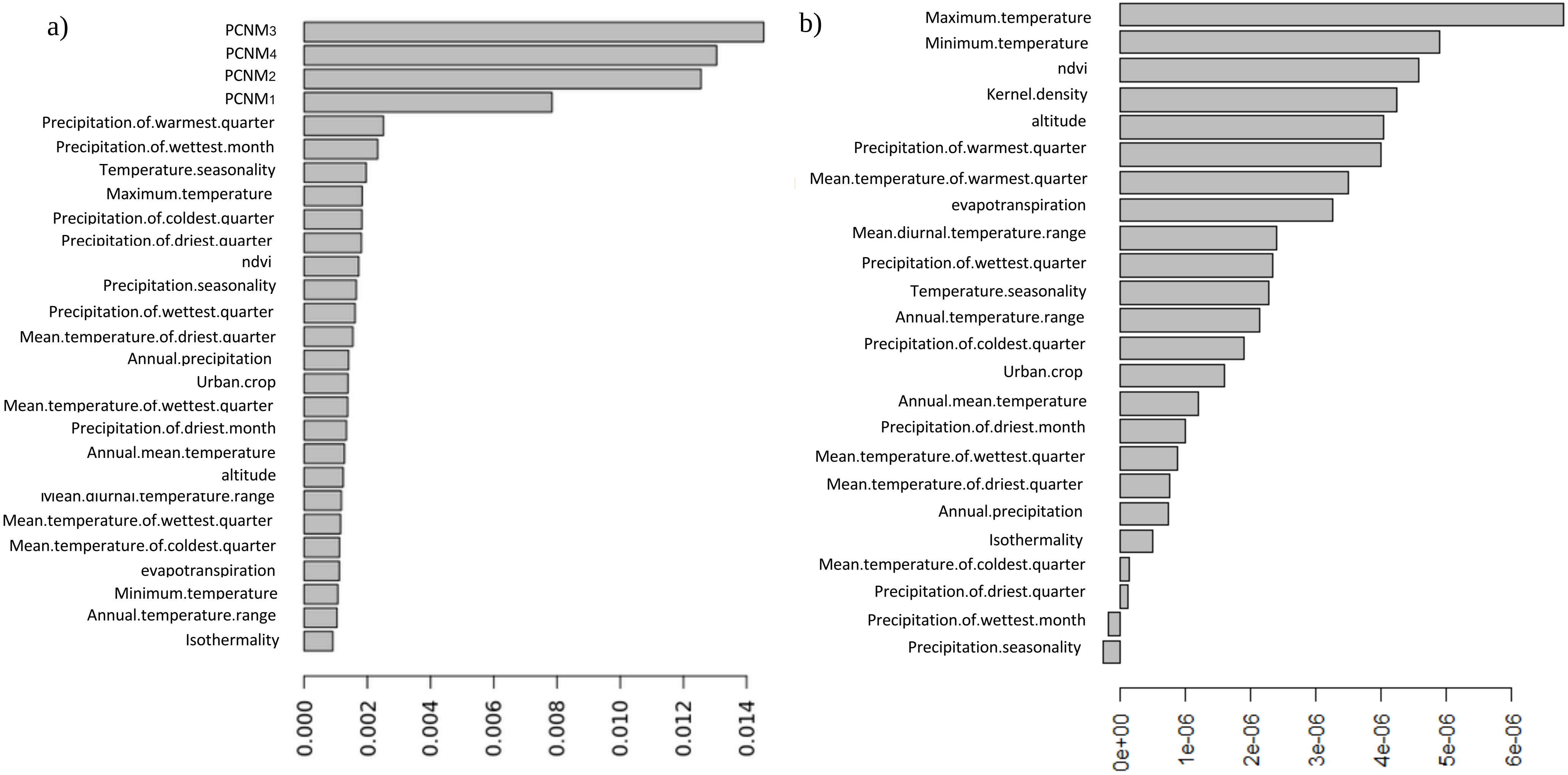
Table 3 – Contribution of each surface to the final multivariate surface using three rasters scenario. In this three-variables scenario, the multivariate scenario contained 70% of pooled allelic frequencies simulated on the fractional brownian motion raster (Raster 1), 30% of the allelic frequencies found in the Gaussian field raster (Raster 2) and no contribution from the random cluster raster (Raster 3). The table presents average results (over 20 iterations) of the variable importance measures for each method and the expected values under this scenario.

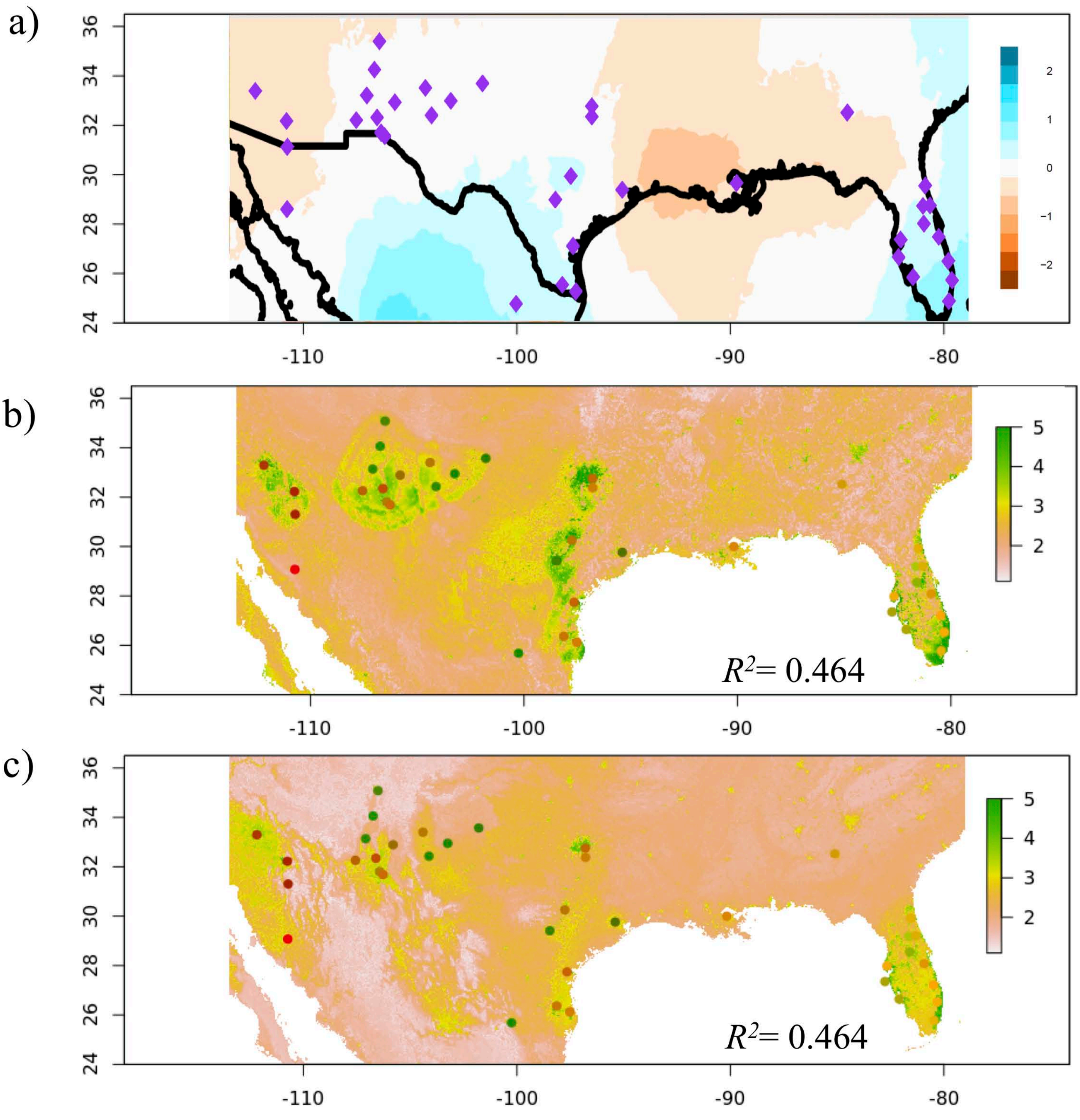
contribution	**Name**	**Expected %**	**ResistanceGA**	**resGF**	**LCTA**
Raster 1	fractional brownian motion	70%	22%	61%	48%
Raster 2	Gaussian field	30%	46%	29%	31%
Raster 3	random cluster	0%	31%	10%	21%


a) Environmental variables
i) Original variables
Var 1
Var 2
Var 3
Var 4
ii) Variable frequencies
Frequency
Var 1
Var 2
Var 3
Var 4
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
Variable values
b) Gradient forest computation
i) Split Density estimation
Density
Var 1
Var 2
Var 3
Var 4
$f(x) = I(x)/d(x)$
ii) Compositional turnover function
Cumulative importance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Var 1
Var 2
Var 3
Var 4
$F(x) = \int_{-\infty}^{x} \hat{f}(y)dy$
0.0
0.2
0.4
0.6
0.8
Predicator
c) Final resistance surface
R^2 weighted importance
Var 1
Var 2
Var 3
Var 4
i) Transformed variable frequencies
Frequency
Var 1
Var 2
Var 3
Var 4
0
100
200
300
400
500
600
0.0
0.2
0.4
0.6
0.8
1.0
ii) Final weighted resistance surface
1.0
0.8
0.6
0.4
0.2


a) Iteration 1
Gaussian Field
ResistanceGA
ResGF
LCTA
SDM
i)
IBD = 0.703
ii)
R^2= 0.691
iii)
R^2= 0.765
iv)
R^2= 0.685
v)
R^2= 0.707
1.0
0.8
0.6
0.4
0.2
0.0
b) Iteration 2
Gaussian Field
ResistanceGA
ResGF
LCTA
SDM
1.0
0.8
0.6
0.4
0.2
0.0
i)
IBD = 0.613
ii)
R^2= 0.577
iii)
R^2= 0.595
iv)
R^2= 0.561
v)
R^2=0.567


a)
2
1
0
−1
−2
b)
1.0
0.8
0.6
0.4
0.2
0.0
R^2= 0.504
c)
1.0
0.8
0.6
0.4
0.2
0.0
R^2= 0.489


a)
PCNM3
PCNM4
PCNM2
PCNM1
Precipitation.of.warmest.quarter
Precipitation.of.wettest.month
Temperature.seasonality
Maximum.temperature
Precipitation.of.coldest.quarter
Precipitation.of.driest.quarter
ndvi
Precipitation.seasonality
Precipitation.of.wettest.quarter
Mean.temperature.of.driest.quarter
Annual.precipitation
Urban.crop
Mean.temperature.of.wettest.quarter
Precipitation.of.driest.month
Annual.mean.temperature
altitude
Mean.diurnal.temperature.range
Mean.temperature.of.wettest.quarter
Mean.temperature.of.coldest.quarter
evapotranspiration
Minimum.temperature
Annual.temperature.range
Isothermality
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.014
b)
Maximum.temperature
Minimum.temperature
ndvi
Kernel.density
altitude
Precipitation.of.warmest.quarter
Mean.temperature.of.warmest.quarter
evapotranspiration
Mean.diurnal.temperature.range
Precipitation.of.wettest.quarter
Temperature.seasonality
Annual.temperature.range
Precipitation.of.coldest.quarter
Urban.crop
Annual.mean.temperature
Precipitation.of.driest.month
Mean.temperature.of.wettest.quarter
Mean.temperature.of.driest.quarter
Annual.precipitation
Isothermality
Mean.temperature.of.coldest.quarter
Precipitation.of.driest.quarter
Precipitation.of.wettest.month
Precipitation.seasonality
0e+00
1e-06
2e-06
3e-06
4e-06
5e-06
6e-06


a)
36
34
32
30
28
26
24
-110
-100
-90
-80
2
1
0
−1
−2
b)
36
34
32
30
28
26
24
-110
-100
-90
-80
5
4
3
2
R^2= 0.464
c)
36
34
32
30
28
26
24
-110
-100
-90
-80
5
4
3
2
R^2= 0.464