Inferring landscape resistance to gene flow when genetic drift is spatially heterogeneous

Paul Savary, Jean-Christophe Foltête, Hervé Moal, Gilles Vuidel, Stéphane Garnier

Inferring landscape resistance to gene flow when genetic drift is spatially heterogeneous

Savary, Paul*[1, 2, 3], Foltête, Jean-Christophe[2], Moal, Hervé[1], Vuidel, Gilles[2], and Garnier, Stéphane[3]

[1] *ARP-Astrance, 9 Avenue Percier, 75008 Paris, France*
[2] *ThéMA, UMR 6049 CNRS, Université de Franche-Comté, 32 Rue Mégevand, 25030 Besançon Cedex, France*
[3] *Biogéosciences, UMR 6282 CNRS, Université Bourgogne-Franche-Comté, 6 Boulevard Gabriel 21000 Dijon, France*

Abstract

In connectivity models, land cover types are assigned cost values characterising their resistance to species movements. Landscape genetic methods infer these values from the relationship between genetic differentiation and cost distances. The spatial heterogeneity of population sizes, and consequently genetic drift, is rarely included in this inference although it influences genetic differentiation. Similarly, migration rates and population spatial distributions potentially influence this inference. Here, we assessed the reliability of cost value inference under several migration rates, population spatial patterns, and degrees of population size heterogeneity. Additionally, we assessed whether considering intra-population variables, here using gravity models, improved the inference when drift is spatially heterogeneous. We simulated several gene flow intensities between populations with varying local sizes and spatial distributions. We then fit gravity models of genetic distances as a function of (i) the 'true' cost distances driving simulations or alternative cost distances, and (ii) intra-population variables (population sizes, patch areas). We determined the conditions making the identification of the 'true' costs possible and assessed the contribution of intra-population variables to this objective. Overall, the inference ranked cost scenarios reliably in terms of similarity with the 'true' scenario (cost distance Mantel correlations), but this 'true' scenario rarely provided the best model goodness-of-fit. Ranking inaccuracies and failures to identify the 'true' scenario were more pronounced when migration was very restricted (< 4 dispersal events/generation), population sizes most heterogeneous and some populations spatially aggregated. In these situations, considering intra-population variables helps identify cost scenarios reliably, thereby improving cost value inference from genetic data.

Keywords: landscape genetics, gene flow, cost distances, connectivity, gravity models, simulation

*Corresponding author: paul.savary@concordia.ca

1 Introduction

Dispersal movements maintain genetic diversity and contribute to species survival in human-shaped landscapes (Frankham, 2005; Spielman et al., 2004). Deriving efficient conservation measures to halt the continuing erosion of biodiversity thus requires knowledge regarding the influence of landscape features on species movements. For that end, landscape ecology studies have provided spatially-explicit models for dispersal paths by quantifying the resistance of landscape features to dispersal (Zeller et al., 2012). This implies assigning a cost value to every landscape feature on a resistance surface in order to identify the most likely dispersal paths, e.g. using least cost path modelling (Adriaensen et al., 2003) or applying circuit theory to ecological connectivity (McRae, 2006). However, these connectivity models are only reliable under the condition that the cost values assigned to each landscape feature realistically depict the behaviour of the species when moving across landscape features. Accordingly, although the choice of cost values on resistance surfaces is often based upon expert opinion, a wide range of biological data can be used to calibrate them so that they somehow fit ecological reality (Zeller et al., 2018).

Following the emergence of landscape genetics (Manel et al., 2003), genetic data have often been used for calibrating cost values because the genetic structure of a set of populations depends upon the structure of the landscape (Keyghobadi, 2007). Indeed, provided enough time has elapsed following population settlement and the most recent landscape changes for the genetic differentiation pattern to reach an equilibrium, we can expect a positive linear relationship between genetic differentiation and effective cost distances between populations (Hutchison and Templeton, 1999; McRae, 2006; Slatkin, 1993). The Isolation By Landscape Resistance (IBLR) model is an extension of the original Isolation By Distance (IBD) model to heterogeneous landscapes in which population spatial distribution is irregular and effective distances are computed as cost-distances or resistance distances rather than geodesic Euclidean distances (Guillot et al., 2009; McRae, 2006). In this context, the inference of cost values from genetic data relies upon the IBLR model and consists in identifying the cost scenario which maximises the strength of the relationship between the corresponding cost-distances and genetic distances among a set of alternative cost scenarios (Cushman et al., 2006; Graves et al., 2013; Peterman, 2018).

These approaches usually assume a preponderant influence of landscape-driven gene flow on genetic differentiation (Richardson et al., 2016), although the latter is also substantially driven by genetic drift. When a population is subdivided into several small populations, especially when their effective sizes are reduced and the migration rate is low, genetic drift is responsible for a loss of genetic diversity which tends to increase genetic differentiation between population pairs (Frankham, 1996; Frankham et al., 2004; Hartl et al., 1997). According to theory, when the size of a population varies over time, genetic drift will be most intense when the population is the smallest. Thus, the harmonic mean of the population sizes over time is a reliable proxy for the intensity of drift over the whole period because it weights smaller populations more heavily (Hartl et al., 1997; Prunier et al., 2017). Applying this same theory to the spatial context of subdivided populations, Serrouya et al. (2012) and Weckworth et al. (2013) showed that the harmonic mean of the population sizes of population pairs was a good predictor of their pairwise genetic differentiation. Therefore, just as gene flow does not affect genetic differentiation between all population pairs in the same way depending on the effective distances between them, genetic drift does not affect genetic differentiation between all pairs in the same way depending on their respective sizes.

Recently, Prunier et al. (2017) introduced the Spatial-Heterogeneity-in-Population-Sizes hypothesis (SHNe) for assessing the contribution of population size spatial heterogeneity to genetic differentiation patterns. Using both simulated and empirical data, they showed that when the migration rate is low and the overall population size heterogeneity is high, pairwise population size heterogeneity contributes to genetic differentiation more significantly than the distance between populations does. These authors developed two pairwise metrics measuring population size heterogeneity that can be included in the analysis of genetic differentiation drivers. This makes it possible to account for drift spatial heterogeneity and to assess more reliably the relationship between i) effective distances and ii) genetic distances, which then directly reflects the spatial drivers of gene flow. Failing to do so may potentially lead to spurious conclusions regarding dispersal patterns (Weckworth et al., 2013; Prunier et al., 2017). Accordingly, metrics quantifying population size heterogeneity could be variables as important as the effective distances between populations under the IBD or IBLR hypotheses for explaining the spatial genetic structure (Prunier et al., 2017). Yet, whether variables accounting for population size heterogeneity also improve cost value inference remains to be investigated.

Estimating population effective sizes is a requisite for taking their heterogeneity into account in the analyses, but is undoubtedly a difficult task (Wang, 2005). Yet, provided that effective sizes are somehow proportional to census sizes, they can be approximated with environmental proxies for the carrying capacities of habitat patches occupied by populations (Prunier et al., 2017), thereby saving costly field work. Furthermore, environmental variables computed at the population level may reflect not only population sizes but also local incentives to departure or establishment (Baguette et al., 2013; Bonte et al., 2012). Such variables have already been shown to influence significantly genetic structure (Murphy et al., 2010; Wang, 2013; Wang et al., 2013), though seldom considered in landscape genetic analyses (Pflüger and Balkenhol, 2014), and could positively contribute to cost value inference.

Finally, Graves et al. (2013) suggested that the spatial aggregation of individuals could prevent gene flow between clumps of individuals thereby preventing gene flow from compensating for drift effects. The influence of the spatial distribution pattern of individuals on the spatial genetic structure has already been evidenced (Ueno et al., 2000). However, it is not known whether the population spatial distribution pattern could influence cost value inference when populations are the focus of the analysis.

The first objective of this study was to assess the reliability of cost value inference from genetic data under several migration rates, population spatial distribution patterns and degrees of population size heterogeneity. We expected the quality of the inference to be reduced when migration rates are limited, some populations are spatially aggregated, and population sizes are spatially heterogeneous. The second objective was to identify situations where the inclusion of intra-population variables in the models improves this inference. When population sizes are heterogeneous, we expected the inclusion of intra-population variables, i.e., either population sizes or patch areas, to move the results of the analyses closer to the ecological reality. Gravity models (Anderson, 1979; Fotheringham and O'Kelly, 1989) have already been used in landscape genetics and allow for the test of these hypotheses because these models enable researchers to assess the influence of intra- and inter-population variables on measures of genetic differentiation (Murphy et al., 2010; Robertson et al., 2018; Watts et al., 2015; Zero et al., 2017). When patch capacities or population sizes and inter-patch distances are the predictor variables of the genetic distance between populations, several models including different predictor variables can be compared on the basis of a same measure of goodness-of-fit, which makes it potentially

possible to identify the most realistic cost value scenario while accounting for population size heterogeneity. Accordingly, we used a factorial design to simulate several intensities of gene flow between sets of populations with varying levels of population size heterogeneities and spatial distribution patterns. We then fit gravity models explaining simulated genetic distances as a function of the cost distance driving the simulation as well as other alternative cost distances, and of intra-population variables, i.e. population sizes and patch areas. We could thus identify the conditions making cost value inference possible and situations where the inclusion of intra-population variables helped identifying the 'true' cost scenario driving the gene flow simulations.

2 Methods

2.1 Overall methodological approach

The basic framework for inferring landscape resistance from the relationship between landscape distances and genetic distances consists in computing cost-distances (CD), or other resistance distances, from a land cover map and several cost scenarios, i.e., sets of cost values assigned to each land cover type. These cost-distances are included in models explaining genetic distances, and the scenario providing the best model goodness-of-fit is supposed to best reflect the effect of the landscape on dispersal (Peterman, 2018). To assess the reliability of this type of inference (objective 1), we simulated the genetic differentiation pattern emerging through gene flow over several generations in a species with limited dispersal capacities (Figure 1). We knew the 'true' cost values associated with land cover types, arbitrarily fixed prior to simulations, and the resulting CD driving dispersal in the simulated landscapes. We then assessed the capacity of landscape genetic models to identify this 'true' cost scenario among a range of 'alternative' cost scenarios diverging more or less from the 'true' cost scenario (Figure 2). Next, using regression trees, we tried to delineate the range of situations over which gravity models including both inter-population CD and intra-population variables improved cost value inference, when population size are heterogeneous (objective 2).

2.2 Simulations

2.2.1 Landscape and population simulations

When simulating landscapes, we ensured that patches were sufficiently large for cost-distances to vary substantially according to the cost value scenario. Indeed, landscape fragmentation is known to affect CD variability when using different cost scenarios (Cushman et al., 2013; Rayfield et al., 2010). This variability ensures that several alternative scenarios lead to different cost-distance matrices, thereby making the inference possible. To this end, we simulated 200 landscapes with four land cover types by discretizing spatially correlated Gaussian random fields models (Schlather et al., 2015). We used a level of land cover auto-correlation leading to variable cost-distances across the cost scenarios (`autocor=30` in the `nlm_gaussianfield()` function from the `NLMR` R package (Sciaini et al., 2018)).

We simulated the movement of a forest specialist species with limited dispersal capacities avoiding anthropogenic land cover types when dispersing. Accordingly, forests covered 20 % of the simulated landscapes and were the most permeable areas for dispersal (cost: 1). Cost values and proportions of the other land cover types were set to reflect the dispersal constraints of a forest specialist species in a heterogeneous landscape: grasslands (cost: 10, proportion: 27%), crops (100, 27%) and artificial areas (1000, 26%). Similar cost values have already been employed to analyse ecological connectivity for forest species (Gurrutxaga et al., 2010; Schadt et al., 2002) and their range (1-1000) matches that inferred from field data in other studies

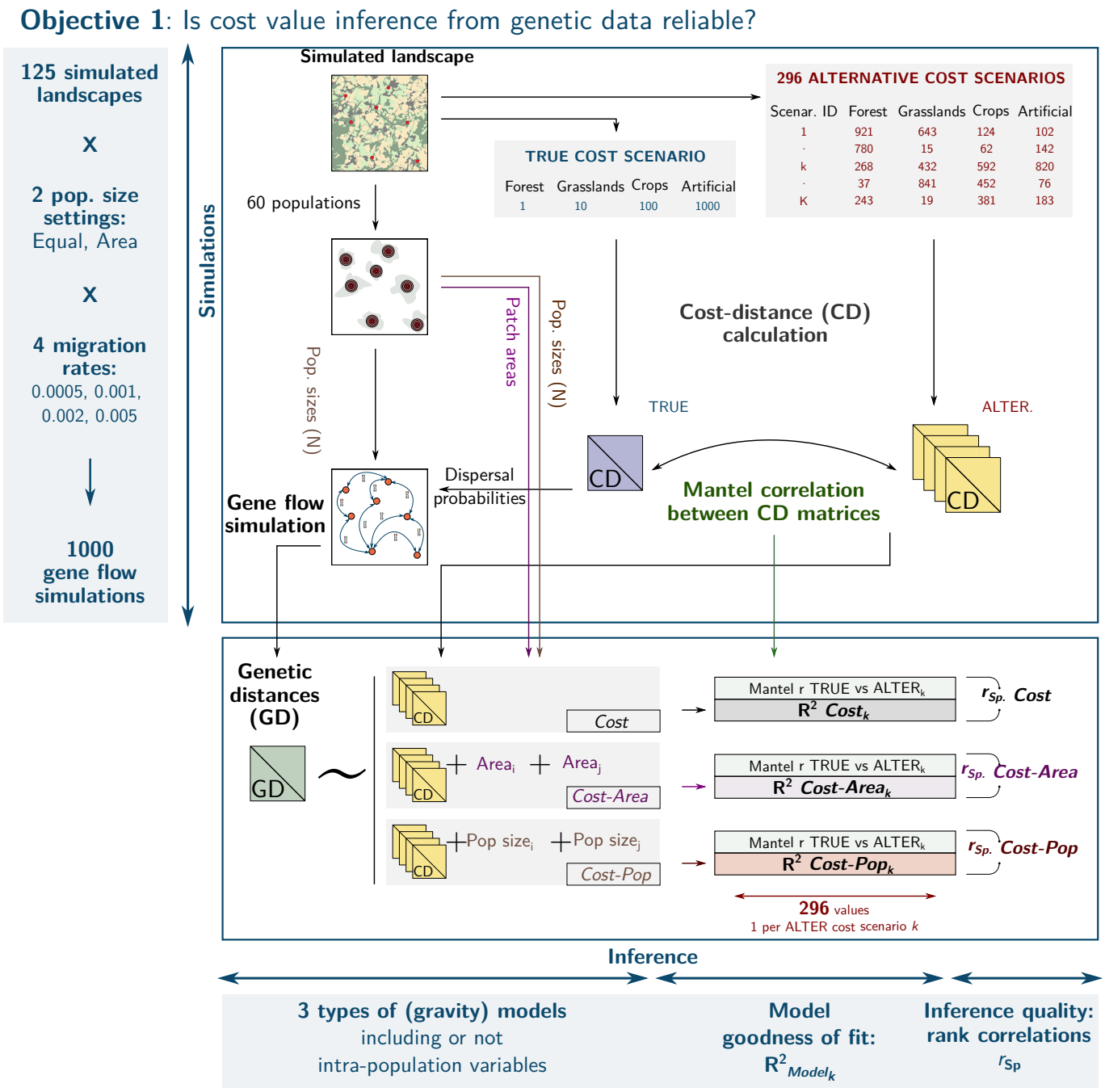

Figure 1: Overall methodology of the gene flow simulations performed for assessing the ability of several types of models to identify the 'true' cost scenario driving gene flow in a set of alternative cost scenarios diverging more or less from this 'true' scenario (Objective 1).

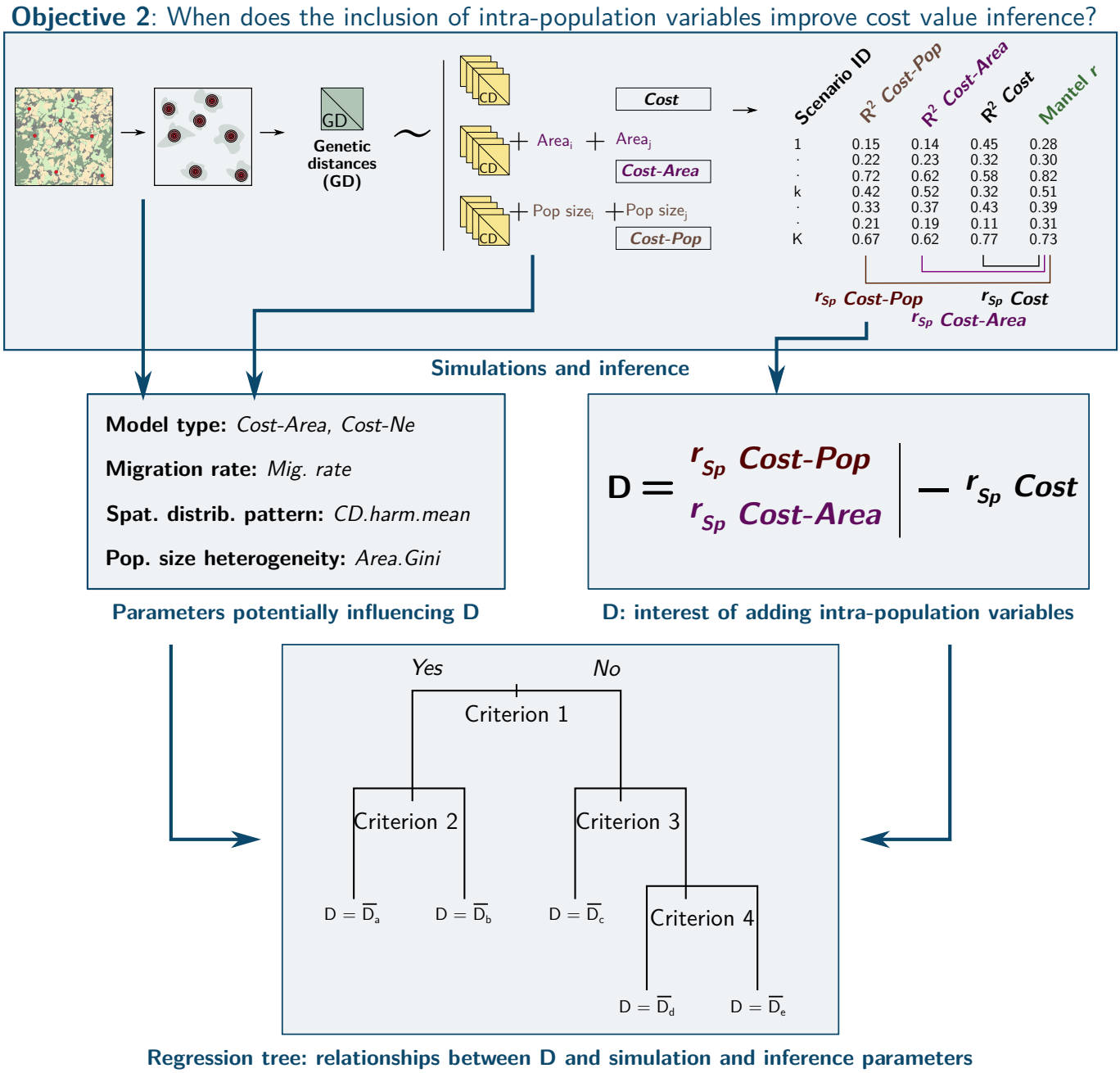

Figure 2: Overall methodology implemented for identifying the situations where the inclusion of intra-population variables in (gravity) models improves the cost value inference (Objective 2). Spat. distrib. pattern: spatial distribution pattern of the simulated population, *CD.harm.mean*: harmonic mean of the cost-distances between populations, *Area.Gini*: Gini index computed from the capacity of the habitat patches occupied by a population. In every leaf of the regression tree delineated from binary criteria applied to the predictor variables, D values are equal to their mean value over the observations included in the leaf ($\bar{D}_a$ for leaf a as an example).

(Khimoun et al., 2017; Pérez-Espona et al., 2008; Ruiz-González et al., 2014; Wang et al., 2008). The resulting landscapes were raster grids of 60 × 60 km with a cell resolution of 100 m.

The spatial heterogeneity of population sizes has been shown to explain a significant share of genetic differentiation variation (Prunier et al., 2017) and one of the aims of our study was to assess whether this heterogeneity could influence cost value inference. Besides, this spatial heterogeneity is partly dependent upon the spatial distribution of the habitat patches as patch area is often positively related with population sizes. Accordingly, we simulated population size distributions with varying degrees of heterogeneity (measured using the Gini inequality index, see below) and matched their heterogeneity with that of habitat patch areas in the simulated landscapes (Figures S1 and S2). We maintained the total population size constant in all cases and we also carried out the simulations without population size variation for comparison purposes.

For that purpose, we randomly sampled 60 cells within the forest patches of every landscape, separated by a distance larger than 1000 m. Habitat patches occupied by every population consisted of a buffer made of the forest pixels located around each sampled cell at a distance less than 500 m (Supporting information, Figure S1). This allowed us to vary the area of the habitat patches according to the landscape structure. The 60 discrete populations (demes) were located in the 60 sampled cells and each was considered to be panmictic at that scale.

We distinguished two spatial distributions of population sizes, totalling 3300 individuals in every case:

1. Equal population sizes (Equal): all 60 populations contained 55 individuals.
2. Area-dependent population sizes (Area): population sizes were spatially heterogeneous. They ranged from 10 to 100 individuals, depending on the area of the habitat patch occupied by the population (rank-correlation = 1).

The 'Equal' setting constituted the reference baseline making it possible to assess the effect of population size heterogeneity on the quality of the inference by comparison with the 'Area' setting (objective 1). For the 'Area' setting, we aimed at covering a wide gradient of population size heterogeneity while mimicking realistic conditions in which population sizes are driven by patch areas. For that purpose, we randomly generated series of 60 values between 10 and 100 making a total of 3300 ± 50. We then classified these series according to their heterogeneity using the Gini inequality index (Gini, 1912)(Supporting information, Figure S2). In parallel, we computed the Gini index describing the heterogeneity of the sampled patch areas in every landscape. We then associated each landscape/population distribution with the population size distribution corresponding to a degree of heterogeneity equivalent to that of patch area. Largest population sizes were therefore assigned to populations located in the largest patches. We then calculated the accumulated cost along the least-cost path between each pair of populations, thereafter referred to as cost-distance (CD).

Apart from simulating a large range of population size and patch area heterogeneity patterns, we also ensured that we simulated diverse patterns of population spatial distributions in order to test for the influence of the population spatial pattern on the inference. Indeed, the presence of clumps of populations exchanging more frequently between themselves than with other populations could influence our capacity to infer landscape resistance to gene flow (Graves et al., 2013). To test for this potential influence, we measured the degree of spatial aggregation of the populations with the harmonic mean of the whole set of CD values between populations.

This index reflects the frequency of small CD values which should favour short distance dispersal as a consequence of population spatial aggregation.

2.2.2 Alternative cost-scenarios

Identifying situations in which landscape genetic models are able to identify the 'true' cost-value scenario among alternative cost-scenarios requires that the CD values resulting from all these cost scenarios are not too highly correlated so that they can somehow be distinguished (Cushman et al., 2013). Rayfield et al. (2010) have shown that least-cost paths were sensitive to the range of relative cost values. Accordingly, alternative CD distributions resulted from 296 randomly generated alternative cost-scenarios. We used the approach of Shirk et al. (2010) to set alternative cost values using the following function:

$$C_i = (\frac{Rank_i}{Rank_{max}})^x \times C_{max}$$

where C_i is the cost value between 1 and C_{max} associated with the i-th land cover type. $Rank_i$ is the rank of the land cover type i between 1 and $Rank_{max} = 4$. Because the maximum cost value provides insight into the contrast between the most and least favourable land cover type for species dispersal, we used C_{max} values equal to 100, 1000 (maximum value of the 'true' cost scenario) and 10,000. We used x values equal to 1, 2, 4, 8 or 16. We therefore obtained 5 series of values for every maximum cost value. Using each of them, we randomly assigned cost values to the four land cover types and randomly selected 296 alternative cost scenarios among these combinations. We then used these cost scenarios to compute the 296 alternative CD distributions in every landscape and we computed the Mantel r correlation coefficient between each alternative CD matrix and the 'true' one (Figure 1). This setting provided us with alternative cost-scenarios covering a gradient of similarity with the 'true' cost-scenario.

2.2.3 Gene flow simulation

We used the CDPOP software program (Landguth and Cushman, 2010) for simulating gene flow and individual allelic state resulting from it. Population sizes and sex-ratio (equal to 1) remained constant throughout the simulations, which lasted for 500 generations to ensure that the equilibrium genetic differentiation pattern had been reached. At each generation, individuals mated in their own population and juveniles could disperse for establishing themselves in other populations. The number of offspring per female followed a Poisson distribution ($\lambda = 3$). Once every population was occupied by a number of individuals equal to its specific size, remaining individuals died. Generations were non-overlapping and mating was done with replacement for males only. Individual genotype was simulated for 20 independent loci (no linkage disequilibrium) with 30 alleles per locus because high allelic richness is known to limit the risk of size homoplasy (Estoup et al., 1995, 2002). Initial genotypes were randomly created at generation 0 by assigning each individual two alleles randomly chosen among the 30 alleles for the 20 loci. There was no selection pressure but mutations could occur (k-alleles mutation model, $\mu = 0.0005$).

According to the concept of dispersal kernel, dispersal probability decreased quickly as inter-population CD_{ij} between populations i and j increased, even if long distance dispersal remained possible (Clobert et al., 2012). Therefore, the dispersal probability between populations i and j was proportional to p_{ij}, which was computed as the negative exponential of the cost-distance CD_{ij}, such that $p_{ij} = e^{-\beta CD_{ij}}$, following Urban and Keitt (2001). β values were calculated such that the CD for which the dispersal probability was equal to 0.01 was equivalent to 1000 cost units, imposing the simulated species constant dispersal limitations over the range of cases.

Prunier et al. (2017) showed that the contribution of population size heterogeneity to the spatial pattern of genetic differentiation also depended upon the migration rate and we therefore carried out these simulations with migration rates equal to 0.0005, 0.001, 0.002 and 0.005 to identify the influence of this parameter on the cost value inference and on the effect of population size heterogeneity. Preliminary analyses showed that migration rates above 0.005 (i.e., > 17 dispersal events per generation at the landscape scale, cf. Results below) led to situations in which gene flow was too strong for heterogeneous drift effects to influence the inference whereas migration rates below 0.0005 (i.e., < 2 dispersal events/generation) led to situations in which drift effects were too strong for inference to be possible, whatever the landscape and population parameters. At every generation, once a fraction of individuals equal to the migration rate was selected in a given population, they dispersed to other populations with probabilities depending on the cost-distances to the other populations, as specified above, rescaled so that they sum to 1. In total, after 125 landscapes were selected (cf. Results), 1000 simulations were performed (125 landscapes $\times$ 2 distributions $\times$ 4 migration rates).

After the simulations, we used population genotypes at generation 500 to compute the pairwise D_{PS} between populations, i.e. the population-based version of a genetic distance equal to 1 - the proportion of shared alleles (Bowcock et al., 1994). This genetic distance has been shown to reflect well landscape resistance influence on genetic differentiation patterns in previous simulation analyses using similar settings (Savary et al., 2021b).

2.3 Gravity models

Gravity models have been initially used in geography and economics (Anderson, 1979; Fotheringham and O'Kelly, 1989; Schneider et al., 1998) to model various types of spatial interactions. Their application in ecology (Bossenbroek et al., 2001, 2007; Ferrari et al., 2006; Kong et al., 2010; Xia et al., 2004) and in landscape genetics (DiLeo et al., 2014; Moran-Lopez et al., 2016; Murphy et al., 2010; Robertson et al., 2018; Watts et al., 2015; Zero et al., 2017) is more recent. They model spatial interactions or fluxes as a function of both the variables characterising the objects involved in the interaction and of the distance between them (masses and distance in Newton's gravity theory, respectively). In landscape genetics, these models have mostly been used to infer the joint effect of local environmental variables and inter-population distances on the different stages of dispersal events leading to gene flow (e.g., Murphy et al. (2010)). In contrast, we here used them to consider both CD values between populations and population sizes as predictor variables in models explaining genetic differentiation. Indeed, genetic differentiation results from (i) dispersal movements leading to gene flow, supposed to be reflected by CD values, and (ii) genetic drift, which occurs at the population level and depends on population sizes. Our approach is novel in that it considered gravity model predictors not only as drivers of the stages of dispersal depending on local patch conditions, but more importantly as drivers of genetic drift intensity (alike predatory fish presence in Murphy et al. (2010)). We hypothesized that these models could improve how we infer cost values from genetic distances reflecting differentiation, by explicitly considering the two genetic processes acting at the "node" and "link" levels (i.e., drift and gene flow). We used gravity models to model the genetic distance G_{ij} between populations i and j (response variable, link-level) as a function of several predictors computed at two levels:

- At the population-level (node level):
  - Habitat patch areas (a_i, a_j)
  - Population sizes (N_i, N_j)

- Between populations (link-level):
  - Cost-distance CD_{ijk} between populations i and j in the cost scenario k

We computed three types of models of the following form in order to assess the quality of cost value inferences (objective 1) and to identify cases in which gravity models improve it when population sizes are heterogeneous (objective 2)(Figures 1 and 2):

'Cost' model: $G_{ij} \sim CD_{ijk}^{m}$

'Cost-Area' model: $G_{ij} \sim CD_{ijk}^{m} \times a_i^n \times a_j^o$

'Cost-Pop' model: $G_{ij} \sim CD_{ijk}^{m} \times N_i^p \times N_j^q$

with m, n, o, p and q being constant. We computed these three models using the CD values obtained with every 'true' or alternative cost scenario. The 'Cost' model is not a gravity model as it does not include any local variable. It was used as a reference for comparing the performance of gravity models with distance-based models commonly used for cost value inference or optimization (Peterman, 2018; Shirk et al., 2017). The 'Cost-Area' and 'Cost-Pop' models include local variables and allowed us to test for the relevance of using gravity models when patch areas or population sizes can be obtained, respectively (objective 2). Because gravity models have a multiplicative form, a natural log was applied to these formulas to obtain the classical formula of a multiple regression model whose parameters (m, n, o, p and q in our case) can be estimated. To account for the non-independence inherent to distance matrices, we used constrained linear mixed effect models by adding a random effect corresponding to the identity of the populations (MLPE models, Clarke et al. (2002)).

2.4 Assessment of model performance

We assessed the quality of the cost inference in the different situations and identified the situations in which the models including intra-population variable improved this inference (Figure 1). From these results, we aimed at deriving general guidelines for cost value inference in landscape genetics.

We first assessed the goodness of fit of the models using Edward's R^2_β (Edwards et al., 2008), which is a reliable model selection criterion when fitting mixed models with residual maximum likelihood estimation (Van Strien et al., 2012). Under our settings, if a given type of model ('Cost', 'Cost-Area' or 'Cost-Pop') performs well in distinguishing among cost scenarios, R^2_β values obtained by including alternative CD values in these models should reflect the correlation degree of every alternative CD matrix with the 'true' CD matrix driving the simulation. In contrast, a given type of model leads to lower quality inferences if the ranks of the models obtained with every alternative cost scenarios according to R^2_β values and their correlation with the 'true' cost-scenario are independent. Therefore, for quantifying the performance of the different types of models, we computed the Spearman rank correlation coefficient r_{Sp} between the series of R^2_β values obtained for each cost scenario and the Mantel r correlation coefficient measuring the similarity of each cost scenario to the 'true' one. In a given case, we expected the difference D between the r_{Sp} value associated with the 'Cost-Area' or 'Cost-Pop' models and the r_{Sp} value associated with the 'Cost' model to take positive values if the inclusion of intra-population variables in the model improves the cost-value inference (Figure 2).

Finally, we assessed the influence of every simulation parameter on the additional performance of the models including intra-population variables, measured by D values, using regression trees built with the CART algorithm (Breiman et al., 1984)(Figure 2). This method involves splitting the predictor space into a limited number of regions called leaves in which the response variable is predicted to take its mean value within the leaf (James et al., 2013). It can take both continuous and categorical predictor variables. Apart from performing better than linear models (e.g. ANOVA) in our case due to non-linear relationships, it provided us with a decision tree showing situations in which including intra-population variables helps identifying cost values. For that purpose, the response variable was D and we used the migration rate, the type of models, the Gini index of patch areas and the harmonic mean of the 'true' CD between populations as predictor variables. Regression trees were pruned with a criterion ensuring that at least 40 landscape and population configurations were included in every leaf, to prevent from overfitting. This minimal sample size allowed us to perform a one-side Student test to test for significant positive values of D in each leaf.

We carried out our analyses in R using `graph4lg` package (Savary et al., 2021c) to sample populations, compute cost-distances, genetic distances and patch areas, `nlme` (Pinheiro et al., 2013), `lme4` (Bates et al., 2007) and `r2glmm` (Jaeger, 2017) packages to fit gravity models and assess their goodness of fit and `rpart` package (Therneau et al., 2010) to fit regression trees.

3 Results

3.1 Simulation results

Overall, the landscape simulation settings allowed us to vary the degree of patch area heterogeneity and population spatial aggregation. We selected 125 landscapes maximising their contrasts in order to test for their respective influence on the inference. The Gini indices measuring the contrasts in population size distributions in the 'Area' setting ranged from 0.170 to 0.290 (median: 0.232). The spatial aggregation of populations also varied substantially with harmonic means of 'true' CD values ranging from 170 to 430 CD units (median: 297). Besides, these simulated landscapes were sufficiently heterogeneous for the CD matrices derived from alternative cost scenarios to exhibit a wide range of correlations with the true CD matrix, with Mantel correlation coefficients between the 'true' CD matrix and the alternative ones ranging from -0.350 to 0.999 (median: 0.628).

During the gene flow simulations, the mean number of migrants between the 60 populations per generation was equal to 2.0, 3.7, 7.1 and 17.3 with migration rates of 0.0005, 0.001, 0.002 and 0.005, respectively. Individuals could potentially disperse between 1770 population pairs ($\frac{60\times(60-1)}{2}$) with a probability depending on cost-distances. However, lower migration rates made long distance dispersal events less likely, and the number of different dispersal paths actually followed across 50 generations averaged 84 ($\pm$ 15), 145 ($\pm$ 16), 240 ($\pm$ 26) and 423 ($\pm$ 68) with migration rates of 0.0005, 0.001, 0.002 and 0.005 respectively. Accordingly, the respective influence on genetic differentiation of gene flow relative to genetic drift, as well as the number of long distance dispersal events substantially increased with migration rates.

3.2 Gravity models

R^2_β values obtained for a given simulation and model with different CD values exhibited large variations (coefficients of variation ranging from 0.49 to 0.78), meaning that the models were able to distinguish cost scenarios among them (Table 1). For all models, population size

heterogeneity settings (Equal, Area) and cost scenarios, the lowest model goodness of fit were obtained for the lowest migration rate (0.0005) and the largest values for the highest migration rate (0.005), relatively to the set of tested rates. Although the median R^2_β values over the 125 landscapes were always larger when including the 'true' CD values in the models rather than the alternative CD values, we observed the opposite trend when considering the maximum R^2_β values (Table 1). This means that in every case, the model with the best goodness of fit was computed with CD values not obtained from the 'true' cost scenario driving dispersal in our simulation. The few alternative scenarios responsible for these results differed from the 'true' cost scenario by their absolute cost values, by how the different land cover types were ordered according to these values or by both criteria (e.g. $[1, 4, 1000, 101]$, $[1, 40, 1002, 10000]$, $[1, 40, 10000, 1002]$, $[40, 625, 10000, 3165]$ instead of $[1, 10, 100, 1000]$). They often assigned low values to forest and grasslands but tended to assign a higher cost to crops than to artificial areas.

Pop. sizes	Mig. rate	Model	Median R^2_β		Max R^2_β	
			TRUE	ALTER	TRUE	ALTER
Equal	0.0005	Cost	0.053	0.036	0.117	0.525
Equal	0.0005	Cost-Area	0.055	0.039	0.117	0.547
Equal	0.001	Cost	0.147	0.107	0.294	0.463
Equal	0.001	Cost-Area	0.151	0.111	0.294	0.475
Equal	0.002	Cost	0.337	0.227	0.519	0.584
Equal	0.002	Cost-Area	0.344	0.233	0.521	0.590
Equal	0.005	Cost	0.558	0.380	0.774	0.776
Equal	0.005	Cost-Area	0.562	0.388	0.774	0.776
Area	0.0005	Cost	0.087	0.063	0.210	0.432
Area	0.0005	Cost-Area	0.100	0.074	0.239	0.439
Area	0.0005	Cost-Pop	0.099	0.073	0.236	0.442
Area	0.001	Cost	0.182	0.133	0.344	0.446
Area	0.001	Cost-Area	0.194	0.142	0.367	0.458
Area	0.001	Cost-Pop	0.193	0.142	0.372	0.471
Area	0.002	Cost	0.334	0.234	0.546	0.543
Area	0.002	Cost-Area	0.345	0.242	0.548	0.545
Area	0.002	Cost-Pop	0.345	0.242	0.547	0.545
Area	0.005	Cost	0.539	0.366	0.732	0.735
Area	0.005	Cost-Area	0.554	0.372	0.734	0.736
Area	0.005	Cost-Pop	0.554	0.373	0.733	0.737

Table 1: Goodness of fit of the gravity models as measured with R^2_β according to the heterogeneity of population size settings (Equal, Area), the migration rate (0.0005, 0.001, 0.002, 0.005), the variables included in the models (Cost, Cost-Pop, Cost-Area) and the cost scenarios corresponding to the CD values included in the models. TRUE means that the models include the 'true' CD values driving the simulations whereas ALTER means that the models include the alternative CD values. For the columns corresponding to the ALTER case, median and maximum reported values were computed for each line from 37000 values (125 landscapes $\times$ 296 scenarios), whereas for the columns corresponding to the TRUE case, they were each computed from 125 values.

The Spearman rank correlation coefficients r_{Sp} between the R^2_β values obtained using alternative CD values in the model and the correlation coefficients between the 'true' CD values and these alternative CD values took large values, with mean values ranging from 0.782 to 0.944 (Table 2). This means that models with the best goodness of fit were obtained when considering cost scenarios similar to the 'true' cost scenario. Therefore, the models performed well in inferring cost values and it was true even in cases where overall R^2_β values were low (Tables 1 and 2).

When population sizes were heterogeneous and depended on patch area, especially when

Pop. sizes	Mig. rate	Model	Min. r_{Sp}	Median r_{Sp}	Mean r_{Sp}	Max r_{Sp}
Equal	0.0005	Cost	-0.526	0.838	0.782	0.980
Equal	0.0005	Cost-Area	-0.606	0.848	0.790	0.982
Equal	0.001	Cost	-0.265	0.910	0.874	0.983
Equal	0.001	Cost-Area	-0.392	0.919	0.878	0.984
Equal	0.002	Cost	0.259	0.938	0.915	0.988
Equal	0.002	Cost-Area	0.186	0.943	0.917	0.984
Equal	0.005	Cost	0.628	0.957	0.944	0.989
Equal	0.005	Cost-Area	0.646	0.961	0.944	0.988
Area	0.0005	Cost	-0.452	0.816	0.727	0.971
Area	0.0005	Cost-Area	-0.397	0.872	0.798	0.974
Area	0.0005	Cost-Pop	-0.431	0.871	0.797	0.974
Area	0.001	Cost	-0.046	0.889	0.837	0.981
Area	0.001	Cost-Area	-0.149	0.907	0.866	0.982
Area	0.001	Cost-Pop	-0.227	0.910	0.865	0.982
Area	0.002	Cost	0.118	0.932	0.904	0.986
Area	0.002	Cost-Area	0.021	0.943	0.914	0.986
Area	0.002	Cost-Pop	-0.027	0.940	0.912	0.986
Area	0.005	Cost	0.656	0.960	0.943	0.989
Area	0.005	Cost-Area	0.661	0.960	0.944	0.988
Area	0.005	Cost-Pop	0.666	0.960	0.944	0.988

Table 2: Spearman rank correlation coefficients (r_{Sp}) between the R^2_β of the models and the correlation coefficients between the 'true' CD values and each alternative CD values, according to the heterogeneity of population size settings (Equal, Area), the migration rate (0.0005, 0.001, 0.002, 0.005) and the variables included in the models. The larger the r_{Sp} values, the better the models are able to identify the cost scenarios most similar to the 'true' one. Minimum (Min.), Median, Mean and Maximum (Max.) r_{Sp} values were each computed from 125 values.

the migration rate was low (0.0005 or 0.001), r_{Sp} values were more variable (Table 2). They often took slightly lower values than when using similar migration rates and models ('Cost', 'Cost-Area') but with equal population sizes, meaning that population size heterogeneity had overall a negative influence on the reliability of cost value inference in these cases. Besides, the differences between r_{Sp} values obtained with the 'Cost' model and either the 'Cost-Area' or 'Cost-Pop' models in the 'Area' case were larger with the lowest migration rates (Table 2). In particular, although maximum r_{Sp} values obtained using the three types of models ('Cost', 'Cost-Area', 'Cost-Pop') were relatively similar, respective median and mean r_{Sp} values were more different ; the values obtained with the 'Cost-Area' and 'Cost-Pop' models being larger than the values obtained with the 'Cost' model (Table 2). This means that although in some landscapes, including intra-population variables provided a very slight advantage, there were some landscapes in which it improved the quality of the inference more significantly. In the next section we therefore focus on the results obtained with the two lowest migration rates to explain the differences of model performance in some landscapes with a regression tree considering landscape characteristics.

3.3 Regression trees

When population sizes depended on patch areas ('Area'), the difference of performance D between models including CD values only ('Cost') and gravity models including both CD values and intra-population variables such as patch areas or population sizes ('Cost-Area', 'Cost-Pop') averaged 0.050 overall when the migration rate was either 0.0005 or 0.001 (Figure 3) and ranged from -0.330 to 0.590 (see figures S3 and S4 for similar results when considering all the migra-

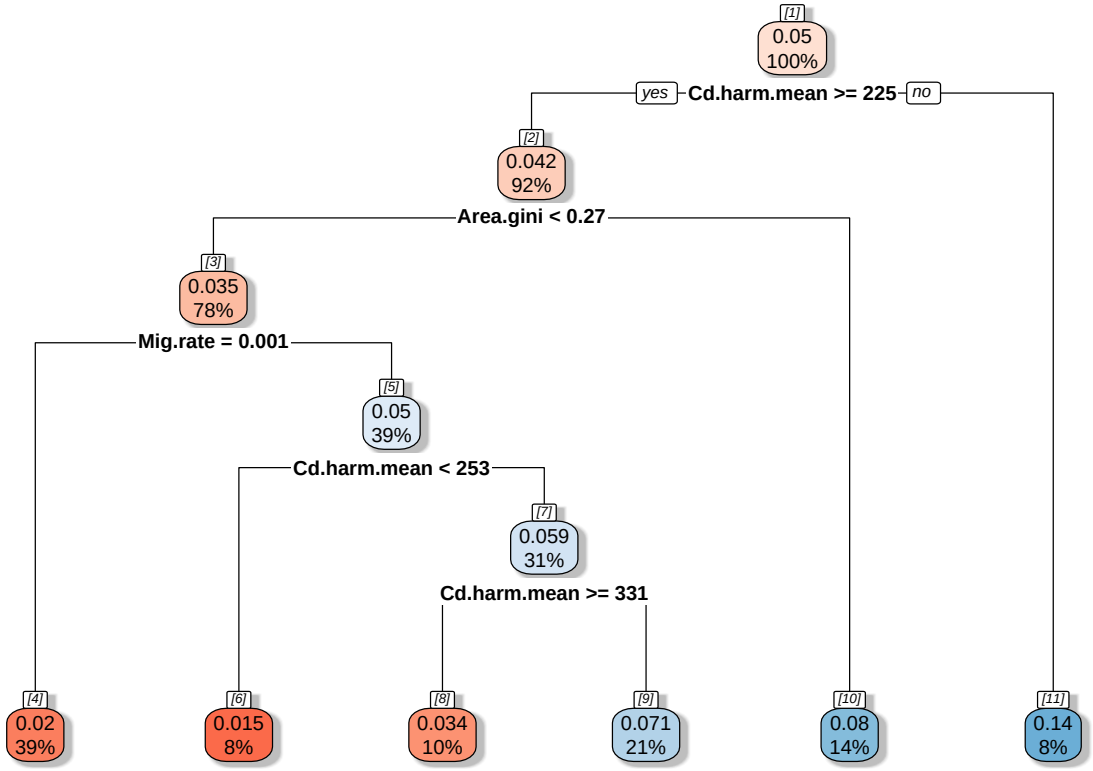

Figure 3: Regression tree obtained when considering four predictors (Model, Mig.rate, CD.harm.mean, Area.gini) to explain the response variable D. Only the cases corresponding to the scenario in which population size and habitat patch areas are rank-correlated (i.e., 'Area' setting), and migration rates are equal to 0.0005 or 0.001, are considered. The tree was pruned in order to have at least 40 observations in every leaf. The total number of observations is 500. The numbers in the boxes refer to the mean values of D for each leaf of the tree. The percentages refer to the proportions of the 500 observations included in each leaf.

tion rates). The best pruned regression tree explaining D contained migration rate, CD value harmonic mean and patch area Gini index as predictor variables but did not include the type of model. There were indeed negligible differences of D values between the 'Cost-Area' and 'Cost-Pop' models (Table 2). This regression tree explained 86 % of the variation in D and was made of six leaves corresponding to different regions of the predictor space (Figures 3 and 4). Values of D were significantly different from 0 in five of these leaves and positively in all five cases (one-side Student tests, $\alpha = 0.05$, with Bonferroni p-value adjustments)(Figure 4).

According to the splitting rules of the regression tree (Figure 3), when small CD values were frequent (Cd.harm.mean < 225), adding intra-population variables improved cost value inference as D values reached an average of 0.140 in these cases. The second splitting rule evidenced that the advantage provided by the inclusion of intra-population variables in cost value inference was larger when the patch areas were the most heterogeneous. Indeed, when the Gini index was larger than 0.27, D values averaged 0.080 whereas they were halved for lower degrees of patch area heterogeneity. In the latter case, mean D values were equal to 0.020 and 0.050 for migration rates equal to 0.001 and 0.0005, respectively (Figure 3), meaning that gravity models improved more substantially the inference when migration rate was the lowest. Then, when the migration rate was equal to 0.0005, although cases in which the harmonic mean of CD values was lower than 253 led to small D values (0.015), cases in which this index was between 253 and 331 led to much larger values (0.071)(Figure 3). Besides, when this harmonic mean was larger

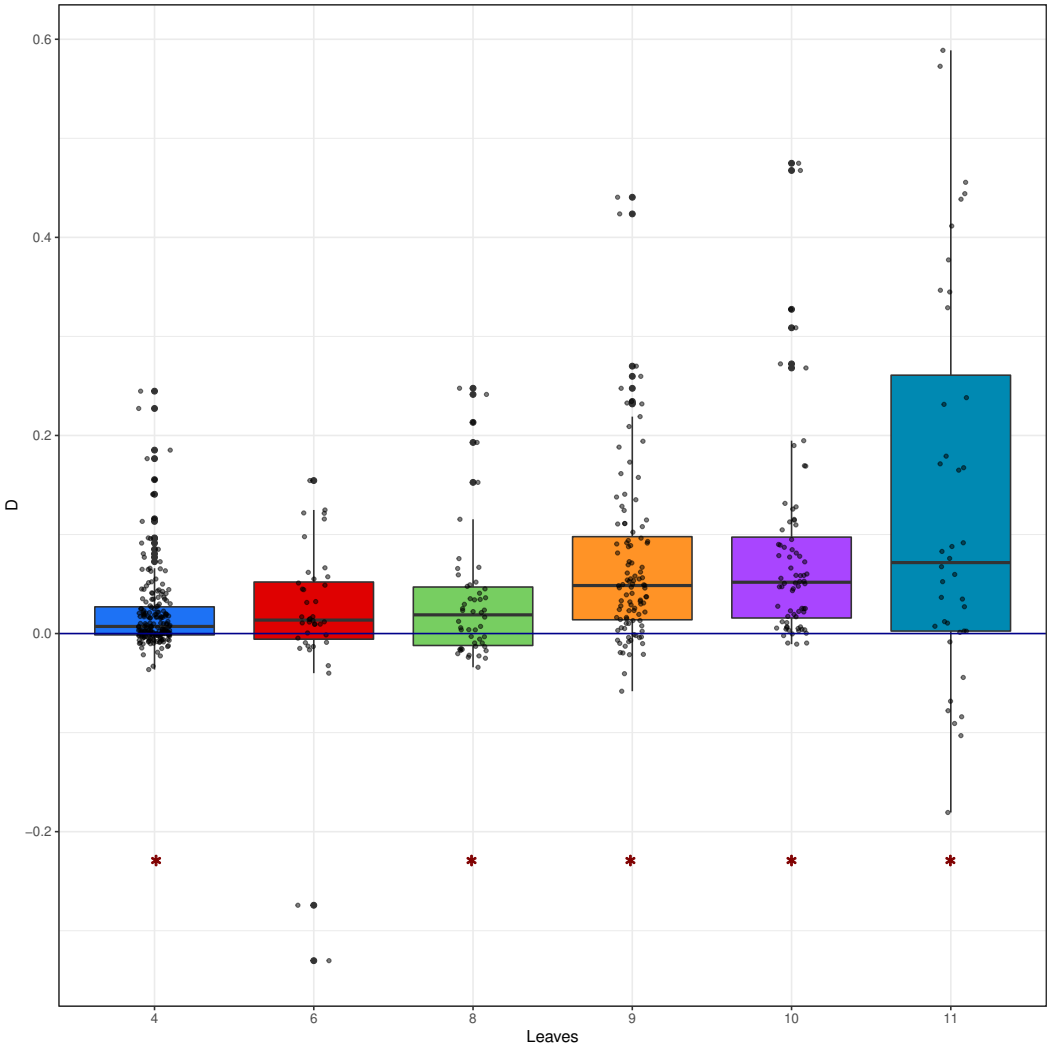

Figure 4: Distribution of D in each leaf of the regression tree displayed on figure 3(refer to this figure for the leaf numbers). Only the cases corresponding to the scenario in which population size and habitat patch areas are rank-correlated (i.e., 'Area' setting), and migration rates are equal to 0.0005 or 0.001, are considered. The red stars indicates that the one-side Student test testing whether the mean was different from 0 was significant (α =0.05).

than 331, D values averaged 0.034, meaning that the inclusion of intra-population variables seemed to improve cost-value inference in several cases where the populations were spatially aggregated according to this index. Note that the interpretation of the regression tree obtained when considering all the migration rates was similar, although the first splitting rules separated cases corresponding to the two largest migration rates (Figures S3 and S4).

4 Discussion

4.1 Is cost value inference from genetic data reliable?

Overall, the models performed well in inferring cost values, which confirms the relevance of genetic data for inferring cost values, as suggested by Beier et al. (2008) and empirically validated by Zeller et al. (2018). Models with the lowest goodness of fit (low R^2_β values) and capacity to identify the most realistic cost scenarios were obtained when migration rates were

the lowest (relatively high r_{Sp} values in comparison). This may stem from the stronger influence of drift relative to gene flow on genetic differentiation in these cases (Hutchison and Templeton, 1999). Yet, even with these low migration rates, the different models still performed relatively well in ranking several cost value scenarios according to their similarity with the 'true' cost scenario (high median and mean r_{Sp} values). This means that even when the ratios between the signal due to gene flow effect and the noise due to random drift effect are low, using genetic data makes it possible to rank different scenarios of landscape resistance to gene flow in a reliable way.

Nonetheless, the 'alternative' cost scenarios leading to the CD values most correlated to the 'true' CD values were often identified as the 'best' ones according to a model goodness-of-fit criterion because they provided a better explanatory power of genetic distances than did the 'true' CD values. These scenarios were different from the 'true' scenario in both their absolute cost values and the relative ranking of these values. This calls for caution when inferring cost values on the sole basis of the goodness-of-fit of models linking genetic distances and CD values, especially when using optimization methods to identify a unique best cost scenario within a set of potentially highly correlated scenarios (e.g., Peterman (2018)). Such an erroneous output could lead to wrong conservation measures for several reasons. First, if the relative ranking of the inferred cost values is not reflecting the actual dispersal behaviour of the study species, it could lead to wrongly consider some landscape features as easy to cross or, conversely, as landscape barriers. Second, if the inferred cost values are used for the spatial modelling of potential dispersal paths although they do not accurately reflect dispersal paths, it could decrease the effectiveness of conservation measures. Indeed, it has been shown by Savary et al. (2021a) that two closely correlated cost-distance values can correspond to least-cost paths that are spatially distinct. For instance, this study showed that the Mantel correlation coefficient between two cost-distance matrices can be up to 0.9, while corresponding least-cost paths only overlap on less than 20 % of their length.

However, retaining a set of several cost scenarios resulting in high values of goodness-of-fit and deriving a set of least-cost paths from them could be a way to account for the uncertainty of the inference when competing cost distances matrices are highly correlated (Rayfield et al., 2010). Besides, in our study, landscapes were simulated so that cost-distance matrices obtained with different cost scenarios ranged widely in terms of similarity, from being positively to negatively correlated. Yet, we can expect this variability to be lower in many landscapes and for many sampling designs, which could compromise the reliability of the inference, even in situations supposed to be optimal based on the goodness-of-fit of the best model. Similarly, Cushman et al. (2013) and Graves et al. (2013) observed that when different cost scenarios lead to highly correlated cost-distance matrices, cost value inference is more difficult. Therefore, it would be first useful to consider the landscape properties responsible for the similarity of cost scenarios, such as land cover diversity or patch configuration (see Savary et al. (2021a)), in order to select study areas in which cost-distances vary substantially. Assessing the correlations between cost-distance matrices deriving from different cost scenarios prior to the inference would also be a way to determine which precision can be reliably expected from the results. When the inference might not be reliable because some CD matrices obtained from different cost scenarios are highly correlated, our results indicate that the relative ranking of a large set of scenarios could overall be reliable, whereas the qualitative interpretation of the unique 'best' scenario and its use for locating dispersal paths should be avoided.

4.2 Do the population size heterogeneity, spatial distribution of populations and migration rates influence the inference?

On the one hand, in accordance with our hypothesis, population size heterogeneity tended to lessen the quality of cost value inference from genetic data when migration rates were the lowest (i.e., 5×10^{-4}, 10×10^{-4}, corresponding to approximately 2 and 4 dispersal events per generation at the landscape scale, respectively). This inference relies upon the assumption that gene flow is sufficient relative to local genetic drift so that genetic differentiation reflects landscape influence on gene flow at the scale of the study area (Savary et al., 2021b). Accordingly, the inference is complicated by the fact that genetic drift adds random noise to the gene flow signal in genetic differentiation, especially when migration rates are low and populations are small (Frankham, 1996). We here found evidence that an additional difficulty arises when population sizes are spatially heterogeneous because this random noise is not homogeneously distributed, which makes it even more difficult to infer landscape resistance to gene flow from genetic differentiation.

On the other hand, also consistent with our prediction, the spatial aggregation of the populations tended to affect the quality of cost value inference. We used the harmonic mean of CD values for distinguishing landscapes in which dispersal events frequently occurred at a restricted scale because populations tended to form spatial aggregates. When population sizes are heterogeneous and dependent upon habitat patch areas, the spatial aggregation of populations in the most favourable areas of the landscapes increases the frequency of dispersal events between neighbour populations of large sizes. This could in turn increase their genetic differentiation from both i) other small and isolated populations and ii) large populations from other 'clusters' of populations. It then makes it more difficult to relate the overall genetic differentiation pattern with landscape matrix resistance. The latter point had been suggested by Graves et al. (2012) and Graves et al. (2013) but was not specifically investigated in the context of cost value inference.

Both population size heterogeneity and spatial aggregation are parameters directly related with the amount and configuration of the habitat and with the spatial distribution of the populations in this habitat. They influenced significantly the cost value inference. This means that independently from the study species and its specific migration rate, landscape structure and in particular habitat spatial distribution are parameters to consider when planning a study aiming at inferring landscape influence on gene flow, as pointed out by Cushman et al. (2013).

We acknowledge that the consideration of our results in empirical landscape genetic analyses is limited by the fact that the variables included in the regression tree (Gini index, CD harmonic mean) hardly allow us to get a representation of similar real situations. However, if the inference of cost values from genetic data can possibly be carried out in several areas, our results indicate that the area where (i) population sizes are expected to be the least heterogeneous and (ii) where the spatial distribution of the sampled populations is the most regular should be chosen. In such situations, using a modelling approach including proxy variables for the local intensity of genetic drift (e.g., gravity models in our study, among other possibilities) would not substantially affect the inference. Besides, although we simulated gene flow in a virtual forest species, these guidelines should apply for every study species forming discrete populations and dispersing in heterogeneous landscapes.

4.3 When should intra-population variables be included in models for cost value inference?

Our second objective was to assess the interest of including intra-population variables for taking population size heterogeneity into account in the cost value inference. For that purpose, we fit gravity models in which both CD values and intra-population variables such as patch areas ('Cost-Area') and population sizes ('Cost-Pop') were predictor variables explaining pairwise genetic distances between populations. In cases where population size heterogeneity influenced cost value inference, the inclusion of intra-population variables in these models improved the quality of the cost value inference (positive D values) in accordance with our hypothesis; although slightly. Our results therefore extend to the specific context of cost value inference the recommendation of Prunier et al. (2017) to account for population size heterogeneity in landscape genetic analyses.

The interest of including intra-population variables in gravity models for inferring cost values was not only dependent upon the migration rates and the heterogeneity of population sizes, it also depended upon the degree of this population size heterogeneity and of the spatial aggregation pattern of the populations. On the one hand, D values were larger in cases where patch areas, and related population sizes, were most heterogeneous according to the Gini index of inequality. This result is similar to that of Prunier et al. (2017), who showed that population size heterogeneity significantly explains genetic differentiation patterns provided a substantial heterogeneity and low migration rates. Yet, these authors quantified overall population size heterogeneity using the coefficient of variation of these sizes, instead of the Gini index here used.

In addition, we showed that landscape variables computed from the habitat spatial pattern at the population level could improve the cost value inference when included in gravity models. Indeed, considering either patch area or population sizes in the gravity models led to similar results in cases where including intra-population variables improved cost value inference. These two variables were rank-correlated but not directly proportional in our settings. This situation is likely to be met in most real cases when patch area drives their carrying capacity and subsequently their population size. Thus, including environmental proxies for population size in gravity models could improve cost value inference in many situations. This result reinforces that of Prunier et al. (2017) which used river width and home-range sizes as environmental proxies for gudgeon (*Gobio occitaniae*) population sizes and this way estimated a significant share of population size heterogeneity effects on genetic differentiation. It also means that costly estimations of population sizes through field works could be saved when there is a close relationship between some environmental variables and population sizes.

4.4 Limits and perspectives

The migration rates for which we observed a significant influence of population size heterogeneity on cost value inference were rather low. However, they reflect realistic situations given that inferred genetic migration rates are often much lower than inter-patch movement rates (0.5 % versus 7-32 % respectively in Riley et al. (2006) study) and very low migration rates have often been inferred from genetic data (Meirmans, 2014). In addition, this result is consistent with that of Prunier et al. (2017) even if we here used migration rates in the lower end of the migration rates these authors used. However, our results show that intra-population variables help inferring cost values when gene flow is very reduced but they do not mean that population size heterogeneity is not substantially affecting genetic differentiation for larger migration rates.

In our simulations, we considered that we knew and sampled the exhaustive set of pop-

ulations. In practice, exhaustive sampling is rarely possible, although strongly recommended (Van Strien, 2017), and we can wonder to what extent our results would be affected by considering only a subset of the populations. Yet, when sampling is not exhaustive, gravity models could reveal helpful for predicting genetic distance between non-sampled populations provided they have a high goodness of fit and can be reliably extrapolated. Given the gain in performance provided by these models in certain situations, this would make these models relevant tools for deriving predictions in landscape genetics. However, in most situations where adding intra-population variables may be interesting for predicting more reliably genetic differentiation, drift effects are very strong and may generate high variability in genetic differentiation, thereby making predictions highly variable and potentially imprecise. The predictive use of these models thus deserves further investigation and would probably be more relevant when intra-population variables reflect processes affecting both population size and dispersal (Pflüger and Balkenhol, 2014; Watts et al., 2015). In this context, landscape graphs, which are commonly used for modelling connectivity and include both patch (node) and potential dispersal paths (links) characteristics, would be an adequate tool as their structure directly provides the inputs of gravity models. Besides, some variables influencing cost value inference such as patch area heterogeneity or patch spatial aggregation could be computed from different landscape graphs before hand as a way to identify contexts in which inference would be most reliable when the effects of population size heterogeneity are at play. Finally, although we simulated dispersal and gene flow in a forest specialist species, our results apply to a large range of species having patchy distributions. Future research aiming at improving cost value inference should be carried out for a set of taxa and landscapes reflecting the diversity of landscape genetic analyses.

5 Conclusion

Considering matrix heterogeneity when inferring landscape resistance to gene flow has been a common feature of many landscape genetic studies. In contrast, they have rarely considered the simultaneous influence of migration rates, population size heterogeneity and population spatial aggregation on this type of inference. Here, we showed that cost value inference from genetic data is reliable in a wide range of conditions but is hampered when migration is very restricted, population size is heterogeneous and populations are not regularly distributed in the landscape. Our study further demonstrates the interest that intra-population variables, such as population sizes or their proxies, represent for genetic differentiation analyses. It extends it to the context of cost value inference and shows that gravity models can be helpful for the inclusion of these variables in the inference of cost values associated with land cover types.

6 Acknowledgements

We thank Marie-Josée Fortin for her constructive comments all along this project. This study was part of a PhD project supported by the ARP-Astrance company under a CIFRE contract supervised and partly funded by the ANRT (Association Nationale de la Recherche et de la Technologie). This work is also part of the project CANON that was supported by the French "Investissements d'Avenir" program, project ISITE-BFC (contract ANR-15-IDEX-0003). We are particularly grateful to ARP-Astrance team for its constant support along the project. Simulations and analyses were carried out on the calculation "Mésocentre" facilities of the University of Bourgogne-Franche-Comté.

References

Adriaensen, F., Chardon, J., De Blust, G., Swinnen, E., Villalba, S., Gulinck, H., and Matthysen, E. (2003). The application of least-cost modelling as a functional landscape model. *Landscape and Urban Planning*, 64(4):233–247.

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *The American Economic Review*, 69(1):106–116.

Baguette, M., Blanchet, S., Legrand, D., Stevens, V. M., and Turlure, C. (2013). Individual dispersal, landscape connectivity and ecological networks. *Biological Reviews*, 88(2):310–326.

Bates, D., Sarkar, D., Bates, M. D., and Matrix, L. (2007). The lme4 package. *R package version*, 2(1):74.

Beier, P., Majka, D. R., and Spencer, W. D. (2008). Forks in the road: choices in procedures for designing wildland linkages. *Conservation Biology*, 22(4):836–851.

Bonte, D., Van Dyck, H., Bullock, J. M., Coulon, A., Delgado, M., Gibbs, M., Lehouck, V., Matthysen, E., Mustin, K., Saastamoinen, M., et al. (2012). Costs of dispersal. *Biological Reviews*, 87(2):290–312.

Bossenbroek, J. M., Johnson, L. E., Peters, B., and Lodge, D. M. (2007). Forecasting the expansion of zebra mussels in the United States. *Conservation Biology*, 21(3):800–810.

Bossenbroek, J. M., Kraft, C. E., and Nekola, J. C. (2001). Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes. *Ecological Applications*, 11(6):1778–1788.

Bowcock, A. M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R., and Cavalli-Sforza, L. L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Clarke, R. T., Rothery, P., and Raybould, A. F. (2002). Confidence limits for regression relationships between distance matrices: estimating gene flow with distance. *Journal of agricultural biological and environmental statistics*, 7(3):361–372.

Clobert, J., Baguette, M., Benton, T. G., and Bullock, J. M. (2012). *Dispersal ecology and evolution*. Oxford University Press.

Cushman, S. A., McKelvey, K. S., Hayden, J., and Schwartz, M. K. (2006). Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. *The American Naturalist*, 168(4):486–499.

Cushman, S. A., Shirk, A. J., and Landguth, E. L. (2013). Landscape genetics and limiting factors. *Conservation Genetics*, 14(2):263–274.

DiLeo, M. F., Siu, J. C., Rhodes, M. K., López-Villalobos, A., Redwine, A., Ksiazek, K., and Dyer, R. J. (2014). The gravity of pollination: integrating at-site features into spatial analysis of contemporary pollen movement. *Molecular Ecology*, 23(16):3973–3982.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., and Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in medicine*, 27(29):6137–6157.

Estoup, A., Jarne, P., and Cornuet, J.-M. (2002). Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Molecular Ecology*, 11(9):1591–1604.

Estoup, A., Tailliez, C., Cornuet, J.-M., and Solignac, M. (1995). Size homoplasy and mutational processes of interrupted microsatellites in two bee species, *Apis mellifera* and *Bombus terrestris* (Apidae). *Molecular Biology and Evolution*, 12(6):1074–1084.

Ferrari, M. J., Bjørnstad, O. N., Partain, J. L., and Antonovics, J. (2006). A gravity model for the spread of a pollinator-borne plant pathogen. *The American Naturalist*, 168(3):294–303.

Fotheringham, A. S. and O'Kelly, M. E. (1989). *Spatial interaction models: formulations and applications*, volume 1. Kluwer Academic Publishers Dordrecht.

Frankham, R. (1996). Relationship of genetic variation to population size in wildlife. *Conservation Biology*, 10(6):1500–1508.

Frankham, R. (2005). Genetics and extinction. *Biological Conservation*, 126(2):131–140.

Frankham, R., Ballou, J. D., and Briscoe, D. A. (2004). *A primer of conservation genetics*. Cambridge University Press.

Gini, C. (1912). Variabilità e mutabilità. *Memorie di metodologica statistica*, 10.

Graves, T. A., Beier, P., and Royle, J. A. (2013). Current approaches using genetic distances produce poor estimates of landscape resistance to interindividual dispersal. *Molecular Ecology*, 22(15):3888–3903.

Graves, T. A., Wasserman, T. N., Ribeiro, M. C., Landguth, E. L., Spear, S. F., Balkenhol, N., Higgins, C. B., Fortin, M.-J., Cushman, S. A., and Waits, L. P. (2012). The influence of landscape characteristics and home-range size on the quantification of landscape-genetics relationships. *Landscape Ecology*, 27(2):253–266.

Guillot, G., Leblois, R., Coulon, A., and Frantz, A. C. (2009). Statistical methods in spatial genetics. *Molecular Ecology*, 18(23):4734–4756.

Gurrutxaga, M., Lozano, P. J., and del Barrio, G. (2010). GIS-based approach for incorporating the connectivity of ecological networks into regional planning. *Journal for Nature Conservation*, 18(4):318–326.

Hartl, D. L., Clark, A. G., and Clark, A. G. (1997). *Principles of population genetics*, volume 116. Sinauer associates Sunderland, MA.

Hutchison, D. W. and Templeton, A. R. (1999). Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of gene flow and drift on the distribution of genetic variability. *Evolution*, 53(6):1898–1914.

Jaeger, B. (2017). Package 'r2glmm'. *R package version*, 3429.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Keyghobadi, N. (2007). The genetic implications of habitat fragmentation for animals. *Canadian Journal of Zoology*, 85(10):1049–1064.

Khimoun, A., Peterman, W., Eraud, C., Faivre, B., Navarro, N., and Garnier, S. (2017). Landscape genetic analyses reveal fine-scale effects of forest fragmentation in an insular tropical bird. *Molecular Ecology*, 26(19):4906–4919.

Kong, F., Yin, H., Nakagoshi, N., and Zong, Y. (2010). Urban green space network development for biodiversity conservation: Identification based on graph theory and gravity modeling. *Landscape and Urban Planning*, 95(1-2):16–27.

Landguth, E. L. and Cushman, S. (2010). CDPOP: a spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, 10(1):156–161.

Manel, S., Schwartz, M. K., Luikart, G., and Taberlet, P. (2003). Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, 18(4):189–197.

McRae, B. H. (2006). Isolation by resistance. *Evolution*, 60(8):1551–1561.

Meirmans, P. G. (2014). Nonconvergence in Bayesian estimation of migration rates. *Molecular Ecology Resources*, 14(4):726–733.

Moran-Lopez, T., Robledo-Arnuncio, J., Diaz, M., Morales, J., Lazaro-Nogal, A., Lorenzo, Z., and Valladares, F. (2016). Determinants of functional connectivity of holm oak woodlands - fragment size and mouse foraging behavior. *Forest Ecology and Management*, 368:111–122.

Murphy, M. A., Dezzani, R., Pilliod, D. S., and Storfer, A. (2010). Landscape genetics of high mountain frog metapopulations. *Molecular Ecology*, 19(17):3634–3649.

Pérez-Espona, S., Pérez-Barbería, F., McLeod, J., Jiggins, C., Gordon, I., and Pemberton, J. (2008). Landscape features affect gene flow of Scottish Highland red deer (*Cervus elaphus*). *Molecular Ecology*, 17(4):981–996.

Peterman, W. E. (2018). ResistanceGA: An R package for the optimization of resistance surfaces using genetic algorithms. *Methods in Ecology and Evolution*, 9(6):1638–1647.

Pflüger, F. J. and Balkenhol, N. (2014). A plea for simultaneously considering matrix quality and local environmental conditions when analysing landscape impacts on effective dispersal. *Molecular Ecology*, 23(9):2146–2156.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Team, R. C., et al. (2013). nlme: Linear and nonlinear mixed effects models. *R package version*, 3(1):111.

Prunier, J. G., Dubut, V., Chikhi, L., and Blanchet, S. (2017). Contribution of spatial heterogeneity in effective population sizes to the variance in pairwise measures of genetic differentiation. *Methods in Ecology and Evolution*, 8(12):1866–1877.

Rayfield, B., Fortin, M.-J., and Fall, A. (2010). The sensitivity of least-cost habitat graphs to relative cost surface values. *Landscape Ecology*, 25(4):519–532.

Richardson, J. L., Brady, S. P., Wang, I. J., and Spear, S. F. (2016). Navigating the pitfalls and promise of landscape genetics. *Molecular Ecology*, 25(4):849–863.

Riley, S. P., Pollinger, J. P., Sauvajot, R. M., York, E. C., Bromley, C., Fuller, T. K., and Wayne, R. K. (2006). FAST-TRACK: A southern California freeway is a physical and social barrier to gene flow in carnivores. *Molecular Ecology*, 15(7):1733–1741.

Robertson, J. M., Murphy, M. A., Pearl, C. A., Adams, M. J., Páez-Vacas, M. I., Haig, S. M., Pilliod, D. S., Storfer, A., and Funk, W. C. (2018). Regional variation in drivers of connectivity for two frog species (*Rana pretiosa* and *R. luteiventris*) from the US Pacific Northwest. *Molecular Ecology*, 27(16):3242–3256.

Ruiz-González, A., Gurrutxaga, M., Cushman, S. A., Madeira, M. J., Randi, E., and Gómez-Moliner, B. J. (2014). Landscape genetics for the empirical assessment of resistance surfaces: the European pine marten (*Martes martes*) as a target-species of a regional ecological network. *PLoS ONE*, 9(10):e110552.

Savary, P., Foltête, J. C., and Garnier, S. (2021a). Cost distances and least cost paths respond differently to cost scenario variations: a sensitivity analysis of ecological connectivity modeling. *International Journal of Geographical Information Science*, pages 1–25.

Savary, P., Foltête, J.-C., Moal, H., Vuidel, G., and Garnier, S. (2021b). Analysing landscape effects on dispersal networks and gene flow with genetic graphs. *Molecular Ecology Resources*, 21(4):1167–1185.

Savary, P., Foltête, J.-C., Moal, H., Vuidel, G., and Garnier, S. (2021c). graph4lg: a package for constructing and analysing graphs for landscape genetics in R. *Methods in Ecology and Evolution*, 12(3):539–547.

Schadt, S., Knauer, F., Kaczensky, P., Revilla, E., Wiegand, T., and Trepl, L. (2002). Rule-based assessment of suitable habitat and patch connectivity for the Eurasian lynx. *Ecological Applications*, 12(5):1469–1483.

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., Strokorb, K., et al. (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, 63(8):1–25.

Schneider, D. W., Ellis, C. D., and Cummings, K. S. (1998). A transportation model assessment of the risk to native mussel communities from zebra mussel spread. *Conservation Biology*, 12(4):788–800.

Sciaini, M., Fritsch, M., Scherer, C., and Simpkins, C. E. (2018). NLMR and landscapetools: An integrated environment for simulating and modifying neutral landscape models in R. *Methods in Ecology and Evolution*, 9(1):2240–2248.

Serrouya, R., Paetkau, D., McLELLAN, B. N., Boutin, S., Campbell, M., and Jenkins, D. A. (2012). Population size and major valleys explain microsatellite variation better than taxonomic units for caribou in western Canada. *Molecular Ecology*, 21(11):2588–2601.

Shirk, A., Wallin, D., Cushman, S., Rice, C., and Warheit, K. (2010). Inferring landscape effects on gene flow: a new model selection framework. *Molecular Ecology*, 19(17):3603–3619.

Shirk, A. J., Landguth, E. L., and Cushman, S. A. (2017). A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Molecular Ecology Resources*, 18(1):55–67.

Slatkin, M. (1993). Isolation by distance in equilibrium and non-equilibrium populations. *Evolution*, 47(1):264–279.

Spielman, D., Brook, B. W., and Frankham, R. (2004). Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences*, 101(42):15261–15264.

Therneau, T. M., Atkinson, B., and Ripley, M. B. (2010). The rpart package.

Ueno, S., Tomaru, N., Yoshimaru, H., Manabe, T., and Yamamoto, S. (2000). Genetic structure of *Camellia japonica* L. in an old-growth evergreen forest, Tsushima, Japan. *Molecular Ecology*, 9(6):647–656.

Urban, D. and Keitt, T. (2001). Landscape connectivity: a graph-theoretic perspective. *Ecology*, 82(5):1205–1218.

Van Strien, M. J. (2017). Consequences of population topology for studying gene flow using link-based landscape genetic methods. *Ecology and Evolution*, 7(14):5070–5081.

Van Strien, M. J., Keller, D., and Holderegger, R. (2012). A new analytical approach to landscape genetic modelling: least-cost transect analysis and linear mixed models. *Molecular Ecology*, 21(16):4010–4023.

Wang, I. J. (2013). Examining the full effects of landscape heterogeneity on spatial genetic variation: a multiple matrix regression approach for quantifying geographic and ecological isolation. *Evolution*, 67(12):3403–3411.

Wang, I. J., Glor, R. E., and Losos, J. B. (2013). Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology Letters*, 16(2):175–182.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society B - Biological Sciences*, 360(1459):1395–1409.

Wang, Y.-H., Yang, K.-C., Bridgman, C. L., and Lin, L.-K. (2008). Habitat suitability modelling to correlate gene flow with landscape connectivity. *Landscape Ecology*, 23(8):989–1000.

Watts, A. G., Schlichting, P. E., Billerman, S. M., Jesmer, B. R., Micheletti, S., Fortin, M.-J., Funk, W. C., Hapeman, P., Muths, E., and Murphy, M. A. (2015). How spatio-temporal habitat connectivity affects amphibian genetic structure? *Frontiers in Genetics*, 6:275.

Weckworth, B. V., Musiani, M., DeCesare, N. J., McDevitt, A. D., Hebblewhite, M., and Mariani, S. (2013). Preferred habitat and effective population size drive landscape genetic patterns in an endangered species. *Proceedings of the Royal Society B*, 280(1769):20131756.

Xia, Y., Bjørnstad, O. N., and Grenfell, B. T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164(2):267–281.

Zeller, K. A., Jennings, M. K., Vickers, T. W., Ernest, H. B., Cushman, S. A., and Boyce, W. M. (2018). Are all data types and connectivity models created equal? validating common connectivity approaches with dispersal data. *Diversity and Distributions*, 24(7):868–879.

Zeller, K. A., McGarigal, K., and Whiteley, A. R. (2012). Estimating landscape resistance to movement: a review. *Landscape Ecology*, 27(6):777–797.

Zero, V. H., Barocas, A., Jochimsen, D. M., Pelletier, A., Giroux-Bougard, X., Trumbo, D. R., Castillo, J. A., Evans Mack, D., Linnell, M. A., Pigg, R. M., et al. (2017). Complementary network-based approaches for exploring genetic structure and functional connectivity in two vulnerable, endemic ground squirrels. *Frontiers in Genetics*, 8:81.

7 Data accessibility

No empirical data were used for this study. Simulated landscapes, R codes and final results are available online: https://dx.doi.org/10.6084/m9.figshare.21587130.

8 Author contributions

J.C.F., S.G. and H.M. designed the project and obtained the funding. P.S. and G.V. performed the simulations. P.S. designed the simulation study and analysed the data. P.S wrote the manuscript with significant contributions and remarks from all co-authors.