Thèse de doctorat de l'établissement
Université Bourgogne Franche-Comté

École Doctorale n° 554 Environnements - Santé

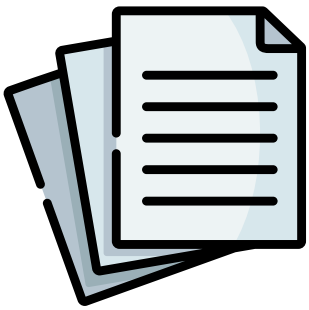
Spécialité : Biologie des populations et écologie

Par

Paul Savary

Utilisation conjointe de graphes génétiques et paysagers pour l'analyse de la connectivité écologique des habitats

Annexes A - Articles scientifiques



ARP-Astrance, 75008 Paris, France
UMR 6049 ThéMA, Université Bourgogne Franche-Comté - CNRS, 25000 Besançon, France
UMR 6282 Biogéosciences, Université Bourgogne Franche-Comté - CNRS, 21000 Dijon, France

Table des matières

Page

Table des matières **vi**

Annexe A1 Coupling landscape graph modeling and biological data : a review **1**

1 Introduction . . . 2
2 Data and methods . . . 3
2.1 The corpus of landscape graph applications . . . 3
2.2 Systematic description of the papers . . . 4
2.3 Statistical analysis of the corpus . . . 6
3 Results . . . 7
3.1 Are biological data specific to the geographical and biological framework ? . . . 7
3.2 Does the use of biological data depend on the objective of the graph analysis ? 8
3.3 The connectivity analyses . . . 10
4 Discussion . . . 12
5 Conclusion . . . 17

Annexe A2 Combining landscape and genetic graphs to address key issues in landscape genetics **19**

1 Introduction . . . 20
2 Comparing graphs to answer current landscape genetic questions . . . 23
2.1 Disentangling the complex relationships between the components of landscape structure and genetic structure . . . 24
2.2 Assessing scale effects in landscape genetics . . . 26
2.3 Identifying barriers in the landscape . . . 27
3 Integrating the graphs to benefit from their complementarity . . . 27
4 Further steps towards a better joint use of landscape and genetic graphs . . . 29

Annexe A3 Analysing landscape effects on dispersal networks and gene flow with genetic graphs **31**

1 Introduction . . . 32
2 Material & Methods . . . 34
2.1 Landscape data . . . 34
2.2 Gene flow simulation . . . 34
2.3 Genetic graphs . . . 35
2.4 Graphs analyses . . . 37
2.5 Simulation results ordination . . . 38
3 Results . . . 40

3.1 Simulation results . . . 40
3.2 Genetic graphs . . . 42
4 Discussion . . . 48
4.1 When and why to prune a genetic graph ? . . . 48
4.2 How to prune a genetic graph to identify the precise topology of the dispersal network ? . . . 49
4.3 How to prune a genetic graph to infer landscape resistance to dispersal ? . . . 50
4.4 Which genetic distance to use and how ? . . . 51
4.5 Limits and perspectives . . . 51
5 Acknowledgements . . . 54
6 Data accessibility . . . 54
7 Author contributions . . . 54
A - Supplementary figures . . . 55
B - Glossary of acronyms . . . 60
C - Mathematical background : independence graphs . . . 60
D - Computation of the DPS genetic distance . . . 65

Annexe A4 `graph4lg` : a package for constructing and analysing graphs for landscape genetics in R 67

1 Introduction . . . 68
2 Workflow . . . 69
2.1 Input data processing . . . 69
2.2 Genetic graph construction and analysis . . . 72
2.3 Landscape graph construction and analysis . . . 73
2.4 Landscape and genetic graph comparisons . . . 73
3 Export facilities and included data . . . 77
4 Limits and conclusion . . . 77
5 Acknowledgements . . . 77
6 Data availability . . . 77
7 Authors' contributions . . . 78
A - Supplementary tables . . . 79

Annexe A5 Assessing the influence of the amount of reachable habitat on genetic structure using landscape and genetic graphs 83

1 Introduction . . . 84
2 Material & Methods . . . 87
2.1 Study species and sampling area . . . 87
2.2 Genetic structure indices . . . 88
2.3 Habitat metric calculations . . . 89
2.4 Analyses of the relationship between habitat metrics and genetic structure indices 92
3 Results . . . 93
3.1 Landscape and genetic graphs . . . 93
3.2 Correlations between ARH metrics and genetic responses . . . 93
3.3 Partial Least Squares regressions . . . 94
4 Discussion . . . 98

4.1 Are ARH metrics relevant predictors of genetic structure ? . . . 99
4.2 Does the ARH influence genetic diversity and genetic differentiation to the same degree and at the same spatial scale ? . . . 99
4.3 Does the resistance of the matrix affect genetic diversity and genetic differentiation in the same way ? . . . 100
4.4 Limits and perspectives . . . 101
A - Supplementary figures . . . 103
B - Supplementary tables . . . 108
C - Rationale behind the use of the Q2 to analyse PLS regression results . . . 109

Annexe A6 Validating graph-based connectivity models for a forest tropical bird species using independent presence and genetic datasets 111
1 Introduction . . . 112
2 Material & Methods . . . 113
2.1 Study area . . . 114
2.2 Connectivity modeling . . . 114
2.3 Genetic data analysis . . . 119
2.4 Validation of landscape graph modeling . . . 120
3 Results . . . 121
3.1 Landscape graphs . . . 121
3.2 Genetic structure . . . 123
3.3 Cost scenario and graph link validation . . . 123
3.4 Metric validation . . . 125
4 Discussion . . . 127
4.1 Landscape graphs are empirically validated by genetic data . . . 127
4.2 Relationship between graph ecological relevance, data requirements and construction and analysis methods . . . 128
4.3 Implications for biodiversity conservation . . . 129
4.4 Limits and perspectives . . . 130
A - Supplementary tables . . . 131
B - Jacobs' specialization index . . . 133
C - Validation R-squared calculation . . . 133
D - Supplementary figures . . . 134

Annexe A7 Cost distances and least cost paths respond differently to cost scenario variations - A sensitivity analysis of ecological connectivity modeling 139
1 Introduction . . . 140
2 Methods . . . 142
2.1 Landscape sampling . . . 142
2.2 Cost scenario creation . . . 142
2.3 Least cost path modeling . . . 143
2.4 Spatial and distance-based comparisons of LCPs . . . 143
2.5 Landscape structure and cost scenario descriptors . . . 143
2.6 Statistical analyses of the drivers of LCP modeling sensitivity to cost scenarios 145
3 Results . . . 146

3.1 Structure of the sampled landscapes . . . 146
3.2 Relative influences of cost values and landscape structure on the sensitivity of LCP modeling outputs . . . 148
3.3 Landscape structure influence on the sensitivity of LCP modeling outputs . . . 148
3.4 Cost scenario characteristics influence on the sensitivity of LCP modeling outputs 150
3.5 Relationship between the spatial overlap of LCP and the correlation between CD matrices . . . 151
4 Discussion . . . 154
4.1 Sensitivity of LCP and CD to cost scenarios . . . 154
4.2 Implications for cost value inference and LCP modeling . . . 156
A - Supplementary figures . . . 159

Annexe A8 Inferring landscape resistance to gene flow using gravity models 161
1 Introduction . . . 162
2 Methods . . . 164
2.1 Overall methodological approach . . . 164
2.2 Simulations . . . 164
2.3 Gravity models . . . 168
2.4 Assessment of model performance . . . 168
3 Results . . . 169
3.1 Simulation results . . . 169
3.2 Gravity models . . . 170
3.3 Regression trees . . . 171
4 Discussion . . . 175
4.1 Is cost value inference from genetic data reliable ? . . . 175
4.2 Does SHNe influence cost value inference ? . . . 175
4.3 Can we improve cost value inference by taking SHNe into account in gravity models ? . . . 176
4.4 When should intra-population variables be included in gravity models for cost value inference ? . . . 176
4.5 Limits and perspectives . . . 177
5 Conclusion . . . 177
A - Supplementary figures . . . 179

Bibliographie 183

Annexe A1

Coupling landscape graph modeling and biological data : a review

Abstract

Context

Landscape graphs are widely used to model networks of habitat patches. As they require little input data, they are particularly suitable for supporting conservation decisions (and decisions about other issues as e.g. disease spread) taken by land planners. However, it may be problematic to use these methods in operational contexts without validating them with empirical data on species or communities.

Objectives

Since little is known about methodological alternatives for coupling landscape graphs with biological data, we have made an exhaustive review of these methods to analyze links between the main purposes of the studies, the way landscape graphs are constructed and used, the type of field data, and the way these data are integrated into the analysis.

Methods

We systematically describe a corpus of 71 scientific papers dealing with terrestrial species, with particular emphasis on methodological choices and contexts of the studies.

Results

Despite a great variability of types of biological data and coupling strategies, our analyses reveal a dichotomy according to the objective of the studies, between (i) approaches aimed at improving ecological knowledge, mainly based on land-cover maps and using biological data to test the influence of landscape connectivity on biological responses, and (ii) approaches with an operational aim, in which biological data are directly integrated into the graph construction and assuming a positive effect of connectivity.

Conclusions

Beyond these main contrasts, the review shows that landscape graphs can benefit from field data of different types at varying scales. The great variability of approaches adopted reveals the flexible nature of these tools.

Cet article a été publié dans *Landscape Ecology* en mars 2020 :

Foltête, J. C., Savary, P., Clauzel, C., Bourgeois, M., Girardet, X., Sahraoui, Y., Vuidel, G. & Garnier, S. 2020. Coupling landscape graph modeling and biological data : a review. *Landscape Ecology*, 35(5), 1035-1052

1 Introduction

Landscape connectivity modeling is a powerful tool for analyzing the movements of species living in fragmented habitats (Correa Ayram *et al.*, 2016 ; Zeller *et al.*, 2018). Connectivity models integrate information about species behavior and landscape structures (Cushman *et al.*, 2013a). However, designing realistic models of the ecological processes involved while providing tools likely to be used by landscape managers remains a major issue (Beier *et al.*, 2008). Even if data acquisition is difficult and local records of species are scarce (Pressey, 2004 ; Pe'er *et al.*, 2005), it is essential to integrate field data in landscape connectivity modeling (Crooks et Sanjayan, 2006 ; Kadoya, 2009).

Graph-theoretic approaches are considered both promising (Calabrese et Fagan, 2004 ; Minor et Urban, 2008 ; Kadoya, 2009 ; Urban *et al.*, 2009) and of uncertain ecological relevance (Moilanen, 2011) in landscape connectivity modeling. Among the spatial graphs used in ecology and conservation (Dale et Fortin, 2010 ; Fall *et al.*, 2007), landscape graphs (also named habitat networks) have emerged in the last 15 years, following the seminal paper of Urban et Keitt (2001). They are mainly used to model networks of discrete habitat patches in many geographical contexts and for numerous species. In these graphs, nodes usually represent habitat patches of the species under study while links represent potential movements (Galpern *et al.*, 2011 ; Urban *et al.*, 2009). As initially designed, landscape graphs are constructed from a landscape map defined in accordance with species' habitat requirements and movement abilities. They are an interesting compromise among other modeling approaches given the little input data needed and their capacity to represent ecological fluxes (Calabrese et Fagan, 2004). This makes them particularly suitable for supporting conservation decisions taken by land planners (Foltête *et al.*, 2014 ; Zetterberg *et al.*, 2010). However, it has been recognized that combining them with empirical data about species would improve their current implementation (Kadoya, 2009). Indeed, when landscape graphs are constructed from land cover maps without field data on species, their ability to represent ecological networks relies entirely on the assumption that the land-cover types identified as potential habitats or corridors are actually suitable for a species to settle in or disperse through. Since this assumption is not always confirmed (Clevenger *et al.*, 2002 ; Shirk *et al.*, 2010 ; Wasserman *et al.*, 2010), it may be problematic to use these methods to support decisions in operational contexts without considering empirical data on species (Cushman *et al.*, 2013a ; Beier *et al.*, 2008). This issue is even more acute when the outcomes of connectivity modeling lead to significant funds being committed to concrete operations of conservation, compensation, or restoration.

While many studies using landscape graphs are based on land-cover maps alone (e.g. Avon et Bergès (2016) ; Dondina *et al.* (2018) ; Martensen *et al.* (2017) ; Poor *et al.* (2019) ; Tannier *et al.* (2016)), others include field data on species. These data may be of various types (presence records, genetic data, movement monitoring, etc.), they may characterize several biological levels (populations, species, communities) and may be integrated at different stages of modeling (Correa Ayram *et al.*, 2016). For example, Estrada-Peña (2005) constructed a graph in which the nodes were defined from a species distribution model (SDM), directly using the presence records to calibrate the connectivity model. This approach, followed by other researchers since then, was recently summarized by Duflot *et al.* (2018). Another way of including field data has been experimented by O'Brien *et al.* (2006) who made use of telemetry data to calibrate the cost values assigned to the inter-patch links. In a quite different approach, other researchers have correlated local connectivity metrics and presence or abundance data, to investigate species' responses to habitat accessibility (Clauzel *et al.*, 2013 ; Foltête

et al., 2012b ; Ribeiro *et al.*, 2011). While the number of studies based on landscape graphs has increased in ecology and conservation (Correa Ayram *et al.*, 2016 ; Fletcher *et al.*, 2016), favored by the diffusion of open access specialized software applications (Csardi et Nepusz, 2006 ; Foltête *et al.*, 2012a ; Saura et Torne, 2009), it is becoming difficult to have a clear view of the methodological options improving the ecological relevance of landscape graphs. Some reviews have already been published about graph construction methods (Galpern *et al.*, 2011), connectivity metrics (Rayfield *et al.*, 2011), and types of operational applications (Bergsten et Zetterberg, 2013 ; Foltête *et al.*, 2014 ; Zetterberg *et al.*, 2010), but little is known about the combination of these spatial graphs with empirical data on species. Therefore, practitioners involved in landscape management as well as researchers working in ecology and conservation would benefit greatly from a comprehensive state-of-the-art review of the current implementation of such combinations.

In this paper, we propose a systematic review of the coupling of landscape graphs with field data on species (limited to terrestrial species). As connectivity analyses can be conducted in studies aimed at (i) improving our theoretical knowledge in ecology or alternatively at (ii) implementing operational approaches in landscape planning and management, we question the link between the main objective of the studies, the way landscape graphs are built and used, the type of field data and the way these data are integrated into the analysis. From a corpus of scientific papers, our aim is to take stock of the methods of coupling in relation to characteristics of the context of the studies. As we suspect these characteristics to be interrelated, we seek to identify the main rationales behind the use of field data, to finally define profiles of landscape graph applications. Our main hypothesis is that the purpose of the studies determines the way biological data are considered, thereby making a contrast between operational applications where landscape graphs are constructed in a simple, time-saving and cost-efficient way (i.e. nodes and links directly delineated from a land-cover map), and scientific applications where researchers integrate field data in complex ways to maximize the fit between model and reality.

2 Data and methods

2.1 The corpus of landscape graph applications

We used the online database Scopus to gather the scientific literature dealing with landscape graphs combined with field data on species. We restricted our search to scientific publications in English. As the terminology of landscape connectivity may vary among authors (Gippoliti et Battisti, 2017 ; Moilanen, 2011), we defined a final request after several tests, by evaluating the efficiency of each request by the percentage of inclusion of a checklist of 33 papers meeting our specifications. Four sets of criteria were defined and combined into a single request (see appendix 1) :

— The first criterion required the presence of "ecological networks" or equivalent terms in the title, the abstract, or the key-words : (landscape OR habitat OR ecological OR patch) AND (graph OR network).

— The second set of criteria was related to the use of field data on species. It was mainly represented by words expressing types of data or techniques of data acquisition in the title, the abstract, or the key-words : ("field data" OR population OR demographic OR occurrence OR abundance

OR presence OR richness OR suitability OR "radio-tracking" OR telemetry OR GPS OR CMR OR genetic OR roadkill).

— The third set of criteria required the papers to focus on terrestrial biodiversity and to remove the papers on hydrographic networks that are too specific to be mixed with non-aquatic networks. Papers whose title, abstract, or key-words include (maritime OR marine OR dendritic OR riverscape OR "river network" OR stream) were excluded.

— The fourth criterion was the presence of at least one key paper on landscape graphs in the references. Based on the number of citations given by Scopus, we listed six key papers cited more than 300 times : Urban et Keitt (2001), Saura et Pascual-Hortal (2007), Urban *et al.* (2009), Bunn *et al.* (2000), Minor et Urban (2008) and Pascual-Hortal et Saura (2006).

2.2 Systematic description of the papers

Each article was summarized using a systematic grid including first the type of biological data, corresponding to the following non-exclusive items : presence, abundance, species diversity, telemetry, genetic diversity, other.

Five topics were also described : (1) geographical and biological framework, (2) purpose of the study, (3) involvement of stakeholders, (4) graph construction, (5) connectivity analysis.

(1) The geographical framework was documented by the country and the region of the study. The biological characteristics were the species under consideration, its habitat, and the type of movement represented in the graph (daily movement, dispersal). To perform the statistical analyses, the geographical locations were clustered by continent, and the species were grouped into the following taxa : mammals, birds, amphibians, insects (or arachnids), chiropterans, plants and "other taxa" in other specific cases.

(2) From reading the title, the abstract, and the introduction, the purpose of the study was categorized into one or more of the following items :

— Understanding : the aim is to understand the link between a biological feature of the population, the species, or the community, and the functional connectivity modeled by the landscape graph.

— Prioritization : the aim is to identify key elements of the network likely to be protected or monitored.

— Impact evaluation : the aim is to assess the impact of a potential or actual landscape change on the network functionality.

— Network design : the aim is to define new components of the ecological network to improve its global connectivity.

— Method design : the aim is to improve methods of ecological network modeling.

The purpose of each article was also described by the global design of the analysis, concerning temporality and a posteriori analysis. Temporality (i.e. the way the temporal dimension was managed) included static, retrospective, and prospective approaches. Analyses linked biological data and connectivity metrics to each other as follows :

— No link between biological data and metrics in analyses.

— Biological data correlated with metrics, i.e. in a descriptive approach based on a visual or statistical investigation.

— Biological data modeled with metrics, i.e. used as the target variable in a statistical model in which metrics are considered as explanatory variables of the biological response.

(3) The level of involvement of stakeholders (e.g., land-planning practitioners) was documented by three non-exclusive possibilities corresponding to a growing order of involvement :

— Acknowledgment of stakeholders : the acknowledgments at the end of the paper mention a land-planning practitioner or a non-academic organization.

— Data from stakeholders : in the method section, the authors explicitly mention a non-academic organization (and non-national, i.e. not a national topographic service) as a data source.

— Stakeholders as co-authors : some of the authors of the papers come from a non-academic area.

(4) Items related to the graph construction itself concerned mainly the patch and link definitions. By investigating the method section of the articles, we listed three approaches concerning patches :

— Land-cover-based patch : the patches are designed from land-cover classes only.

— Suitability-based patch : they are defined from a suitability map or another SDM output.

— Protected areas : patches are a set of protected areas or zones designed by field experts.

In landscape graphs, each patch may have its own weight in the connectivity analysis. This weight is usually considered as a proxy of its demographic or carrying capacity (Urban et Keitt, 2001). We observed three possibilities :

— Uniform weight patch : all patches have the same weight.

— Area-based weighting : each patch is weighted by its area, which is a priori the most common option. In some specific cases, the patch is weighted by the area of a nearby resource patch (e.g. Tournant *et al.* (2013)).

— Suitability-based weighting : the patches are weighted by a statistical indicator (specifically sum or average) computed from the pixel values of a suitability map.

The links are also characterized by a weight which is most of the time the edge-to-edge distance separating the patches connected by these links. Focusing on the data used to compute these distances rather than on the computation details, we found three types of distance weighting :

— Euclidean link.

— Land-cover based link, resulting from least-cost distances (or resistance distances derived from circuit theory) where a cost value is assigned to each land-cover class.

— Suitability-based link, i.e. least-cost distances where the costs are derived from a suitability map, for instance the inverse of the presence probability.

(5) The connectivity analysis was first described by the level of analysis among the following non-exclusive possibilities :

— Network-level : the analysis deals with the entire graph's connectivity.

— Component-level : the analysis deals with the comparison of connectivity between components (i.e. sub-graphs resulting from link pruning) or between clusters, i.e. compartments resulting from a clustering method.

— Patch-level : the analysis is focused on the local connectivity computed for each patch.

— Link-level : the analysis is focused on the links' attributes (e.g. potential fluxes).

The graph analysis was then characterized by the type of connectivity metric computed from the graph. We listed eight metrics among the most used ones, plus a class "other" corresponding to more specific measures, and a class "no metric" when no measure was computed. The eight metrics are :

— Probability of Connectivity (PC), a global metric integrating both patch and link weights in a measure of spatial interaction (Saura et Pascual-Hortal, 2007).

— Integral Index of Connectivity (IIC), a global metric of spatial interaction integrating patch weight and a topological distance between patches (Pascual-Hortal et Saura, 2006).

— Delta Probability of Connectivity (dPC), a local metric used to identify the key patches, computed from the iterative removal of each patch with PC as the global reference (Saura et Pascual-Hortal, 2007). We also included in this item the decomposition of this metric into three fractions (Saura et Rubio, 2010).

— Delta Integration Index of Connectivity (dIIC), a local metric similar to dPC but using IIC as a global reference (Pascual-Hortal et Saura, 2006). As previously, the decomposition of dIIC into three fractions was included in this item.

— Betweenness Centrality Index (BC), applied with (Foltête *et al.*, 2012a) or without (Zetterberg *et al.*, 2010) weighting of the graph elements. It represents the theoretical level of local transit.

— Flux (F), also named Area Weighted Flux (Foltête *et al.*, 2012a), a local metric expressing the potential of dispersal from a given patch.

— Degree (Dg), a topological measure derived from the global framework of graph theory and equivalent to the number of links connected to a given patch.

— Expected Cluster Size (ECS), a landscape level metric corresponding to the area-weighted mean cluster size (O'Brien *et al.*, 2006).

2.3 Statistical analysis of the corpus

After the review stage, the topics of the previous grid were analyzed statistically by combining the related variables and the types of biological data. Because of the qualitative nature of all the variables, we performed multiple correspondence analyses (MCA). The principle of a MCA is to define orthogonal factors synthesizing the variance of a qualitative dataset (Tenenhaus et Young, 1985). All categories and all individuals (here the articles) are given coordinates in this multidimensional space. We considered only the first two factors. When strong relationships were found between biological data and the variables of a given topic, we expected biological data categories to be widely distributed across the factorial space. Otherwise, this would mean no significant link existed between biological data and the topic.

The link between biological data and the first topic (geographical and biological framework) was analyzed separately using a first MCA. The topics 2 (purpose of the study), 3 (involvement of stakeholders), and 4 (graph construction) were grouped because they form a consistent set to be compared with biological data. As this part of the analysis includes numerous variables, the second MCA was followed by Hierarchical Clustering (HC) applied using the Ward criterion to summarize information and identify the main types of articles. The resulting typology was mapped to investigate the possible geographical effect in the use of landscape graphs. Finally, the last topic (connectivity analysis) was investigated separately by means of a third MCA.

3 Results

The request implemented on 12 December 2018 returned 338 articles. As several methods of landscape connectivity modeling rely on a partially similar vocabulary (e.g. approaches focused on spatial genetics), many of the 338 articles were outside the scope of our investigation. After a systematic reading to finalize the selection, our corpus included 71 articles (see list in Appendix 2). The articles were published from 2005 to 2018, although a large majority (82 %) was published after 2012 (Fig. 1).

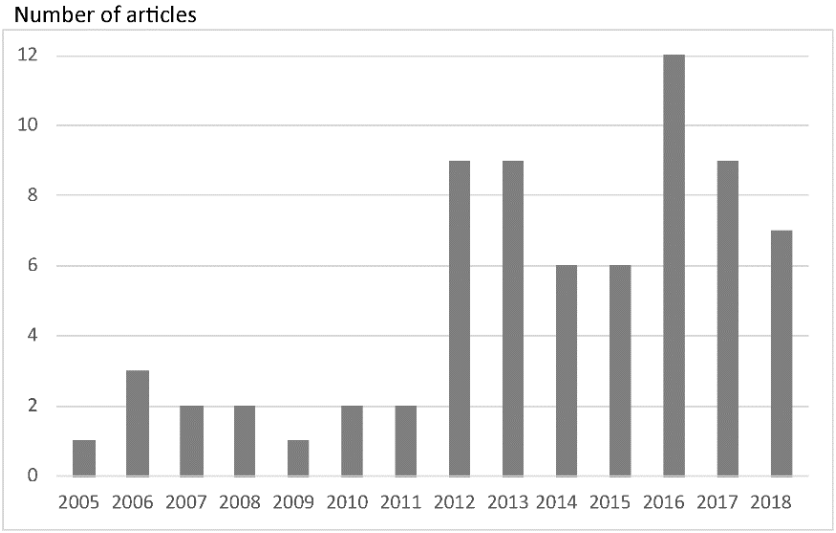


FIGURE 1 – Number of articles combining landscape graphs and biological data published every year from 2005 to 2018

Among the five types of data, presence data are used in 61 % of the articles, followed by abundance data (18 %), genetic diversity and GPS-track data (11 % for each type), and species diversity data (8 %). Other data amount to 7 % and include for example evidence of breeding, density, or roadkill data.

3.1 Are biological data specific to the geographical and biological framework ?

Studies are unequally distributed over the five continents, with 52.1 % of them in Europe, 23.9 % in North America, 12.7 % in South America, 11.3 % in Asia, and only 1.4 % in Africa. The two dominant taxa are mammals (35.2 %) and birds (23.9 %), the other main taxa being plants (16.9 %), insects (11.3 %), and amphibians (8.5 %). The studied habitats are mainly forests (57.7 %), followed by grasslands and agricultural habitats (14.4 %), and wetlands and aquatic habitats (11.3 %). A large proportion of the articles (26.8 %) dealt with more complex habitats such as hedgerows or urban areas (hereafter "other habitats"). The type of movement analyzed is mainly dispersal (87.3 %) and secondarily daily movements (12.7 %).

The MCA combining biological data and geographical and biological characteristics primarily highlights studies concerning wetlands and amphibians, without evidencing any connection with a particular type of biological data (Fig. 2). The location of the items "plants", "insects", and "grasslands" close to "species diversity", "genetic diversity", and "abundance" also suggests a link between them, in studies more frequently conducted in Europe. However, the investigation of cross frequencies (Appendix 3) shows that these associations are only partial. For example, genetic data never come

from grasslands in these studies but concern plants and insects in 37.5 % and 12.5 % of the cases. Over half of the abundance estimations concern insect populations. Almost all the species diversity measures concern plants (five out of six studies), a third of them in grasslands. The location of the other items in the factorial space suggests that studies using presence or telemetry data frequently concern forest species of mammals and birds. Among these former studies some deal with daily movements on other continents than Europe.

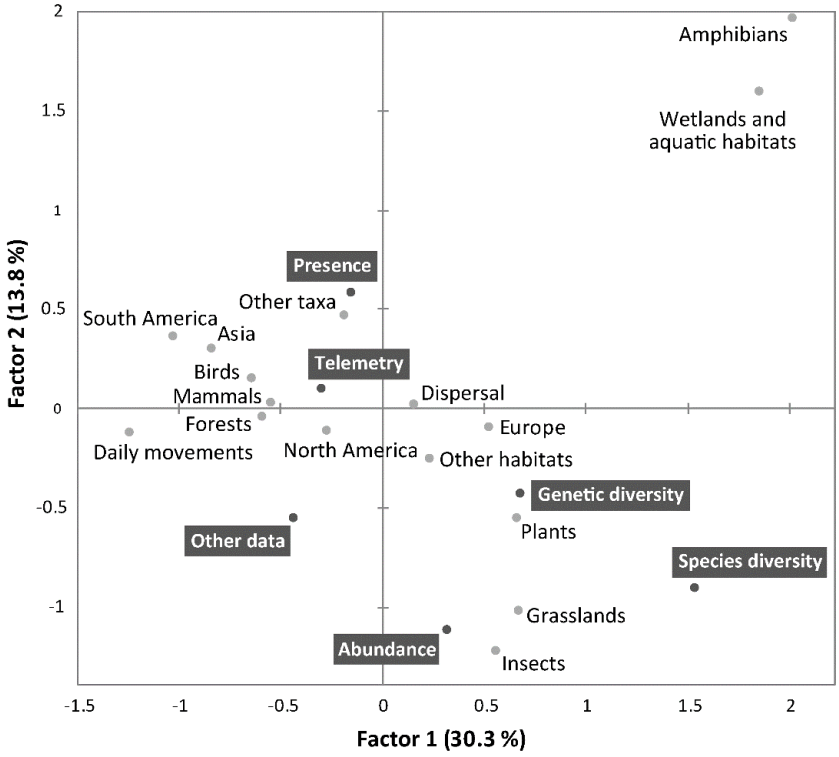


FIGURE 2 – Factorial space of the variables describing the geographical and biological framework of the articles. The terms in the grey rectangles represent the types of biological data. Note that the single study applied to Africa was removed to avoid a category associated with too low a frequency

3.2 Does the use of biological data depend on the objective of the graph analysis ?

The studies aimed at understanding the relationships between landscape connectivity and populations, species, or communities amount to not quite half of the corpus (46.5 %). The prioritization of elements or components of the ecological networks is present in 25.3 % of the articles. Then comes the impact evaluation (16.9 %) followed by the network design (12.7 %) and the methodological approaches (9.9 %). Most of the studies rely on a static approach (80.3 %) as only 12.7 % and 7.0 % of them are based on prospective and retrospective approaches respectively. Among the 71 articles, 57.7 % integrate stakeholders through acknowledgements and/or by the use of their data. Stakeholders are among the authors in only 16.9 % of the articles.

Globally, biological data are included in graph construction (i.e. before the connectivity analysis) in 45 articles (63.4 %). This integration more frequently concerns the definition of patches, delineated (36.6 %) and/or weighted (21.1 %) by using a suitability map or another SDM output. Biological data can also be used to define and weight the links (23.9 %), either by converting presence probabilities into a set of cost values (15.5 %) or less frequently following other approaches from telemetry data or genetic differentiation measures (8.4 %). Conversely, biological data are used after the graph analysis

in 42.2 % of the articles, correlated with connectivity metrics (14.1 %) and more frequently used as target variables in explanatory models (28.2 %).

The MCA applied to all these criteria results in a marked decrease in inertia from the first factor containing 46.9 % of variance (Fig. 3). Given the position of categories along the x-axis, the first factor contrasts (1) studies where biological data are integrated in graph construction, and (2) studies where biological data are a posteriori correlated or modeled with connectivity metrics. In the first case (left side of Fig. 3), more related to presence and other data, patches and links are preferably defined and weighted from suitability maps or other outcomes of SDM. This approach is often adopted when pursuing operational objectives such as prioritization and network design. In the second case (right side of Fig. 3) where species diversity and abundance data are preferably used, graph elements are more frequently defined from land-cover maps only. This approach tends to be implemented when the main purpose is to understand the relationships between the species' ecological response and the connectivity levels inferred from metrics. Beyond this main contrast, the second factorial axis provides complementary information by emphasizing a gradient of stakeholder involvements, that is more prominent in approaches of method design and using telemetry data and secondarily genetic data (high side of Fig. 3), but is less frequent in studies dealing with impact evaluation either retrospectively or prospectively (low side of Fig. 3).

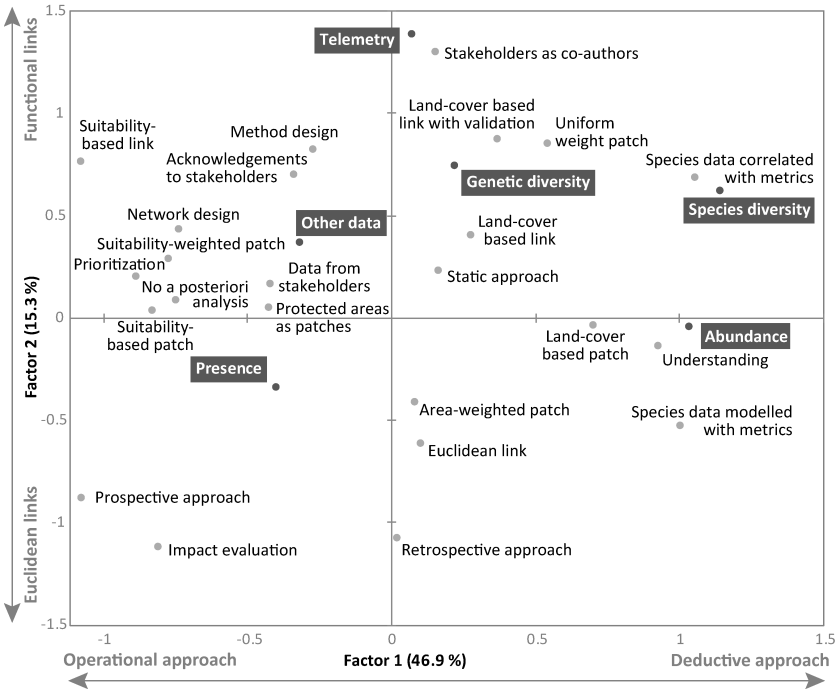


FIGURE 3 – Factorial space of the variables describing the aim of the articles, the graph applications, and the involvement of the stakeholders. The terms in the grey rectangles represent the types of biological data.

The application of the HC to the factorial axes provided the dendrogram we cut to define four classes (Fig. 4). These classes have relatively homogeneous frequencies and can be nested into a higher-level where classes 1 and 2 contrast with classes 3 and 4. According to their position in the factorial space and to the frequencies of each category (see Appendix 3), these classes can be summarized as follows :

— Class 1 "deductive approach, Euclidean links" : studies aimed at understanding the relationships between populations, species, or communities and habitat connectivity, by correlating biological data with metrics after the graph analysis. The data used are mainly presence (55 %), abundance (40 %), and/or species diversity (15 %). In these studies, patches are mainly defined from land-cover maps (90 %) whereas links are most often Euclidean (80 %). They rarely mention non-academic stakeholders.

— Class 2 "deductive approach, functional links" : studies directed at similar objectives as those from class 1 but using a larger panel of data types. Patches (85 %) as well as links (77 %) are mainly defined from land-cover maps. A strong link with stakeholders is mentioned in the papers (role in data acquisition and/or presence as co-author).

— Class 3 "operational approach, impact evaluation" : studies aimed at evaluating the impact of a past or future landscape change. Presence data are used to define the patches from an SDM (62 %) whereas links are mainly Euclidean (76 %). Few links with stakeholders are mentioned.

— Class 4 "operational approach, network design" : studies with an operational aim in which presence data are integrated in graph construction via an SDM to define patches (65 %) and to weight links (47 %). The connectivity analysis does not lead to a statistical investigation with biological data. An explicit link with stakeholders is frequent.

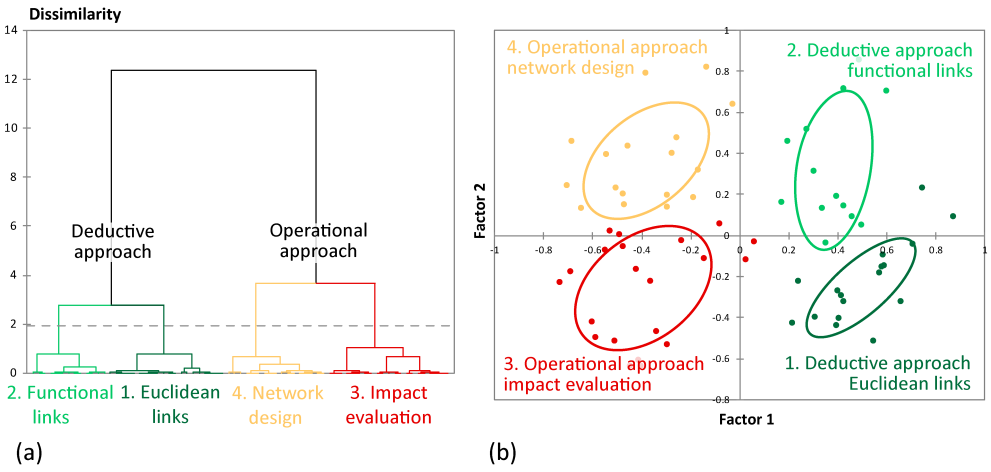


FIGURE 4 – Hierarchical Ascendant Classification applied to the first two factors of MCA. (a) The dotted line represents the cutoff defining four classes. (b) These classes are represented by individual positions and 50 % ellipses in the factorial space.

The map of field locations shows that the four classes are distributed worldwide (Fig. 5). But, deductive approaches tend to be more commonly adopted in Europe (specifically in Spain and France) than in North America.

3.3 The connectivity analyses

The connectivity analyses sometimes rely on a visual interpretation of maps without computation of connectivity metrics (eight articles). Overall, patch-level analyses are the most common (71.8 %), being much more frequent than global-level (25.4 %), component-level (15.5 %), and link-level analyses (11.3 %). The dominance of patch-level analyses is supported by the frequent occurrence of the metrics dPC (26.8 %), F (16.9 %), dIIC (15.5 %), BC (15.5 %), and Dg (9.9 %). Global-level metrics such as PC and IIC are used only eight (11.3 %) and seven (9.9 %) times respectively. The most frequent

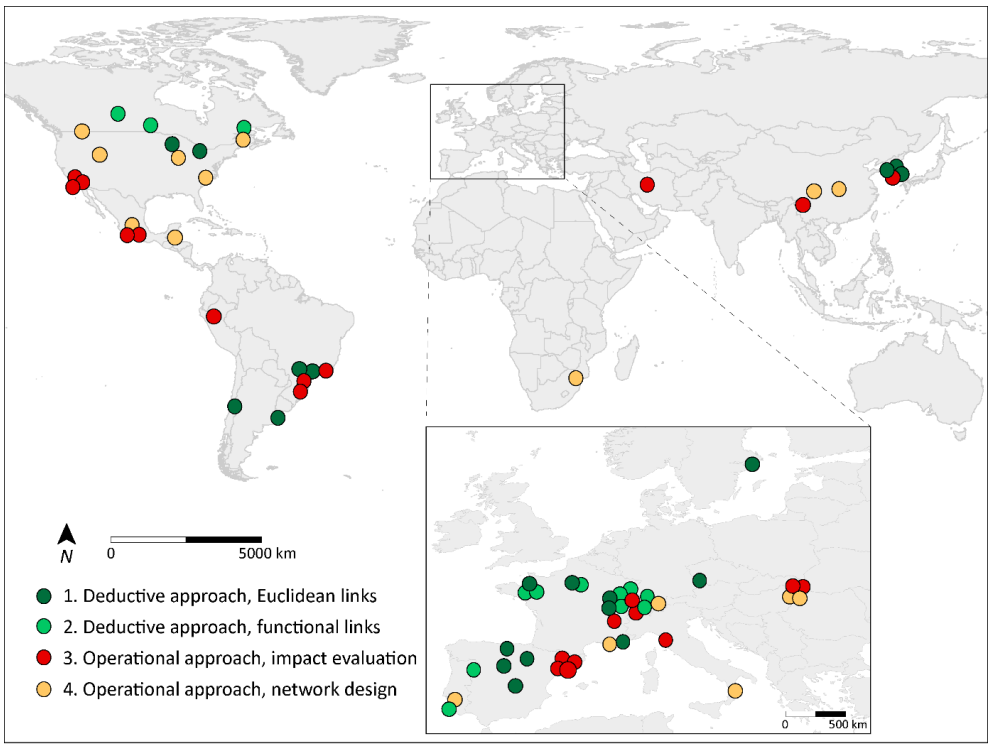


FIGURE 5 – Map of the typology of use of biological data in landscape graphs.

component-level metric is the ECS (5.6 %). It should be noted that a total of 20 other metrics (28.2 %) are used in only one or two articles.

An analysis combining the biological data and the graph analysis outcomes is carried out in 42.2 % of cases. Among the articles in which these data are analyzed jointly, we can distinguish (i) explanatory approaches based on statistical modeling (e.g. regression) where connectivity metrics are used as explanatory variables of biological responses measured by biological data (28.1 %), and (ii) descriptive approaches in which biological data and graph analysis are simply correlated, statistically or visually (14.1 %).

In the factorial space combining connectivity analyses and the types of biological data, genetic data prove to be specific to particular approaches (right side of Fig. 6). These data, sometimes used in patch-level analyses, are more frequently associated either with link-level analyses conducted without metric computation, as for example in Galpern et al. (2012) and Keller et al. (2013), or with analyses performed at the component-level as in Moran-Lopez et al. (2016). These specific approaches contrast with common patch-level analyses linked to presence data and based on a series of local metrics such as delta PC, delta IIC, BC index, or Degree (left side of Fig. 6). The categories positioned at both extremities of axis 2 represent two specific cases encountered in the articles : (1) the studies in which a global-level analysis is conducted, frequently based on PC or IIC metrics, but without any particular link with a type of biological data, (2) the studies dealing with species diversity and abundance data, more often based on the metric flux.

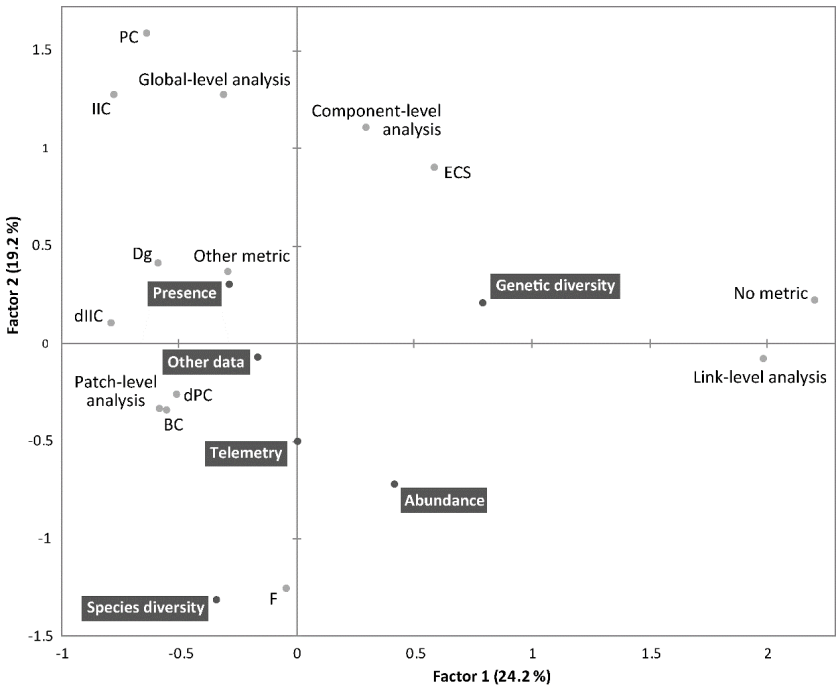


FIGURE 6 – Factorial space of the variables describing the connectivity analyses. The terms in the dark grey rectangles represent the types of biological data.

4 Discussion

The number of studies of landscape connectivity has been rising at a sustained pace since 2005 (Correa Ayram *et al.*, 2016 ; Fletcher *et al.*, 2016). Among these studies, approaches based on landscape graphs have been employed frequently, especially since 2012. The success of these approaches may be explained by the publication of reviews about graph construction (Galpern *et al.*, 2011) and graph-theoretic metrics (Rayfield *et al.*, 2011), by empirical research into connectivity metrics (e.g. Saura et Rubio (2010)), by the availability of software to perform graph-based analyses (Foltête *et al.*, 2012a ; Saura et Torne, 2009), and by the rising interest of stakeholders and policy-makers in connectivity conservation (Correa Ayram *et al.*, 2016). Despite growing interest in connectivity modeling, and considering the large number of references reviewed in the syntheses by Fletcher *et al.* (2016) and Correa Ayram *et al.* (2016) about landscape connectivity (370 and 162 respectively), we found relatively few articles (i.e. 71) using biological data in combination with landscape graphs. This is all the more unfortunate because many studies assume (1) that connectivity has a positive influence on biodiversity or (2) that the connectivity model is a reliable representation of the actual habitat network, but without validating these assumptions from field data. Although the small number of articles using biological data was quite limiting for carrying out statistical analyses, our review clearly highlighted a preponderance of articles dealing with mammals and birds in forests of Europe and North America. This result recalls that obtained from a larger pool of articles in the review by Correa Ayram *et al.* (2016), and this should encourage a greater diversity in the studied topics as connectivity conservation is a pervasive issue across taxa and regions.

A main contrast in the use of biological data : Prior coupling *vs* a posteriori analysis

Our review shows that biological data are integrated into the analysis at different stages depending on the objective, and that this integration is performed in many different ways. Specifically, the second

MCA encompassing variables describing the purpose and application of landscape graphs and the involvement of the stakeholders revealed a main contrast between two objectives (Fig. 3). First, articles from classes 1 and 2 (Fig. 4) aim at providing knowledge about relationships between biological responses and landscape connectivity. In these articles, graph nodes are identified from relatively simple data, e.g. a land-cover map, without integrating biological data in the initial graph's construction stage. Once the connectivity metrics have been computed, biological data are considered as target variables in a statistical analysis in which connectivity metrics are used as explanatory variables. Examples of such a hypothetico-deductive approach can be seen in Martín-Queller et Saura (2013) or Mony *et al.* (2018). Separating the connectivity modeling step and the use of biological data amounts to considering connectivity, measured by metrics, as a potential driver of the biological response. This deductive reasoning aims at (i) providing new knowledge about the influence of landscape connectivity on species, (ii) evaluating the strength of this relationship, and (iii) identifying the spatial scale at which it occurs. It is worth noting that such an approach requires land-cover data that are suitable for mapping the habitat of the focal species and its dispersal paths. It also assumes that graph-based modeling captures connectivity patterns well so that connectivity metrics actually account for the influence of landscape connectivity on biological responses.

We return to the initial data by focusing on classes 1 and 2 to specify how these deductive approaches were conducted. The most frequent design is to analyze a given species at patch-level. This consists in explaining species presence (e.g. Andersson et Bodin (2009) ; Awade *et al.* (2012) ; Foltête *et al.* (2012b) ; Melles *et al.* (2012) ; Song et Kim (2016)) or abundance (e.g. Estrada-Peña (2005) ; Betbeder *et al.* (2017)) in habitat patches by including connectivity metrics in a regression model. In some cases concerning birds or plants, regression or correlation analyses performed at the community level attest to the effect of connectivity either on patch richness (Mony *et al.*, 2018) or on an inter-patch dissimilarity measure (Muratet *et al.*, 2013). Of the papers referring to the deductive approach, only a smallish proportion (27 %) includes a sensitivity analysis to evaluate how statistical dependence is influenced by cost values (Foltête et Giraudoux, 2012), minimum patch size (Andersson et Bodin, 2009), the distance of graph pruning (Koh *et al.*, 2013 ; Muratet *et al.*, 2013) or the setting of the dispersal kernel in the metric calculation (Martín-Queller *et al.*, 2017 ; Gil-Tena *et al.*, 2014). Finally, several other papers report more specific protocols adapted to particular contexts.

Conversely, the opposite objective (represented by classes 3 and 4, Fig. 4) aims at prioritizing key habitat patches or supporting ecological network design. In this case, biological data, most of the time presence data, are directly embedded in the definition of patches from an SDM or another output of suitability modeling, following procedures synthesized in Duflot *et al.* (2018). In these approaches, a first step of modeling precedes the connectivity analysis, leading to map presence probabilities. These probabilities are thresholded to define the habitat patches, and sometimes used to set the cost values defining the links. The status given to the quantification of connectivity makes a big difference with the previous approach. Here, the connectivity computations derived from the landscape graph are not used to test a specific assumption about the landscape's influence. In contrast, they serve as a decision support tool in a rationale of action. This means that the positive influence of connectivity on species is not questioned but it is rather considered as an initial postulate.

Although percentages of articles corresponding to these contrasted objectives (66.2 % including biological data in the graph construction compared with 42.2 % analyzing biological data a posteriori) show that these rationales are in most cases exclusive, we found 12 articles (8.4 %) in which both approaches are present. They correspond to various specific cases integrating different biological data. For example, Ribeiro *et al.* (2011) conducted a field campaign to select the patches with actual presence of amphibians. They then assessed the correlation between connectivity metrics and species richness in a second step. Other researchers have included field data in the definition of links, as in O'Brien *et al.* (2006) and Galpern *et al.* (2012) who brought into play telemetry data for defining the cost values used to compute the links. Galpern *et al.* (2012) finally compared genetic data with the outcomes of connectivity analyses, whereas O'Brien *et al.* (2006) used independent telemetry data to validate the cost values definition. In the same vein, Bergerot *et al.* (2013) carried out an individual release procedure to set up an inter-patch movement model. It was used to construct the landscape graph before comparing the graph's components with the results of a mass release-recapture operation.

The relationship between the purpose of the study and the region where it was carried out was not straightforward, but we observed a trend whereby studies with operational objectives were often performed in North America (Fig. 5). The involvement of public agencies in the research on connectivity such as the USDA, the US Fish and Wildlife service, or Parks Canada, already noticed by Correa Ayram *et al.* (2016) and apparent in some of the articles we reviewed, may be a reason for such a result.

A secondary contrast related to functional connectivity

Apart from this dichotomy between operational and academic objectives, we observed in the same analysis (factor 2 in Fig. 3) a gradient in the way graph links are defined. This gradient contrasts Euclidean links (i.e. uniform matrix) on the one hand and, on the other hand, suitability-based links and land-cover based links (i.e. weighted by least-cost distances) that are validated or not by biological data. This gradient additionally suggests that studies aiming at modeling functional connectivity rather than structural connectivity make wider use of telemetry data. Indeed, eight articles from classes 2 and 4 are based upon this kind of data whereas no articles from classes 1 and 3 use them. Including telemetry data in the definition of links is associated here with the greater involvement of stakeholders. Such a result may not be universal but more probably stems from the frequent use of telemetry in studies carried out in North America, where investment in the acquisition of data reflecting functional connectivity may be higher.

Focusing on the studies in which the functional aspect of connectivity is better considered (positive coordinates on factor 2 in Fig. 3), classes 2 and 4 differ both in their aims (providing knowledge *vs* supporting action, respectively) and the main types of data they include (54 % using genetic or telemetry data vs 71 % using presence data, respectively). Presence data can be easily and affordably obtained compared with genetic or telemetry data. The latter are technically more sophisticated and demand a more intensive collection effort in the field and/or in the lab. This probably explains the dominance of presence data (61 %) in the literature we reviewed. This difference may also explain why studies with operational purposes are mainly based on presence data, because the time available for decision-making is often much more constrained than the time available for pure research. However, a direct link between the type of data and ecological processes underlying connectivity is critical to the graph's relevance whatever its purpose. In this context, several papers highlight that habitat suitabi-

lity is often a poor predictor of dispersal or gene flow for numerous reasons, and that data directly related to movement or gene flow should be considered instead (e.g. Spear *et al.* (2010) ; Peterman *et al.* (2014) ; Khimoun *et al.* (2017)). It does not mean that SDMs are pointless in graph construction, as they could be particularly relevant for delineating patches. But their utility for calibrating the cost values used in the least-cost paths raises questions. In addition, designing movement corridors or reserves for conservation aims not only at satisfying biological needs in different places for a species of concern over its life-cycle, but it should also maintain some genetic exchanges between populations to warrant their long-term viability and adaptive potential (Allendorf *et al.*, 2012 ; Lindenmayer et Fischer, 2006). While movements and gene flow are both important and complementary for conservation purposes, connectivity modeling should integrate different ecological data depending on the precise biological response that is considered.

Two key-levels of integration for biological data

Although a large range of biological data were used in the studies we reviewed, different types of biological data were rarely used in the same study. Due to the influence of landscape structure on both extinction/colonization and migration/drift equilibria driving species and genetic diversities, respectively (Vellend et Geber, 2005), diversity data have considerable potential for improving landscape graph-based approaches. It is noteworthy that these data were used in very different ways in papers coupling them with landscape graphs. Genetic data were mainly used at link level to analyze genetic differentiation between locations (seven out of eight articles, *versus* two articles only using within-location genetic diversity), whereas species diversity data were almost exclusively used at patch level to derive species diversity values within locations (i.e. alpha diversity ; five out of six articles, versus one article only using between-locations diversity, i.e. beta diversity ; see Figure 5). In addition, species alpha diversity was always used as a biological response to be explained statistically from some connectivity measures obtained from the landscape graph, whereas genetic data were used either as an input to help graph construction or as a biological response that was correlated with some connectivity metrics (six and four articles, respectively). This difference in treatment between species diversity and genetic diversity has a long history despite their theoretical connections as responses to similar processes (Vellend et Geber (2005), but see Taberlet *et al.* (2012)). Indeed, populations occurring in small and isolated habitat patches are exposed to a higher genetic drift and a lower gene flow than large and continuous populations, and both processes are expected to lead simultaneously to lower within-population genetic diversity and greater genetic divergence between populations (Keyghobadi, 2007). In the same way, communities occurring in small and isolated habitat patches are expected to harbor a lower species richness and to differ more widely from each other, because of a higher extinction rate and a reduced colonization rate (MacArthur et Wilson, 1967). Thus, whether diversity is measured within or between entities (patches, populations, communities, etc.), and whether it is assessed at the genetic or species level, the value obtained always results from (at least) two processes operating within and between the entities considered. In this framework, graph theory offers a flexible and valuable landscape modeling approach, as metrics can be extracted from graphs to reflect processes operating at the patch level only, at the link level only, or to integrate both kinds of processes in the same connectivity measure (Pascual-Hortal et Saura, 2006 ; Saura et Pascual-Hortal, 2007). However, this opportunity is still under-used because even in articles coupling species or genetic diversity with a landscape graph, we found only two articles (Schoville *et al.*, 2018 ; Ribeiro *et al.*, 2011) and a single article (Neel, 2008) trying to explain levels of species diversity and genetic diversity, respectively, by

connectivity metrics integrating both patch and link properties.

Which strategy should we adopt in applying landscape graphs ?

In light of the findings of our review, one may ask which is the more relevant approach to landscape graph modeling. This mainly relates to the opposition discussed above between prior coupling and a posteriori analysis. Should we directly incorporate biological data in graph construction, or use it once the graph has been constructed, to validate it or to verify the influence of connectivity ? As researchers involved in scientific approaches, our initial reflex would be to advocate the second proposition, where the role of habitat connectivity and the reliability of the model have to be questioned before being applied in an operational rationale. Theoretically, if the connectivity metric assumed to represent the process under consideration proves to be significantly linked to biological data, the graph could be used as a decision support tool, but otherwise not. A positive statistical test provides reassurance about the approach, but such validation should not be overestimated, especially when biological data are limited to presence data. A significant relationship between a graph metric and biological data only provides a global validation of the entire modeling approach ; it does not inform us about the details of the relationship and the strengths and weaknesses of the model. Indeed, uncertainty may remain with respect to (1) the definition of patches, (2) the definition of links, (3) the choice and the setting of a metric for quantifying connectivity, and finally (4) the role of connectivity. Only a sensitivity analysis applied to the first three parameters could help to clarify their area of validity. Thus, the deductive approach seems to be preferable because it enhances the model's overall legitimacy, but any application should be subject to a sensitivity analysis to justify and specify the modeling choices.

Let us turn now to the direct incorporation of biological data in graph construction. Our review shows that the more frequent approach is to delineate habitat patches from outputs of a species distribution model. Since such models are usually the output of statistical approaches including a validation step, it can be taken that the patches are validated by empirical data. The interest of this approach is to provide more realistic habitat patches, consistent with the concept of ecological niches, and not based on a land-cover map alone. However, the absence of *a posteriori* validation concerns the subsequent graph construction choices as to the definition and weighting of the links and the parameterization of the other criteria mentioned above. Consequently, to be reliable, this approach should preferably be supplemented by a validation process focused on the links and the connectivity metrics. In this perspective, the "gold standard" approach should ideally include additional data to calibrate or validate the links. Zeller *et al.* (2018) show that the type of biological data best reflecting movement patterns are genetic and telemetry data. Because genetic structure mainly results from multi-generational dispersal movements (Keyghobadi, 2007), genetic data are therefore a reliable proxy for the dispersal movements which matter most for species conservation (Zeller *et al.*, 2012). However, genetic data provide limited information about current habitat connectivity in cases where recent changes have not yet affected genetic structure due to time lags (Landguth *et al.*, 2010). In such cases, telemetry data are probably more helpful. In sum, we think that genetic and telemetry data should be used to validate landscape graph models, especially when they rely only upon land use and presence data.

5 Conclusion

This review has shown a major contrast in studies coupling landscape graphs and biological data depending on their main objective. In approaches aimed at providing knowledge, patches are defined from land-cover maps and biological data are correlated a posteriori with connectivity metrics. Conversely, in operational approaches, patches are more frequently derived from SDMs while directly including biological data in graph construction. The second contrast concerns the more or less functional nature of links and highlights the role of telemetry and genetic data in validating them.

Beyond these main contrasts, the review shows that landscape graphs can benefit from field data of different types at varying scales. The great variability of approaches adopted in the articles we have reviewed reveals the flexible nature of these tools. Since field data allow us to understand empirically ecological processes such as dispersal and its dependence on landscape connectivity, we encourage others to multiply studies coupling landscape graphs and field data. While biological data may reflect a functional biological response to landscape connectivity, they are usually gathered from a limited set of locations. Therefore, their complementarity with landscape graphs, which represent the exhaustive set of potential habitat patches, is an additional reason for encouraging this coupling. This may theoretically improve the reliability of connectivity analyses and the way they are carried out. Nevertheless, we do not discard the results from all studies performed without field data as they indirectly benefit from methodological improvements and ecological knowledge acquired in studies based on field data. For example, Clevenger et al. (2002) showed that connectivity models based on information derived from the literature were a better proxy of empirical models than those designed exclusively from expert opinion. Finally, whether ecological data are used as input in the graph construction process or as a biological response that has to be explained statistically by certain landscape properties, it is crucial to choose both the appropriate data and graph metrics in accordance with the biological process under consideration.

Annexe A2

Combining landscape and genetic graphs to address key issues in landscape genetics

Abstract

Context

All the components of landscape and genetic structures can be associated with the nodes and links of landscape graphs and genetic graphs. Yet, these graphs have long been used separately despite the potential for their combined use in landscape genetics.

Objectives

First, comparing these graphs could be an effective way to disentangle the influence of intra-patch features from that of inter-patch connectivity on genetic structure or to assess whether intra-population genetic diversity and inter-population genetic differentiation are sensitive to the same landscape influences.

Methods

Moreover, because graph pruning determines which connections between nodes are considered in calculating neighbourhood-based metrics or graph-based distances, comparing the metrics or distances derived from differently pruned graphs can be an effective way to identify the scale of landscape effects or the scale at which both gene flow and drift determine genetic differentiation. Similarly, comparing node partitions in both types of graphs could strengthen the validity of barrier identifications.

Results

Second, beyond mere comparisons, the integration of landscape and genetic graphs through gravity models can further enhance their joint use for theoretical and applied objectives alike.

Conclusion

We thus believe that future research could illustrate and enhance the relevance of these methods for a wider range of applications in landscape genetics.

Keywords : landscape genetics, graph theory, habitat connectivity, dispersal, gene flow

Cet article a été soumis à *Landscape Ecology* en novembre 2020 :

Savary, P., Foltête, J. C., Moal, H., & Garnier, S. Combining landscape and genetic graphs to address key issues in landscape genetics. Submitted to *Landscape Ecology*

1 Introduction

Understanding how species move and settle within landscapes is key to designing conservation programmes to counter the continuing erosion of biodiversity (Barton *et al.*, 2015 ; Bennett *et al.*, 2006 ; Jeltsch *et al.*, 2013 ; Kool *et al.*, 2013). For the species whose populations are scattered over discontinuous habitat patches embedded in the landscape matrix (Bowne et Bowers, 2004), modelling habitat connectivity involves studying the properties of a spatial network. Accordingly, among the wide range of methods for modelling connectivity (Correa Ayram *et al.*, 2016), those derived from graph theory prove relevant (Dale et Fortin, 2010 ; Galpern *et al.*, 2011).

Landscape graphs are used in landscape ecology to represent habitat networks (Urban et Keitt, 2001). A landscape graph is a set of habitat patches (nodes) connected by a set of potential dispersal paths (links)(Box 1). Several factors lie behind the success of landscape graphs (Galpern *et al.*, 2011). First, the habitat network topology they bring out often provides key insights into population spatial structure and dispersal patterns (Brooks, 2003 ; Keitt *et al.*, 1997 ; Ortiz-Rodríguez *et al.*, 2019). Then, they are widely used because of their computational potential, providing a suitable framework for calculating a broad range of connectivity metrics, e.g. as a decision-making aid to indicate which patches should be conserved as a priority (Foltête *et al.*, 2014). Further, landscape graphs are of great cartographic interest, enabling land planners to fully grasp the variations in connectivity in the regions they manage (Bergsten et Zetterberg, 2013). All these applications are made possible because landscape graphs rely upon an exhaustive representation of the potential habitat patches within the landscape.

The way the basic elements (i.e. nodes and links) of landscape graphs are defined is a critical issue that challenges their ecological validity (Box 1). At this stage, assumptions have to be made about species' habitat preferences and dispersal capacities, but these assumptions are rarely tested and validated with biological data. This is probably the biggest pitfall with graph-based habitat connectivity modelling (Correa Ayram *et al.*, 2016 ; Kadoya, 2009 ; Moilanen, 2011).

Several types of biological data can be used in graph-based connectivity modelling (Foltête *et al.*, 2020). When focusing on single species, genetic data or movement data obtained by telemetry reflect landscape resistance to individual movements better than simple presence data do (Diniz *et al.*, 2020 ; Zeller *et al.*, 2018). When using movement data, it is hardly possible to distinguish home-range movements from dispersal movements (Koenig *et al.*, 1996), potentially skewing estimates of landscape resistance to dispersal (Horskins *et al.*, 2006 ; Mateo-Sánchez *et al.*, 2015 ; Zeller *et al.*, 2012). In contrast, genetic data reflect movements of successful breeders over several generations, making neutral genetic diversity structure a reliable proxy for the dispersal movements that matter most for species conservation (Koenig *et al.*, 1996 ; Zeller *et al.*, 2012). When individuals can easily disperse between habitat patches, they spread their genes thereby maintaining genetic diversity within patches and preventing any increase in genetic differentiation between patches (Keyghobadi, 2007). Therefore, genetic data can be used to infer landscape influences on genetic structure (Manel *et al.*, 2003 ; Storfer *et al.*, 2007). Although it used to be costly and difficult to obtain genetic data, they are now increasingly available for a wide range of species (Miraldo *et al.*, 2016) and of direct benefit for landscape genetic studies.

In landscape genetics, studying gene flow events between discrete populations also involves studying the properties of a spatial network (Murphy *et al.*, 2016). The nodes of such a network, hereafter called a genetic graph, represent populations or individuals (Box 2). The links represent substantial gene flow between them and are weighted by genetic distances. Analyses of genetic graphs contribute to a better understanding of how landscape influences genetic response because the graphs are based on empirical biological data and are free from any prior assumptions. The main drawback with genetic graphs is that they are often constructed from small population samples, especially when the study area is too large to be sampled exhaustively because data acquisition is time-consuming and expensive. Only in rare instances have populations been sampled exhaustively (Keller *et al.*, 2013 ; Murphy *et al.*, 2010a ; Watts *et al.*, 2015) although it has been pointed out that missing populations can substantially affect inferences (Albert *et al.*, 2013 ; Koen *et al.*, 2013 ; Naujokaitis-Lewis *et al.*, 2013).

Landscape graphs and genetic graphs depict the same ecological reality but they have mainly been used in two separate research fields : landscape ecology and population genetics, respectively. Landscape graphs are derived from the assumed influence of landscape connectivity on biological fluxes. They cover the exhaustive set of potential habitat patches but they may lack ecological validity. Conversely, genetic graphs empirically express the outcome of biological fluxes but they are restricted to samples of populations or individuals. Both types of graph have been used in many studies (Galpern *et al.*, 2011 ; Greenbaum et Fefferman, 2017)(Online Resource 1) but for the most part separately. Previous works have called for the use of genetic data in connectivity modelling with landscape graphs (Foltête et Vuidel, 2017 ; Luque *et al.*, 2012) but without mentioning explicitly the complementarity between landscape and genetic graphs. Similarly, Garroway *et al.* (2011), Manel et Holderegger (2013) and Murphy *et al.* (2016) have suggested that they could be advantageously compared or even integrated but they failed to detail such an approach as it was not the main focus of their work. We believe that identifying methods for comparing and integrating landscape and genetic graphs could further extend the contribution of graph-theoretic approaches to landscape genetics.

As a multidisciplinary field (Dyer, 2015a), landscape genetics inherits both its methods and its research questions from landscape ecology and population genetics. Whereas a landscape ecologist asks whether the amount of habitat matters more than habitat connectivity for biological responses (Lindenmayer *et al.*, 2020), a population geneticist asks whether landscape structure influences local genetic diversity are genetic differentiation similarly (Keyghobadi *et al.*, 2005). Furthermore, landscape ecology studies investigating the 'scale of effect' of the landscape on biological responses (Jackson et Fahrig, 2012 ; Miguet *et al.*, 2016) mirror population genetics studies seeking to identify either the neighbourhood size or the scale at which both gene flow and drift determine genetic differentiation at equilibrium (Hardy et Vekemans, 1999 ; Van Strien *et al.*, 2015). Other parallel questions are to be found and we believe that each of them can be addressed in analyses involving the node and link properties of landscape and genetic graphs. Therefore, combining landscape graphs and genetic graphs could be one more step towards the true interdisciplinary connection between landscape ecology and population genetics that landscape genetics has been looking for. The objective of this paper is to provide an outline answer to a rather broad question : 'Why and how should we use genetic graphs and landscape graphs conjointly to address key issues in landscape genetics ?' We begin by discussing the theoretical questions that could be answered by a comparison of these graphs. We then go beyond

mere comparisons to imagine actually integrating these tools in order to benefit from their respective advantages. We end with perspectives for future investigation.

Box 1 : Landscape graphs

Landscape graphs are used for analysing networks of habitat patches and have grown in popularity following the seminal paper of Urban et Keitt (2001). In these spatially-explicit graphs, nodes are habitat patches and links represent potential dispersal paths between them (Figure 7). They are built from geographical layers of landscape features potentially influencing species movements and distribution. Habitat patches (nodes) are delineated according to land cover criteria (Foltête *et al.*, 2014) or habitat suitability thresholds (Duflot *et al.*, 2018). Links are spatially delimited paths that take into account landscape feature resistance through least-cost path calculation (Foltête *et al.*, 2012a) or use of circuit theory (Brodie *et al.*, 2016).

Node and link weights often depict the heterogeneity of habitat properties and accessibility (Dale et Fortin, 2010). Link weights correspond to 'landscape distances' such as resistance, least-cost, or geodesic distances. Similarly, habitat patches are often weighted by a proxy of their demographic capacity, usually their area.

The graph can be pruned to select a subset of links between habitat patches. Pruning can speed up the computations (Foltête *et al.*, 2012a) and is useful for selecting the links corresponding to direct movements between patches, e.g. by removing links corresponding to distances greater than the species maximum dispersal distance.

Then, connectivity metrics can be computed on the scale of the entire graph, a component of the graph or an individual patch. Largely inspired by the metapopulation framework, graph-theoretic connectivity metrics take into account patch area as well as dispersal probabilities between patches (e.g. PC, IIC)(Saura et Pascual-Hortal, 2007). So-called 'delta metrics' indicate the unique contribution of each node to the connectivity metric computed for the entire graph and can even be broken down into several fractions for a better understanding of their functional role (Saura et Rubio, 2010). To help pick among the myriad different metrics, several authors propose synthetic metric classifications (Baranyi *et al.*, 2011 ; Calabrese et Fagan, 2004 ; Rayfield *et al.*, 2011).

Landscape graph nodes may also be subjected to modularity analyses revealing the existence of well connected clusters of habitat patches (Foltête et Vuidel, 2017). Besides, landscape distance matrices taking into account the topology of the habitat network can be derived from the graphs and used in functional connectivity analyses (Etherington, 2012 ; Pinto et Keitt, 2009).

Landscape graphs are therefore useful for supporting the prioritisation of patches and corridors for conservation or restoration measures, for assessing the impacts of specific infrastructures (Foltête *et al.*, 2014) and also for deriving relevant explanatory variables for subsequent analyses (Foltête *et al.*, 2012b ; Pereira *et al.*, 2011). Unfortunately, without validation of hypotheses about species habitat distribution and dispersal capacities made for constructing the graph, the results

of these approaches are questionable. Yet the use of empirical data in graph-based connectivity analyses is rarely directed at 'validating' graph construction (Foltête *et al.*, 2020).

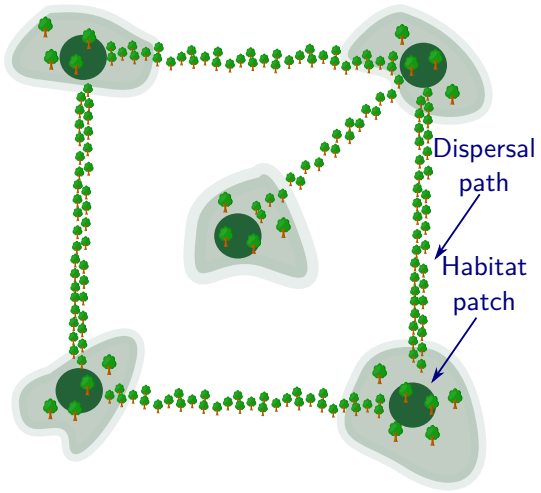


FIGURE 7 – Schematic representation of a landscape graph

Box 2 : Genetic graphs

Genetic graphs represent the genetic structure of a network of sampled populations (i.e. groups of individuals sampled at the same site (Dyer et Nason, 2004) or individuals (Castillo *et al.*, 2016)). Input data are therefore multilocus individual genotypes. The links are represented by straight lines, not depicting physical paths, and are weighted by genetic distances such as F_{ST} (Munwes *et al.*, 2010), D_{PS} (Naujokaitis-Lewis *et al.*, 2013), Euclidean genetic distance (Dyer et Nason, 2004) or others.

Analyses derived from these graphs can be directed at identifying direct dispersal paths followed by propagules between populations (Dyer et Nason, 2004) or key populations for the genetic connectivity of the network (Cross *et al.*, 2018). Such analyses are based on visual inspection of these spatially-explicit graphs and on metric calculation. Alternatively, the graph links may form the basis for inferring landscape resistance (Garroway *et al.*, 2011). In most cases, the complete set of links should be pruned and the objective of the analyses is to determine which pruning method to use. To study direct dispersal networks, maximum dispersal distance thresholds, topological constraints or the conditional independence principle (Dyer et Nason, 2004) can be powerful tools although they depend on previous knowledge of the study species and/or gene flow frequency (Savary et al., in correction). Because dispersal events leading to gene flow occur in a multigenerational time frame, greater distance thresholds are used for selecting population pairs in order to infer landscape resistance to dispersal (Boulanger *et al.* (2020) ; Savary et al., in correction).

2 Comparing graphs to answer current landscape genetic questions

Embedding landscape predictors and genetic responses in both node and link elements of graphs of the same nature opens the door to several types of landscape genetic analyses at the node-, neighbourhood-, link- and boundary-levels (Wagner et Fortin, 2013). Graph comparisons at these levels could be useful for addressing current issues in landscape genetics (Manel *et al.*, 2003 ; Storfer

et al., 2010), including i) the complex relationships between landscape structure and genetic structure, ii) the intricacy of scale effects in landscape genetics and iii) the identification of dispersal barriers (Figure 8).

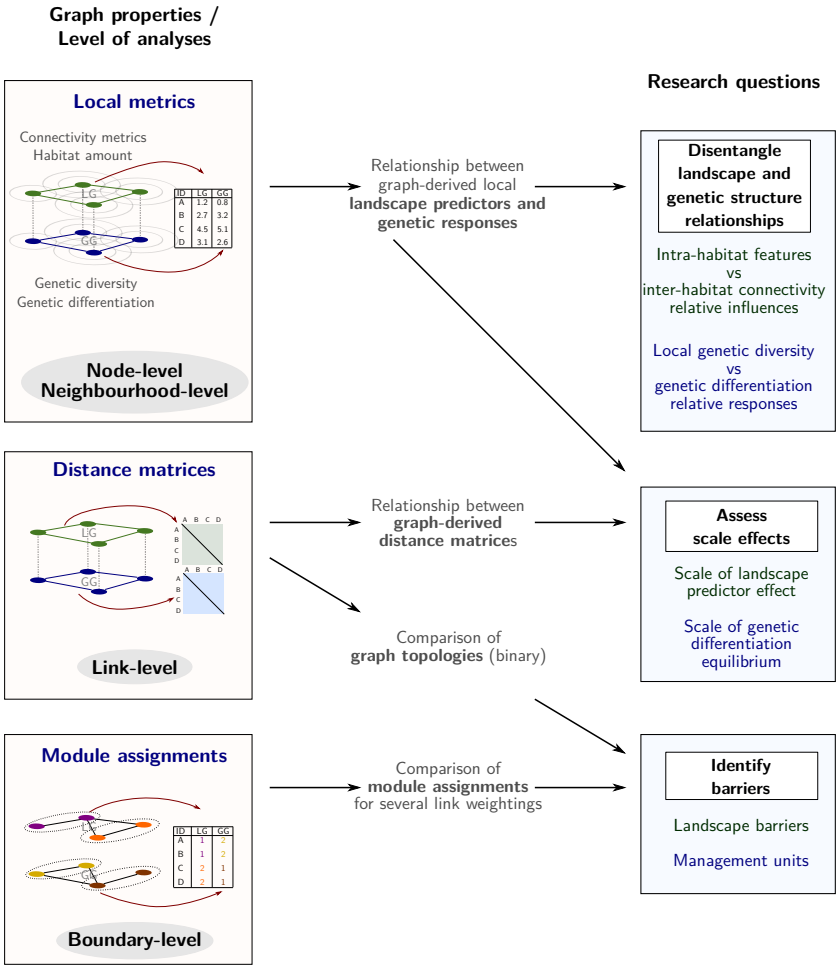


FIGURE 8 – Joint use of landscape graphs (LG) and genetic graphs (GG) to address landscape genetic questions. Populations and habitat patches (A, B, C, D) are the same in both types of graph.

2.1 Disentangling the complex relationships between the components of landscape structure and genetic structure

In landscape genetics, both genetic response and landscape predictors may be i) 'node-level' variables describing the population or the habitat in isolation or ii) 'neighbourhood-level' variables depicting the function of each node given its location in the network by taking into account its links with other populations or habitat patches (Wagner et Fortin, 2013). Graphs more than any other objects make it possible to compute both node-level and neighbourhood-level variables to characterise the different components of landscape structure and genetic structure. In depth assessment of the links between these variables computed in landscape and genetic graphs provides the opportunity to disentangle the complex relationships between landscape structure and genetic structure.

Theories derived from population genetics and landscape ecology point to an influence of both intra-habitat patch features and inter-habitat patch connectivity on genetic structure. The neutral genetic structure of species occupying habitat patches is the result of both drift (a demographic effect

driven by patch quality and area) and gene flow (which depends on inter-patch connectivity). Therefore, genetic structure can be explained by both intra-patch (node) and inter-patch (link) landscape predictors computed from landscape graphs. Landscape genetic studies have evidenced the influence of habitat connectivity on genetic structure more frequently than that of local landscape features (DiLeo et Wagner (2016), but see Ezard et Travis (2006)), although landscape ecology studies have reported strong habitat amount effects on a large range of biological responses (Fahrig, 2003, 2017). This is probably due to the predominance of link-level analyses in landscape genetics (DiLeo et Wagner, 2016). Besides, assigning to each type of landscape predictor its respective influence on genetic structure can be complicated by the interdependence between habitat amount and fragmentation (Didham *et al.*, 2012).

Apart from terminological aspects, it is not easy to identify the respective influence of these spatial habitat properties. This has triggered heated debates (Fahrig, 2013 ; Fletcher Jr *et al.*, 2018 ; Haddad *et al.*, 2017 ; Hanski, 2015) in part because the quantification of habitat amount in the landscape around a sampling site is dependent upon the habitat spatial configuration and fragmentation (Saura, 2021) and because habitat connectivity depends on habitat amount (Saura et Rubio, 2010). Given that it turns out to be hardly possible to totally isolate these variables, a variable quantifying the Amount of Reachable Habitat (ARH) could explain how both habitat amount and configuration influence biodiversity patterns in heterogeneous landscapes (Blazquez-Cabrera *et al.*, 2014 ; Martensen *et al.*, 2017 ; Villard et Metzger, 2014). Several metrics derived from landscape graphs, such as the delta variant of the 'Equivalent Connectivity' (dEC) as well as local metrics such as Flux (F) and Interaction Flux (IF) are examples of such neighbourhood-level local metrics. These metrics make it possible to vary the weights of both patch capacities (different capacity measures or associated exponents) and inter-patch distances (different dispersal probability functions) in their calculation, providing a way to test for the relative influence of local habitat quality and matrix resistance on genetic responses.

The genetic response can also be described at the node- and neighbourhood-level in a genetic graph to depict both intra- and inter-population genetic structure. Intra-population genetic diversity has been shown to depend more strongly on habitat amount than on habitat connectivity (Jackson et Fahrig, 2015), because the former determines population size and drift intensity. Similarly, Bruggeman *et al.* (2010) and (Cushman *et al.*, 2012) showed that inter-population genetic differentiation depended more on habitat configuration than on habitat amount. However, DiLeo et Wagner (2016) showed that most landscape genetic studies focused on habitat configuration and matrix resistance effects on genetic differentiation at the link-level whereas intra-population diversity was rarely included as a response variable. Characterising genetic structure at both the node- and neighbourhood-levels of a genetic graph could be relevant in that context. Node-level genetic variables include population specific genetic diversity indices such as allelic richness or heterozygosity rates. Neighbourhood-level genetic responses are graph-theoretic metrics taking into account node connections with other nodes (Wagner et Fortin, 2013). For example, Koen *et al.* (2016) evidenced the linear relationship between the average genetic distances (and inverse genetic distances) weighting the links connected to a population and the simulated connectivity between that population and the others. Hence, although genetic differentiation is usually assessed between population pairs, neighbourhood-level genetic indices make it possible to measure genetic differentiation at the population level. Moreover, neighbourhood-level genetic indices could also estimate local genetic diversity by inspiring from the rationale behind metrics

quantifying the amount of reachable habitat in landscape graphs. Shirk et Cushman (2011) took the first step towards computing spatially-explicit genetic diversity indices by considering sampling areas defined from genetic neighbourhood sizes in their calculations. Taking graph topology into account to compute genetic diversity indices could be the next step in deriving a metric quantifying the potential genetic diversity in a population given potential dispersal events from other connected populations.

Once these variables have been computed from landscape and genetic graphs, we end up with landscape predictors and genetic responses in both cases computed at the node- and neighbourhood-level. It then remains to integrate these variables in analyses to reach conclusions. In order to compare different landscape pattern effects on genetic response or different genetic responses to landscape patterns on the same basis, it is important to integrate the relevant variables within a common statistical framework (Leroux *et al.*, 2017). For that purpose, neighbourhood-level variables derived from graphs offer the advantage of integrating inter-patch or inter-population distances in the computation of local metrics. This makes it possible to use models in which statistical individuals are patches or populations, while considering distances between them. It also saves using widely criticised distance-based analyses such as Mantel tests (Balkenhol *et al.*, 2009b ; Legendre et Fortin, 2010).

2.2 Assessing scale effects in landscape genetics

Results obtained by comparing landscape and genetic graphs at the node- and neighbourhood-level will probably depend on the size of the neighbourhood considered, due to so-called scale effects, which have been shown to be intricate in landscape ecology (Brooks, 2003 ; Jackson et Fahrig, 2012 ; Miguet *et al.*, 2016) and population genetics (Hutchison et Templeton, 1999).

In landscape graphs, neighbourhood size is embedded in the computation of connectivity metrics derived from the Probability of Connectivity Index (Saura et Pascual-Hortal, 2007). When calculating these metrics, paths are computed from each patch in turn, using a specific dispersal kernel to weight the influence of the surrounding patches. The dispersal kernel is defined by converting the inter-patch distances (i.e. the cumulated length of the links separating patches) into a dispersal probability, according to a function determining the shape of the kernel (usually a negative exponential function). Identifying the dispersal kernel maximising the strength of the relationship between the connectivity metric and the genetic response computed from a genetic graph would be another way to identify the scale of effect of the habitat spatial pattern on genetic structure. Because the genetic variable computed from the graph at the population level can measure either the genetic diversity or the genetic differentiation (neighbourhood-level), scales of effects influencing these two components of genetic structure could be compared. This approach also has the advantage of making it possible to integrate cost-distances in analyses of scales of effects, whereas the use of Euclidean geodesic distances implying isotropic neighbourhoods has been the norm so far (but see Miguet *et al.* (2017)).

On the other hand, the way the graphs have been pruned determines which connections are taken into account in calculating neighbourhood-level variables. Comparing the strength of the relationship between predictors and response variables derived from landscape and genetic graphs for different pruning methods can provide insight into the most likely topology of the propagule dispersal events.

Similarly landscape and genetic graphs can be compared by assessing the relationship between matrices of graph-based distances separating similar nodes derived from these graphs with methods and models commonly used in link-based landscape genetic analyses (Mantel correlations, Multiple Regressions on Distance Matrices, linear mixed models), with varying graph pruning methods. On a pruned landscape graph, inter-patch distances are computed as the shortest distances along the links and can be used to test for the likelihood of stepping-stone dispersal models (Saura *et al.*, 2014). In contrast, in genetic graphs, pruning removes distance values associated with some population pairs from link-level analyses which could improve the power of link-level analyses, as suggested by Wagner et Fortin (2013). We also think that pruning a genetic graph could provide insight into the scale at which genetic differentiation at equilibrium depends both on gene flow and drift, a question long asked in population genetics (Hardy et Vekemans, 1999 ; Hutchison et Templeton, 1999).

2.3 Identifying barriers in the landscape

Another goal of landscape genetics analyses is to identify landscape barriers isolating groups of populations from each other (Manel *et al.*, 2003 ; Segelbacher *et al.*, 2010 ; Storfer *et al.*, 2007). Populations located on each side of these unbridgeable barriers cannot exchange migrants directly. If a landscape graph successfully represents the habitat network, it should not include links between these populations, thereby evidencing the absence of connectivity between the habitat patches. When genetic graphs are pruned based on a genetic criterion, such as the conditional independence principle used in 'population graphs' (Dyer et Nason, 2004), their topology provides a reference free from any hypothesis regarding landscape feature influence that can be compared to the topology of the landscape graph to test for its ecological validity. This 'topological congruence' analysis (Dyer, 2015b) is one way to identify dispersal barriers and to assess the maximum dispersal capacities of the study species. When graphs share the same nodes, it basically consists of a comparison of two binary classifications (presence-absence of links) which can be performed with classification assessment indices (Fletcher *et al.*, 2011 ; Matthews, 1975).

Landscape and genetic graphs can also be compared through boundary-based analyses in which link weights are taken into account to quantify the strength of the interaction between groups of nodes (Wagner et Fortin, 2013). Both types of graphs can be subjected to modularity analyses in order to define clusters (modules) of nodes within which connections are stronger than with nodes from other clusters (e.g. Fortuna *et al.* (2009) ; Foltête et Vuidel (2017)). Two partitions are similar if two nodes from the same cluster in the partition of a graph are also in the same cluster in the partition of the other graph. This similarity can be assessed with indices comparing partitions such as the Adjusted Rand Index (Hubert et Arabie, 1985) or the Normalized Mutual Information (Danon *et al.*, 2005 ; Reichert *et al.*, 2016). The similarity between the partitions of a landscape graph and a genetic graph could indicate that the spatial structure of the genetic variation is due to the existence of barriers resistant to dispersal. It could also validate the identification of significant management units performed with landscape graphs (Foltête et Vuidel, 2017).

3 Integrating the graphs to benefit from their complementarity

The comparison of these two types of graph is promising and has been poorly exploited so far (but see Castillo *et al.* (2016), Creech *et al.* (2014), Draheim *et al.* (2016) and Schoville *et al.* (2018) for

inspiring approaches). Nevertheless, combining genetic graphs with landscape graphs is useful for more than just validating landscape graphs ecologically. Genetic graphs usually have few nodes, which limits the assessment of ecological connectivity whereas landscape graphs consider an exhaustive set of potential habitat patches (Figure 9). Landscape graphs and genetic graphs are thus truly complementary and connectivity modelling could benefit from a two-way interaction between these tools.

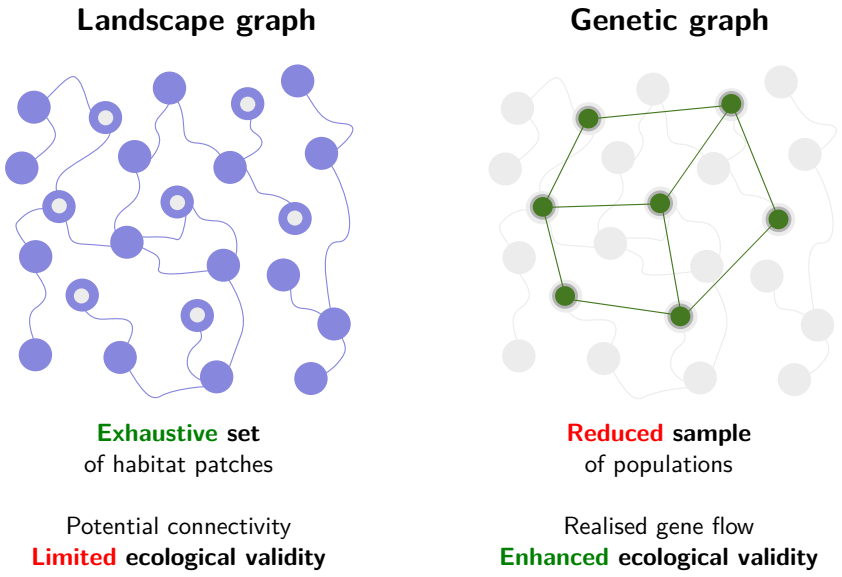


FIGURE 9 – Complementarity between genetic graphs and landscape graphs. Habitat patches in which genetic sampling occurs are displayed by a white dot on the left panel.

Integrating these graphs could enable predictive analyses in landscape genetic studies, which are currently very scarce although they could benefit to conservation practitioners (but see Van Strien *et al.* (2014)). If robust models linking landscape graph and genetic graph properties are calibrated, then an extrapolation could indicate the potential role for the genetic connectivity of a large number of both sampled and non-sampled habitat patches. For example, Creech *et al.* (2014) extrapolated a model linking landscape distance and expected gene flow using a graph-theoretical approach in order to model the differences between gene flow and colonization network topologies due to sex-biased dispersal in Bighorn sheep.

This integration could go even further through the use of gravity models (Murphy *et al.*, 2010a). In these models, both intra-patch and inter-patch landscape predictors could be included in the same framework. Model fit estimates could also assess the relevance of such a model for predictive purposes. In addition, including estimates of population size in these models could be a way to partial out the effect of the spatial heterogeneity of population sizes on drift intensities, and thus on genetic differentiation (Prunier *et al.*, 2017). Gravity models have already been used in landscape genetics (Murphy *et al.*, 2010a ; Robertson *et al.*, 2018a ; Zero *et al.*, 2017) but they have rarely been combined with both landscape and genetic graph modelling or used for predictions, although we think this could reinforce both their ecological relevance and their value for conservation.

4 Further steps towards a better joint use of landscape and genetic graphs

Further research is still needed to broaden the potential of the joint use of landscape and genetic graphs and go beyond current limits in landscape genetic studies. We think that statistical methods commonly used in graph theory could be applied when analysing genetic graphs or landscape graphs to overcome statistical limits often pointed out (Balkenhol *et al.*, 2009b). For example, permutation and randomization methods (Farine et Whitehead, 2015 ; Reichert *et al.*, 2016) should be used more often to test for the significance of network properties. Besides, recent advances regarding statistical methods used for graph pruning, whether from genetic data (Greenbaum *et al.*, 2016 ; Kuismin *et al.*, 2017, 2020 ; Neuditschko *et al.*, 2012 ; Peterson *et al.*, 2019) or landscape data (Fletcher *et al.*, 2011 ; Serrano *et al.*, 2009) do not find sufficient applications in both fields although graph pruning is key to identifying landscape barriers to dispersal.

Finally, connectivity modelling most often assumes symmetrical exchanges between habitat patches although several works on dispersal reveal their asymmetrical nature (Baguette *et al.*, 2013 ; Bonte *et al.*, 2012). Directed graphs have rarely been built with neither landscape data nor genetic data (Holderegger et Gugerli, 2012) but initial attempts to build these types of graphs in these fields appeared useful for understanding source-sink dynamics (Jordán *et al.*, 2007). Further work on this type of graph should therefore be encouraged.

In conclusion, combining landscape and genetic graphs paves the way for a wide range of analyses which could both shed light on complex landscape genetic relationships and support decision-making with empirically grounded arguments.

Acknowledgements

This study is 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). We are particularly grateful to ARP-Astrance team for its constant support along the project. This work is part of the project CANON that was supported by the French "Investissements d'Avenir" program, project ISITE-BFC (contract ANR-15-IDEX-0003). We thank Christopher Sutcliffe for revising the English manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

The first draft of the manuscript was written by Paul Savary and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Annexe A3

Analysing landscape effects on dispersal networks and gene flow with genetic graphs

Abstract

Graph-theoretic approaches have relevant applications in landscape genetic analyses. When species form populations in discrete habitat patches, genetic graphs can be used i) to identify direct dispersal paths followed by propagules or ii) to quantify landscape effects on multi-generational gene flow. However, the influence of their construction parameters remains to be explored. Using a simulation approach, we constructed genetic graphs using several pruning methods (geographical distance thresholds, topological constraints, statistical inference) and genetic distances to weight graph links (F_{ST}, D_{PS}, Euclidean genetic distances). We then compared the capacity of these different graphs to i) identify the precise topology of the dispersal network and ii) to infer landscape resistance to gene flow from the relationship between cost-distances and genetic distances. Although not always clear-cut, our results showed that methods based on geographical distance thresholds seem to better identify dispersal networks in most cases. More interestingly, our study demonstrates that a sub-selection of pairwise distances through graph pruning (thereby reducing the number of data points) can counter-intuitively lead to improved inferences of landscape effects on dispersal. Finally, we showed that genetic distances such as the D_{PS} or Euclidean genetic distances should be preferred over the F_{ST} for landscape effect inference as they respond faster to landscape changes.

Keywords : landscape genetics, ecological connectivity, graph theory, simulation, dispersal

Cet article a été publié dans *Molecular Ecology Resources* en janvier 2021 :

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

1 Introduction

Landscape connectivity is defined as the degree to which the landscape facilitates or impedes movement among resource patches (Taylor *et al.*, 1993). Such dispersal events reduce metapopulation extinction risk (Den Boer, 1968 ; Hanski, 1998) and give rise to gene flow, thereby preventing inbreeding depression and maintaining local adaptation potential (but see Crispo *et al.* (2011), Richardson *et al.* (2016) and Lenormand (2002)). Therefore, understanding dispersal patterns is crucial for biodiversity conservation.

Landscape genetic approaches have been increasingly used to assess landscape influence on dispersal (Balkenhol *et al.*, 2016 ; Dyer, 2015a ; Manel *et al.*, 2003 ; Storfer *et al.*, 2007) because genetic data based inferences provide insights into effective movements that led to reproduction when inferences drawn from mark-recapture data or GPS tracks mostly identify current movements (Mateo-Sánchez *et al.*, 2015 ; Zeller *et al.*, 2018). Although advances have been achieved in landscape genetics in the last 15 years (Manel et Holderegger, 2013 ; Storfer *et al.*, 2010), there are still methodological and theoretical challenges, to analysing and interpreting genetic data especially (Balkenhol *et al.*, 2009a,b ; Dyer, 2015a).

Graph-theoretic approaches are particularly relevant when dispersal events occur between patchy populations forming a network (Greenbaum et Fefferman, 2017). A genetic graph is made of i) a set of nodes corresponding to gene pools sampled from different sites, and ii) a set of links connecting them through gene flow. The graph is basically a pairwise adjacency matrix with 0 and 1 reflecting absence or presence of links between populations, but the links can also be weighted by measures of genetic differentiation. In this case, it is often recommended to prune the complete graph, in other words to remove links between some node pairs, e.g. indirectly connected through intermediate nodes, to make the topology easier to visualise and to keep only the most relevant links in light of the study aim.

Genetic graphs are flexible tools that can be used in multiple fashions in landscape genetic studies, offering a great potential for inferring models of network flow (Murphy *et al.*, 2016). Indeed, a certain level of gene flow between two populations can result from direct exchanges of propagules and/or indirect exchanges through intervening populations in a stepwise way over several generations. Although considering only the genetic distance between two populations does not indicate whether gene flow occurred in a direct or indirect way, estimating genetic differentiation between a population pair conditionally upon other populations should make it possible to disentangle direct versus indirect gene flow between them (Dyer, 2015b). Hence, using the conditional independence principle (Magwene, 2001 ; Whittaker, 2009) can be a way to identify the precise topology of the dispersal network (i.e. the set of links depicting dispersal of propagules between populations), in other words identifying the set of edges that represents contributing connections among nodes (Murphy *et al.*, 2016).

Alternatively, a genetic graph can be used for quantifying landscape feature resistance to gene flow through distance-based analyses (Garroway *et al.*, 2011). Assessing the correlation between genetic distances and geographical or effective landscape distances is a way to identify the hypothesis that best fits the genetic data, and thus reflects landscape influence on gene flow, among several hypotheses of landscape feature resistance (Cushman *et al.*, 2006 ; Khimoun *et al.*, 2017 ; Peterman, 2018 ; Ruiz-Gonzalez *et al.*, 2015). Although such inferences are usually based upon complete matrices of

distance, several authors have suggested that reducing these matrices to a subset of population pairs may improve their robustness (Van Strien *et al.*, 2015 ; Van Strien, 2017 ; Wagner et Fortin, 2013 ; Zeller *et al.*, 2016). Graph pruning precisely involves selecting a subset of population pairs which therefore makes genetic graphs particularly relevant in this context.

Reducing the dataset by removing population pairs in order to improve inferences of landscape resistance is somehow counter-intuitive, but it lies on the following rationale. Assuming that dispersal is generally spatially limited, several theoretical models of populations genetics predict that measures of genetic differentiation are linearly and positively correlated with geographical distance, provided enough time has elapsed for this equilibrium pattern to become established (Guillot *et al.*, 2009 ; Kimura et Weiss, 1964 ; Slatkin, 1993 ; Wright, 1943). Models also predict that the spatial scale over which this pattern of Isolation by Distance (IBD) has reached its stationary state should increase with time following the establishment of populations (Slatkin, 1993). In other words, before complete equilibrium has been reached, IBD is only observed between nearby populations but not between more distant ones. Note that all these models assume that the landscape exerts an homogeneous effect on dispersal, and most of them exclude spatial variation in population density (Guillot *et al.*, 2009). However, most real landscapes are heterogeneous, and a common way to consider landscape feature suitability for dispersal is to replace Euclidean distances by landscape distances (e.g. cost-distances or resistance distances) in the analysis of population genetic structure (Balkenhol *et al.*, 2016 ; Coulon *et al.*, 2004 ; Peterman, 2018). If the isolation by landscape resistance (IBLR) model extends the IBD model in heterogeneous landscapes, its theoretical expectations have been less strongly investigated. Nevertheless, a model developed by McRae (2006) predicts a linear positive relationship between genetic differentiation and landscape distances. Here again, some time is needed for patterns of differentiation to reflect the influence of landscape features on dispersal, and model assumptions are more likely to be verified at shorter landscape distances before the complete equilibrium has been reached (McRae, 2006). Hence, better inferences of landscape resistance to gene flow may be obtained when selecting the subset of populations pairs that are within a certain spatial distance. This issue is critical as landscape genetic studies are frequently performed in human-shaped landscapes which have undergone recent modifications potentially affecting demography (Manel et Holderegger, 2013 ; Storfer *et al.*, 2010), but the relevance of the different graph pruning methods in this context has been rarely investigated.

In this study, we used a simulation approach to compare the relative efficiency of several graph pruning methods, genetic distances and analysis parameters of a genetic graph regarding two objectives in inferring network flow : i) identifying the precise topology of the dispersal network and ii) assessing the capacity of landscape distances to predict genetic distances. First, we assessed the efficiency of three kinds of criteria used for excluding graph links : geographical distance thresholds (leading to the exclusion of links corresponding to geographical distances larger than a threshold value), topology (involving topological constraints in graph pruning), and statistical inference of conditional independence based on genetic data (Dyer et Nason, 2004). Second, we compared some of the numerous genetic distances used to weight graph links (Murphy *et al.*, 2016) : F_{ST} (Keller *et al.*, 2013 ; Munwes *et al.*, 2010), D_{PS} (Naujokaitis-Lewis *et al.*, 2013 ; Keller *et al.*, 2013), genetic Euclidean distance (Excoffier *et al.*, 1992). Finally, we compared two common practices in distance-based analyses. The first one relies on the correlation between genetic and landscape distances corresponding with population pairs

that are directly connected in the genetic graph. The second one is based on the same correlation, but considering all pairwise genetic distances (between population pairs directly connected or not), by summing genetic distances along the shortest direct or indirect path between these populations on the graph. Dyer *et al.* (2010) revealed a higher correlation of this conditional genetic distance (cGD) than pairwise F_{ST} with landscape distance. Yet, these two matrices (cGD *vs* complete F_{ST} matrix) involved two different genetic distances (Euclidean genetic distance and F_{ST}) and two kinds of links (paths made of direct dispersal paths only *vs* direct plus indirect paths) at the same time, thereby introducing a confounding factor in the comparison. Therefore, the ability of these practices in distance-based analyses to infer landscape effects on dispersal still needs investigation.

2 Material & Methods

2.1 Landscape data

We simulated 10 landscapes using spatially correlated Gaussian random fields models (autocorrelation range : 10)(Schlather *et al.*, 2015) with NLMR package in R (Sciaini *et al.*, 2018). Land cover proportions were close to those encountered in agricultural landscapes dominated by crops and grasslands with small remaining forest fragments. Cost values were assigned to five cover types to simulate the dispersal capacities of a forest specialist species. These cost values and land cover proportions were the following : crops (cost : 60, proportion : 35 %), grassland (40, 35 %), forest (1, 15 %), shrubland (5, 7.5 %) and artificial areas (1000, 7.5 %). We based these costs on values already employed to analyse ecological connectivity in forest species (Gurrutxaga *et al.*, 2010 ; Schadt *et al.*, 2002), and their range (1-1000) matches that inferred from field data in other empirical studies on a wide range of taxa with contrasted dispersal capacities (Khimoun *et al.*, 2017 ; Pérez-Espona *et al.*, 2008 ; Ruiz-González *et al.*, 2014 ; Wang *et al.*, 2008).

The resulting landscapes were square raster grids of 3600 square kilometres with a resolution of 100 m. We randomly selected 50 population locations within the forest patches, separated by a distance larger than 3 km from one another. Ten population location distributions were created for each landscape in order to vary the cost-distance value distribution. Each population contained 30 individuals during the simulation.

2.2 Gene flow simulation

We used CDPOP (Landguth et Cushman, 2010) to simulate gene flow. Population size and sex-ratio (equal to 1) remained constant throughout the simulation of 500 generations. At each generation, individuals mate in their own population and juveniles may disperse to establish in other populations. The number of offspring per female follows a Poisson distribution ($\lambda = 3$). Once every population is occupied by 30 native or dispersing individuals, following individuals immigrating die. Mating is done with replacement for males only, and generations are non-overlapping. Individual genotypes were simulated for 20 loci with 30 alleles per locus, thereby emulating the frequent use of microsatellites in landscape genetic studies (Storfer *et al.*, 2010). Initial genotypes were assigned randomly at generation 0 as starting allele frequencies do not affect the overall final pattern of genetic differentiation (Graves *et al.*, 2013). There was no selection but mutations could occur (k-alleles mutation model, $\mu = 0.0005$).

Gene flow depended on simulated landscape resistance. With respect to the second objective, i.e. assessing the capacity of landscape distances to predict genetic distances, we aimed at simulating contrasted patterns of genetic structure in terms of spatial scale at which IBLR was observed. We first explored several simulation settings before retaining the following one. For every 100 combinations of landscape and distribution of populations, a landscape graph with 50 nodes was built. Each node corresponded to a habitat patch occupied by one population. Cost-distances (CD) between habitat patches were calculated following Adriaensen *et al.* (2003) as the accumulated cost along the least-cost path between each pair of habitat patches, using Dijkstra's algorithm on Graphab software (Foltête *et al.*, 2012a). Then, these CD values were used to weight the links of the graph, which initially had a complete topology. Using the edge-thinning method (Urban et Keitt, 2001), we removed links one by one in descending order of CD until we identified the link whose removal would have disconnected the graph into two components. The CD associated to this link was the "percolation threshold" (Rozenfeld *et al.*, 2008). During gene flow simulations, dispersal probabilities associated with links whose CD values were above $1.1 \times percolation.threshold$ were set to 0. $1.1 \times percolation.threshold$ is therefore the maximum dispersal distance. The resulting population networks were made of the set of direct dispersal paths which could possibly be followed by individuals and thus represented the potential dispersal network. It had a single component, thereby preventing single populations from being totally isolated, which is theoretically necessary for populations to survive (Allendorf *et al.*, 2007 ; Frankham *et al.*, 2004).

The decrease of individual dispersal probability according to CD was modeled by a negative exponential function (Clobert *et al.*, 2012 ; Hanski *et al.*, 2000 ; Urban et Keitt, 2001), such that : $p(CD) = e^{-\beta CD}$. β values were calculated such that the CD associated with a dispersal probability of 0.01 was equivalent to 5 % of the percolation threshold. Preliminary tests revealed that these settings resulted in proportions of migrants akin to those empirically described by Bowne et Bowers (2004).

For each simulation scenario (*i.e.* combination of landscape and distribution of populations), gene flow was simulated 10 times (1000 simulations in total). We used genotypes from generations 50 and 500 to construct genetic graphs. After each simulation, a "realised dispersal graph" was built. Its links were all the links that had been followed by at least one individual during the simulation. Genetic graphs built in order to recover the topology of the dispersal network were supposed to reproduce the topology of this realised dispersal graph.

2.3 Genetic graphs

We constructed genetic graphs using several pruning methods and genetic distances to weight the links (see table 1 for the list of combinations).

2.3.1 Pruning method

We pruned the genetic graphs using nine pruning methods based upon three criteria : i) geographical distance thresholds, ii) topology and iii) statistical inference.

First, we pruned graphs by removing all the links between nodes separated by a geographical distance larger than a given threshold. We used 4 thresholds : 10, 15, 20 and 30 km (GEO-10, GEO-15, GEO-20 and GEO-30, respectively). We chose this range of values to keep most graphs connected

(Naujokaitis-Lewis *et al.*, 2013) and because above 30 km, the resulting graphs were complete graphs given the size of the landscapes. We used thresholds in geographical distance units instead of cost-distance units because in practice researchers are not supposed to have previous knowledge of cost values associated with land cover types.

The second pruning criterion aimed at constructing graphs with a specific topology, in agreement with the species dispersal pattern hypothesised *a priori*. Genetic graphs were first given the topology of a Gabriel graph (GAB)(Arnaud, 2003). This type of graph in which only neighbouring populations are connected assumes a stepping-stones migration model. We also created minimum spanning trees (MST)(Naujokaitis-Lewis *et al.*, 2013) as they reflect the "backbone" of the dispersal network (Bunn *et al.*, 2000). Here again, we assumed that cost values of landscape features are unknown and GAB and MST connections were computed based on geographical distances as in Arnaud (2003) and Keller *et al.* (2013).

Third, graph pruning was based on a statistical procedure selecting the minimal set of links explaining population genetic structure. Based upon the conditional independence principle (Magwene, 2001 ; Whittaker, 2009), it is supposed to select links corresponding with direct dispersal paths and discard links associated with genetic similarities due to stepping-stones dispersal. We used the original method of Dyer et Nason (2004) but we also modified some of the calculation steps implemented in the popgraph package (cf. section C of supporting information). This method involves the calculation of a genetic covariance matrix from a genetic distance matrix that must have Euclidean properties, following Gower (1966). Therefore, we used a PCA-derived Euclidean genetic distance as well as the Euclidean genetic distance computed by default when using the `popgraph` package. From a strict mathematical point of view, the formula used to calculate the covariance c_{ij} from the distance d_{ij} between populations i and j is the following : $c_{ij} = -\frac{1}{2} \times (d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2)$ (Everitt et Hothorn, 2011 ; Smouse et Peakall, 1999), although the formula implemented in `popgraph` package is : $c_{ij} = -\frac{1}{2} \times (d_{ij} - d_{i\bullet} - d_{\bullet j} + d_{\bullet\bullet})$ (Dyer et Nason, 2004)($d_{i\bullet}$ and $d_{\bullet j}$ correspond respectively to the sum of distances over a column/row of the distance matrix). In our modified version (CI), we used the former formula while we also implemented the latter for comparative purposes (CI2). We also added a p-value adjustment, following sequential Bonferroni procedure (Holm, 1979), to limit type-I errors. In sum, we constructed genetic independence graphs relying upon the conditional independence principle using either our modified method (CI) or the original method of Dyer et Nason (2004) (CI2), with either PCA-derived Euclidean distance (PCA, cf section C of supporting information) or popgraph derived Euclidean genetic distance (PG), and either adjusting (ADJ) p-values or not (Table 1).

Finally, we constructed complete genetic graphs (COMP) because graph topologies sometimes include all the potential links between nodes (Naujokaitis-Lewis *et al.*, 2013). Besides, these complete graphs constituted a baseline to assess the relevance of graph pruning.

2.3.2 Genetic distance

Four genetic distances were used to weight the graph links. First, we used the linearised F_{ST} (*i.e.* $F_{ST}/(1\text{-}F_{ST})$), hereafter noted F_{ST} (Rousset, 1997). Second, we also used the "inter-population version" of D_{PS} (DPS), a genetic distance based on the dissimilarities of population allele pools computed as 1 - the proportion of shared alleles (Bowcock *et al.* (1994), cf. section D of supporting information).

This commonly used genetic distance is supposed to reflect recent gene flow changes (Murphy *et al.*, 2010b, 2016 ; Naujokaitis-Lewis *et al.*, 2013). Third, we computed a Euclidean genetic distance by first performing a PCA of the matrix of allelic frequencies and then computing the Euclidean distance between populations in the space defined by all independent principal components to derive a PCA-based Euclidean genetic distance (PCA), following Paschou *et al.* (2014) and Shirk *et al.* (2017a). Finally, we used the Euclidean genetic distance computed by default in `popgraph` package (PG).

Genetic independence graph links were weighted only with the two Euclidean genetic distances. The links of the other genetic graphs were weighted using the F_{ST}, the D_{PS} and the PCA-derived genetic distance. Every genetic distance, including that computed with the original `popgraph` method, was used to weight the links of the complete graphs in order to provide a baseline for the comparison of all pruning methods. In sum, 30 genetic graphs were constructed at generations 50 and 500 for every simulation (Table 1).

2.4 Graphs analyses

The dispersal pattern of the simulated species is reflected by the realised dispersal graph topology, and simulated gene flow was driven by the cost-distance values between populations. Hence, a genetic graph can be considered accurate if i) its topology reflects well the direct paths of the realised dispersal graph or if ii) the genetic distances derived from its links are highly correlated to the cost-distance values between populations.

2.4.1 Topology similarity analyses

We assessed the topological similarity between realised dispersal graphs and genetic graphs. To that purpose, we created contingency tables classifying the potential links of both types of graphs into two categories : absence or presence (see Fletcher *et al.* (2011)). Then, we calculated the Matthews correlation coefficient (Matthews, 1975), considered as a reliable index of binary classification quality because it takes into account all the elements of the contingency table and is calculated with respect to a random baseline (Baldi *et al.*, 2000). A Matthews correlation coefficient of 1 is reached when both graphs are identical, whereas a 0 value means that they are no more similar than if they were built by selecting links randomly. In our case, a large value indicates that a genetic graph recovers well the realised dispersal graph topology.

2.4.2 Distance-based analyses

We calculated the Mantel correlation coefficients r (Mantel, 1967) between genetic distances and CD values. For each simulated genetic dataset, we considered three sets of genetic distances : i) the subset of "raw" genetic distances associated with population pairs directly connected in genetic graphs (see Van Strien *et al.* (2015)), ii) the graph-based genetic distances between every population pair, calculated as the sum of link weights along the shortest path between nodes (an extended use of the "cGD" introduced by (Dyer *et al.*, 2010) to other types of genetic graphs) and iii) the full set of "raw" genetic distances between every population pair derived from complete graphs. Large r values indicate that the set of genetic distances derived from genetic graphs reflects well the simulated landscape effects on gene flow. This approach is commonly used in landscape genetics (Graves *et al.*, 2013 ; Shirk *et al.*, 2017b ; Van Strien *et al.*, 2015 ; Zeller *et al.*, 2016) as the use of Mantel correlation coefficients is

relevant when the hypothesis can only be formulated in terms of distances (Legendre et Fortin, 2010). We focused on the correlation coefficient values rather than on statistical significance because it has been shown to provide reliable results when few hypotheses are compared (Shirk *et al.*, 2017b ; Zeller *et al.*, 2016). Besides, type-I error rate is high with Mantel tests (Balkenhol *et al.*, 2009b), which limits their relevance.

2.5 Simulation results ordination

We performed a large number of simulations by varying landscapes and population locations. Given our objectives, we intended to reproduce in our simulations the cases I and IV from the hypothetical classification of the relationship between genetic and geographical distances proposed by Hutchison et Templeton (1999). Although case I corresponds to an equilibrium between gene flow and drift over the whole region, case IV corresponds to a transient situation where this equilibrium has been reached at a smaller spatial scale because of the time lag of the genetic response. The case I is characterised by a linear increase of genetic differentiation with increasing geographical distances over the whole region considered. In contrast, the case IV depicts this positive correlation up to a certain geographical distance threshold above which the relationship flattens out. This distance threshold was defined by Van Strien *et al.* (2015) as the distance of maximum correlation (DMC), i.e. the geographical distance threshold below which the subset of population pairs maximises the linear correlation between genetic and geographical distances. The highest DMC occurs for case-I patterns of IBD as it should be equal to the maximum inter-population distance whereas it decreases when case-IV patterns of IBD are observed. Thereby, we used the DMC as a proxy of the spatial scale above which equilibrium has not been reached, and below which genetic structure depends both on gene flow and drift, which does not necessarily mean that equilibrium has been reached. Considering the linear positive relationship between genetic differentiation and landscape distances expected under an IBLR model (McRae, 2006), we aimed at reproducing the cases I and IV defined by Hutchinson and Templeton but considering cost-distances instead of geographic distances.

We determined the DMC by iteratively computing the Mantel correlation coefficients between i) the CD values driving the simulation and ii) the F_{ST} and the D_{PS}, using increasing threshold values. We also visualised scatter plots of the relationship between genetic distances and CD to identify the type of IBLR pattern corresponding to each simulation and time step. We could thus check for potential biases in Mantel r values. Most graph analyses were performed using `graph4lg` package in R (Savary *et al.*, 2021b).

To extract the main trend among the results of 1000 simulations, we applied a Principal Component Analysis to eight variables describing the simulation parameters (proportion of migrants per population, CD threshold used to build the potential dispersal graph, number of links in the realised dispersal graph, mean CD covered by migrants) and their genetic output (DMC computed at generation 50 and 500 for F_{ST} and D_{PS}). These variables were averaged over the 10 runs for each configuration combining a landscape and a population spatial distribution. We carried out a hierarchical clustering from the PCA factors in order to distinguish the main trend in the PCA results.

Graph name	Pruning method	Genetic distance
COMP-FST	No pruning	F_{ST}
COMP-DPS	No pruning	D_{PS}
COMP-PCA	No pruning	PCA-derived Eucl. dist.
COMP-PG	No pruning	Eucl. gen. dist. (from `popgraph`)
GEO-10-FST	Geo. dist. threshold (10-km)	F_{ST}
GEO-10-DPS	Geo. dist. threshold (10-km)	D_{PS}
GEO-10-PCA	Geo. dist. threshold (10-km)	PCA-derived Eucl. dist.
GEO-15-FST	Geo. dist. threshold (15-km)	F_{ST}
GEO-15-DPS	Geo. dist. threshold (15-km)	D_{PS}
GEO-15-PCA	Geo. dist. threshold (15-km)	PCA-derived Eucl. dist.
GEO-20-FST	Geo. dist. threshold (20-km)	F_{ST}
GEO-20-DPS	Geo. dist. threshold (20-km)	D_{PS}
GEO-20-PCA	Geo. dist. threshold (20-km)	PCA-derived Eucl. dist.
GEO-30-FST	Geo. dist. threshold (30-km)	F_{ST}
GEO-30-DPS	Geo. dist. threshold (30-km)	D_{PS}
GEO-30-PCA	Geo. dist. threshold (30-km)	PCA-derived Eucl. dist.
GAB-FST	Topological (Gabriel graph, geo. dist.)	F_{ST}
GAB-DPS	Topological (Gabriel graph, geo. dist.)	D_{PS}
GAB-PCA	Topological (Gabriel graph, geo. dist.)	PCA-derived Eucl. dist.
MST-FST	Topological (MST, geo. dist.)	F_{ST}
MST-DPS	Topological (MST, geo. dist.)	D_{PS}
MST-PCA	Topological (MST, geo. dist.)	PCA-derived dist.
CI-PCA	Condit. indep.	PCA-derived Eucl. dist. (covar. from squared dist.)
CI-ADJ-PCA	Condit. indep.	PCA-derived Eucl. dist. (covar. from squared dist.) with Holm-Bonferroni adjustment
CI-PG	Condit. indep.	Eucl. gen. dist. (from `popgraph`, covar. from squared dist.)
CI-ADJ-PG	Condit. indep.	Eucl. gen. dist. (from `popgraph`, covar. from squared dist.) with Holm-Bonferroni adjustment
CI2-PCA	Condit. indep.	PCA-derived Eucl. dist. (covar. from dist.)
CI2-ADJ-PCA	Condit. indep.	PCA-derived Eucl. dist. (covar. from dist.) with Holm-Bonferroni adjustment
CI2-PG	Condit. indep.	Original `popgraph` method
CI2-ADJ-PG	Condit. indep.	Original `popgraph` method with Holm-Bonferroni adjustment

Table 1 – Genetic graph construction parameters. Cf. section B of supporting information for a glossary of the acronyms

3 Results

3.1 Simulation results

For each simulation, the realised dispersal graph was connected meaning that each population exchanged migrants with at least another population during the first 50 generations. The overall proportion of dispersing individuals over 500 generations ranged from 13.3 % to 24.1 %. Although all the landscapes were simulated with the same parameters and populations were located randomly in habitat patches, values of the maximum dispersal distance exhibited substantial variations (from 1321 to 3564 CD units). Consequently, the number of links in dispersal graphs ranged from 155 to 858 links (Figure 10), depicting a wide range of gene flow patterns.

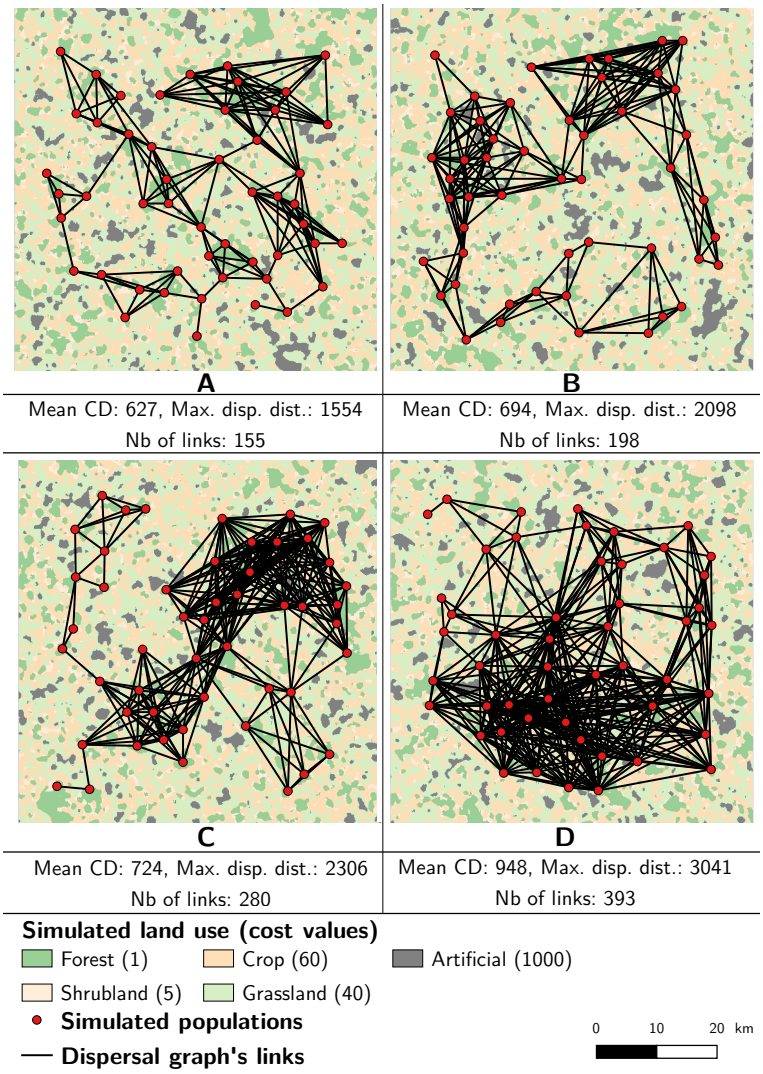


FIGURE 10 – Four contrasted landscape/distribution of populations configurations exhibiting large differences in the number of links in the dispersal graph. Mean CD between populations, maximum dispersal distance in CD units and number of links followed by individuals are indicated in each landscape.

Although a case-IV pattern of IBLR was often observed at generation 50 (Figure 12), DMC values increased from generation 50 to 500 suggesting that genetic structure reached its stationary state at increasing spatial scale over time. Note that DMC values were always larger than the maximum CD over which dispersal was possible. PCA results evidenced these variations (Figure 11). The first

principal component (56.6 % of the variance) was positively correlated with the DMC (based on F_{ST} and D_{PS} values at generation 50 in particular), the maximum CD threshold, the number of links in the realised dispersal graph and to a lesser extent with the mean CD covered by individuals and the proportion of migrants. The second principal component explained a lower proportion of variance (24.9 %) and mainly reflected differences between simulations due to the interplay between the number of links, the proportion of dispersing individuals (negatively correlated) and the mean CD covered by dispersing individuals (positively correlated).

Three main clusters of landscape/distribution of populations configurations were identified through the hierarchical clustering of the PCA results (Figure 11). The first cluster is characterised by low numbers of links in dispersal graphs because of low maximum dispersal distances and by low DMC at generation 50 while the third cluster is characterised by high DMC, high numbers of links and high maximum dispersal distance. In the second cluster, dispersal graphs counted many links, the proportions of migrants were high and the DMC took intermediate values.

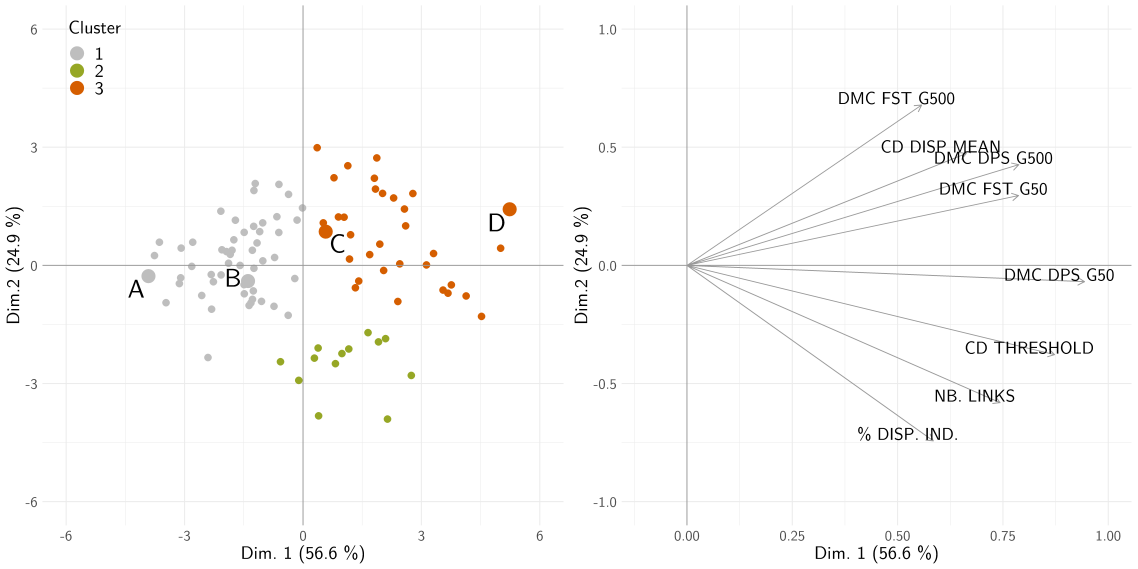


FIGURE 11 – Principal Components Analysis of eight variables (100 observations) describing the simulation results. Configurations A to D are also displayed in figure 10.

The first and third clusters included configurations in which case-IV and case-I patterns of IBLR take form at generation 50, respectively. One objective of this study was to compare the usefulness of genetic graphs when gene flow influences genetic structure at the complete landscape scale (case I) or at smaller scale (case IV). In addition, the relative performance of graph construction and analysis methods exhibited marginal variation along the second principal component. Thus, for the sake of brevity, we chose to describe the results of the subsequent analyses based on four configurations (A, B, C, D ; displayed on the figures 10, 12 and 11) along the first principal component which defines a gradient between these two opposite patterns. Configuration A was typical of a case-IV pattern of IBLR (at generation 50 in particular) and D of a case-I pattern (Figure 12). Configurations B and C corresponded to intermediate situations.

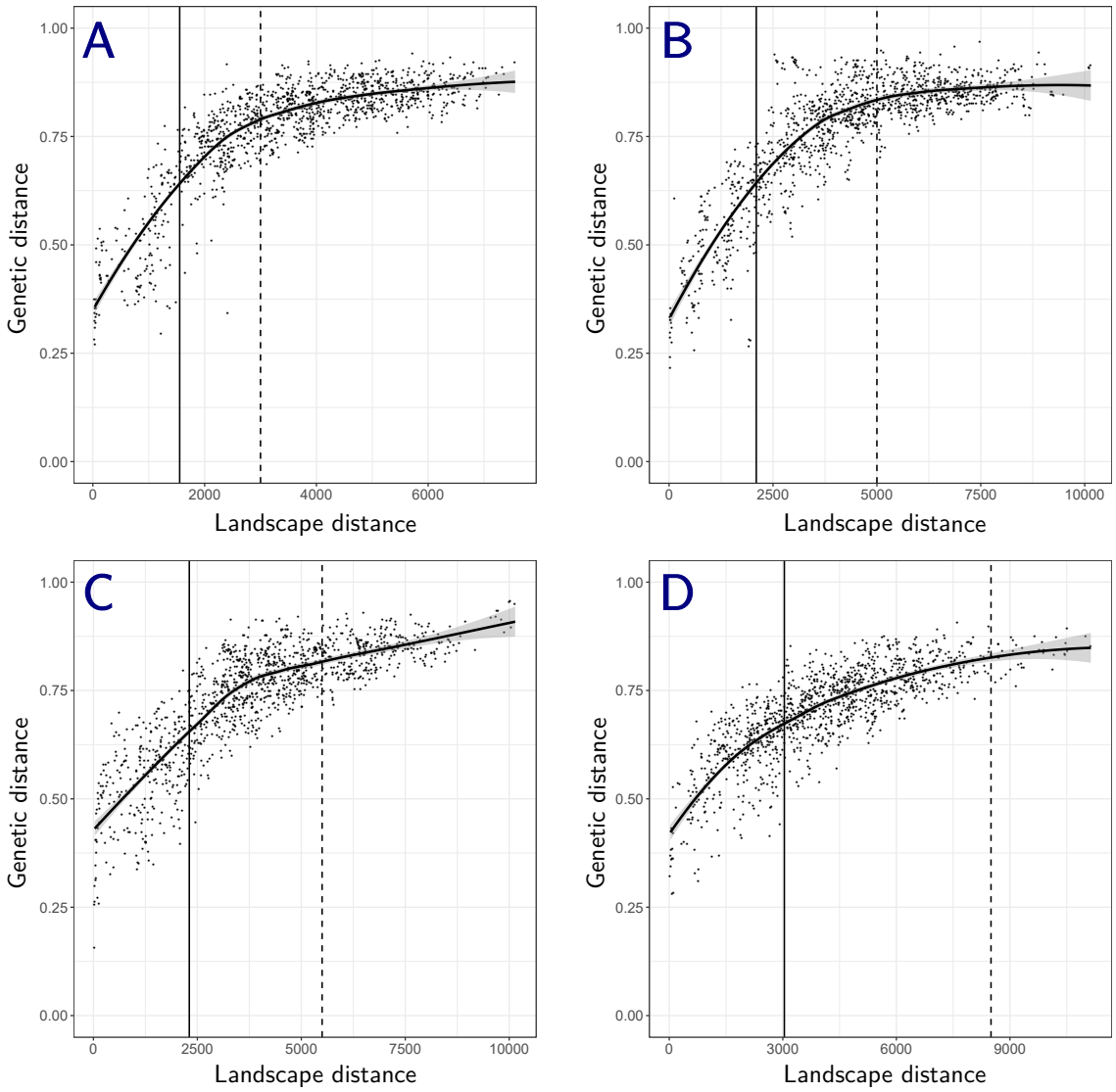


FIGURE 12 – Scatter plots of the genetic distance (D_{PS}) plotted against cost-distance at generation 50. Cases A to D illustrate the gradient of IBD patterns (from type-IV to type-I). Solid vertical lines indicate the maximum dispersal distance, dashed lines indicate the DMC. See figure A1 for the same figure with the F_{ST}.

3.2 Genetic graphs

3.2.1 Topology similarity analyses

Depending on the pruning method used, the mean number of links in the genetic graphs was highly variable as it ranged from 49 (MST) to 802 (GEO-30)(Table 2 and figure A3). In contrast, the number of links in the realised dispersal graphs, which genetic graph topologies were supposed to reproduce, were 155, 198, 280 and 393 in the configurations A to D, respectively.

When the graphs were pruned with methods based on geographical distance thresholds or on topological constraints, their number of links was stable over generations given these methods do not rely on genetic data. The number of links of MST and Gabriel graphs (49 and around 90 links, respectively) was also highly stable among configurations and much lower than the number of links in realised dispersal graphs (Table 2). As a consequence, these topological pruning methods never performed well in reflecting realised dispersal graph topology (correlation values from 0.29 to 0.54 ; Table 2).

On the contrary, as pruning based on conditional independence takes into account genetic data, the number of links in the genetic independence graphs varied strongly, from 58.5 to 519.7 in average. The number of links in these graphs tended to increase from generation 50 to generation 500 even if the number of realised direct dispersal paths was stable, but this trend was much lower when using genetic distances (CI2), as in the original `popgraph` method, than squared genetic distances (CI), as in our modified version. As a consequence, the ability of a genetic graph topology to reflect the topology of the realised dispersal graph was fairly stable between generations G50 and G500 when using genetic distances (CI2), whereas it decreases between G50 and G500 when using squared genetic distances (CI ; Table 2). Adjusting the p-values to assess the significance of the partial correlations almost reduced the number of links by a factor of 2. When the covariance between allelic frequencies was calculated using squared genetic distances (CI), the number of links was consistently larger than when using genetic distances (CI2). In some cases, graphs obtained using the latter formula were not connected, especially when p-values were adjusted to assess the significance of partial correlation values.

Genetic graphs pruned with geographical distance thresholds presented the topology closest to that of realised dispersal graphs in all configurations (correlation values above 0.6) except the least connected one (*i.e.* configuration A)(Table 2). The closest the geographical distance threshold (GEO) from the maximum dispersal distance (CD threshold converted into Euclidean distance) used in the simulations, the better the genetic graph reflects the topology of the realised dispersal graph. In contrast, for the dispersal graphs created in configuration A, which counted fewer links (Figure A3), the highest correlation values were reached with pruning methods based on conditional independence. Correlation values above 0.6 were reached every time covariance was computed from genetic distances (CI2), and only at generation 50 with p-value adjustment when covariance was computed from squared genetic distances (CI) with our modified method. For the configuration B, correlation values above 0.6 were also reached when independence genetic graphs were pruned by computing the covariance from genetic distances (CI2) whatever the type of genetic distance used (PCA or PG). For the configuration C, the original `popgraph` method (CI2-PG) enabled to reach a correlation value of 0.6. Note that when computing the covariance from squared genetic distances (CI), the genetic graphs included links between population pairs not connected in the dispersal graph (Figure A3). p-value adjustment reduced the number of these false long-distance links.

Overall, genetic graphs which succeeded in accurately reproducing the topology of the realised dispersal graphs counted much the same number of links as the dispersal graph (Table 2). However, this condition is not sufficient to explain the correlation values given that in some cases, relatively low correlation values were obtained with a similar number of links to the realised dispersal graph (e.g. CI-ADJ-PG, configuration A at G500 : Matthews correlation coefficient = 0.52, with a difference in the number of links between graphs equal to 2.4 ; Table 2).

3.2.2 Distance-based analyses

The correlation coefficients between genetic distances and CD separating population pairs which were directly connected in the genetic graphs were highly variable as they ranged from 0.47 to 0.86 in average at generation 50 (Figure 13, see figure A4 for generation 500). In all cases, Mantel correlation coefficients between genetic distances and geographical distances were lower than those between genetic distances and CD, showing that the isolation by landscape resistance model better explained

genetic structure than the isolation by distance model did, as expected from our simulations.

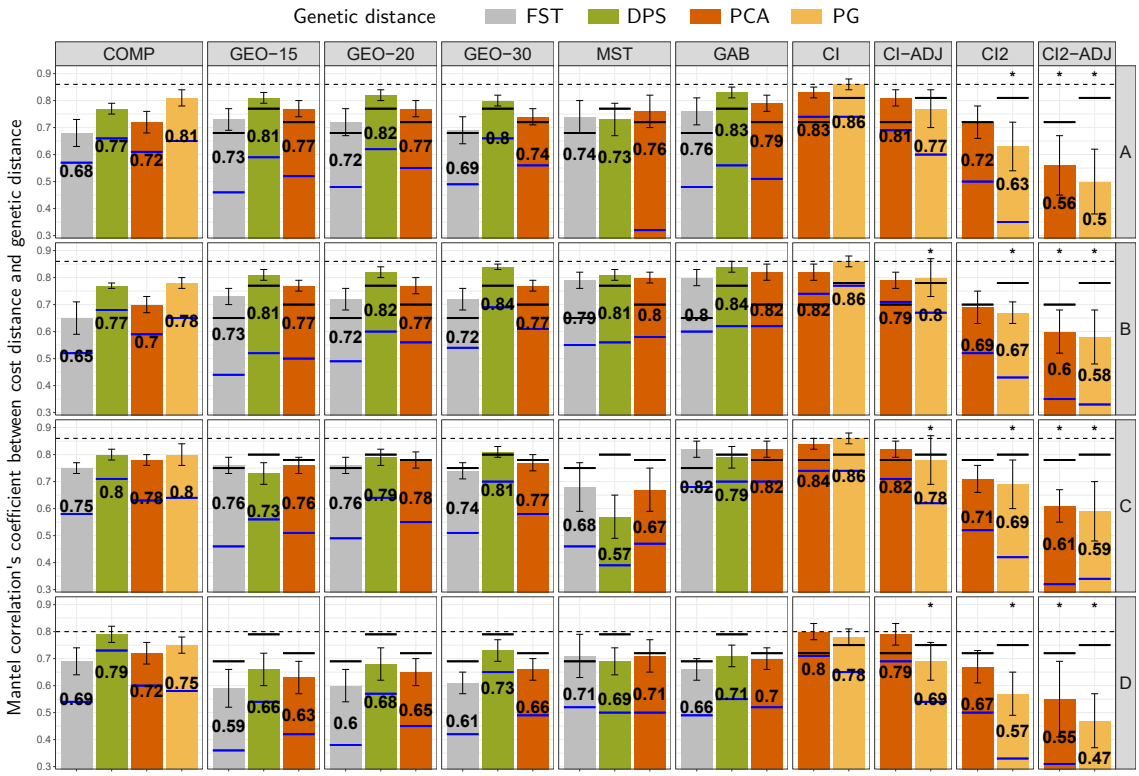


FIGURE 13 – Mantel correlation between genetic distances and cost-distances separating nodes directly connected on the genetic graphs, according to the type of genetic distance and the pruning method at generation 50 (see table 1 for the graph names). Mean ± SD values were computed for the 10 runs simulated in each scenario. Blue bars refer to the correlation coefficient between genetic distance and geographical distance, when it is above 0.3. Black bars refer to the correlation coefficient obtained using every population pair to compute the correlation. When black and blue bars overlap, the bar is black. Stars indicate graphs counting several components. The dashed line indicates the maximum r value obtained for each configuration.

When a case-I pattern of IBLR was observed at generation 50 (configurations C and D), larger correlation coefficients were always those obtained when using genetic distance matrices derived from complete graphs instead of pruned graphs, except for genetic independence graphs built with our modified method (CI). Conversely, when a case-IV pattern of IBLR was observed at generation 50 (configurations A and B), correlation coefficients were almost always larger when genetic distance values were those associated with the links of a pruned graph, whatever the pruning method, than when they were associated with all population pairs (Figure 13). At generation 500, there were few differences between configurations (A to D) given that the relationship between genetic differentiation and cost-distance almost linearised over time in all cases, and higher correlation between genetic distances and CD were observed with a complete graph, compared with a pruned graph, except for genetic independence graphs based on squared genetic distance (CI) which still performed better (Figure A4).

The largest correlation coefficients were always reached when selecting genetic distance values from genetic independence graphs built without p-value adjustment and based on the computation of the covariance from squared genetic distances (CI). When computing covariance from genetic distances (CI2), as in the original `popgraph` method, correlation coefficients were much lower. This method never strengthened the correlation obtained with the corresponding complete genetic distance matrices (COMP-PG or COMP-PCA), and it provided the lowest correlation values (Figure 13). For

configurations A and B (case IV), correlation coefficients obtained when selecting population pairs from a Gabriel graph were slightly larger than correlation obtained selecting population pairs from an MST or by using geographical distance thresholds (Figure 13). In most cases, correlation coefficients between genetic distances and CD were lower when the genetic distance was the F_{ST} rather than the D_{PS} or Euclidean genetic distances.

When we computed the Mantel correlations between CD and graph-based genetic distances, correlation coefficients values ranged from 0.57 to 0.93 at generation 50 (Figure 14, see figure A5 for a similar variation at G500). However, differences between configurations were less pronounced when analysing the correlation this way, even at generation 50.

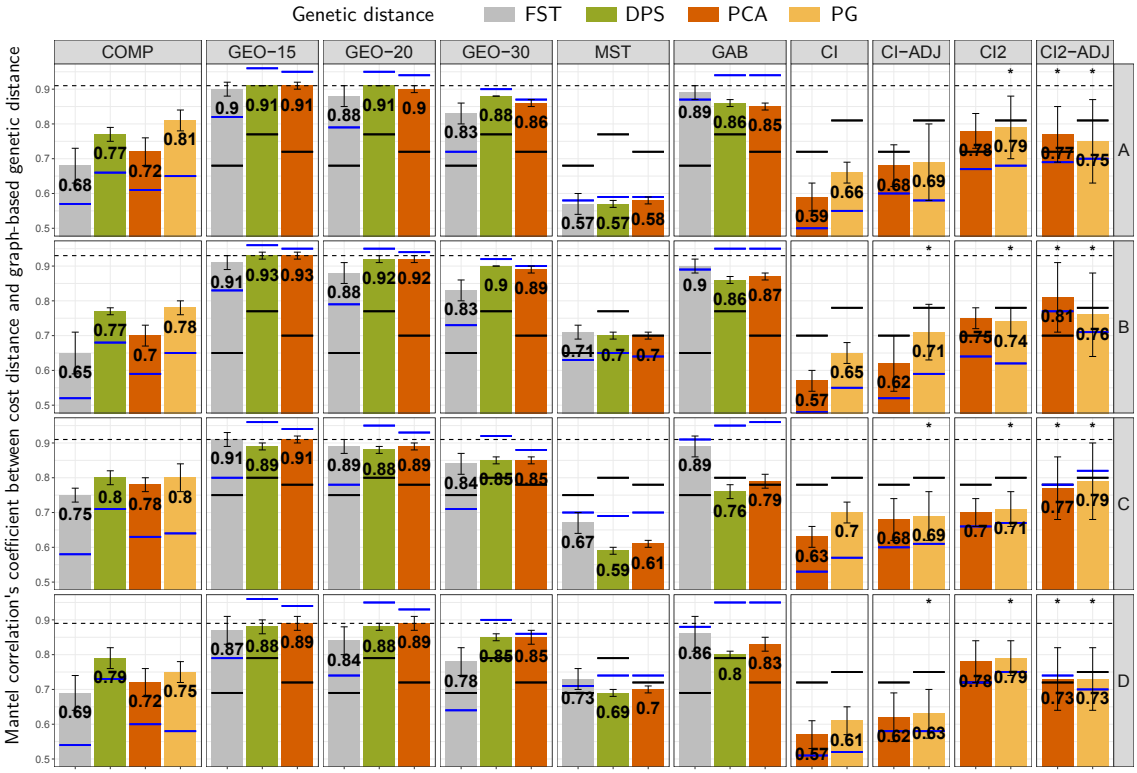


FIGURE 14 – Mantel correlation between conditional genetic distances and cost-distances separating nodes on the genetic graphs, according to the type of genetic distance and the pruning method at generation 50 (see table 1 for the graph names). Mean $\pm SD$ values were computed for the 10 runs simulated in each scenario. Blue bars refer to the correlation coefficient between genetic distance and geographical distance, when it is above 0.5. Black bars refer to the correlation coefficient obtained using every population pair to compute the correlation. When black and blue bars overlap, the bar is black. Stars indicate graphs counting regularly several components. The dashed line indicates the maximum r value obtained for each configuration.

The correlation coefficient took its largest values when the graphs were pruned using a geographical distance threshold or a topological constraint. However, when computing graph-based genetic distances from these graphs, the correlation coefficients between these genetic distances and geographical distances were higher than those computed between the same genetic distances and CD values, supporting an IBD model over an IBLR model despite our simulation settings. The only exception was when this distance was computed from F_{ST} values (Figure 14).

Conversely, when computing graph-based genetic distances from independence graphs, these distances were more correlated to CD than to geographical distances (Figure 14). The correlation co-

efficients were higher in that case when the pruning relied on the calculation of the covariance from genetic distances (CI2) rather than from squared genetic distances (CI). However, the correlation between Euclidean genetic distances and CD was higher when considering the complete graph instead of graph-based genetic distances, except when computing covariance from PCA-based genetic distances (CI2-PCA). Besides, we reproduced the result described by Dyer *et al.* (2010) who showed that landscape influence on gene flow was frequently better recovered when using these graph-based genetic distances derived from independence graphs (CI2) than when using the complete matrix of F_{ST}. Moreover, scatter plots created using graph-based genetic distance values revealed that summing genetic distances in case-IV pattern of IBLR tends to mask the fact that the relationship between genetic differentiation and CD flattens out beyond a CD threshold (Figure A2).

Combination	**A**		**B**		**C**		**D**	
a) Dispersal graphs								
Max. disp. dist.	1554		2098		2306		3041	
Max.disp.dist.(km)	13.1		15.1		18.2		20.8	
Nb. disp. paths	155		198		280		393	
b) Number of links								
Generation	G50	G500	G50	G500	G50	G500	G50	G500
GEO-10	**127.0**	***127.0***	**120.0**	***120.0***	120.0*	*120.0**	112.0*	*112.0**
GEO-15	274.0	*274.0*	**237.0**	***237.0***	**253.0**	***253.0***	**246.0**	***246.0***
GEO-20	455.0	*455.0*	**384.0**	***384.0***	**408.0**	***408.0***	**413.0**	***413.0***
GEO-30	802.0	*802.0*	693.0	*693.0*	746.0	*746.0*	757.0	*757.0*
GAB	92.0	*92.0*	89.0	*89.0*	96.0	*96.0*	92.0	*92.0*
MST	49.0	*49.0*	49.0	*49.0*	49.0	*49.0*	49.0	*49.0*
CI-PCA	274.4	*427.7*	269.7	*412.1*	283.8	*435.0*	305.7	*458.8*
CI-ADJ-PCA	**108.2**	*246.5*	107.7	*234.6*	109.0	*251.5*	114.8	*275.8*
CI-PG	325.5	*441.3*	330.9	*397.9*	342.8	*447.2*	375.3	*519.7*
CI-ADJ-PG	**96.8**	*152.6*	101.2*	*152.5*	95.7*	*149.6**	102.2*	*165.3*
CI2-PCA	**132.6**	***158.9***	**134.6**	***161.2***	141.8	*176.3**	146.9	*204.9*
CI2-ADJ-PCA	**76.7***	***85.8****	71.3*	*81.4**	65.0*	*81.5**	60.9*	*70.0**
CI2-PG	**135.4***	***148.1***	**134.3***	***149.2****	137.5*	***155.4***	140.5*	*153.9*
CI2-ADJ-PG	**77.8***	***80.5****	72.5*	*75.6**	63.7*	*72.9**	59.4*	*58.5**
c) MCC								
GEO-10	**0.64**	***0.64***	**0.68**	***0.68***	0.56*	*0.56**	0.44*	*0.44**
GEO-15	0.59	*0.59*	**0.73**	***0.73***	**0.68**	***0.68***	**0.62**	***0.62***
GEO-20	0.47	*0.47*	**0.60**	***0.60***	**0.68**	***0.68***	**0.69**	***0.69***
GEO-30	0.28	*0.28*	0.38	*0.38*	0.44	*0.44*	0.51	*0.51*
GAB	0.54	*0.54*	0.54	*0.54*	0.41	*0.41*	0.39	*0.39*
MST	0.49	*0.49*	0.43	*0.43*	0.36	*0.36*	0.29	*0.29*
CI-PCA	0.42	*0.25*	0.39	*0.24*	0.33	*0.19*	0.26	*0.14*
CI-ADJ-PCA	**0.60**	*0.34*	0.53	*0.33*	0.43	*0.25*	0.34	*0.18*
CI-PG	0.41	*0.34*	0.39	*0.37*	0.35	*0.32*	0.27	*0.21*
CI-ADJ-PG	**0.64**	*0.52*	0.55*	*0.48*	0.45*	*0.41**	0.35*	*0.29*
CI2-PCA	**0.77**	***0.69***	**0.69**	***0.65***	0.58	*0.55**	0.46	*0.40*
CI2-ADJ-PCA	**0.65***	***0.65****	0.55*	*0.57**	0.42*	*0.46**	0.33*	*0.33**
CI2-PG	**0.80***	***0.75***	**0.72***	***0.69****	0.59*	***0.60***	0.48*	*0.45*
CI2-ADJ-PG	**0.66***	***0.65****	0.56*	*0.56**	0.42*	*0.44**	0.32*	*0.31**

TABLE 2 – Topologies of the dispersal graphs and genetic graphs. a) Topologies of the dispersal graphs. Maximum dispersal distances are given in CD units and in kilometres (conversion obtained after performing the linear regression of CD values against geographical distances values). b) Number of links in the genetic graphs. c) Matthews correlation coefficients assessing the topology similarity of both types of graphs (genetic and dispersal), according to the type of genetic distance and the pruning method in the four landscape/distribution of populations configurations and at two generations (see table 1 for the graph names). Mean values and standard deviations were computed for the 10 runs simulated in each scenario but standard deviations are not displayed because they were negligible. Matthews correlation coefficients above 0.6 and corresponding numbers of links in the genetic graphs are displayed in bold. Values referring to generation 500 are displayed in italics. Stars indicate that some of the ten graphs created for each combination were not connected.

4 Discussion

In this study, we demonstrated that the ability of different pruning methods to identify the precise topology of dispersal networks is highly variable, especially between methods based either on geographical distance or genetic independence criteria. In addition, we highlighted the importance of graph pruning for assessing landscape effects on gene flow in non-equilibrium situations. We provide users with rough guidelines that are schematically illustrated in Figure 15.

4.1 When and why to prune a genetic graph ?

On the one hand, graph pruning is hardly avoidable when the objective is to identify the topology of the direct dispersal network followed by individuals (Figure 15). Indeed, except in very rare panmictic configurations or when the study area is very small, dispersal events are not expected between all population pairs (Kimura et Weiss, 1964). On the other hand, our results show that the relevance of graph pruning for inferring landscape resistance to gene flow depends on the scale at which gene flow effect on differentiation is detectable. When an IBLR pattern is observed at the scale of the entire landscape, graph pruning hardly ever improved the inference made from a complete graph, except when pruning relied on the conditional independence principle and squared genetic distances (CI). In contrast, when this pattern is observed up to a limited scale, graph pruning strengthened the linear correlation between genetic distances and cost-distance values driving the simulation, suggesting that graph pruning is useful to infer landscape resistance to gene flow in this situation.

Migration-drift equilibrium is less likely to be reached for the complete set of sampled population pairs when dispersal distances are short regarding the study area and/or landscapes have undergone recent modifications. Such non-equilibrium situations correspond to the case-IV pattern of IBD proposed by Hutchison et Templeton (1999). It has been observed in several theoretical (Slatkin, 1993) and empirical studies (Ciofi *et al.*, 1999 ; Clegg et Phillimore, 2010 ; Hänfling et Weetman, 2006 ; Hutchison et Templeton, 1999 ; Kuehn *et al.*, 2003 ; Méndez *et al.*, 2011) and is expected to be frequent in landscape genetic studies dealing with dynamic human-shaped landscapes (Manel et Holderegger, 2013 ; Storfer *et al.*, 2010). In such situations, not pruning a genetic graph might be problematic if the objective is to infer landscape resistance to gene flow. Indeed, such inferences may involve genetic distances that do not reflect the long-term effect of landscape on genetic structure. Wagner et Fortin (2013) suggested that considering a subset of population pairs could increase the power of distance-based analyses in landscape genetics. Indeed, a few studies reported stronger relationships between landscape structure and population genetic structure using this approach (Angelone *et al.*, 2011 ; Coster *et al.*, 2015 ; Jaquiéry *et al.*, 2011 ; Keller *et al.*, 2013 ; Van Strien *et al.*, 2015). Most of them used geographical thresholds somehow linked to maximum dispersal abilities and they considered between populations distances while ignoring their spatial arrangement (but see Keller *et al.* (2013) for an explicit graph-based approach). However, Van Strien (2017) argued that population topology (*i.e.*, the arrangement of populations throughout a landscape) should be better incorporated in link-based landscape genetic studies. In this context, graph-theoretic methods offer great opportunities in link selection (Dyer, 2015b), but to date, the relative performance of the wide range of graph pruning methods had not been assessed.

4.2 How to prune a genetic graph to identify the precise topology of the dispersal network ?

Our results show that pruning methods based on topological constraints are rarely suitable for recovering the topology of the dispersal network. Indeed, minimum spanning trees do not include any cycles and Gabriel graphs cannot take into account the presence of some long-distance dispersal events between population pairs. Because topological constraints generally impose a constant number of links given the number of populations, they lack ecological significance (Serrano *et al.*, 2009).

When dispersal capacities of the study species are precisely known, pruning based on geographical distance thresholds always makes it possible to recover well the topology of the realised dispersal graphs, and this method is often the best one. Of course, estimating dispersal capacities is a difficult task (Schneider, 2003 ; Van Dyck et Baguette, 2005), and even if several thresholds can be tested in empirical studies, it is impossible to determine which genetic graph reflects best the true dispersal pattern. As expected, our results showed that a serious underestimation of maximum dispersal distance led to a disconnected graph, thus wrongly suggesting the existence of landscape barriers to dispersal. The similarity between genetic graphs pruned using distance thresholds and realised dispersal graphs may also depends on the correlation strength between cost distances (CD) and geographical distances, which is sometimes high (Marrotte et Bowman, 2017). However, geographical distance is not always a good proxy of CD (Balkenhol *et al.*, 2009b), for instance when a barrier prevents dispersal between close populations or less frequently when large geographical distances are covered by dispersing individuals because they correspond to low CD values. Ignoring these rare long distance dispersal event may be problematic given their ecological and evolutionary consequences (Clobert *et al.*, 2012 ; Greenbaum et Fefferman, 2017 ; Nathan *et al.*, 2003).

Our results also suggest that building genetic independence graphs is a suitable option to recover dispersal network topology when dispersal distances are unknown, especially in less connected configurations (A and B, Table 2). In the latter case, these results can be as satisfactory as when dispersal distance is known. The topology of the dispersal graphs is better recovered when the covariance is computed with the original `popgraph` method from genetic distances (CI2) instead of squared genetic distances (CI). In the latter case, the presence of links in genetic graphs that were never followed by dispersing individuals during the simulations indicates that it does not identify direct dispersal paths reliably. The variability in the number of links among independence graphs was mainly due to the covariance formula, the genetic distance, the p-value adjustment, and to a lesser extent the generation. In contrast, the expected large difference in the number of links between very different connectivity configurations (A and D) was not observed.

It may appear puzzling to infer single-generation dispersal events from genetic structure shaped by multi-generational dispersal. However, it seems to be the promise behind the genetic independence graphs as the conditional independence is supposed to recover the actual route of propagules (Dyer, 2015b). Though our results seem to support this idea in some conditions, further research is needed on this pruning method. For instance, we expect this method to perform poorly when sampling is incomplete, which is often the rule in empirical studies, but the potential bias this introduces in the inferences remains to be estimated.

4.3 How to prune a genetic graph to infer landscape resistance to dispersal ?

Genetic graphs reflecting precisely the dispersal network topology are not necessarily those that enable to quantify well landscape effects on dispersal. Indeed, the distance of maximum correlation (DMC) was always larger than the maximum dispersal distance in the simulation (Figure 12), suggesting that the set of genetic distance values to include in link-based analyses, should not be restricted to direct dispersal paths. Migration-drift equilibrium may become established between populations separated by distances beyond dispersal capacities, as expected under the stepping-stones model of Slatkin (1993), because they may exchange genes over several generations even if not connected by direct dispersal paths. We suggest including such population pairs in link-based inferences because their genetic divergence should reflect landscape influence on gene flow. Our view contrasts with the exclusive use of population pairs that are within migration range of each other recommended by others when assessing the effect of landscape on gene flow (Keller *et al.*, 2013 ; Van Strien *et al.*, 2015 ; Van Strien, 2017). Therefore, a reliable pruning method to estimate landscape resistance to gene flow should identify population pairs whose genetic differentiation reflects the long term gene flow between them.

In this context, we do not advise using pruning methods based on fixed criteria (*i.e.* geographical distances or topological constraints), even if they provided correlations between genetic distances and CD that were slightly lower than the maximum correlation obtained for a given configuration at generation 50, especially when using D_{PS} (Figure 13). Indeed, these methods seem inappropriate because the spatial scale of IBLR changes over time (McRae, 2006). Pruning methods relying on genetic data and statistical inference seem to provide the best inference of landscape resistance as they can account for the dynamic nature of IBLR. Indeed, in case-I and case-IV patterns of IBLR, the correlation between genetic distances associated with the genetic graph links and CD values was maximised when using pruning methods based upon the conditional independence principle. However, this result only holds when computing the covariance from squared genetic distances to stick with mathematical requirements (Everitt et Hothorn, 2011 ; Magwene, 2001 ; Smouse et Peakall, 1999). Although the original `popgraph` method reproduced the dispersal pattern quite well, it often produced the lowest correlation between genetic distances and CD. Nevertheless, these methods deserve further investigation because some connected population pairs in our independence graphs (using our modified method) were separated by CD values larger than the DMC. Even if the use of the DMC to determine the spatial scale at which genetic structure depends on both gene flow and drift needs stronger theoretical support, this suggests that genetic differentiation between these populations may still need time before stabilising.

We believe that an essential but tricky issue remains the identification of population pairs matching migration-drift equilibrium. Ciofi *et al.* (1999) developed a likelihood-based approach to assess whether population structure is best explained by a model of migration-drift equilibrium or by a model of pure drift. However, it seems that this approach fails to detect case-IV patterns of IBD (Hänfling et Weetman, 2006). Assuming that the DMC may be used as a proxy of the spatial scale of migration-drift equilibrium, a promising approach would consist in pruning the genetic graphs with a CD threshold equal to the DMC. This approach requires knowledge of the cost values associated with landscape features to estimate the CD between population pairs. However, assessing cost scenarios is often the aim of empirical link-based analyses (Balkenhol *et al.*, 2016). Recent methods for optimisation of landscape

resistance surfaces have been developed (Peterman, 2018), and a potential improvement would consist in using genetic graphs pruned with different CD thresholds in such optimisation procedures.

4.4 Which genetic distance to use and how ?

Weighting graph links using the D_{PS} or Euclidean genetic distances always produced a better inference of landscape influence on gene flow than using F_{ST}, even if the difference in performance of these genetic distances decreased as pairwise genetic differentiation tends to reach its equilibrium level (*i.e.* from G50 to G500). Though F_{ST} is an excellent measure of genetic differentiation, using it for particular demographic inferences (e.g. the number of migrants entering a population every generation) requires assumptions of migration-drift equilibrium to be met (Neigel, 2002 ; Whitlock et Mccauley, 1999). Further theoretical work is required to assess the sensitivity of F_{ST}-based inferences of landscape resistance to equilibrium conditions. Besides, D_{PS} had already been shown to better reflect recent landscape changes than other genetic distances, including F_{ST} (Landguth *et al.*, 2010 ; Robin *et al.*, 2015). We would expect other genetic distances such as the Chord distance (Cavalli-Sforza et Edwards, 1967) to provide similar results as those obtained with D_{PS}.

Once the genetic graph has been pruned, we discourage summing genetic distances along shortest paths to create a complete matrix of graph-based genetic distances. This approach led to spurious conclusions by detecting an isolation by distance pattern instead of the true isolation by landscape resistance pattern when graph pruning was based on geographical thresholds or topological constraints (Figure 14). Interestingly, landscape influence on dispersal was frequently better recovered when using these graph-based genetic distances derived from independence graphs (CI2) than when using the complete matrix of F_{ST} (Figure 14). This result has been previously used to evidence the value of this graph-based genetic distance (Dyer *et al.*, 2010). However, the correlation between genetic distances and the driver of dispersal (*i.e.* CD) was lower when considering these graph-based genetic distances than when using the complete matrix of corresponding raw genetic distances.

4.5 Limits and perspectives

Our simulations produced contrasted patterns of connectivity, but we acknowledge that our results are limited to cases where a single functional unit of populations that can somehow exchange migrants (i.e. a dispersal network made of a single component) is considered. We still need to investigate relative performances of graph-theoretic methods in landscapes with complete barriers isolating population clusters. The differences we detected between the compared methods were informative and promising, yet sometimes subtle. Although the migration rates we obtained were similar to those reported by Bowne et Bowers (2004), they were larger than those reported from other empirical data (Meirmans, 2014) or from simulated data reproducing case-IV patterns (Van Strien *et al.*, 2015). Given that case-IV patterns of IBLR were observed in situations where dispersal was most constrained, repeating our simulations with more limited dispersal would have produced stronger contrasts between complete and pruned graphs in their ability to infer landscape resistance to gene flow and might have made the discrimination between pruning methods even more straightforward. Considering that low migration rates are probably the norm, this reinforces the relevance of genetic graph pruning in empirical studies. However, it remains to be determined whether there is a threshold below which dispersal has only a marginal effect on genetic differentiation as compared with genetic drift. If this

case could reveal a complete barrier to dispersal, it would prevent from inferring the relative resistance of the different landscape features surrounding populations.

Our results challenge a common practice in landscape genetics consisting in using the complete matrix of genetic distance to infer the resistance of landscape features. Consequently, it must be further examined whether and how graph-theoretic methods may improve calibration of resistance surfaces. Here, we did not compare different cost scenarios as we knew the "true" cost values driving the simulation. We therefore assumed that maximising the linear correlation (Mantel r) between genetic distances and landscape distances measured for a subset of population pairs allows for reliably identifying the best graph construction method, *i.e.* the one that selects the best subset of population pairs for the analysis (but see Graves *et al.* (2013)). As we did not aim at performing a fine-tuned calibration of cost values, we consider our approach as suitable (Shirk *et al.*, 2017b ; Zeller *et al.*, 2016).

In our simulations, we assumed all populations of the study area were sampled. Such a sampling intensity is rarely achieved in practice although it is often recommended (Keller *et al.*, 2013 ; Van Strien, 2017). Assessing how partial sampling of populations affects our conclusions needs further investigation (as in Koen *et al.* (2013) and Naujokaitis-Lewis *et al.* (2013)). In these situations, the complementarity between genetic graphs and landscape graphs needs to be explored, because the nodes of the latter are the exhaustive set of potential habitat patches in the study area (Foltête et Vuidel, 2017). In addition, a growing set of studies in landscape genetics now use individual-based sampling schemes. Though the conditional independence principle evaluated in our study is not applicable when nodes are individuals, a few studies have applied genetic graphs to individuals (Draheim *et al.*, 2016 ; Greenbaum *et al.*, 2016). This possibility offers a great potential and deserves further investigation.

Lastly, gene flow and drift were the main processes driving genetic differentiation in our simulations. Drift strength depends on population sizes, which were maintained equal and constant over generations. Although this choice allowed us to keep drift constant among our simulations in order to focus only on the effect of landscape on dispersal and to substantially reduce computation times, we acknowledge that it strongly simplifies the reality. Indeed, landscape changes also create spatial heterogeneity in effective population sizes, which can be a strong driver of genetic differentiation (Prunier *et al.*, 2017). Besides, local features such as patch size or habitat quality can affect gene flow between populations (Pflüger et Balkenhol, 2014 ; Robertson *et al.*, 2019 ; Weckworth *et al.*, 2013), though we did not make it possible in our simulations. In this context, gravity models seem particularly relevant as they can be based on genetic graphs and additionally include local variables (Murphy *et al.*, 2010a ; Watts *et al.*, 2015 ; Zero *et al.*, 2017).

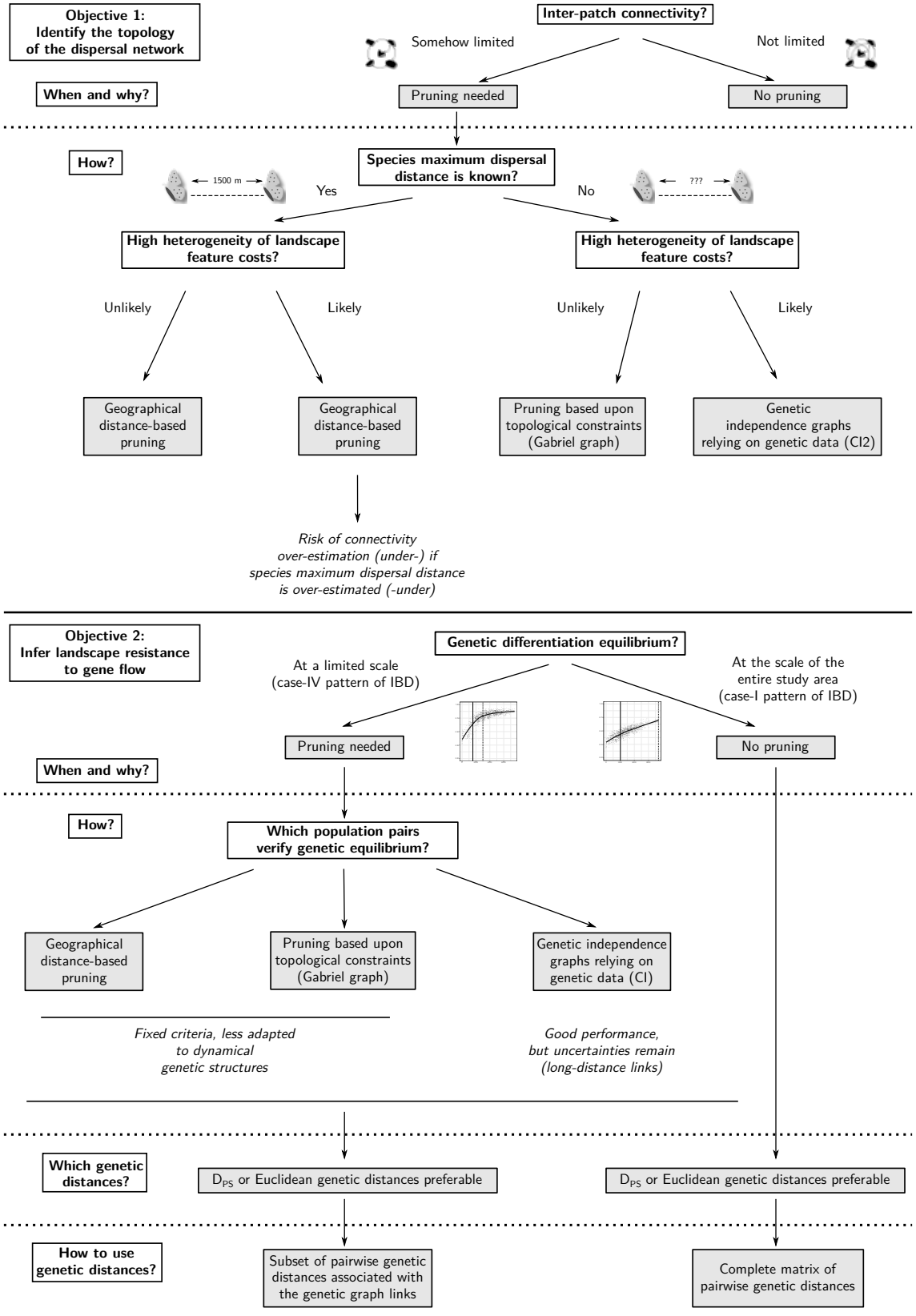


FIGURE 15 – Guidelines based on our results to build and analyse genetic graphs for i) identifying the topology of a dispersal network and ii) inferring landscape resistance to gene flow. CI2 (CI) : Conditional independence assessed with (squared) genetic distances.

5 Acknowledgements

We thank the editor and referees, as well as Aurélie Khimoun and Maarten van Strien, for their very constructive comments that improved our manuscript. This study is 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. We thank Ahmed Jebrane, Catherine Labruère and Catherine Larédo for their help on mathematical aspects. We thank Christopher Sutcliffe for revising the English manuscript. Simulations and analyses were carried out on the calculation "Mésocentre" facilities of the University of Bourgogne-Franche-Comté.

6 Data accessibility

Simulated genotypes and R codes are available online at : Dryad DOI :10.5061/dryad.6q573n5xr

7 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 and S.G. wrote the manuscript with significant contributions and remarks from all co-authors.

A - Supplementary figures

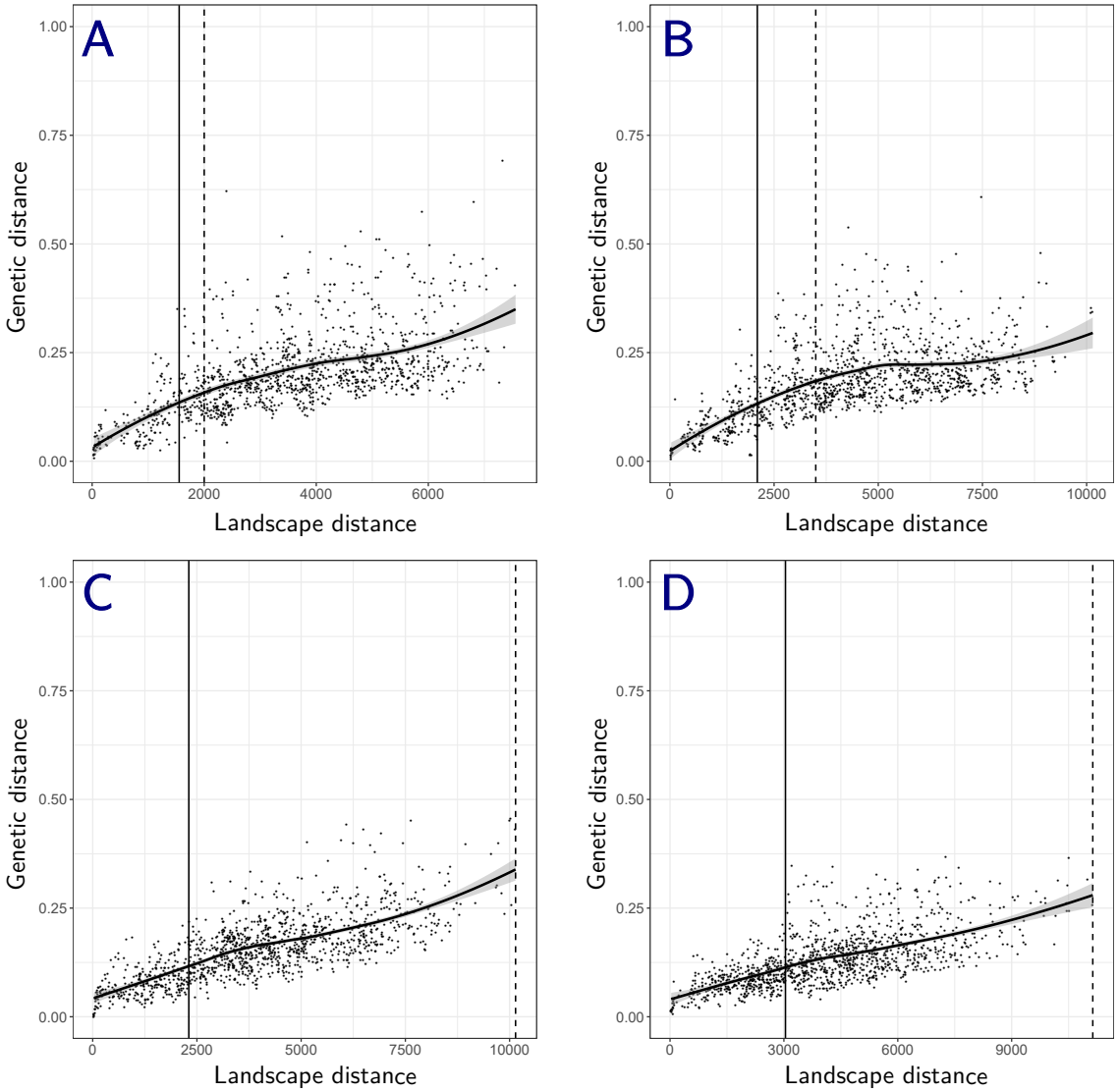


FIGURE 16 – Scatter plots of the genetic distance (F_{ST}) plotted against cost-distance at generation 50. Cases A to D illustrate the gradient of IBD patterns (from type-IV to type-I) along the first component of the PCA. Solid vertical lines indicate the maximum dispersal distance, dashed lines indicate the DMC.

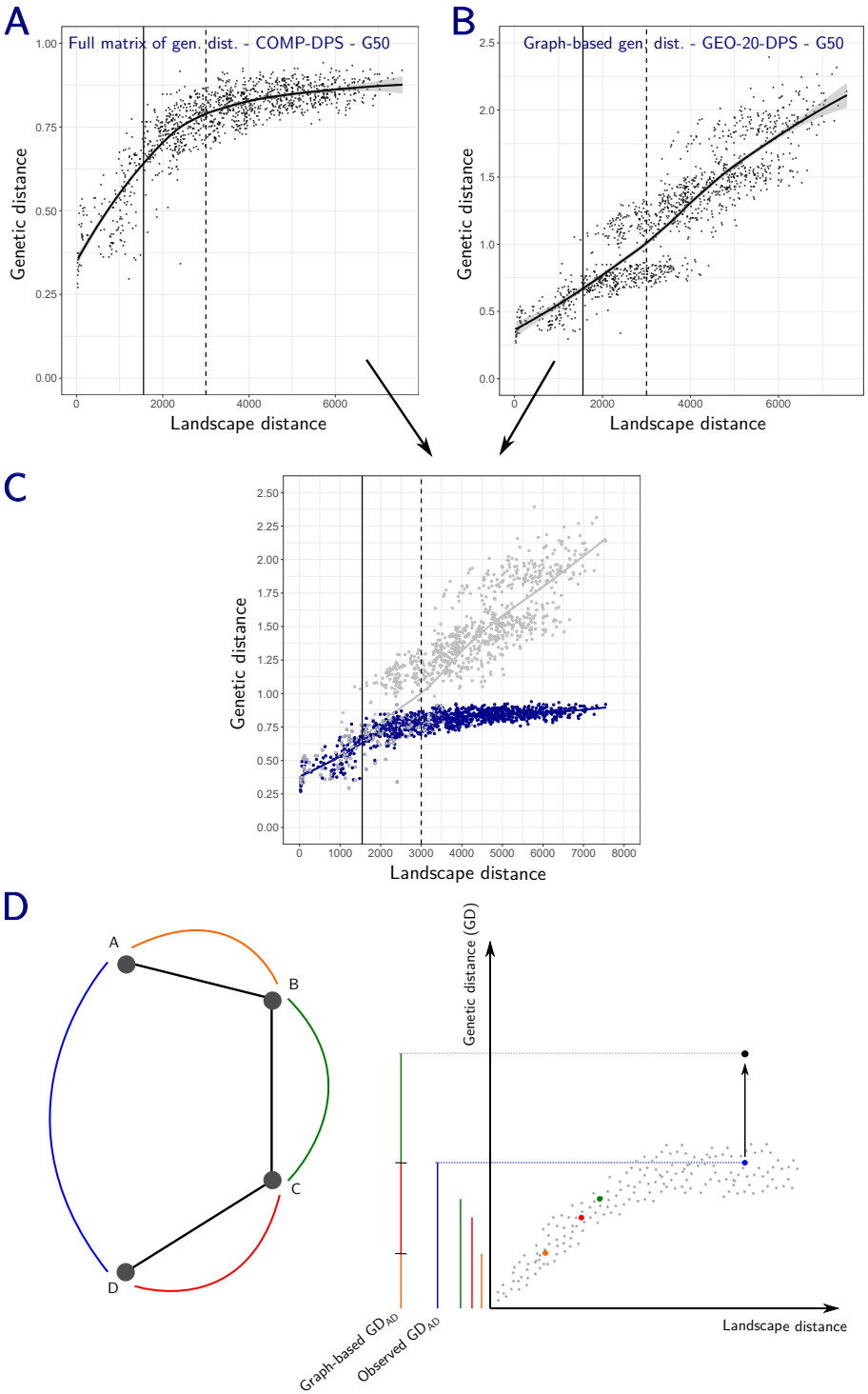


FIGURE 17 – Illustration of the potential bias induced by the use of the graph-based genetic distance. Scatter plots of the complete matrix of D_{PS} (A) or the graph-based D_{PS} from the graph GEO-20-DPS (B) against the CD. Both scatter plots are combined in panel C to display genetic distance value differences induced by the sum of genetic distances. Panel D illsutrates the mechanism behind this. Populations A and D cannot directly exchange propagules while dispersal occurs in a stepwise way between population pairs A-B, B-C and C-D. Although the genetic distance between A and D is high, it is not directly proportional to the distance between and A and D because the relationship between genetic and landscape distance flattens out, as expected under a case-IV pattern of IBLR. Considering that the genetic distance GD_{AD} between A and D is equal to $GD_{AB} + GD_{BC} + GD_{CD}$ therefore over-estimates it. Example data come from the simulation A (generation 50). Solid vertical lines indicate the maximum dispersal distance, dashed lines indicate the DMC computed with the D_{PS} at generation 50.

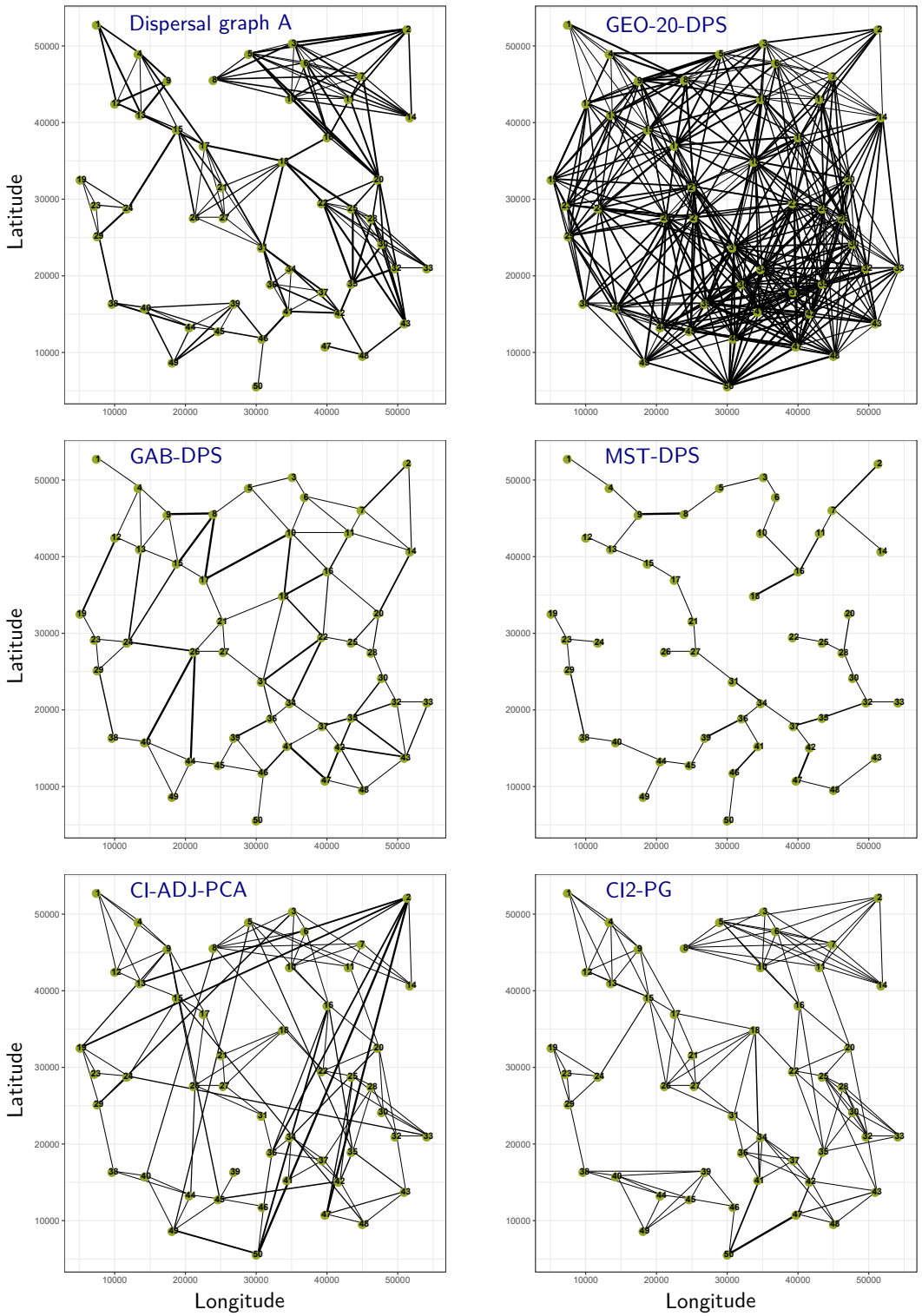


FIGURE 18 – 6 examples of graphs (from left to right and top to bottom) : a dispersal graph, a graph pruned at a geographical distance of 15 km, a Gabriel graph, a Minimum Spanning Tree and two independence graphs (see table 1 for the graph names). Example data come from a run of the simulations performed for configuration A (generation 50). Link width is proportional to genetic distance values.

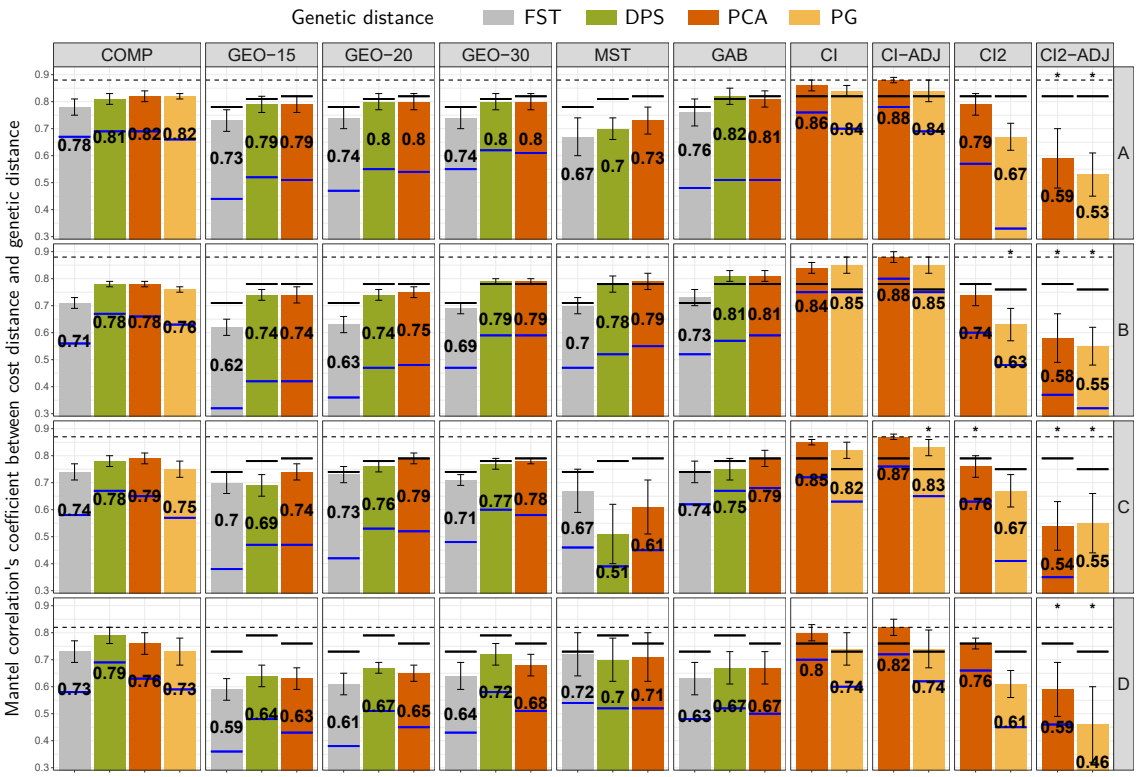


FIGURE 19 – Mantel correlation between genetic distances and cost-distances separating nodes directly connected on the genetic graphs, according to the type of genetic distance and the pruning method at generation 500 (see table 1 for the graph names). Mean ± SD values were computed for the 10 runs simulated in each scenario. Blue bars refer to the correlation coefficient between genetic distance and geographical distance, when it is above 0.3. Black bars refer to the correlation coefficient obtained using every population pair to compute the correlation. When black and blue bars overlap, the bar is black. Stars indicate graphs counting several components. The dashed line indicates the maximum r value obtained for each configuration.

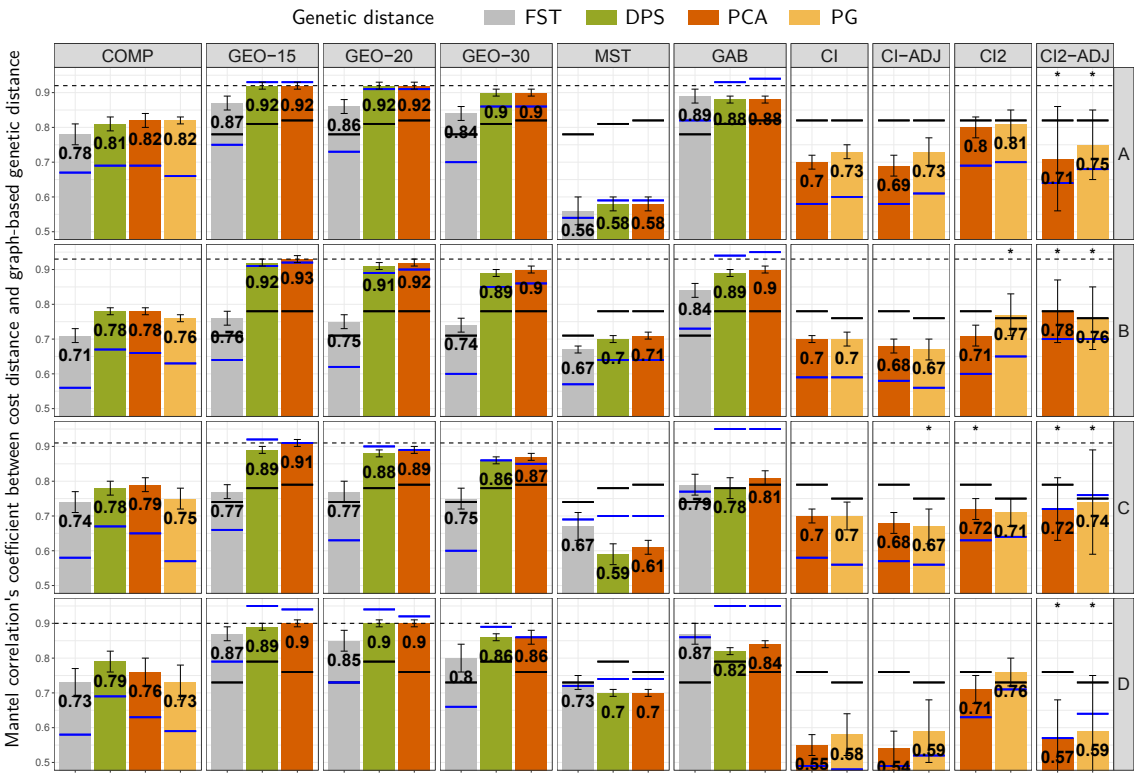


FIGURE 20 – Mantel correlation between graph-based genetic distances and cost-distances separating nodes on the genetic graphs, according to the type of genetic distance and the pruning method at generation 500 (see table 1 for the graph names). Mean $\pm SD$ values were computed for the 10 runs simulated in each scenario. Blue bars refer to the correlation coefficient between genetic distance and geographical distance, when it is above 0.5. Black bars refer to the correlation coefficient obtained using every population pair to compute the correlation. When black and blue bars overlap, the bar is black. Stars indicate graphs counting regularly several components. The dashed line indicates the maximum r value obtained for each configuration.

B - Glossary of acronyms

ADJ Designates a genetic graph pruned using the conditional independence principle and p-value Adjustment

CD Cost Distance

cGD Conditional Genetic Distance

CI Designates a genetic graph pruned using the Conditional Independence principle and computing the covariance using squared genetic distances between populations

CI2 Designates a genetic graph pruned using the Conditional Independence principle and computing the covariance using squared genetic distances between populations

COMP Designates a Complete genetic graph (not pruned)

DMC Distance of Maximum Correlation. Landscape distance threshold below which the subset of population pairs maximize the linear correlation between genetic and landscape distances

G50 (G500) Generation 50 of the simulation (Generation 500)

GAB Designates a genetic graph with the topology of a Gabriel graph

GEO Designates a genetic graph pruned using a geographical distance threshold

IBD Isolation By Distance

IBLR Isolation By Landscape Resistance

MST Designates a genetic graph with the topology of a Minimum Spanning Tree

PCA Principal Component Analysis. Also designates a genetic graph whose links are weighted using a Euclidean genetic distance computed after the Principal Component Analysis of the population allelic frequencies.

PG Designates a genetic graph whose links are weighted using a Euclidean genetic distance computed using the same formula as that used in `popgraph` package.

C - Mathematical background : independence graphs

Two events or independent variables are conditionally independent if they are statistically independent after accounting for a third event or variable (Magwene, 2001). An independence graph is a graph that summarizes conditional independence relationships between a set of variables (Magwene, 2001). Genetic independence graphs were first used by Dyer et Nason (2004) in population genetics. In this case, the "variables" are populations and the series of values of each "variable" are allelic frequencies. Creating a genetic independence graph is tantamount to identifying pairs of populations that can be considered independent once all relationships with other populations have been taken into account.

Let $\mathbf{Y}$ be a set of p variables following a normal multivariate distribution : $\mathbf{Y} = \{y_1, y_2, \ldots, y_p\}$. The three following assumptions are equivalent (Krzanowski et Marriott, 1995, in Magwene, 2001) :

— y_1 and y_2 variables are independent, conditionally to $\mathbf{Y}_K$, with $\mathbf{Y}_K$ every subset of $\mathbf{Y}$ excluding y_1 and y_2.

— Partial correlation between y_1 and y_2 is null : $\rho_{ij.\{K\}} = 0$

— If $\mathbf{C}$ is the covariance matrix of the set of variables $\mathbf{Y}$, then the element π_{ij} of the inverse covariance matrix $\mathbf{\Pi} = \mathbf{C}^{-1}$ (precision matrix), is null.

Therefore, to assess conditional independence between a set of populations, partial correlation matrix or precision matrix have first to be calculated from genetic data. In population genetics, the multilocus genotypes of individuals from populations are frequently coded as a matrix with alleles as columns and individuals as rows. The absence of an allele is coded as a 0. The presence of 1 or 2 copies of an allele in the genotype of an individual are coded respectively as a 0.5 or a 1. If these data are coded with 0, 1 and 2 values, as in Fortuna *et al.* (2009) and Smouse et Peakall (1999), it does not affect the calculation.

First, mean allelic frequencies in each population are computed. These frequencies are elements of a matrix **F** counting as many columns as alleles and as many rows as populations. The allele frequencies are the series of values characterizing each population, considered as variables in the construction of the genetic independence graph. The next step consists in computing the covariance between populations (betweens rows of **F**). Dyer et Nason (2004) calculates this covariance by first calculating a matrix of Euclidean genetic distance between populations and then following Gower (1966), who demonstrated the duality between distance and covariance.

To that purpose, the matrix **F** of mean allelic frequencies by population has to be centered both by rows and by columns for the calculation of covariance from genetic distance in subsequent steps to be correct. However, in this particular case, this step is not mandatory given 1) the row sums are all equal to the number of loci because the allelic frequencies sum to 1 for each locus, and 2) the centering by columns does not affect the Euclidean distance between populations (rows). Without the double-centering, the between populations covariance matrix calculated from the genetic distances is however equivalent to the matrix of covariance between the columns of the transpose **X** of the double-centered matrix **F** of allelic frequencies. We demonstrate why thereafter. We also demonstrate why the covariance has to be calculated from the squared distances and not from distances, from a strict mathematical point of view, following Everitt et Hothorn (2011)(page 107), Gower (1966) and Smouse et Peakall (1999)(equation 13).

The Euclidean genetic distance d_{ij} between populations i and j is calculated from the transpose matrix **X** of the matrix **F** of allelic frequencies. **X** is of dimension $n \times p$, with n the number of alleles and p the number of populations. The genetic distance is computed with the following formula :

$$d_{ij} = \sqrt{\sum_{k=1}^{n} (x_{ki} - x_{kj})^2} \tag{3.1}$$

The sample covariance c_{ij} between variables/populations i and j is :

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j}) \tag{3.2}$$

As **F** has been centred both by rows and by columns, $\bar{x_i} = \bar{x_j} = 0$. Then, the sample covariance between variables/populations i and j is simply :

$$c_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{ki} x_{kj} \tag{3.3}$$

And consequently :

$$\begin{aligned} c_{ii} &= \frac{1}{n}\sum_{k=1}^{n} x_{ki}^2 \\ c_{jj} &= \frac{1}{n}\sum_{k=1}^{n} x_{kj}^2 \end{aligned} \tag{3.4}$$

Then, the covariance matrix $\mathbf{C}$ is :

$$\mathbf{C} = \frac{1}{n}\mathbf{X^T X} \tag{3.5}$$

such that $\mathbf{X^T}$ is of size $p \times n$, $\mathbf{X}$ of size $n \times p$ and $\mathbf{C}$ of size $p \times p$.

The sum of the elements of each row of $\mathbf{C}$ is :

$$\begin{aligned} \sum_{j=1}^{p} c_{ij} &= \sum_{j=1}^{p} \frac{1}{n}\sum_{k=1}^{n} x_{ki}x_{kj} \\ &= \frac{1}{n}\left[(\sum_{k=1}^{n} x_{ki}x_{k1}) + (\sum_{k=1}^{n} x_{ki}x_{k2}) + \ldots + (\sum_{k=1}^{n} x_{ki}x_{kp})\right] \\ &= \frac{1}{n}\left[(x_{1i}x_{11} + x_{2i}x_{21} + \ldots + x_{ni}x_{n1}) + \ldots + (x_{1i}x_{1p} + x_{2i}x_{2p} + \ldots + x_{ni}x_{np})\right] \\ &= \frac{1}{n}\left[x_{1i} \times (\sum_{j=1}^{p} x_{1j}) + x_{2i} \times (\sum_{j=1}^{p} x_{2j}) + \ldots + x_{ni} \times (\sum_{j=1}^{p} x_{nj})\right] \\ &= \frac{1}{n}\left[x_{1i} \times 0 + x_{2i} \times 0 + \ldots + x_{ni} \times 0\right] \\ &= 0 \end{aligned} \tag{3.6}$$

as the row sums of $\mathbf{X}$ are null since $\mathbf{F}$ was centered by rows and by columns.

The trace T of $\mathbf{C}$ is :

$$T = \sum_{i=1}^{p} c_{ii} \tag{3.7}$$

Let express d_{ij}^2 in function of the elements of $\mathbf{C}$:

$$\begin{aligned} d_{ij}^2 &= \sum_{k=1}^{n}(x_{ki} - x_{kj})^2 \\ &= \sum_{k=1}^{n}(x_{ki}^2 - 2x_{ki}x_{kj} + x_{kj}^2) \\ &= \sum_{k=1}^{n} x_{ki}^2 + \sum_{k=1}^{n} x_{kj}^2 - 2\sum_{k=1}^{n} x_{ki}x_{kj} \\ &= n \times (c_{ii} + c_{jj} - 2c_{ij}) \end{aligned} \tag{3.8}$$

We then have :

$$\begin{aligned} \sum_{i=1}^{p} d_{ij}^2 &= \sum_{i=1}^{p} n \times (c_{ii} + c_{jj} - 2c_{ij}) \\ &= n \times (\sum_{i=1}^{p} c_{ii} + \sum_{i=1}^{p} c_{jj} - 2\sum_{i=1}^{p} c_{ij}) \end{aligned} \tag{3.9}$$

As $\sum_{j=1}^{p} c_{ij} = 0$ and $\mathbf{C}$ is a symmetric matrix, $\sum_{i=1}^{p} c_{ij} = 0$. We then have :

$$\begin{aligned}
\sum_{i=1}^{p} d_{ij}^2 &= n \times (T + pc_{jj} - 2 \times 0) \\
&= n \times (T + pc_{jj}) \\
\sum_{j=1}^{p} d_{ij}^2 &= n \times (T + pc_{ii})
\end{aligned} \tag{3.10}$$

We calculate $\sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2$:

$$\begin{aligned}
\sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2 &= \sum_{i=1}^{p} \sum_{j=1}^{p} n \times (c_{ii} + c_{jj} - 2c_{ij}) \\
&= n \times (\sum_{i=1}^{p} \sum_{j=1}^{p} c_{ii} + \sum_{i=1}^{p} \sum_{j=1}^{p} c_{jj} - 2 \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij}) \\
&= n \times (pT + pT - 2 \times 0) \\
&= n \times 2pT
\end{aligned} \tag{3.11}$$

We then calculate $d_{i\bullet}^2$, $d_{\bullet j}^2$ and $d_{\bullet\bullet}^2$:

$$\begin{aligned}
d_{i\bullet}^2 &= \frac{1}{p} \sum_{j=1}^{p} d_{ij}^2 \\
&= \frac{1}{p} \times n \times (T + pc_{ii}) \\
&= n \times (\frac{T}{p} + c_{ii}) \\
&= n \times (\frac{1}{p} \sum_{i=1}^{p} c_{ii} + c_{ii}) \\
d_{\bullet j}^2 &= n \times (\frac{1}{p} \sum_{i=1}^{p} c_{ii} + c_{jj}) \\
d_{\bullet\bullet}^2 &= \frac{1}{p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} d_{ij}^2 \\
&= \frac{n}{p^2} \times 2pT \\
&= 2 \times \frac{n}{p} \sum_{i=1}^{p} c_{ii}
\end{aligned} \tag{3.12}$$

Because of the formula used to calculate the Euclidean distance, we have :

$$
\begin{aligned}
d_{ij}^2 &= n \times (c_{ii} + c_{jj} - 2c_{ij}) \\
c_{ij} &= -\frac{1}{2}(\frac{d_{ij}^2}{n} - c_{ii} - c_{jj}) \\
&= -\frac{1}{2n}(d_{ij}^2 - n \times c_{ii} - n \times c_{jj}) \\
&= -\frac{1}{2n}(d_{ij}^2 - n \times c_{ii} - n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii} - n \times c_{jj} - n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii} + 2n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii}) \\
&= -\frac{1}{2n}\left[d_{ij}^2 - n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{ii}) - n \times (\frac{1}{p}\sum_{i=1}^{p} c_{ii} + c_{jj}) + 2n \times \frac{1}{p}\sum_{i=1}^{p} c_{ii}\right] \\
&= -\frac{1}{2n}(d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2)
\end{aligned}
\tag{3.13}
$$

Hence, to conform with the covariance definition, c_{ij} has to be calculated from squared distances although Dyer et Nason (2004) in `popgraph` package use the following formula :

$$c_{ij} = -\frac{1}{2}(d_{ij} - d_{i\bullet} - d_{\bullet j} + d_{\bullet\bullet}) \tag{3.14}$$

The division by n in (3.13) does not have any influence on subsequent computation steps, given the covariance matrix $\mathbf{C}$ is then standardised into a correlation matrix $\mathbf{R}$. This correlation matrix is inverted into the inverse correlation matrix $\mathbf{\Omega}$, which is also standardised. The non-diagonal elements ω_{ij} of $\mathbf{\Omega}$ are multiplied by -1 to obtain the partial correlation matrix $\mathbf{P}$ such that (Magwene, 2001) :

$$\rho_{ij} = \frac{-\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}} \tag{3.15}$$

Finally, to determine if populations i and j are independent conditionally to all other populations, we have to test if each element ρ_{ij} is significantly different from 0. To that purpose, the Edge Exclusion Deviance criterion (EED) is calculated following Whittaker (2009) as :

$$EED = -N \ln(1 - \rho_{ij}^2) \tag{3.16}$$

with N the total number of observations (total number of individuals, as implemented by Dyer et Nason (2004)).

We assumed that an independence genetic graph should have links between populations positively correlated if it is to represent direct gene flow between populations. Therefore, we converted negative elements of $\mathbf{P}$ into 0 before the calculation of EED, although it was not the case in the original method of Dyer et Nason (2004).

EED has an asymptotic χ^2 distribution with one degree of freedom (Whittaker, 2009). This property allows to test the significance of every EED value and thereby to test the hypothesis $H_0 : \rho_{ij} = 0$ against $H_1 : \rho_{ij} \neq 0$. When H_0 is rejected, there is a link between populations i and j in the resulting graph.

The 0.05 level is commonly used to test the significance of EED, without p-values adjustment in the original method. However, we adjusted p-values using Holm (1979) method to limit the risk of type-I error because $\frac{p(p-1)}{2}$ tests are carried out to build a graph.

D - Computation of the DPS genetic distance

The D_{PS} is a genetic distance that relies upon the dissimilarities between the allele pools of different populations. It was initially developed as an inter-individual genetic distance (Bowcock *et al.*, 1994). An "inter-population version" exists and has been used repeatedly in landscape genetics (Murphy *et al.*, 2016).

To compute it, we used the formula used in MSA software :

$$D_{PS} = 1 - \frac{\sum_{d}^{D} \sum_{k}^{K} min(f_{a_{kd,i}}, f_{a_{kd,j}})}{D}$$

such as a_{kd} is the allele k at locus d, $f_{a_{kd,i}}$ is the frequency of a_{kd} in population i, D is the total number of loci and K is the allele number at each locus.

This genetic distance can be computed in R with the function `mat_gen_dist()` in `graph4lg` package.

Annexe A4

`graph4lg` : a package for constructing and analysing graphs for landscape genetics in R

Abstract

1 - In landscape genetics, habitat connectivity and population genetic structure have been analysed using graph-theoretic approaches to understand how landscape features influence demography (i.e. dispersal and population size).
2 - Despite substantial advances in enhancing both genetic and landscape graph use, a software tool bringing together a large range of construction and analysis parameters for these two types of graphs was lacking in the landscape genetic toolbox. Moreover, although these two types of graphs appear complementary for answering landscape genetic questions, methods for comparing them have not been forthcoming.
3 - We have developed an R package to improve and encourage the use of these graphs. It includes functions for converting and importing genetic data and for genetic distance computing. It also implements time-efficient geodesic and cost-distance calculations from spatial data. A large range of parameters can be used to create genetic and landscape graphs from these data, including several graph pruning methods. We made available to R users the command-line facilitaties of Graphab software to easily model landscape graphs in R. The package functions perform preliminary analysis to adapt methodological choices to research questions. Landscape and genetic graphs created can be analysed with node-level metrics as well as link-level and modularity analyses. Users can compare and visualise these graphs and export them to shapefiles to facilitate interpretation and subsequent analyses.
4 - `graph4lg` contributes to expanding landscape and genetic graph potential for analysing ecological connectivity while encouraging further investigations on methodological implications related to these tools.

Keywords : ecological connectivity, dispersal, graph theory, landscape genetics, R

Cet article a été publié dans *Methods in Ecology and Evolution* en novembre 2020 :

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

1 Introduction

Landscape genetic studies aim at understanding how landscape characteristics such as habitat spatial distribution and matrix quality shape population genetic structure (Balkenhol *et al.*, 2016 ; Manel *et al.*, 2003). In this recent field, new methods have been developing at a sustained pace to describe landscape structure (e.g. Galpern *et al.* (2012)), population genetic structure (e.g. Al-Asadi *et al.* (2019) ; Prunier *et al.* (2017) ; Shirk et Cushman (2011)) and to bring landscape and genetic data together in multi-level analyses (Hall et Beissinger, 2014 ; Wagner et Fortin, 2013). Among these methods, graph-theoretic approaches were identified as particularly relevant (Manel et Holderegger, 2013) because they grasp interactions between sets of habitat patches or populations in a comprehensive way (Dale et Fortin, 2010 ; Dale, 2017 ; Fortin *et al.*, 2012).

A graph is basically a set of nodes connected by a set of links. In landscape ecology, the use of landscape graphs dates back to the 2000s (Galpern *et al.*, 2011 ; Urban et Keitt, 2001) and has developed through the availability of free software tools such as Graphab (Foltête *et al.*, 2012a) or Conefor (Saura et Torne, 2009). The nodes of landscape graphs are habitat patches and their links correspond to potential dispersal paths, e.g. identified by computing least-cost paths across resistance surfaces. From these graphs, a large range of connectivity metrics can be computed (Rayfield *et al.*, 2011) and used for inference (Pereira *et al.*, 2011) or conservation-oriented decision making (Foltête *et al.*, 2014). Besides, landscape graph nodes can be partitioned through modularity analyses to identify management units or perform analyses at a coarser grain (Fletcher *et al.*, 2013 ; Foltête et Vuidel, 2017).

Although landscape graphs enable close investigation of habitat connectivity, their construction often relies upon expert-based opinion and combining them with biological data can improve this approach (Foltête *et al.*, 2020). Landscape graph modelling software tools already make possible the import of biological data (Foltête *et al.*, 2012a) and genetic data are a relevant candidate for such a combination (Luque *et al.*, 2012). Thus, this approach would benefit from being performed in a statistical software where genetic data processing and complex statistical analyses can both be done.

Similarly, migration models theorised in population genetics (Kimura et Weiss, 1964 ; Wright, 1931) often rely upon topological network representations and even if these models rarely reflect real situations (Greenbaum et Fefferman, 2017 ; Milligan *et al.*, 2018), population geneticists have developed a large range of genetic graph construction methods that can potentially fit all the observed migration networks (Greenbaum et Fefferman, 2017). Thus, population genetic structure has been frequently represented as genetic graphs in which nodes correspond to sampled populations and links to substantial gene exchanges between them (Arnaud, 2003 ; Excoffier *et al.*, 1992).

When building a genetic graph, the construction method should always be guided by the specific research question (Miele *et al.*, 2019). For example, an important step in this process is graph pruning, which consists in removing some links and should be performed differently if the aim of the analysis is (i) to identify single generation (direct) dispersal paths (Boulanger *et al.*, 2020 ; Dyer, 2015b) or (ii) to infer landscape effects on dispersal from the genetic differentiation measurements between populations connected on the graph (Savary, P. et al., in correction ; Van Strien (2017)). Indeed, in the first case, paths that are not within reach of individuals given their dispersal capacities should be removed in order to represent the dispersal network topology. In the second case, links corresponding

to multi-generational indirect dispersal can be conserved given that they reflect the genetic connectivity emerging over generations due to stepping-stone dispersal (Boulanger *et al.*, 2020 ; Saura *et al.*, 2014). In both cases, several graph pruning methods can be used and must be chosen accordingly (Greenbaum et Fefferman, 2017 ; Van Strien, 2017). Apart from these link-level analyses, once genetic graphs have been constructed, they can be analysed at the node- and boundary-levels (Wagner et Fortin, 2013), according to the research question. For example, node-level metric calculations and module partitions are possible to assess genetic diversity and relative genetic differentiation at the population level (e.g. Koen *et al.* (2016)) and to identify population clustering patterns (e.g. Fortuna *et al.* (2009)), respectively.

Then, comparing genetic graph characteristics such as node-level indices, graph topology, link weights and module partitions to the exact same characteristics derived from a landscape graph could contribute to a better understanding of the effect of habitat spatial patterns on population genetic structure. It has therefore been claimed that genetic graphs and landscape graphs should complement each other (Foltête et Vuidel, 2017 ; Galpern *et al.*, 2011 ; Manel et Holderegger, 2013 ; Murphy *et al.*, 2016). However, a practical tool for building, analysing and comparing these graphs was still lacking. Accordingly, we have developed the R package `graph4lg` to bridge all these gaps in the implementation of landscape and genetic graphs. It provides graph users with a software tool facilitating their choice and implementation of graph construction and analysis methods and builds on previous developments of R packages for landscape genetic and graph-theoretical analyses such as `igraph` (Csardi et Nepusz, 2006), `gstudio` (Dyer, 2014) or `adegenet` (Jombart, 2008).

2 Workflow

`graph4lg` package functions can be divided into four categories following the steps of landscape genetic analyses :

1. The package allows for genetic and spatial input data processing in preparation for graph construction, calculating intra-population genetic indices and inter-population genetic and landscape distances and performing preliminary analyses through diagnostic plots.
2. It provides users with functions for building genetic graphs and analysis tools for these graphs.
3. In parallel, some 'wrapper functions' run command-line functionalities of Graphab software (Foltête *et al.*, 2012a) directly from R to construct and analyse landscape graphs.
4. Finally, genetic and landscape graphs can be compared, plotted and exported to other formats.

This workflow is described in the following sections and depicted in Figure 21. All the package functions are also listed in Table S1 with their dependencies on other R packages.

2.1 Input data processing

2.1.1 Genetic data

Studies of gene flow pattern and/or intensity and of landscape influence on it rely upon neutral genetic markers which reflect genetic variation due to demographic changes and are supposedly independent from adaptive processes (Holderegger *et al.*, 2006). Microsatellite loci assumed or identified as being neutral are the most frequent type of markers used in landscape genetics (Storfer *et al.*, 2010).

Besides, SNP markers are now becoming widespread and can also be used to perform these analyses provided that loci under selective pressures (identified as outliers) have been discarded (Cushman *et al.*, 2018 ; Foll et Gaggiotti, 2008). Accordingly, `graph4lg` functions use genetic data with 2- or 3-digit allele coding to fit the common microsatellite coding. SNP data can also be used when loaded as `genind` object in R, because R objects with the `genind` class attribute from the `adegenet` package (Jombart, 2008) are the input of genetic data processing functions from `graph4lg`. Landscape genetic analyses performed with `graph4lg` are population-based and like most applications in this field rely on the a priori delineation of populations (Milligan *et al.*, 2018 ; Waits et Storfer, 2015). Populations are identified by the `pop` strata of `genind` objects and usually correspond to sampling units.

We included conversion functions (`gstud_to_genind`, `loci_to_genind`, `structure_to_genind`, `genepop_to_genind`) to easily get `genind` objects from formats used in other software tools such as `gstudio` (Dyer, 2014), `pegas` (Paradis, 2010), STRUCTURE (Pritchard *et al.*, 2000) or GENEPOP (Raymond et Rousset, 1995). We also made possible the creation of external text files in GENEPOP format from `genind` objects (`genind_to_genepop`) for users willing to perform analyses with this commonly used R package and executable software.

The `mat_gen_dist` function computes eight different inter-population genetic distances from `genind` objects (Table 3). However, 'external' genetic distance matrices or `dist` objects imported by users can be the input of the functions described in the next sections.

Genetic distance	**Description**	**Eq.**	**Depend.**	**Ref.**
F_{ST}	Fixation index	Yes	`diveRsity`	Weir et Cockerham (1984)
Linearised F_{ST}	Linearised fixation index	Yes	`diveRsity`	Rousset (1997)
G'_{ST}	Standardised fixation index	Yes	`diveRsity`	Hedrick (2005)
D_{Jost}	Standardised fixation index	Yes	`diveRsity`	Jost (2008)
D_{PS}	1 - proportion of shared alleles	No	None	Bowcock *et al.* (1994), implementation of MSA software formula (Dieringer et Schlötterer, 2003)
Euclidean genetic distance	Computed from allelic frequencies differences	No	None	Excoffier *et al.* (1992)
Weighted Euclidean genetic distance	Computed from allelic frequencies differences giving more weights to rare alleles	No	None	Fortuna *et al.* (2009) ; Greenbaum *et al.* (2016)
PCA-derived Euclidean genetic distance	Inter-population distance in the space defined by the principal components obtained from a PCA of the allelic frequencies table	No	None	Inspired by the distances computed by Paschou *et al.* (2014) and Shirk *et al.* (2017a)
`popgraph`-derived genetic distance	Inter-population distance in the space created after a PCA of the allelic frequencies table	No	None	Dyer et Nason (2004)

TABLE 3 – Inter-population genetic distances computed with the `mat_gen_dist` function. The 'Eq.' column indicates whether the genetic distance implies that migration-drift equilibrium assumptions are made. The 'Depend.' column indicates the R packages on which the function depends for each genetic distance.

2.1.2 Spatial data

Two kinds of pairwise landscape distance matrices can be computed from point spatial coordinates and resistance surface raster layers :

— `mat_geo_dist` function computes geodesic distances from point sets with either projected or polar coordinate reference systems using Euclidean distance or great circle distance formulas, respectively.

— `mat_cost_dist` function computes pairwise cost-distance matrices from a point set, a categorical resistance surface raster layer and a `data.frame` indicating the cost associated with each cell value. This function depends on `gdistance` package (Van Etten, 2012), but can also use an external .jar file which substantially reduces computation times for large rasters (Table S2), providing R users with a time-efficient alternative to `gdistance` for cost-distance computing.

2.1.3 Preliminary analyses

When the study objective is to infer landscape effects on dispersal from the relationship between genetic and landscape distances associated with graph links, visualising a scatterplot using complete distance matrices (`scatter_dist`) can be a first step before graph construction. Isolation by distance patterns due to limited dispersal are common in population genetics (Wright, 1943). However, if the studied species has low dispersal capacities or after a decrease in landscape connectivity, the increase of genetic differentiation with distance is only observed at a small scale, which tends to expand over time (Slatkin, 1993). In that case, drift will be more important than migration as a driver of genetic differentiation between populations separated by large distances. This results in a non-linear relationship between landscape and genetic distances exhibiting a plateau beyond a given landscape distance threshold (Hänfling et Weetman, 2006 ; Hutchison et Templeton, 1999). Conversely, a linear relationship is expected when equilibrium is established at the scale of the study area. Because migration-drift equilibrium is a pre-requisite for genetic differentiation to reflect landscape effects on dispersal, inference ignoring it may be biased (Bradbury et Bentzen, 2007 ; Van Strien *et al.*, 2015). Genetic graph pruning method determines population pairs included in the inference and should therefore be chosen after consideration of the scale at which populations verify this equilibrium. Similarly, genetic distances based on fixation indices require equilibrium assumptions to be confirmed so that derived inferences are reliable (Neigel, 2002 ; Whitlock et Mccauley, 1999).

Van Strien *et al.* (2015) estimated the threshold distance between population pairs maximizing the correlation between genetic differentiation and landscape distance. This distance of maximum correlation (DMC) can be viewed as an estimate of the scale at which populations verify migration-drift equilibrium, i.e. their neighborhood size (Addicott *et al.*, 1987 ; Kierepka *et al.*, 2020). It is computed by the `dist_max_corr` function (Figure 22A), which can help choosing among pruning methods (cf. section 2.2.1).

When the objective is to recover direct dispersal paths by taking into account landscape resistance and maximum dispersal capacities of the study species, it is important to know how species dispersal distances expressed in geodesic distance units convert into cost-distance values, especially if geodesic distance thresholds are used to prune the graphs. To that purpose, the `convert_cd` function performs

a linear or log-log linear regression of cost-distance values against geodesic distance values (Tournant *et al.*, 2013).

2.2 Genetic graph construction and analysis

2.2.1 Genetic graph construction

The `graph4lg` package implements a wide range of pruning methods for constructing genetic graphs. Some of these methods can equally apply to landscape graphs. First, graphs can be pruned by removing the links between populations separated by genetic or landscape distances above or below a specified threshold value (`gen_graph_thr`). Such an approach can be efficient for identifying direct dispersal paths provided maximum dispersal distance is known. It has also been used to select population pairs to include in the inference of landscape effects on dispersal (Angelone *et al.*, 2011 ; Keller *et al.*, 2013). In that case, the distance threshold can be the DMC if a clear plateau is identified in the IBD pattern.

Second, when the study species is assumed to have stepping stone dispersal or when migration-drift equilibrium establishes at short distance, graphs can be pruned depending on topological constraints (`gen_graph_topo`), thereby constructing minimum spanning trees, planar graphs, k-nearest-neighbour graphs and Gabriel graphs (Arnaud, 2003 ; Bunn *et al.*, 2000 ; Keller *et al.*, 2013 ; Naujokaitis-Lewis *et al.*, 2013). The topological constraints can be applied to matrices of landscape distances as well as genetic distances. Similarly, the edge-thinning method, linked to percolation theory, identifies the distance threshold above which graph thresholding breaks the graph into more than one connected component (Urban et Keitt, 2001 ; Rozenfeld *et al.*, 2008) and creates a thresholded graph using this threshold.

Finally, the `gen_graph_indep` function creates genetic graphs directly from genetic data stored in `genind` objects, in the same way as the `popgraph` function from the `popgraph` package. This approach prunes graphs by conserving links between populations that are dependent on each other based on the covariance of their allelic frequencies, after having looked at the covariance with allelic frequencies from all the other populations. This use of the conditional independence principle (Whittaker, 2009) is supposed to conserve links between populations directly exchanging migrants through single generation dispersal events (Dyer et Nason, 2004). This function expands the original `popgraph` function by implementing p-value adjustments (Benjamini et Hochberg, 1995 ; Holm, 1979), among other options compared by Savary et al. (in correction).

2.2.2 Genetic graph analyses

Once genetic graphs have been created, the `compute_node_metric` function computes graph-theoretic metrics such as the degree, closeness and betweenness centrality indices, which identify keystone hubs of genetic connectivity (Cross *et al.*, 2018). It also computes the average and sum of the inverse genetic distance weighting the links. Koen *et al.* (2016) showed that these relative genetic differentiation indices can be good proxies of connectivity. Apart from these metrics depending on graph topology, population-level genetic diversity indices such as allelic richness and heterozygosity rates can be computed with the `pop_gen_index` function from `genind` objects. All these metrics can

be added as node attributes with the `add_nodes_attr` function.

Link weights are used to partition nodes into modules (`compute_graph_modul`) and identify population clusters possibly delineated by sharp dispersal limitations (Fortuna *et al.*, 2009 ; Garroway *et al.*, 2008). We implement several modularity algorithms : `fast greedy` (Clauset *et al.*, 2004), `louvain` (Blondel *et al.*, 2008), `optimal` (Brandes *et al.*, 2008) and `walktrap` (Pons et Latapy, 2006) from the `igraph` package. Link weights can also be exported into data frames for subsequent link-level analyses using the `graph_to_df` function.

2.3 Landscape graph construction and analysis

The `graph4lg` package integrates the graph construction and analysis options of Graphab software (Foltête *et al.*, 2012a) by implementing its command-line functionalities. Thus, both the package documentation and Graphab software manual can be of substantial help for users.

First, the `graphab_project` function creates a Graphab project from a categorical raster layer. It defines habitat patches as contiguous cells with a given cell value and creates a directory containing this project in the user's machine. Then, the least cost paths between these habitat patches are computed (`graphab_link`). Once the Graphab project and link set have been created, users can create complete, thresholded or planar graphs (`graphab_graph`). A large range of connectivity metrics can be computed at the graph or node levels (`graphab_metric`). Delta-metrics can also be computed, e.g. for prioritisation analyses. These metrics have been extensively tested and compared in the literature (Baranyi *et al.*, 2011 ; Rayfield *et al.*, 2011).

Users can either import planar graphs created in Graphab as `igraph` objects (`graphab_to_igraph`) in order to compute metrics in R, or only import link weights or node-level metrics computed with Graphab (`get_graphab_metric`, `get_graphab_linkset`). In cases when users want to relate punctual field observations to connectivity metrics, they can get the metrics of the nearest habitat patches from a set of points (`graphab_pointset`). Finally, users can partition habitat patches through modularity analyses in Graphab (`graphab_modul`).

2.4 Landscape and genetic graph comparisons

Although landscape and genetic graphs have each been repeatedly used, their direct comparison has rarely been performed (Draheim *et al.*, 2016 ; Schoville *et al.*, 2018). To facilitate the interpretation of their respective topology and the formulation and test of hypotheses regarding their similarities, the `plot_graph_lg` function allows users to visualise the topology and connectivity of the created graphs (Figure 22D). It maps graphs in a spatially-explicit way or implements an attraction-repulsion algorithm based on link weights (Fruchterman et Reingold, 1991) to assess whether nodes cluster together independently from their spatial locations. Node metrics, link weights and module partitioning can also be visualised with this function. Moreover, the link weight distribution can be plotted as a histogram (`plot_hist_w`)(Fig. 22C) and the pruning intensity can be visualised by representing population pairs in a different color on the scatter plot relating genetic distance with landscape distance (`scatter_dist_g`).

To test the formulated hypotheses, when landscape and genetic graphs share the same nodes, the correlation between population- or patch-based indices can be assessed (`graph_node_compar`) in order to understand the relationship between i) landscape attributes (habitat surface area or connectivity) and ii) genetic attributes (genetic differentiation or local diversity).

Similarly, the `graph_topo_compar` function compares the topologies of two graphs (link-based analysis) by creating a contingency table (based on the presence/absence of corresponding links in both graphs) and computing indices commonly used to assess classifications (e.g. Matthews' correlation coefficient, false discovery rate, accuracy). The congruence of two graph topologies can be visualised by plotting them on the same map while colouring their links to indicate whether they occur in both graphs or just one of them (`graph_plot_compar`). Mismatches between genetic and landscape graph topology can provide insights regarding the realised connectivity in the study area or the modelisation method itself.

Finally, quantifying how many node pairs classified in the same module in one graph are also classified together in the modules created from another graph indicates us how far these graphs reflect the same real-world situation. This boundary-based analysis is made possible by the `graph_modul_compar` function which computes the Adjusted Rand Index (Hubert et Arabie, 1985) to compare partitions.

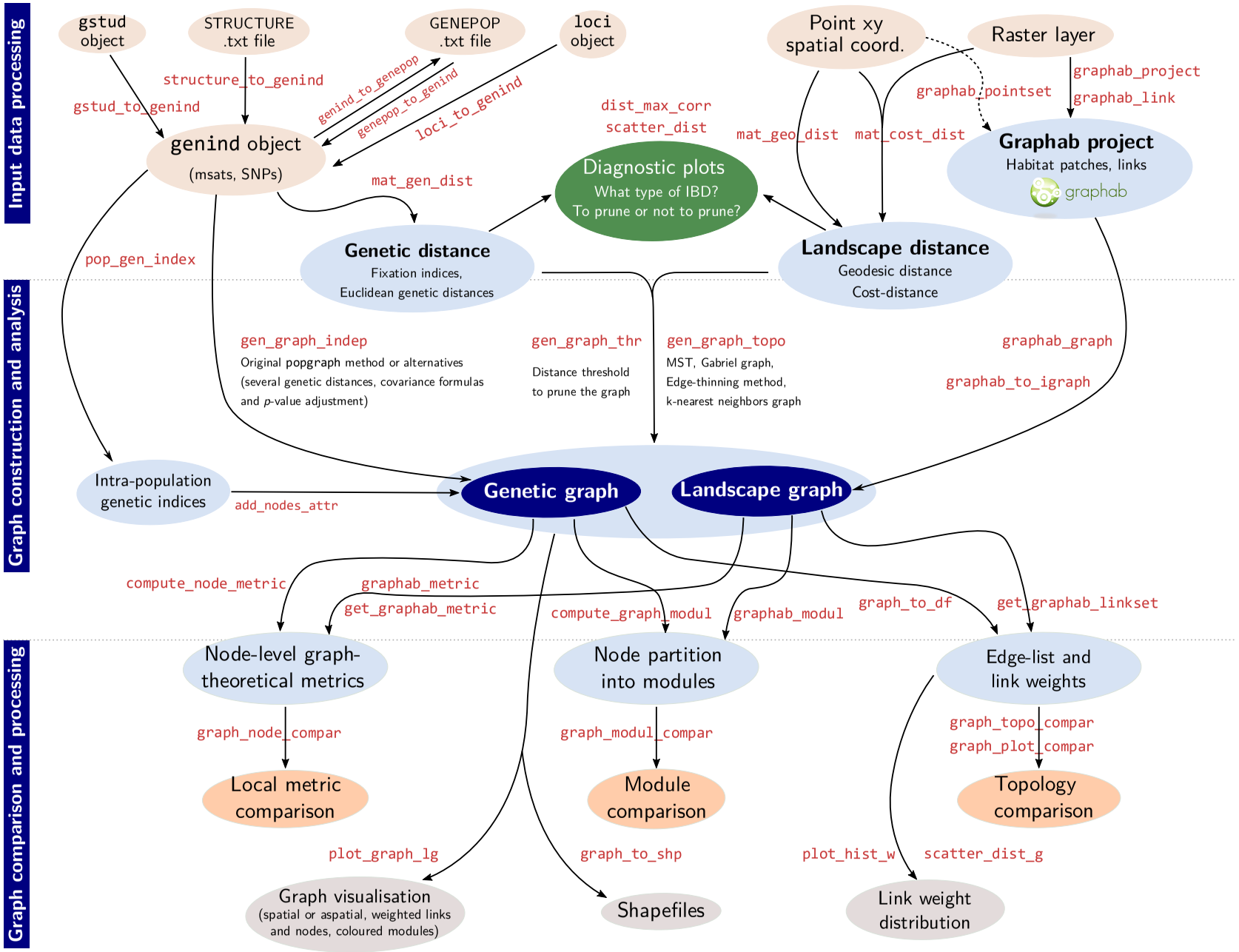


FIGURE 21 – **graph4lg** workflow. Functions are represented by arrows linking their arguments to their output. Their names are written in red. Light beige boxes correspond to input data whereas grey boxes correspond to final output that can be exported to figures or files. Light blue boxes indicate analysis output that are used to construct graphs or derived from them. Dark blue boxes correspond to the two types of graphs that can be constructed with the package. The green box corresponds to diagnostic plots that can be assessed to help choose among the construction methods. Orange boxes indicate the analyses that can be carried out at several levels using both landscape and genetic graphs.

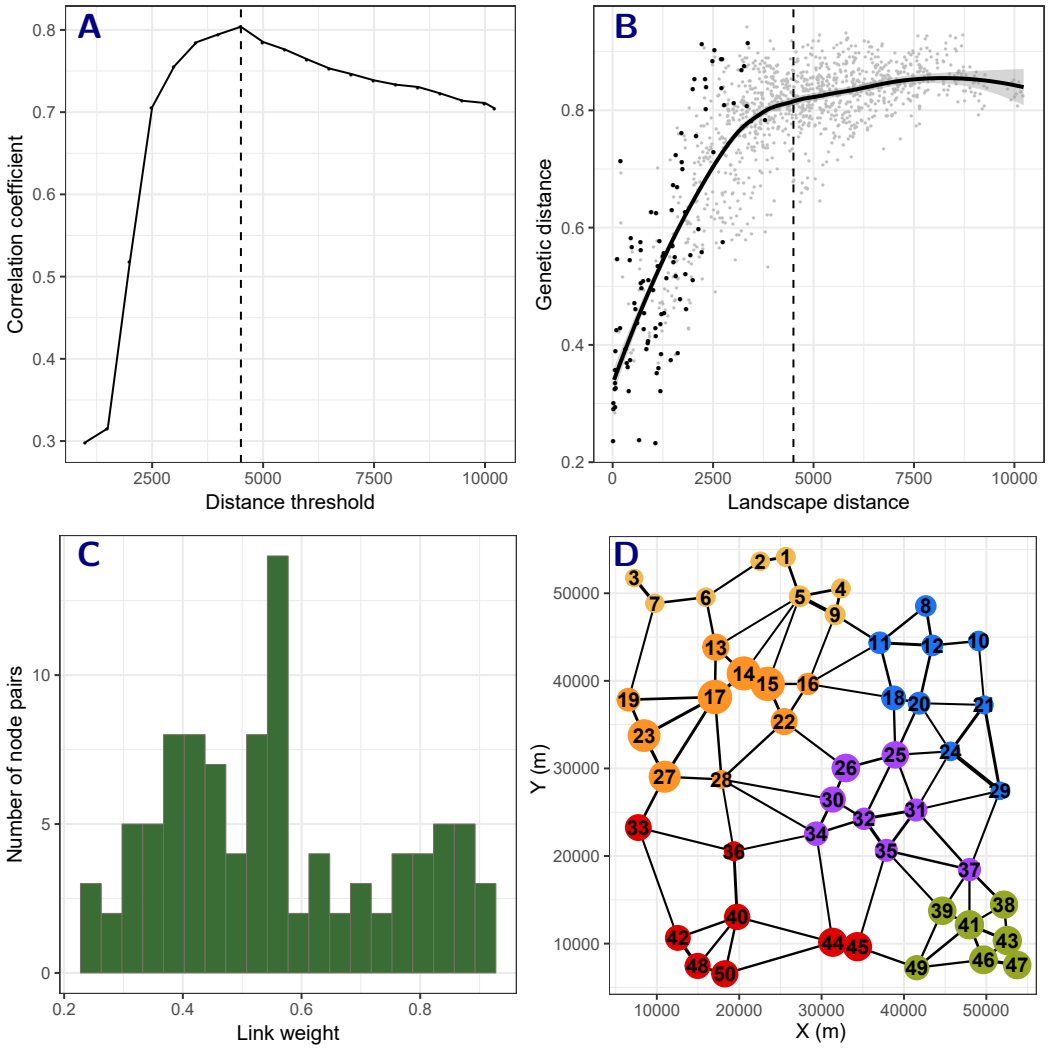


FIGURE 22 – Diagnosis plots (A, B) and genetic graph analysis plots (C, D) produced with the datasets `data_simul_genind` and `pts_pop_simul` obtained after a simulation on CDPOP(Landguth et Cushman, 2010). A) Identification of the distance of maximum correlation (DMC) (Van Strien *et al.*, 2015) with the `dist_max_corr` function. Here, the DMC corresponds to the dashed vertical line on plots A and B. B) Scatter plot produced with the `scatter_dist_g` function representing the relationship between genetic distance (D_{PS}) and cost-distance. The grey shaded region around the smoothed line corresponds to the 95 % confidence interval of the smoothing function. The black dots represent population pairs connected in the Gabriel graph. C) Histogram of the genetic distances separating the population pairs connected in the Gabriel graph produced with the `plot_w_hist` function. D) Gabriel graph mapped with the `plot_graph_lg` function. Link width is inversely proportional to the genetic distances weighting the links. Node size is proportional to the connectivity metric Flux derived from the corresponding landscape graph. Node color indicates the module to which pertains every node after a modularity analysis using `compute_graph_modul`.

3 Export facilities and included data

The graphs can be exported to shapefile layers for integration into a GIS (`graph_to_shp`) and analyses involving other types of geographical data. Included genetic and spatial data can be used to discover the functionalities of the package. We included a data set created from simulations with CDPOP(Landguth et Cushman, 2010) on a simulated landscape (`data_simul_genind`). It consists of 1500 individuals from 50 populations genotyped at 20 microsatellite loci. This dataset exhibits a typical type-IV pattern of IBD and was used to create Figure 22 and the tutorial presenting the package.

4 Limits and conclusion

Landscape and genetic graphs have great potential for analysing ecological connectivity and we do not claim to have compiled an exhaustive set of graph construction and analysis methods. Other pruning methods have been developing (Brooks, 2006 ; Greenbaum *et al.*, 2016 ; Kininmonth *et al.*, 2010 ; Milligan, in prep ; Peterson *et al.*, 2019) and could be expanded to directed graphs for example. Besides, a large range of local metrics inspired from the metapopulation theory have been developed for landscape graphs and could similarly inspire genetic graph local metrics. Graph-based analyses could also benefit from significance testing approaches through permutations to enhance their robustness. Finally, although Graphab software can efficiently handle very large spatial data sets (Foltête *et al.*, 2012a), genetic graph modelling can involve higher computational costs, thereby limiting this approach to smaller spatial and genetic datasets. Further development of the package could introduce improvements.

Our first goal in developing `graph4lg` is to bring together and make accessible a large range of methods currently used in landscape genetics for constructing and analysing graphs. We hope this package will foster the use of genetic and landscape graphs as well as further investigation regarding theoretical as well as methodological aspects.

5 Acknowledgements

We thank the editor and referees, as well as the CRAN volunteers for their constructive comments. This study is 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). We are particularly grateful to ARP-Astrance teams for their support throughout the project. We thank Ahmed Jebrane, Catherine Labruere and Catherine Laredo for their help with mathematical aspects.

6 Data availability

No empirical data were used in this article. The `graph4lg` package, all the R source codes, documentation files and vignettes can be downloaded by users under the GPL-2 license from the CRAN repository (https://cran.r-project.org/web/packages/graph4lg/index.html).

7 Authors' contributions

J.C.F., S.G. and H.M. obtained the funding for the project. P.S. and G.V. designed the package and developed the codes. P.S, J.C.F. and S.G. wrote the manuscript with significant contributions and remarks from all co-authors.

A - Supplementary tables

TABLE 4 – Functions available in the package and their dependencies on other R packages or software programs

Function name	Description	Dependencies
Input data processing		
`genepop_to_genind`	Convert a GENEPOP file into a `genind` object	`pegas`
`genind_to_genepop`	Convert a `genind` object into a GENEPOP file	None
`structure_to_genind`	Convert a file in STRUCTURE format into a `genind` object	`pegas`
`gstud_to_genind`	Convert a file from a `gstudio` or `popgraph` object into a `genind` object	`pegas`
`loci_to_genind`	Convert a `loci` object into a `genind` object	`pegas`
`pop_gen_index`	Compute population-level genetic indices	`adegenet`
`mat_gen_dist`	Compute a pairwise matrix of genetic distances between populations	`diveRsity` for fixation indices, independent codes for other genetic distances
`mat_geo_dist`	Compute Euclidean geographic distances between points	None
`mat_cost_dist`	Compute cost distances between points on a raster	`gdistance` for option `"gdistance"`, external `costdist-0.3.jar` file for option `"java"`, `raster`, `sp`, `rappdirs`
`reorder_mat`	Reorder the rows and columns of a symmetric matrix	None
`convert_cd`	Fit a model to convert cost-distances into Euclidean distances	`ggplot2`
`kernel_param`	Compute dispersal kernel parameters	None
`pw_mat_to_df`	Convert a pairwise `matrix` into an edge-list `data.frame`	None
`df_to_pw_mat`	Convert an edge-list `data.frame` into a pairwise `matrix`	None
Genetic graph construction and analysis		

Function name	Description	Dependencies
`dist_max_corr`	Compute the distance at which the correlation between genetic distance and landscape distance is maximal	`ggplot2`
`scatter_dist`	Plot scatterplots of genetic distance vs landscape distance	`ggplot2`
`gen_graph_thr`	Create a graph of genetic differentiation using a link weight threshold	`igraph`
`gen_graph_topo`	Create a graph of genetic differentiation with a specific topology	`igraph` for options `"mst"` and `"comp"`, independent codes for options `"gabriel"`, `"percol"` and `"knn"`
`gen_graph_indep`	Create an independence graph of genetic differentiation from a `genind` object	None, adapted from `popgraph`
`compute_node_metric`	Compute graph-theoretic metrics from a graph at the node level	`igraph`
`compute_graph_modul`	Compute modules from a graph by maximising a modularity index	`igraph`
Landscape graph construction and analysis		
`get_graphab`	Download Graphab in the user's machine	`rappdirs`
`graphab_project`	Create a Graphab project	External file `graphab-2.4.jar`
`graphab_link`	Create a link set in the Graphab project	External file `graphab-2.4.jar`
`graphab_graph`	Create a graph in the Graphab project	External file `graphab-2.4.jar`
`graphab_metric`	Compute connectivity metrics from a graph in the Graphab project	External file `graphab-2.4.jar`
`graphab_modul`	Create modules from a graph in the Graphab project	External file `graphab-2.4.jar`
`graphab_pointset`	Create modules from a graph in the Graphab project	External file `graphab-2.4.jar`, `sf`, `sp`
`get_graphab_linkset`	Import to R a link set computed in the Graphab project	None
`get_graphab_metric`	Import to R metrics computed at the node level in the Graphab project	None

Function name	Description	Dependencies
`graphab_to_igraph`	Create in R landscape graphs from a Graphab link set	`igraph`, `sf`, `sp`
`graph_plan`	Create a graph with a minimum planar graph topology from point coordinates	`spatstat`, `igraph`
Graph comparisons		
`graph_node_compar`	Compare the local properties of the nodes from two graphs	`igraph`
`graph_topo_compar`	Compute an index comparing graph topologies	`igraph`
`graph_plot_compar`	Visualise the topological differences between two spatial graphs on a map	`ggplot2`, `igraph`
`graph_modul_compar`	Compare the partitions into modules of two graphs	`igraph`
Graph plotting and export functionalities		
`add_nodes_attr`	Add attributes to the nodes of a graph	`igraph`
`scatter_dist_g`	Plot scatterplots of pairwise genetic and landscape distances to visualise graph pruning intensity	`ggplot2`
`plot_graph_lg`	Plot graphs in spatial or aspatial bidimensional spaces	`ggplot2`, `igraph`
`plot_w_hist`	Plot histograms of graph link weights	`ggplot2`
`graph_to_shp`	Export a spatial graph to shapefile layers	`sf`, `sp`
`graph_to_df`	Convert a graph into an edge list `data.frame`	`igraph`

Raster size	Nb. points	"gdistance"	"java"
50×50 **cells**			
	4	**0.248**	1.457
	8	**0.249**	1.480
	16	**0.289**	1.533
	32	**0.418**	1.652
	64	**0.882**	1.866
100×100 **cells**			
	4	**0.921**	1.508
	8	**0.944**	1.502
	16	**1.043**	1.551
	32	**1.160**	1.687
	64	**1.749**	2.036
200×200 **cells**			
	4	3.903	**2.191**
	8	3.902	**1.601**
	16	3.854	**1.651**
	32	4.244	**1.818**
	64	4.947	**2.288**
$400\times$ **400 cells**			
	4	15.571	**1.627**
	8	15.423	**1.687**
	16	15.283	**1.801**
	32	15.764	**1.992**
	64	16.934	**3.195**
$800\times$ **800 cells**			
	4	58.694	**1.821**
	8	57.970	**1.965**
	16	58.803	**2.137**
	32	59.546	**2.624**
	64	60.123	**7.406**
$1600\times$ **1600 cells**			
	4	250.184	**2.566**
	8	249.261	**2.910**
	16	251.396	**3.816**
	32	252.675	**4.746**
	64	256.906	**28.385**

TABLE 5 – Computation times using the function `mat_cost_dist` with options `method ="gdistance"` or `method ="java"`. A random raster is created with four equiprobable cell classes with costs of 1, 10, 100 and 1000. Points are chosen randomly on these raster surfaces and a pairwise matrix of cost distances between these points is computed. Every combination of raster size, number of points and method is randomly simulated 10 times. Mean computation times in seconds are indicated for each method. The lowest times for each combination are displayed in bold. Computations were performed on 1 core of a personal computer (Lenovo, Windows 10 Pro, Intel Xeon CPU 2 GHz, 8 Go RAM, 64 bits)

Annexe A5

Assessing the influence of the amount of reachable habitat on genetic structure using landscape and genetic graphs

Abstract

Genetic structure, i.e. intra-population genetic diversity and inter-population genetic differentiation, is influenced by the amount and spatial configuration of habitat. Measuring the amount of reachable habitat (ARH) makes it possible to describe habitat patterns by considering intra-patch and inter-patch connectivity, dispersal capacities and matrix resistance. Complementary ARH metrics computed under various resistance scenarios are expected to reflect both drift and gene flow influence on genetic structure. Using an empirical genetic dataset concerning the large marsh grasshopper (*Stethophyma grossum*), we tested whether ARH metrics are good predictors of genetic structure. We further investigated (i) how the components of the ARH influence genetic structure and (ii) which resistance scenario best explains these relationships. We computed local genetic diversity and genetic differentiation indices in genetic graphs, and ARH metrics in the unified and flexible framework offered by landscape graphs, and we tested the relationships between these variables. ARH metrics were relevant predictors of the two components of genetic structure, providing an advantage over commonly used habitat metrics. Although allelic richness was significantly explained by three complementary ARH metrics in the best PLS regression model, private allelic richness and MIW indices were essentially related with the ARH measured outside the focal patch. Considering several matrix resistance scenarios was also key for explaining the different genetic responses. We thus call for further use of ARH metrics in landscape genetics to explain the influence of habitat patterns on the different components of genetic structure.

Keywords : landscape genetics, genetic structure, amount of reachable habitat, graph theory

Cet article est en voie d'être re-soumis après modifications au journal *Heredity* :

Savary, P., Foltête, J. C., van Strien, M.J., Moal, H., Vuidel, G. & Garnier, S. Assessing the influence of the amount of reachable habitat on genetic structure using landscape and genetic graphs. In correction for *Heredity*

1 Introduction

The genetic structure of populations of the same species occupying subdivided habitat patches is characterised by two components : (i) the local genetic diversity within each population and (ii) the genetic differentiation between populations. Genetic drift and gene flow are the main processes influencing these two components when assessed from neutral genetic markers (Hedrick, 2011). Their combined effects depend on the habitat spatial pattern, i.e. the area and the configuration of habitat patches (DiLeo et Wagner, 2016 ; Keyghobadi, 2007). Indeed, on the one hand, when the effective size of a population is limited by the small area or poor quality of a habitat patch, genetic drift tends to erode its local genetic diversity (Frankham *et al.*, 2004), thereby increasing the risk of inbreeding depression and local extinction (Frankham, 2005 ; Spielman *et al.*, 2004). It also increases its genetic differentiation from other populations. On the other hand, if there are other habitat patches within dispersal distance, gene flow events due to dispersal from neighbouring populations can counterbalance this loss of local genetic diversity while limiting genetic differentiation between populations (Frankham, 2015 ; Lehnen *et al.*, 2021 ; Ingvarsson, 2001). Understanding precisely how these two components of the genetic structure are influenced by the habitat spatial pattern is crucial in an era when habitat destruction is globally threatening all biodiversity levels (Díaz *et al.*, 2019).

Describing the spatial pattern of habitat implies taking into account both habitat amount and configuration (Villard et Metzger, 2014), which are largely interdependent (Didham *et al.*, 2012 ; Saura, 2021). For a given amount of habitat in the landscape, the configuration of habitat patches determines how much habitat is reachable from every patch (Saura, 2021 ; Villard et Metzger, 2014). The concept of habitat reachability integrates both habitat amount and configuration and extends that of habitat connectivity by considering both intra-patch and inter-patch connectivity (Pascual-Hortal et Saura, 2006 ; Saura et Rubio, 2010). The Amount of Reachable Habitat (ARH) computed for a patch is made of the area of the patch itself, and of the areas of its neighbouring patches according to species dispersal capacities. In addition, a patch may contribute to the ARH at a large scale by allowing "stepping-stone" dispersal over several generations between patches that are very distant from each other (Saura *et al.*, 2014). To account for the latter situation when computing the ARH for a patch, one must consider the topology of the whole habitat network because it determines the role of that patch for indirect connections between distant patches (Saura et Rubio, 2010). Besides, as soon as the ARH includes habitat areas outside the focal patch, it should best include the resistance exerted by the landscape matrix on individual movements between patches (Andersson et Bodin, 2009 ; Joly *et al.*, 2014). In sum, computing a set of complementary metrics makes it possible to measure the ARH from the species point of view and according to its dispersal capacities through the landscape matrix over large spatial scales and multiple generations (Saura et de la Fuente, 2017).

ARH metrics have been developed from landscape graphs, which represent habitat networks as sets of habitat patches (nodes) connected by potential dispersal paths (links) (Galpern *et al.*, 2011 ; Saura et de la Fuente, 2017 ; Urban et Keitt, 2001). These graphs offer a unified framework for the computation of complementary habitat metrics in a more flexible way than commonly used metrics such as the distance to the nearest patch or the amount of habitat in a circular buffer area (see Figure 23 for background information on habitat metrics). Accordingly, ARH metrics have proven helpful for explaining biological responses such as the composition of species communities (Awade *et al.*, 2012 ; Mony *et al.*, 2018) and are commonly used for conservation purposes (Bergès *et al.*, 2020 ; Saura et

de la Fuente, 2017). They have more rarely been used to explain the genetic structure of populations despite their potential relationships with both genetic drift and gene flow processes (but see Bertin *et al.* (2017), Flavenot *et al.* (2015) and Schoville *et al.* (2018)). Three metrics can be sufficient for describing the habitat pattern properties determining the ARH (Baranyi *et al.*, 2011 ; Rayfield *et al.*, 2011). These metrics should reflect the potential size of the population occupying a patch and the contribution of a patch to dispersal fluxes and to long-distance dispersal events occurring through multiple generations over the whole habitat network. These properties have been named recruitment, flux and traversability by Urban et Keitt (2001), respectively.

In population genetics, the potential advantage of ARH metrics over other habitat metrics lies on the following rationale. Genetic drift depends on population size, which can be approximated by the capacity of a patch (i.e. recruitment component of the ARH, Figure 23C). Besides, even if every suitable habitat patch in the landscape may not be systemically occupied by a population (Pasinelli *et al.*, 2013), we can expect gene flow intensity between a given population and the others to increase with the flux component of the ARH. For a given patch, this component is measured by considering the potential connections to other habitat patches (e.g. with the F metric, Figure 23D). Finally, the relative location of a patch in the topology of the whole network, taken into account in the traversability component of the ARH, is known to be a good predictor of multi-generational gene flow (Boulanger *et al.*, 2020 ; Van Strien *et al.*, 2014 ; Van Strien, 2017)(as reflected for example by the Betweenness Centrality (BC) metric, Figure 23E). In contrast, while the distance to the nearest patch may only partially reflect the contribution of a patch to gene flow events (Figure 23A), the amount of habitat in a buffer area (Figure 23B) may not allow for distinguishing the effect of the habitat pattern on drift *versus* gene flow processes.

The most frequent landscape genetic analyses focus on the relationship between genetic and landscape distances between patches (link-level, *sensu* Wagner et Fortin (2013)) to test for the effect of landscape structure on genetic differentiation (DiLeo et Wagner, 2016). In contrast, landscape influence on local genetic diversity or population-specific indices of genetic differentiation (node- or neighbourhood-level analyses, *sensu* Wagner et Fortin (2013)) have rarely been studied (DiLeo et Wagner, 2016) (see Barr *et al.* (2015), Millette et Keyghobadi (2015) or Toma *et al.* (2015) for examples). In addition, genetic diversity estimates tend to be taken as a result of genetic drift in empirical studies, while genetic differentiation is mainly explained by levels of gene flow. However, genetic diversity and differentiation are both influenced by the interaction of drift and gene flow. Furthermore, node-based studies mostly focus on either genetic diversity or differentiation (Flavenot *et al.*, 2015 ; Toma *et al.*, 2015) and consider simple habitat metrics such as habitat amount in circular neighbourhoods around populations and distances to nearest habitat patches (Hahn *et al.*, 2013 ; Taylor et Hoffman, 2014). Because ARH metrics comprehensively reflect the drivers of both drift and gene flow, they could be relevant predictors of both genetic diversity and differentiation (Foltête *et al.*, 2020). This would help understanding how each response is influenced by the habitat spatial pattern. Computing ARH metrics under different matrix resistance scenarios additionally offers the opportunity to assess the role of matrix resistance in these relationships.

ARH metrics are even more relevant for landscape genetics since the genetic structure of a set of populations can also be represented as a genetic graph in which nodes are sampled populations

whereas links are weighted by genetic distances and represent substantial gene exchanges between populations (Dyer, 2015b ; Greenbaum et Fefferman, 2017 ; Savary *et al.*, 2021a). Their nodes can be weighted by local genetic diversity indices (node-level) as well as indices considering genetic differentiation with other populations (neighbourhood-level)(Koen *et al.*, 2016). In the latter case, the topology of the population network can be taken into account through graph pruning, which removes certain links between populations. It makes it possible to consider gene exchanges at different spatial scales when computing these genetic differentiation indices (Savary *et al.*, 2021a). As evidenced by DiLeo et Wagner (2016), node- and neighbourhood-level approaches are the only landscape genetic approaches making it possible to study the relationships between (i) either genetic diversity or differentiation and (ii) either habitat amount or configuration.

Accordingly, in this study, we aimed at answering the following question : are ARH metrics better predictors of genetic structure than commonly used habitat metrics ? To that purpose, we used an empirical genetic dataset concerning the large marsh grasshopper (*Stethophyma grossum*). This species has limited dispersal capacities and forms discrete populations in small habitat patches, making it a good model for understanding how the spatial patterns of habitat influence genetic structure. We computed local genetic diversity and genetic differentiation indices from genetic graphs. In parallel, we computed three ARH metrics (capacity, F, BC) at different scales in landscape graphs, while taking into account different matrix resistance scenarios. We also computed the distance to the nearest neighbouring patch (DistNN hereafter) and the amount of habitat in a circular buffer (buffer metric hereafter), two commonly used habitat metrics, for comparison purposes. We finally assessed the relationships between these genetic responses and landscape predictors through correlation analyses as well as partial least square regressions. These analyses also allowed us to compare the relationship between ARH metrics and either genetic diversity or differentiation, and the way the spatial scale and the resistance scenario influenced it.

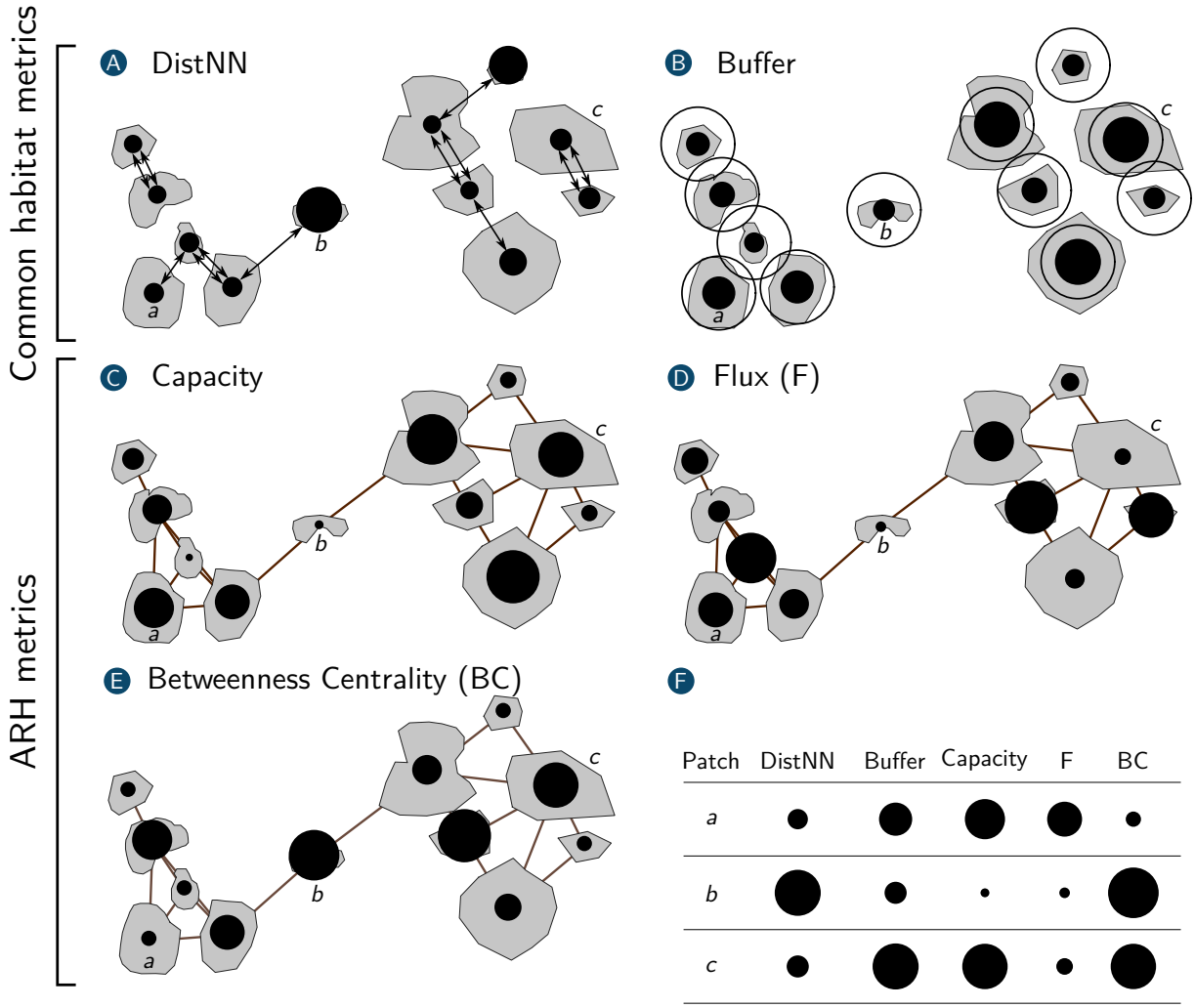


FIGURE 23 – Differences between common habitat metrics computed from a land cover map (A, B) and ARH metrics computed from a landscape graph (C, D, E). Grey areas correspond to habitat. The table (F) illustrates these differences by considering the metrics computed for three habitat patches (a, b, c). (A) The distance to the nearest habitat patch (DistNN) is computed for each habitat patch. (B) The amount of habitat in a circular buffer (so-called 'buffer' metric) is computed as the area of the pixels located within the black circle centered on each patch centroid. (C) The capacity is the area of each patch (node) of the landscape graph. (D) The Flux metric for patch i is the sum of the area of all the habitat patches j of the graph weighted by the dispersal probability between i and every patch j. (E) The Betweenness Centrality metric corresponds to the number of times every patch is located on a least cost path between two other patches of the graph, weighted by the product of the connected patch areas and the dispersal probability between them. Brown lines on panels C, D and E correspond to landscape graph links.

2 Material & Methods

2.1 Study species and sampling area

We analysed an empirical dataset acquired and described by Keller *et al.* (2013) and Van Strien *et al.* (2014). The large marsh grasshopper (*Stethophyma grossum*) is a specialist orthoptera species exhibiting a patchy distribution throughout most of Europe where it finds its habitat in periodically flooded grasslands and open wetlands (Bönsel et Sonneck, 2011 ; Reinhardt *et al.*, 2005 ; Sonneck *et al.*, 2008). In this species, dispersal seems possible even in suboptimal open areas such as dry grasslands (Marzelli, 1994) and the species is able to cross streams but suitable patches surrounded by trees cannot be reached (Reinhardt *et al.*, 2005). Exceptionally, individuals can cover up to 1500 m, as observed by Griffioen (1996) in a permeable landscape.

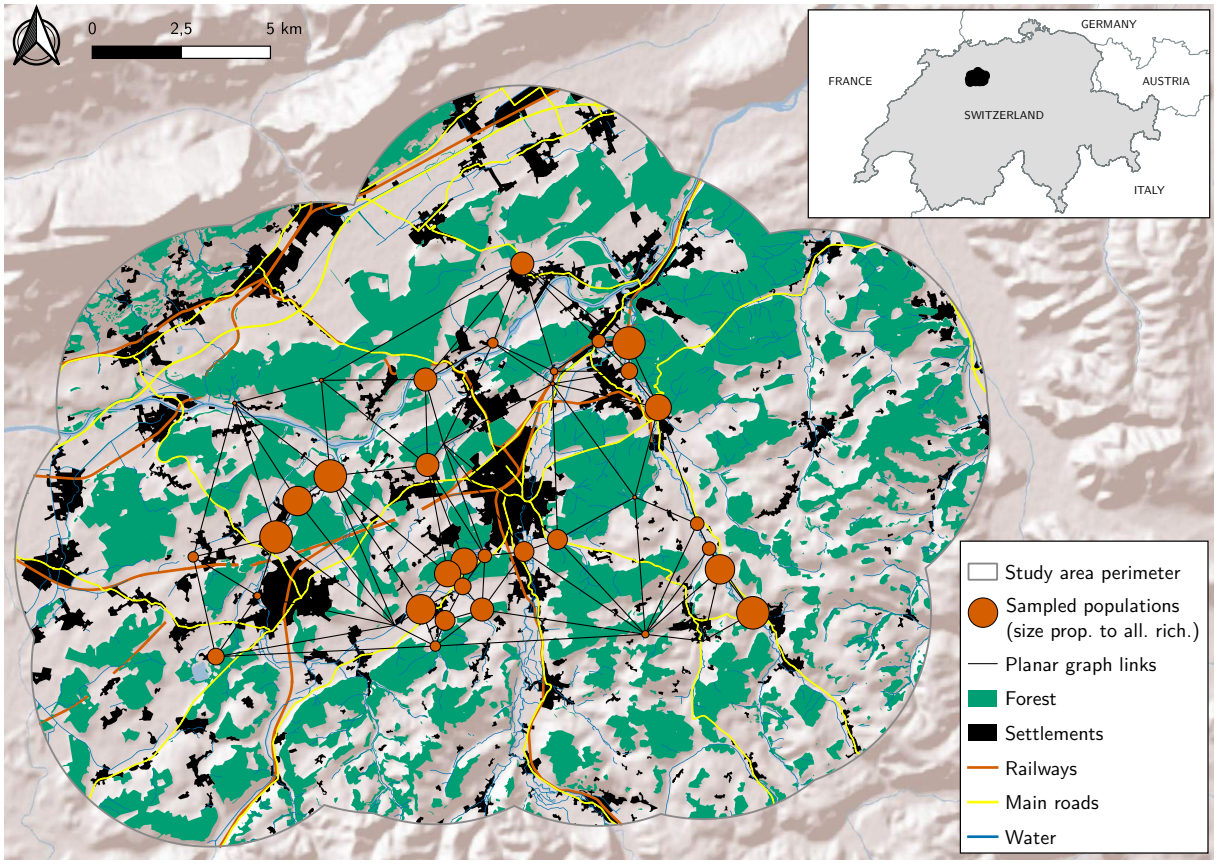


FIGURE 24 – Location of the sampled populations in the surroundings of Langental in the Oberaargau region.

Keller *et al.* (2013) modelled the potential habitat of the large marsh grasshopper in the surroundings of the city of Langenthal in the Oberaargau region of the Swiss plateau. This 180 km^2 area is characterised by intensive agriculture areas with forest patches and settlements. Across the potential habitat areas, thirty-nine large marsh grasshopper populations were sampled exhaustively (Figure 24) in July and August 2010. The tibia and tarsus of a mid-leg of each individual were sampled for genetic data analyses.

The genetic data analyses of eight microsatellite markers are described in Keller *et al.* (2013). Like those authors, we excluded the Sgr14 microsatellite marker from the analyses because of genotyping errors and high null allele frequency. This did not prevent us from detecting significant levels of genetic differentiation. Besides, we excluded two populations located on the eastern margin of the study area, as well as three other populations whose low numbers of individuals would have impaired our rarefied estimations of local genetic diversity (see below). In sum, we considered 34 populations with at least 12 individuals for a total of 886 individuals.

2.2 Genetic structure indices

At the intra-population level, we estimated the total (AR) and private (Priv. AR) allelic richness from rarefaction indices calculated using ADZE (Szpiech *et al.*, 2008) to account for sample size differences. Note that the private allelic richness index indicates the number of alleles found in a given population while absent from all the others (Kalinowski, 2004). Thus, it can be considered as both a local genetic diversity index and a genetic differentiation index. For assessing genetic differentiation

between pairs of populations, we computed the matrix of D_{PS} (calculated as 1 - pairwise proportion of shared alleles)(Bowcock *et al.*, 1994). This distance has been shown to respond quickly to recent landscape changes, making it relevant for estimating contemporary gene flow in landscape genetic analyses (Murphy *et al.*, 2010b). We also computed the matrix of pairwise F_{ST} (Weir et Cockerham, 1984), which is known to reflect historical gene flow (Latta, 2006 ; Murphy *et al.*, 2010b).

We then built genetic graphs whose nodes represented grasshopper populations. Links were weighted with either D_{PS} or F_{ST} values. In the complete graphs, every population was connected to every other population but we also created pruned graphs in which only a subset of links was included. In order to avoid any artefactual correlation between habitat metrics and graph-based genetic indices, we used a pruning method taking only genetic distances into account. To that purpose, we identified the so-called "percolation threshold" using an edge-thinning method (Urban et Keitt, 2001). Following Rozenfeld *et al.* (2008), we computed this threshold from genetic data, searching for the genetic distance associated with the graph link whose removal would break the graph into two components. All the links corresponding to genetic distances larger than this threshold were removed. Gene flow has been shown to be frequent but spatially limited in this area (Keller *et al.*, 2013) and we therefore assumed that above this genetic threshold distance, genetic differentiation between populations poorly reflected landscape effects on gene flow. From these graphs, we computed the mean of the inverse weight of the links connected to each node (thereafter referred to as MIW-D_{PS} and MIW-F_{ST} according to the genetic distance used). High values of MIW indicate a high degree of genetic similarity of a population with the others. This metric has been shown to correlate well with the number of migrants (Koen *et al.*, 2016) and other population-specific genetic differentiation indices have already been recommended and used for landscape genetic analyses (DiLeo et Wagner, 2016 ; Gaggiotti et Foll, 2010 ; Millette et Keyghobadi, 2015 ; Peterman *et al.*, 2015). Genetic graphs were constructed and metrics were computed using the `graph4lg` package in R (Savary *et al.*, 2021b).

2.3 Habitat metric calculations

We used rasterised (resolution : 10 m) land cover data from the sampling year in the area encompassing buffers of 5 km radius around each sampling site. In this area, we described the habitat spatial pattern by computing three ARH metrics (capacity, F, BC) from a landscape graph (Figure 25).

2.3.1 Landscape graph construction

We considered six land cover types : (i) potential habitat areas, (ii) forest areas, (iii) settlements, (iv) agricultural areas, (v) wetlands and water areas, and (vi) railways and roads. Potential habitat areas corresponded to areas close to open water (<= 500 m), within open agricultural areas and where water from the surroundings (500 m radius) can accumulate (Keller *et al.*, 2013). We created a resistance surface by combining these land cover data. The sampling of Keller *et al.* (2013) was exhaustive within the modelled potential habitat. Therefore, we built landscape graphs whose nodes were the 37 sampling sites in which several individuals were observed. The terms nodes, patches and populations are used interchangeably here. We used the resistance surface for computing the cost-distances between the nodes, which were used to weight the graph links.

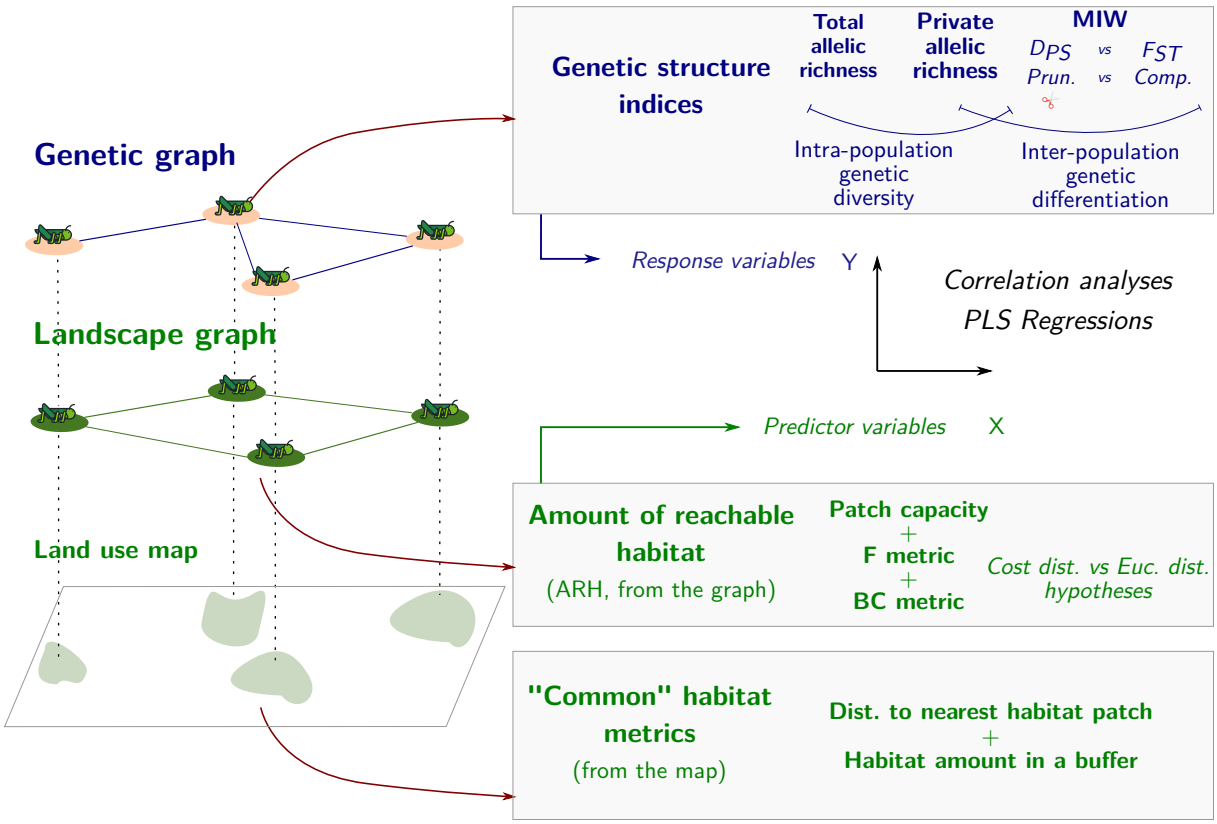


FIGURE 25 – Genetic indices and habitat metrics computed in both types of graphs to depict both genetic structure (diversity, differentiation) and the amount of reachable habitat and to perform correlation and partial least squares analyses. "Prun." : pruned graph, "Comp." : complete graph, "Cost dist." : cost-distances, "Euc. dist." : Euclidean distances

We distinguished several "expert-based" scenarios of landscape matrix resistance when assigning a cost value to every land cover type on the resistance surface. In the first four scenarios, we set the cost values as indicated in table 6. With these four scenarios, we varied the influence of wetlands and water areas (cost values : 50 or 1000, W50 and W1000 respectively) and of the roads and railways (cost values : 50 or 1000, R50 and R1000 respectively) because we wanted to test for the respective influence of these potential linear barriers on gene flow. Cost values associated with other land cover types were set assuming that this species moves easily in potential habitat areas or open areas, whereas it is hardly able to move across forests and anthropogenic areas (Bönsel et Sonneck, 2011 ; Griffioen, 1996 ; Marzelli, 1994).

TABLE 6 – Scenarios of matrix resistance considered for computing cost-distances between habitat patches

Scenario	Potential habitat	Forest	Settlements	Agricultural areas	Wetlands, water	Roads, railways
W50-R50	1	1000	1000	50	50	50
W1000-R50	1	1000	1000	50	1000	50
W50-R1000	1	1000	1000	50	50	1000
W1000-R1000	1	1000	1000	50	1000	1000

We computed the least cost paths between every pair of habitat patches using Dijkstra's algorithm and weighted the links with the corresponding cost distances. In a fifth scenario, we built a graph whose link weights were geodesic Euclidean distances between patches. As this species is assumed to

disperse by stepping stones given its limited dispersal capacities, for every resistance scenario, the landscape graphs were pruned with a Delaunay triangulation resulting in a planar graph (Figure 24).

2.3.2 Amount of reachable habitat metrics

To account for the influence of the ARH on genetic drift and gene flow, we took advantage of the spatial graph approach for computing three complementary ARH metrics. The graph nodes were located at the centroid pixel of every sampled habitat patch and we first computed their capacity as the area of potential habitat reachable at the patch scale. To that end, we assigned to every potential habitat cell surrounding the central pixel of the sampling site a weight that decreases with its cost-distance to this pixel. The weight of the potential habitat cell j located at a cost-distance d_{ij} from the central pixel of site i is equal to $e^{-\alpha d_{ij}}$, such that the $Capacity_i$ of patch i is equal to :

$$Capacity_i = \sum_{j=1}^{N} e^{-\alpha d_{ij}}$$

where N is the total number of potential habitat cells. We set α values such that $p = e^{-\alpha d_{ij}} = 0.05$ at a cost-distance equivalent to 1500 m from the sampling site centroid, because distance-weighting exponential functions assuming that landscape effects on biological responses progressively decay with distance have been shown to outperform weighting functions based on fixed distance thresholds (Miguet *et al.*, 2017). We converted this geodesic metric distance into cost-distance units using a log-log regression, following Tournant *et al.* (2013). After performing the same calculations for distances of 500 and 1000 m with very similar results, we retained 1500 m as the best scale because it is in the same order of magnitude as the maximum dispersal capacity of the species. Given that the large marsh grasshopper occupies small localised habitat patches, this metric reflects the amount of habitat reachable by individuals at the scale of the discrete patch occupied by their population. It is thus a suitable proxy for the effective population size driving genetic drift (DiLeo et Wagner, 2016). It was computed for each resistance surface and cost scenario.

As the capacity reflects the intra-patch component of the ARH, we computed two other metrics reflecting the ARH due to other patches :

— The Flux metric (F) represents the amount of habitat that is reachable when dispersing from a focal patch to other habitat patches. It can also be thought of as the amount of habitat from which migrants can originate. We computed the F using the following formula :

$$F_i = \sum_{j=1;j\neq i}^{n} Capacity_j^{\beta} e^{-\alpha d_{ij}}$$

with i the index of the focal patch and j the index of all the other n patches, d_{ij} the distances (cost-distance or geodesic Euclidean distance) between patches i and j, $Capacity_j$ is the capacity of patch j and β indicates whether the patch capacity is taken into account ($\beta = 1$) or not ($\beta = 0$) in the calculation. Note that when $\beta = 0$, the F metric is essentially a topological metric reflecting the influence of the number and proximity of patches that are reachable. α was computed according to different dispersal kernels in order to test for the influence of the scale at which dispersal takes place. To that purpose, we set α values such that $p = e^{-\alpha d_{ij}} = 0.05$ for distances d_{ij} ranging from 1500 to 7500 m (with steps of 500 m). We thereby considered the ARH

beyond the scale at which patch capacities were computed and until large scales as compared with the species dispersal capacities. For the sake of brevity, we refer to these distances d_{ij} at which $p_{d_{ij}} = 0.05$, either cost-distances or geodesic distances, as maximum dispersal distances (MDD) and express them in equivalent metric units after conversion.

— The Betweenness Centrality metric (BC) represents the number of times a focal patch (node/population) is a step on the indirect least cost path from one patch to another when considering all possible patch pairs, excluding pairs involving the focal patch itself. It therefore reflects the role of that patch for potential dispersal movements at the scale of the whole habitat network and across several generations (traversability). Each term of this metric is weighted by the product of connected patch capacities (if $\beta = 1$) and dispersal probabilities associated with the inter-patch distance such that :

$$BC_i = \sum_j \sum_k Capacity_j^{\beta} Capacity_k^{\beta} e^{-\alpha d_{jk}}$$

$$j, k \in \{1, \ldots, n\}, k < j, i \in P_{jk}$$

where P_{jk} represents the set of patches that are located along the least cost path between patches j and k. We used the same α and β values as for the F index.

Because patches with large BC values may play a key role for dispersal between a large number of habitat patches ($\beta = 0$) and/or a great amount of habitat areas ($\beta = 1$), populations occupying these patches are expected to be genetically similar to the others and to have a high genetic diversity (Zetterberg *et al.*, 2010).

As these three ARH metrics are complementary and make it possible to cover a large range of calculation parameters, other habitat metrics found in the literature (Capurucho *et al.*, 2013 ; Peterman *et al.*, 2015 ; Taylor et Hoffman, 2014) are particular cases of these metrics. Thus, although we aimed at assessing the relevance of the unified and flexible framework of the ARH metrics, we computed buffer metrics and the DistNN, two other habitat metrics, for comparative purposes. We first computed the buffer metrics, which measure the amount of potential habitat in circular neighbourhoods around each sampling sites considering similar scales as for the ARH metrics calculation. When considering small radius (from 100 to 500 m with steps of 100 m), "local buffer" metrics were akin to the capacity metric whereas "large buffer" metrics (from 1000 to 5000 m with steps of 500 m) more closely reflected the F metric calculation. We also computed the amount of potential habitat in non-circular neighbourhoods whose radius depended on cost-distance values according to every cost scenario. We use the terms "Local.Buffer" and "Large.Buffer" hereafter. In the Euclidean resistance scenario, the buffer is circular, and non-circular in the other scenarios. Finally, we computed the distance from each population to the nearest neighbour habitat patch occupied by a sampled population (DistNN) under every cost scenario. We built landscape graphs and computed metrics using Graphab 2.4 software (Foltête *et al.*, 2012a).

2.4 Analyses of the relationship between habitat metrics and genetic structure indices

2.4.1 Correlation analyses

We first assessed the correlations between the habitat metrics and the genetic indices (Figure 25). Because all these variables were not normally distributed, we computed the Spearman rank correlation coefficient and tested for the significance of the correlations. We adjusted the p-values using the Benjamini et Hochberg (1995) method to control for the False Discovery Rate.

2.4.2 Partial Least Squares regressions

Simple correlation analyses allowed us to identify the habitat metrics, spatial scales and matrix resistance scenarios most strongly related with each genetic response. However, they could not depict the complex relationships between genetic indices and our set of complementary ARH metrics. We therefore carried out Partial Least Squares regressions (PLS-R)(Carrascal *et al.*, 2009) in which genetic indices were the response variables whilst ARH metrics were the predictor variables (Figure 25). PLS regressions are an alternative to multiple linear regression and principal component regression (Roy *et al.*, 2015 ; Wold *et al.*, 2001), particularly adapted when predictor variables are collinear. The main difference with Principal Component Regression is that both the response and predictor variables are considered for creating a factorial space (Long, 2013). Response variables were rank-transformed because of departures from normal distributions. We assessed the complementarity of the ARH metrics through multivariate analyses, by testing for all combinations of three predictor variables involving a patch capacity, F and BC metric.

Following Tenenhaus (1998), we computed the Q^2 index to assess the role of every component in improving the prediction of the response variable when performing leave-one-out cross-validation. We only described the results obtained with models in which at least one component significantly improved the prediction of the response variable, i.e. when the Q^2 associated with this component is larger than 0.0975 (Supplementary information 2). We compared these models according to the Q^2 values associated with their significant components. Variable influences were assessed by computing their squared weights on the significant components. Variable weights were validated through bootstrap procedures following Pérez-Rodríguez *et al.* (2018). For every top model, the dataset was sampled with replacement 1000 times and the variable weights were estimated. If the 2.5-97.5 % interval of the series of obtained values did not overlap zero, then we considered that the variable contributed significantly to the construction of the component.

3 Results

3.1 Landscape and genetic graphs

The planar landscape graphs included 37 nodes and 95 links (Figure 24) and the complete genetic graphs included 34 nodes connected by 561 links. The genetic graphs pruned using percolation thresholds computed from D_{PS} or F_{ST} values both included 412 links, although they had slightly different topologies (Figure 27).

3.2 Correlations between ARH metrics and genetic responses

The DistNN metric never significantly correlated with any genetic index (Table 7). Although the Local.Buffer metric consistently exhibited positive correlations with genetic indices (up to $r = 0.347$

with allelic richness), this correlation was never significant. Besides, the Large.Buffer metric was only significantly correlated to the allelic richness when considering a radius equivalent to 1000 m or 4500 m in the cost scenarios assigning water areas a low resistance (W50-R1000 : $r = 0.482$ and W50-R50 : $r = 0.432$, respectively). Overall, these commonly used habitat metrics performed poorly as compared with ARH metrics derived from landscape graphs.

Allelic richness was positively correlated with patch capacity ($r = 0.447$). This correlation was only significant when the capacity was computed under the cost scenario assigning a low resistance to water areas and a high resistance to roads and railways (W50-R1000). Thus, the local genetic diversity of a population is greater when this population occupies a patch with a large habitat surface reachable without crossing roads or railways. In contrast, patch capacities were not significantly correlated with any index of private allelic richness or relative genetic differentiation (MIW) derived from the genetic graphs, whatever the genetic distance and graph topology considered in the calculation.

All genetic indices tended to correlate more strongly with the Flux (F) metrics than they did with the Betweenness Centrality (BC) metrics (Table 7). Allelic richness and private allelic richness were respectively positively and negatively correlated with both metrics (Figures 28 and 29). While allelic richness was more strongly correlated with the F metric computed considering cost-distances, especially when assigning roads and railways a high resistance (W50-R1000 : $r = 0.538$ or $r = 0.566$ when MDD = 1500 and $\beta = 0$ or $\beta = 1$ respectively, Table 7), private allelic richness was more strongly correlated with this metric when computed using Euclidean distances ($r = -0.609$ or $r = -0.593$ when MDD = 5500 or 2500 and $\beta = 0$ or $\beta = 1$, respectively, Table 7). The MDD did not have much influence on the correlation coefficients (Figure 28) and we could not identify a scale of effect. Overall, the correlation values depended only slightly on the weight given to patch capacities (β value) when computing the metrics.

The MIW indices were positively and most often significantly correlated with the F and BC metrics (Table 7), indicating that populations located in habitat patches surrounded by large and nearby habitat patches tended to be genetically more similar to others than populations located in habitat patches isolated from large habitat patches (Figure 26). Overall, the correlations were stronger when computing the MIW indices from pruned graphs rather than from complete graphs (Table 7). This was especially apparent when using the D_{PS} to weight the genetic graph links. However, these correlations were influenced by both the genetic distance used in the calculation and the type of distances (geodesic or cost-distances) used to compute the F and BC metrics. MIW-D_{PS} indices were more strongly correlated with F metrics computed using Euclidean distances whereas MIW-F_{ST} indices were more strongly correlated with F metrics computed using cost-distances under the scenarios W50-R50 or W50-R1000 which both assign a low resistance to water areas (Figure 26). In both cases, correlation coefficients reached their highest values when the MDD was between 2000 and 3000 m.

3.3 Partial Least Squares regressions

Among all combinations of capacity, F and BC metrics, only one component had a significant effect in the PLS-R models explaining one of the genetic indices, except in one case where two components significantly explained the MIW-D_{PS} derived from a pruned graph. Among these combinations, the best models were very similar for a given response variable. Overall, the best model fits were obtained

TABLE 7 – Spearman correlation coefficients between genetic indices and ARH metrics according to the cost scenario used, the MDD considered and the weight given to patch capacities in the metric calculation (β value). The largest correlation coefficient obtained for each genetic index, habitat metric and β value are displayed. The 'Signif.' column indicates whether the correlation is still significant after p-value adjustment (* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$). For the cost scenarios, refer to the table 6. 'DistNN' means 'Distance to the Nearest Neighbour'.

Genetic index	**Habitat metric**				**Correlation**	
	Metric	Cost sc.	MDD	β	$r_{Spearman}$	Signif.
AR	Capacity	W50-R1000	1500	∅	0.447	*
AR	Local.Buffer	Euc.	200	∅	0.347	
AR	DistNN	W50-R1000	∅	∅	-0.057	
AR	F	W50-R1000	1500	0	0.538	*
AR	F	W50-R1000	1500	1	0.566	*
AR	Large.Buffer	W50-R1000	1000	∅	0.482	*
AR	BC	W1000-R1000	7500	0	0.382	
AR	BC	W1000-R1000	7500	1	0.387	
Priv. AR	Capacity	W50-R1000	1500	∅	0.238	
Priv. AR	Local.Buffer	W50-R1000	300	∅	0.289	
Priv. AR	DistNN	W50-R50	∅	∅	-0.181	
Priv. AR	F	Euc.	5500	0	-0.609	**
Priv. AR	F	Euc.	2500	1	-0.593	**
Priv. AR	Large.Buffer	W50-R1000	2000	∅	0.391	
Priv. AR	BC	Euc.	3000	0	-0.437	*
Priv. AR	BC	Euc.	1500	1	-0.387	
$MIW_{comp.DPS}$	Capacity	W50-R1000	1500	∅	0.203	
$MIW_{comp.DPS}$	Local.Buffer	W50-R50	500	∅	0.274	
$MIW_{comp.DPS}$	DistNN	W50-R50	∅	∅	0.126	
$MIW_{comp.DPS}$	F	Euc.	3000	0	0.500	*
$MIW_{comp.DPS}$	F	Euc.	3500	1	0.467	*
$MIW_{comp.DPS}$	Large.Buffer	Euc.	5000	∅	-0.348	
$MIW_{comp.DPS}$	BC	Euc.	1500	0	0.380	
$MIW_{comp.DPS}$	BC	Euc.	1500	1	0.335	
$MIW_{prun.DPS}$	Capacity	W50-R1000	1500	∅	0.176	
$MIW_{prun.DPS}$	Local.Buffer	W50-R50	500	∅	0.279	
$MIW_{prun.DPS}$	DistNN	W50-R50	∅	∅	0.113	
$MIW_{prun.DPS}$	F	Euc.	3000	0	0.632	**
$MIW_{prun.DPS}$	F	Euc.	3500	1	0.597	**
$MIW_{prun.DPS}$	Large.Buffer	Euc.	5000	∅	-0.356	
$MIW_{prun.DPS}$	BC	Euc.	1500	0	0.453	*
$MIW_{prun.DPS}$	BC	Euc.	1500	1	0.411	
$MIW_{comp.FST}$	Capacity	W50-R1000	1500	∅	0.329	
$MIW_{comp.FST}$	Local.Buffer	W50-R50	500	∅	0.303	
$MIW_{comp.FST}$	DistNN	Euc.	∅	∅	-0.053	
$MIW_{comp.FST}$	F	W50-R50	3000	0	0.663	**
$MIW_{comp.FST}$	F	W50-R50	1500	1	0.602	**
$MIW_{comp.FST}$	Large.Buffer	W50-R1000	1000	∅	0.370	
$MIW_{comp.FST}$	BC	W50-R1000	7000	0	0.472	*
$MIW_{comp.FST}$	BC	W1000-R1000	1500	1	0.441	*
$MIW_{prun.FST}$	Capacity	W50-R1000	1500	∅	0.327	
$MIW_{prun.FST}$	Local.Buffer	W50-R50	500	∅	0.310	
$MIW_{prun.FST}$	DistNN	W1000-R1000	∅	∅	0.059	
$MIW_{prun.FST}$	F	W50-R50	2000	0	0.686	**
$MIW_{prun.FST}$	F	W50-R50	1500	1	0.624	**
$MIW_{prun.FST}$	Large.Buffer	W50-R1000	1000	∅	0.358	
$MIW_{prun.FST}$	BC	W50-R1000	7000	0	0.507	*
$MIW_{prun.FST}$	BC	W1000-R1000	1500	1	0.454	*

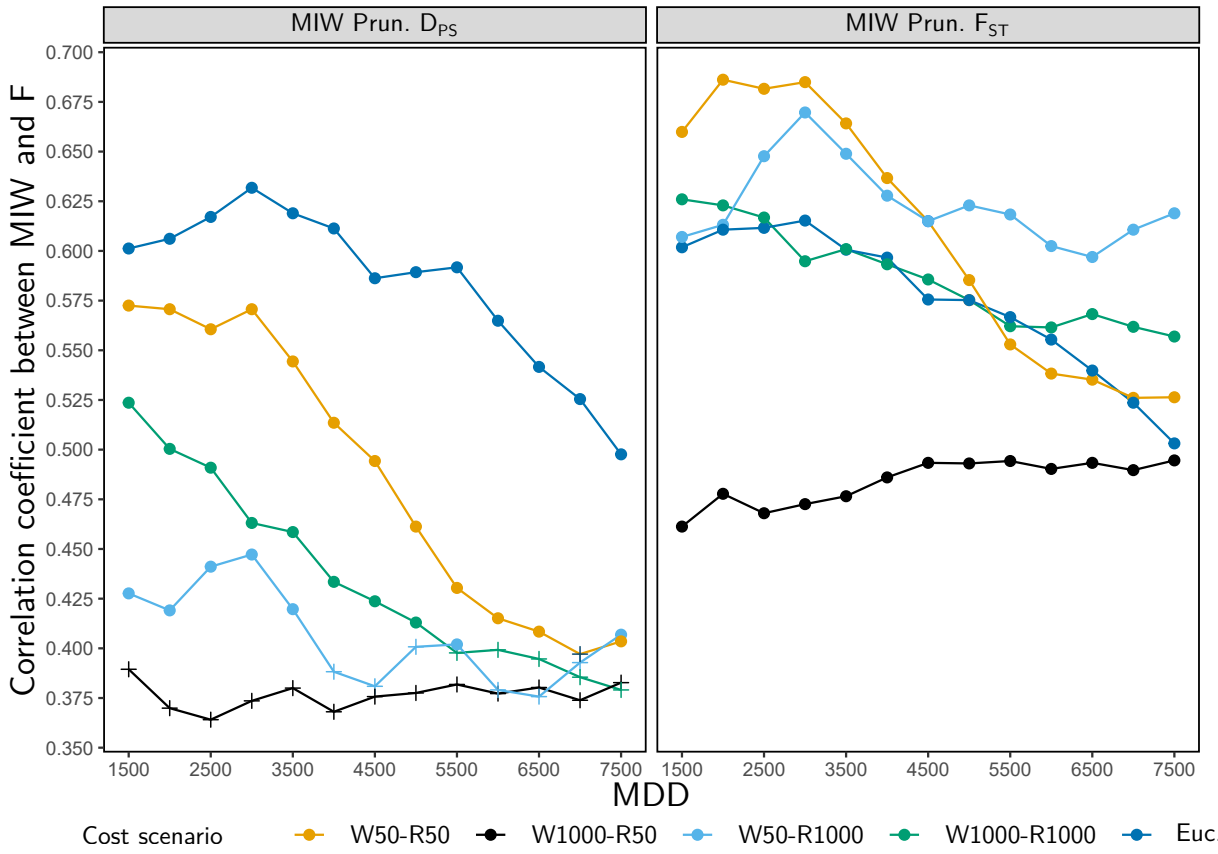


FIGURE 26 – Variation of the Spearman correlation coefficients between the relative genetic differentiation index MIW computed from the pruned genetic graphs and the F metric according to the genetic distance, cost scenarios and dispersal kernels used to compute these indices. The x axis indicates the dispersal kernels used to compute the metrics and corresponds to the MDD (maximum dispersal distances). In this figure, the F metric was computed without weighting patch capacities ($\beta = 0$). Point colours refer to the cost scenario used to compute cost-distances (see Table 6). The left and right panels display the variations observed when computing MIW from a genetic graph weighted with D_{PS} and F_{ST} values respectively. Crosses indicate that the correlation is not significant after p-value adjustment.

when patch capacities were not included in the calculation of the F metric ($\beta = 0$) and, except for the MIW-F_{ST}, included for the calculation of the BC metric ($\beta = 1$). Yet, these differences were most often subtle (Table 8). Accordingly, we only describe the results of the best models created with each response variable (Table 8).

TABLE 8 – Results of the Partial Least Squares regression (PLS-R) of the genetic indices by the capacity, flux and betweenness centrality metrics. For each genetic index and patch capacity weighting parameter for computing F and BC (β value), the best model according to the Q^2 associated with the first PLS component (Q^2.t1) is displayed (largest Q^2 value for each genetic index displayed in italics). When β is equal to 1, patch capacities are included in the metric calculation and not otherwise ($\beta = 0$). MDD indicates the distance at which the dispersal probability is set to 0.05 for the metric calculation. For the cost scenarios, refer to table 6. The r.t1 column gives the Pearson correlation coefficient between the PLS components t1 and the habitat metrics. These values are displayed in bold when the metrics significantly contribute to the construction of the PLS components. R^2.t1 and Q^2.t1 values associated with the first component respectively indicate the proportion of the response variable variance and the cross-validated proportion of this variance explained by each PLS component. Q^2 values above 0.0975 indicate that the PLS component has a significant effect on the response variable and are displayed in bold. Q^2.t2 values associated with the second component are also displayed for information purposes. $MIW_{comp.DPS}$ and $MIW_{prun.DPS}$ refer to the MIW-D_{PS} computed from complete and pruned genetic graphs respectively (similar notation for the MIW-F_{ST}).

	Capacity		**F**				**BC**				**Model fit**		
Gen. index	Cost sc.	r.t1	β	MDD	Cost sc.	r.t1	β	MDD	Cost sc.	r.t1	Q^2.t1	R^2.t1	Q^2.t2
AR	W50-R1000	**0.848**	0	2000	W50-R1000	**0.885**	0	1500	W1000-R50	0.331	**0.273**	0.326	-0.035
AR	*W50-R1000*	***0.829***	*0*	*2000*	*W50-R1000*	***0.860***	*1*	*7500*	*W1000-R1000*	***0.787***	***0.280***	*0.325*	*-0.098*
AR	W50-R1000	**0.847**	1	2000	W50-R1000	**0.916**	0	5500	W1000-R1000	**0.870**	**0.214**	0.246	-0.095
AR	W50-R1000	**0.860**	1	2000	W50-R1000	**0.880**	1	7500	W1000-R1000	**0.807**	**0.230**	0.270	-0.136
Priv. AR	W50-R1000	0.201	0	1500	Euclid.	**-0.947**	0	5500	W1000-R1000	-0.396	**0.317**	0.443	-0.003
Priv. AR	*W50-R1000*	*0.427*	*0*	*2500*	*Euclid.*	***-0.813***	*1*	*6000*	*W1000-R1000*	*0.160*	***0.411***	*0.511*	*-0.147*
Priv. AR	W50-R1000	0.221	1	2000	Euclid.	**-0.943**	0	7000	W1000-R1000	-0.385	**0.309**	0.437	-0.006
Priv. AR	W50-R1000	0.448	1	2500	Euclid.	**-0.790**	1	7000	W1000-R1000	0.182	**0.410**	0.508	-0.141
$MIW_{comp.DPS}$	W1000-R50	-0.216	0	4000	Euclid.	**0.952**	0	2500	Euclid.	**0.880**	**0.148**	0.217	-0.076
$MIW_{comp.DPS}$	*W50-R1000*	*0.390*	*0*	*4500*	*Euclid.*	***0.929***	*1*	*6500*	*W1000-R1000*	*0.492*	***0.178***	*0.291*	*0.063*
$MIW_{comp.DPS}$	W1000-R50	-0.178	1	2500	Euclid.	**0.927**	0	3500	Euclid.	**0.905**	**0.110**	0.182	-0.077
$MIW_{comp.DPS}$	W50-R1000	0.491	1	3500	Euclid.	**0.904**	1	5500	W1000-R1000	0.542	**0.098**	0.211	0.025
$MIW_{prun.DPS}$	W1000-R50	-0.131	0	3000	Euclid.	**0.958**	0	2000	Euclid.	**0.901**	**0.312**	0.366	-0.081
$MIW_{prun.DPS}$	*W50-R1000*	*0.284*	*0*	*3500*	*Euclid.*	***0.974***	*1*	*7500*	*W1000-R1000*	*0.477*	***0.314***	*0.394*	***0.101***
$MIW_{prun.DPS}$	W1000-R50	-0.093	1	3000	Euclid.	**0.938**	0	1500	Euclid.	**0.899**	**0.270**	0.325	-0.077
$MIW_{prun.DPS}$	W1000-R50	-0.084	1	3500	Euclid.	**0.947**	1	1500	Euclid.	**0.891**	**0.241**	0.303	-0.065
$MIW_{comp.FST}$	*W1000-R50*	*0.117*	*0*	*2000*	*W50-R1000*	***0.969***	*0*	*3500*	*Euclid.*	*0.618*	***0.381***	*0.445*	*-0.115*
$MIW_{comp.FST}$	W50-R1000	0.387	0	3500	Euclid.	**0.942**	1	2500	W1000-R50	0.091	**0.355**	0.448	-0.070
$MIW_{comp.FST}$	W1000-R50	0.228	1	2500	W50-R1000	**0.974**	0	3000	Euclid.	0.615	**0.324**	0.393	-0.091
$MIW_{comp.FST}$	W1000-R50	0.234	1	2500	W50-R1000	**0.969**	1	2000	Euclid.	0.701	**0.295**	0.365	-0.065
$MIW_{prun.FST}$	*W1000-R50*	*0.147*	*0*	*2500*	*W50-R1000*	***0.972***	*0*	*3500*	*Euclid.*	*0.643*	***0.443***	*0.504*	*-0.103*
$MIW_{prun.FST}$	W1000-R50	0.164	0	2500	W50-R1000	**0.965**	1	3000	Euclid.	0.714	**0.408**	0.474	-0.076
$MIW_{prun.FST}$	W1000-R50	0.226	1	2500	W50-R1000	**0.970**	0	3500	Euclid.	0.633	**0.368**	0.436	-0.083
$MIW_{prun.FST}$	W1000-R50	0.238	1	2500	W50-R1000	**0.965**	1	2500	Euclid.	0.717	**0.327**	0.402	-0.062

The allelic richness was best explained when fitting a PLS-R model including the three following ARH metrics : capacity computed under the cost scenario W50-R1000, F metric computed under the same scenario with $\beta = 0$ and a MDD of 2000 m, and BC metric computed under the cost-scenario W1000-R1000 with $\beta = 1$ and a MDD of 7500 m. These variables were highly and positively correlated with the first component (Capacity : $r = 0.829$, F : $r = 0.860$, BC : $r = 0.787$ in the best model, Table 8 and supplementary information 1, figure 31A). The R^2 associated with this component was equal to 0.325 in the best model, whereas the corresponding Q^2 was about 0.280 (Table 8).

In contrast, the only variable contributing significantly to the first component derived from the best PLS-R model explaining the private allelic richness was the F computed using Euclidean distances, with $\beta = 0$ and MDD of 1500 or 2000 m (Table 8). This variable was negatively correlated with the first component ($r = -0.816$) indicating that private allelic richness is lower when habitat patches are surrounded by other nearby habitat patches (Supplementary information 1, figure 31B). Both the R^2 and the Q^2 values associated with this first component were larger than in the PLS-R models explaining allelic richness (R^2=0.515, Q^2=0.415).

The best PLS-R models explaining the MIW indices were obtained when computing them from pruned genetic graphs, the pruning step making the greatest differences in model fits when computing the MIW-D_{PS} (Table 8). Model goodness of fit was overall better when modelling the MIW-F_{ST} than the MIW-D_{PS}. The first component alone explained about 40 % of the variance of the MIW index and up to 50 % when modelling the MIW-F_{ST} derived from a pruned graph (Table 8). This share was moderately reduced when performing the cross-validation (Q^2 : from 0.314 to 0.443, Table 8). Here again, only the F contributed significantly to the first component, which was in most cases the only component explaining significantly the MIW (Supplementary information 1, figures 31C and 31D). While the F was computed from Euclidean distances with $\beta = 0$ and MDD of 3500 m in the best model explaining the MIW-D_{PS}, it was computed with cost-distances under the scenario W50-R1000 with $\beta = 0$ and considering dispersal at a smaller scale (MDD = 2500 m) in the best model for MIW-F_{ST} (Table 8). In all cases, the correlation between the first component of the PLS models and the F was strong and positive (r about 0.97).

4 Discussion

We assessed the advantage of using complementary metrics measuring the amount of reachable habitat (ARH) instead of two other commonly used habitat metrics for explaining population genetic structure. The three ARH metrics derived from the unified and flexible framework offered by landscape graphs, i.e. the patch capacity, Flux and Betweenness Centrality metrics, were relevant predictors of the two components of genetic structure, i.e. genetic diversity and genetic differentiation. They provided an advantage over the distance to the nearest neighbour patch (DistNN) and the amount of habitat in buffer areas (Local.Buffer or Large.Buffer) that were poor predictors in this study. Besides, although allelic richness was significantly explained by the three complementary ARH metrics in the best PLS-R model, private allelic richness and MIW indices were essentially related with the ARH measured outside the focal patch. Finally, considering several matrix resistance scenarios for computing ARH metrics was key for evidencing that local genetic diversity seemed to be negatively influenced by transport

infrastructures and positively by water surfaces, whereas these landscape features did not influence genetic differentiation in the same way when measured with either the D_{PS} or the F_{ST}.

4.1 Are ARH metrics relevant predictors of genetic structure ?

All the genetic indices describing the genetic structure of the grasshopper populations were significantly correlated with at least one ARH metric and explained by these metrics in PLS models. In contrast, the two habitat metrics (DistNN and buffer metrics) previously used in landscape genetic analyses were hardly significantly correlated with the genetic indices, and in these rare cases, the correlation was much lower. Our results therefore confirm that the three ARH metrics here considered are relevant for describing the habitat pattern driving both genetic drift (capacity) and gene flow (Flux, BC) processes. Interestingly, our results match the results of Moilanen et Nieminen (2002) regarding the respective performance of several habitat metrics in predicting colonisation events. Although based on different biological responses, their results and ours provided similar evidence for the poor performance of metrics considering habitat amount in neighbourhoods delineated with fixed radius or distances to nearest patches, as compared with metrics considering dispersal probabilities to neighbouring patches.

As the computation of ARH metrics is very flexible, they include habitat metrics already computed in previous studies, as for example the amount of habitat in a circular neighbourhood with a radius of 15 km, identified by Capurucho *et al.* (2013) as the best predictor of genetic diversity in a tropical bird species (see Keyghobadi *et al.* (2005), Millette et Keyghobadi (2015) or Peterman *et al.* (2015) for other examples). Using complementary ARH metrics in this and similar study could thus have provided stronger statistical relationships and complementary insights into drift and gene flow processes driving genetic responses. In sum, although other metrics can explain genetic structure, landscape graphs offer a unified and flexible framework for understanding the influence of habitat patterns on genetic structure.

Including patch capacities in the calculation of the F and BC metrics only marginally influenced our results. Therefore, the number of reachable patches in a habitat network alone was often a good predictor of genetic structure. This recalls the results of Peterman *et al.* (2015) which have identified the isolation of a patch relative to others as the best predictor of population-specific genetic differentiation indices. Thus, the advantage of the landscape graph approach for measuring the ARH could stem from their direct consideration of population topology, already recognised as an important driver of dispersal and gene flow patterns (Saura *et al.*, 2014 ; Van Strien, 2017).

4.2 Does the ARH influence genetic diversity and genetic differentiation to the same degree and at the same spatial scale ?

It has previously been observed that genetic differentiation and local genetic diversity indices were not influenced to the same degree and at the same spatial scale by the habitat pattern (Balkenhol *et al.*, 2013 ; Keyghobadi *et al.*, 2005 ; Kierepka *et al.*, 2020 ; Taylor et Hoffman, 2014). Our results confirm these previous results given that we used a common statistical approach for analysing these two components of genetic structure. On the one hand, allelic richness was significantly correlated with both the F metric and the patch capacity and was the only genetic index significantly explained by

the capacity in the PLS models. On the other hand, private allelic richness and MIW indices appeared to be only related with F and BC metrics. Thus, local genetic diversity was influenced by the ARH at the scale of the focal patch and outside that patch, whereas genetic differentiation was influenced by the ARH outside the focal patch only. While genetic diversity and differentiation are expected to be driven by both gene flow and drift, DiLeo et Wagner (2016) suggested that a stronger effect of the local habitat amount on genetic diversity could stem from the close relationship between habitat amount and population size. In contrast, the effect of the large scale habitat pattern on migration rates seems to influence genetic differentiation more substantially than the effect of habitat area on drift does (Cushman *et al.*, 2012).

The relative genetic differentiation among populations was better explained by the spatial pattern of habitats when computed from pruned genetic graphs. The relevance of graph pruning for landscape genetic analyses has already been suggested by Wagner et Fortin (2013) for link-level analyses and evidenced by Arnaud (2003), Angelone *et al.* (2011) and Savary *et al.* (2021a), among others. Besides, Shirk et Cushman (2011) have highlighted the importance of considering the spatial distribution of populations for computing genetic diversity indices in a genetic neighbourhood including several populations. Here, we further stress the relevance of reducing the set of population pairs considered for computing neighbourhood-level genetic indices from genetic graphs. The stronger relationship between the ARH and the relative genetic differentiation when considering only population pairs connected by frequent gene flow events confirms the result obtained by Keller *et al.* (2013) when analysing this dataset. They showed that the relationship between genetic differentiation and geodesic distance was positive only up to a limited spatial scale, suggesting that the large marsh grasshopper is currently expanding. Indeed, although it has been negatively affected in the past by the reduction of wetland and grassland areas, intensive grassland management and river control reducing periodic flooding (Koschuh, 2004 ; Krause, 1996 ; Malkus, 1997 ; Reinhardt *et al.*, 2005), the species has been recolonising new areas due to wetland conservation programmes and changes in grassland management practices, among others (Trautner et Hermann, 2008). Therefore, genetic differentiation at the scale of the entire study area might not have reached its equilibrium level, as expected from the IBD pattern dynamics theorised by Slatkin (1993). In this context (case-IV IBD *sensu* Hutchison et Templeton (1999)), the genetic differentiation pattern is best explained when considering only a subset of nearby population pairs, reinforcing the interest of genetic graph pruning. In summary, the spatial and temporal scales over which drift and gene flow influence population genetic structure could be identified by jointly using landscape and pruned genetic graphs for relating ARH metrics with genetic indices.

4.3 Does the resistance of the matrix affect genetic diversity and genetic differentiation in the same way ?

The allelic richness and the relative genetic differentiation indices computed using the F_{ST} were most strongly correlated and best explained by ARH metrics computed with cost-distances. In contrast, considering geodesic Euclidean distances was the best option for explaining the private allelic richness and the relative genetic differentiation indices computed using the D_{PS}. These differences might result from i) the different time scales at which genetic diversity and differentiation respond to landscape changes and ii) the ability of genetic differentiation indices to reflect landscape influence on either historical or contemporary gene flow.

First, as expected from theory (Varvio *et al.*, 1986), genetic differentiation reaches its equilibrium level faster than local genetic diversity does. For example, Keyghobadi *et al.* (2005) detected a positive influence of recent forests on genetic differentiation in a butterfly species dispersing through open areas and avoiding forests, while local genetic diversity was best explained by patch isolation metrics taking only geodesic Euclidean distances into account. Accordingly, the results we obtained can be interpreted from the following hypotheses. The closer relationship between local genetic diversity and ARH metrics considering cost-distances instead of geodesic Euclidean distances reflects the past influence of the matrix on dispersal. Second, the closer relationship between private allelic richness and MIW indices computed from graphs pruned with the D_{PS} and ARH metrics considering Euclidean distances instead of cost-distances points towards a lower influence of matrix resistance on contemporary dispersal. These hypotheses are consistent with the current expansion of this species.

Second, previous landscape genetic studies have shown that the D_{PS} reflects recent landscape effects on genetic structure while the F_{ST} should be preferred for reflecting past landscape effects (Holzhauer *et al.*, 2006 ; Murphy *et al.*, 2010b ; Storfer *et al.*, 2010). This could explain why genetic differentiation indices computed using the F_{ST} were most correlated with ARH metrics taking into account the high resistance of some landscape features on dispersal. Although difficult to verify, this explanation would also mean that the landscape matrix have become more permeable for this species in recent years, thereby explaining its expansion.

Finally, Holzhauer *et al.* (2006) observed that roads and railways might be barriers for the large marsh grasshopper while water areas are not. Accordingly, the scenario in which roads and railways had a low resistance to movement and water areas a high resistance (W1000-R50) never provided the best fits when studying local genetic diversity and historical gene flow (F_{ST}). In contrast, the scenario in which transport infrastructures strongly limited dispersal and water areas were relatively permeable (W50-R1000) performed well in explaining these variables. This result is inconsistent with that of Keller *et al.* (2013) showing a positive effect of roads on dispersal in this species. However, these authors only considered a measure of genetic differentiation related to contemporary landscape influence on gene flow (mean assignment probabilities) as a response variable. Similarly, MIW indices based on the D_{PS} were best explained by ARH metrics without considering matrix resistance.

4.4 Limits and perspectives

The relationship between habitat structure and genetic structure is dynamic and takes time to reach an equilibrium (Slatkin, 1993). Besides, the topology of the habitat network has a strong influence on genetic structure, which may be related to the species dispersal pattern (Van Strien, 2017). Even under the hypothesis where only the amount of habitat at a given scale drives diversity patterns (Fahrig, 2013), habitat configuration has been shown to affect them significantly (Saura, 2021). For example, different traversability properties of the habitat network may influence long-distance gene flow patterns over time, which would result in a different genetic structure. We also acknowledge that the relative effects of the ARH on the two components of genetic structure here observed may be specific to the habitat spatial pattern of our case study, but our results encourage using ARH metrics in empirical landscape genetic studies. These aspects could be further investigated using ARH metrics and performing gene flow simulations with varying population sizes, topologies, dispersal capacities,

matrix resistances and habitat patterns.

Finally, our results are hardly comparable with previous ones distinguishing the effects of habitat amount and configuration on genetic structure (Cushman *et al.*, 2012 ; Jackson et Fahrig, 2015 ; Millette et Keyghobadi, 2015). Most of these studies used link-level analyses (DiLeo et Wagner, 2016), whereas here we used a node- and neighbourhood-level approach. We may wonder whether it influences the detection of landscape genetic relationships. Indeed, the MIW index is based on genetic differentiation between one population and all links with neighbouring populations. It therefore averages landscape effects over all these links, which may preclude the possibility of precisely estimating the resistance of every type of landscape feature. Besides, in most previous studies, habitat configuration measures such as inter-patch distances or patch isolation were strongly correlated with habitat amount, which should have ruled out any conclusion that habitat configuration exerts a stronger influence than habitat amount does on genetic structure (Jackson et Fahrig, 2015). Accordingly, we focused here on complementary ARH metrics derived from spatial graphs because they account for the compounded effects of both habitat amount and configuration, which are highly interdependent (Didham *et al.*, 2012). Their use has already been advocated (Saura, 2018) and we showed here that it makes it possible to understand how spatial habitat patterns influence both drift and gene flow at several spatial and temporal scales, while considering matrix resistance.

Acknowledgements

This study is 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 Daniela Keller and Rolf Holderegger for sharing the empirical and geographical data, and to ARP-Astrance team for its constant support along the project. Part of the analyses were carried out on the calculation "Mésocentre" facilities of the University of Bourgogne-Franche-Comté.

Competing interests

The authors declare no conflict of interest.

Data archiving

Genotype data are available from the Dryad repository : doi :10.5061/dryad.17cm4.

A - Supplementary figures

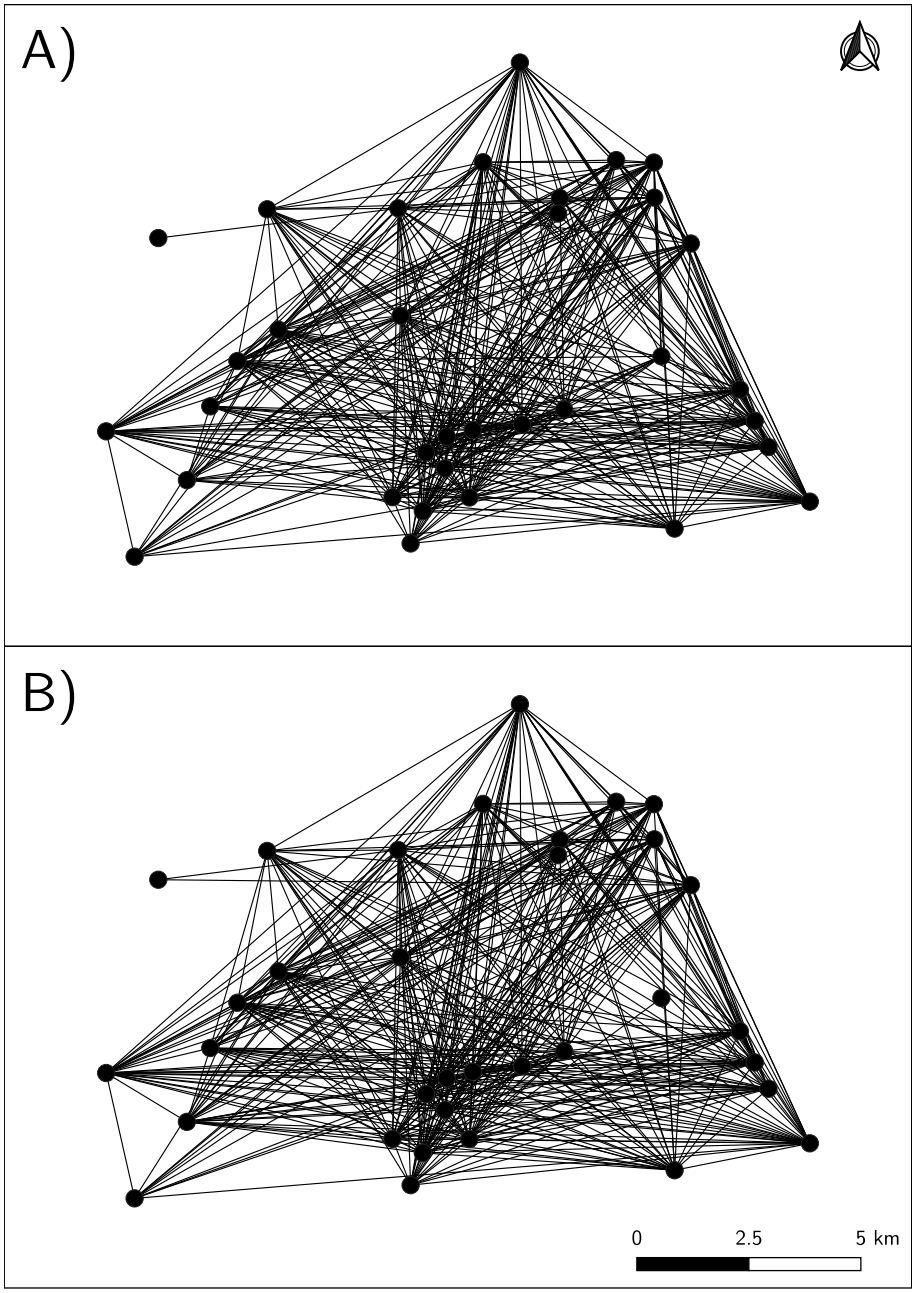


FIGURE 27 – Comparisons of the topology of the genetic graphs pruned using the percolation threshold computed with the D_{PS} (A) or F_{ST} (B). Both graphs include 412 links and 34 nodes.

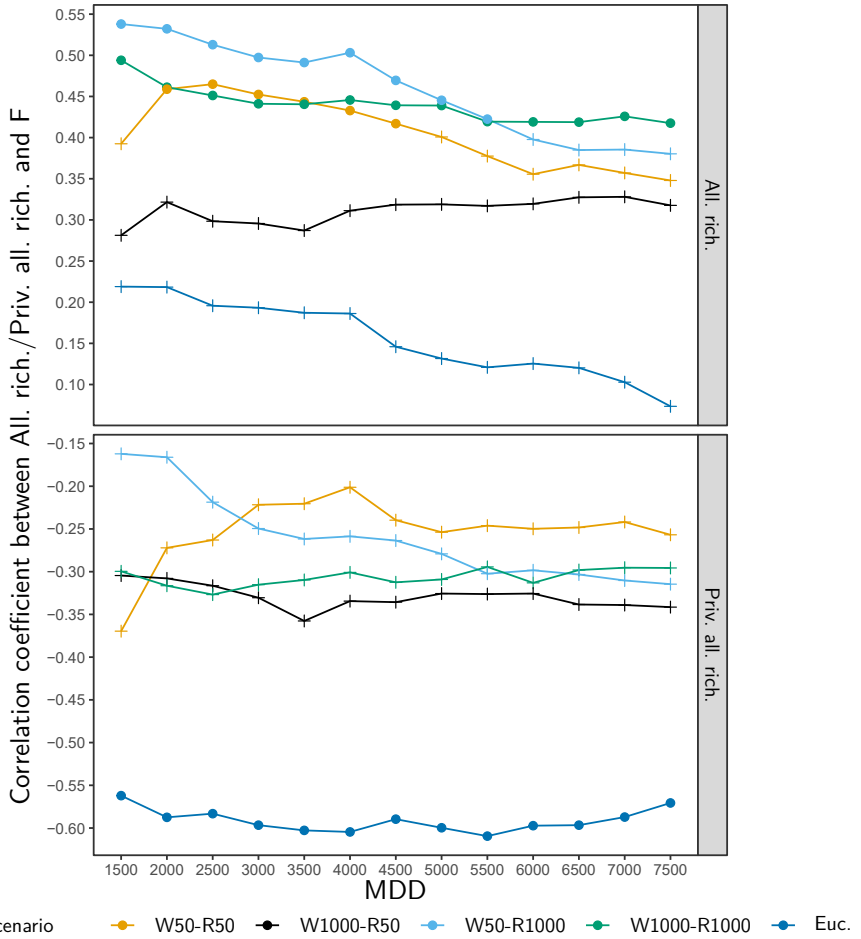


FIGURE 28 – Variation of the Spearman correlation coefficients between total allelic richness (top) or private allelic richness (bottom) and the F metric according to the cost scenarios and MDD used to compute these indices. x axis indicates the dispersal kernels used to compute the metrics and corresponds to the MDD (maximum dispersal distance). In this figure, the F metric was computed without weighting patch capacities ($\beta = 0$). Point colours refer to the cost scenario used to compute cost-distances (see Table 6). Crosses indicate that the correlation is not significant after p-value adjustment.

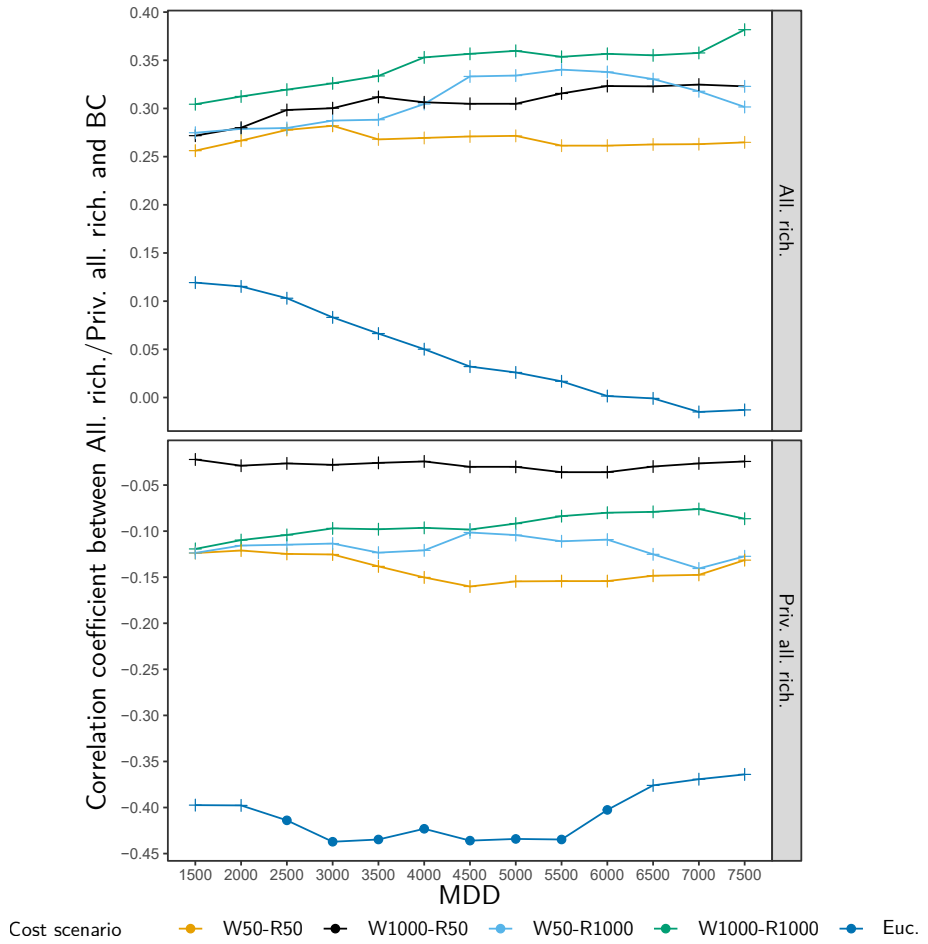


FIGURE 29 – Variation of the Spearman correlation coefficients between total allelic richness (top) or private allelic richness (bottom) and the BC metric according to the cost scenarios and MDD used to compute these indices. x axis indicates the dispersal kernels used to compute the metrics and corresponds to the MDD (maximum dispersal distance). In this figure, the BC metric was computed without weighting patch capacities ($\beta = 0$). Point colours refer to the cost scenario used to compute cost-distances (see Table 6). Crosses indicate that the correlation is not significant after p-value adjustment.

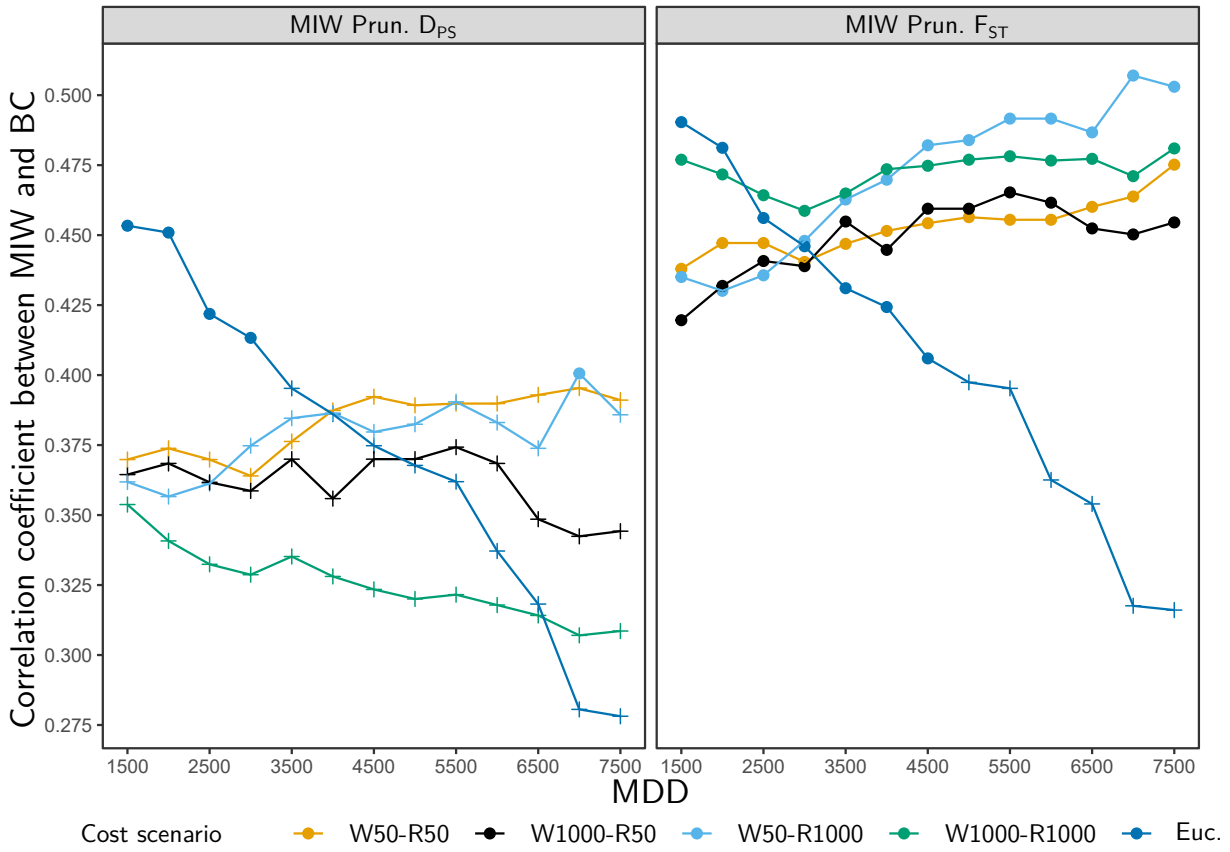


FIGURE 30 – Variation of the Spearman correlation coefficients between the MIW index computed from the pruned genetic graphs and the BC metric according to the genetic distance, cost scenarios and MDD used to compute these indices. x axis indicates the dispersal kernels used to compute the metrics and corresponds to the MDD (maximum dispersal distance). In this figure, the BC metric was computed without weighting patch capacities ($\beta = 0$). Point colours refer to the cost scenario used to compute cost-distances (see Table 6). The left and right panels display the variations observed when computing MIW from a genetic graph weighted with D_{PS} and F_{ST} values respectively. Crosses indicate that the correlation is not significant after p-value adjustment.

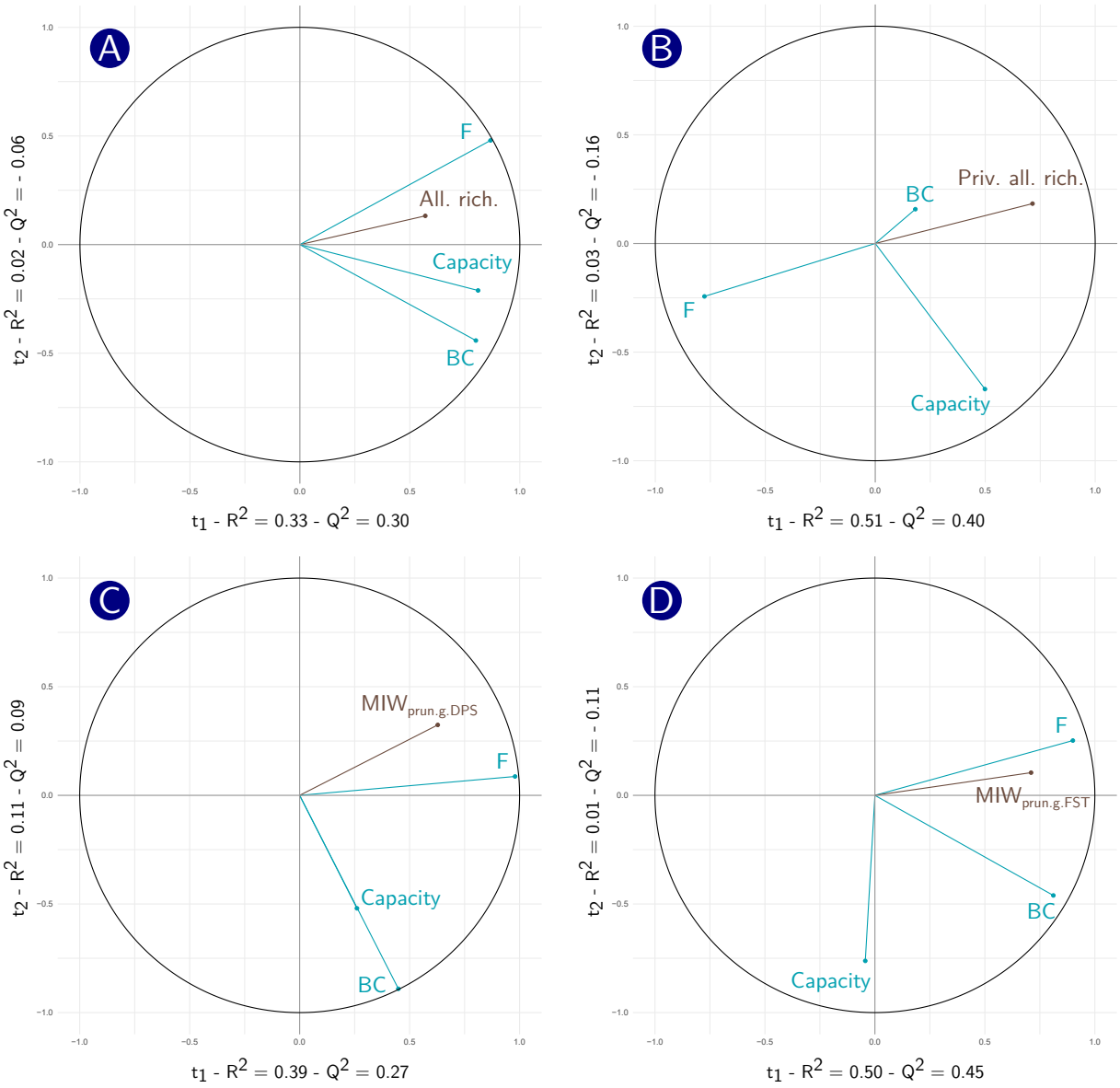


FIGURE 31 – Projection of both the response variable (genetic indices) and the predictor variables (habitat metrics : capacity, F, BC) of the best PLS-R1 regression for each genetic index (according to the Q^2 value) on the obtained factorial space. (A) Response variable : allelic richness. Predictor variables : Capacity computed under the cost scenario W50-R1000, F computed under the cost scenario W50-R1000 with a MDD of 2000 m and $\beta = 0$, BC computed under the cost scenario W1000-R1000 with a MDD of 7500 m and $\beta = 1$ (B) Response variable : private allelic richness. Predictor variables : Capacity computed under the cost scenario W50-R1000, F computed under the Euclidean cost scenario with a MDD of 2500 m and $\beta = 0$, BC computed under the cost scenario W1000-R1000 with a MDD of 6000 m and $\beta = 1$ (C) Response variable : $MIW_{prun.g.DPS}$. Predictor variables : Capacity computed under the cost scenario W50-R1000, F computed under the Euclidean cost scenario with a MDD of 3000 m and $\beta = 0$, BC computed under the cost scenario W1000-R1000 with a MDD of 7500 m and $\beta = 1$ (D) Response variable : $MIW_{prun.g.FST}$. Predictor variables : Capacity computed under the cost scenario W1000-R50, F computed under the cost scenario W50-R1000 with a MDD of 2500 m and $\beta = 0$, BC computed under the Euclidean cost scenario with a MDD of 3000 m and $\beta = 1$

B - Supplementary tables

TABLE 9 – Pearson correlation coefficients between habitat metrics and genetic indices according to the cost scenario used, the MDD considered and the weight given to patch capacities in the metric calculation (β value). The largest correlation coefficient obtained for each genetic index, habitat metric and β value are displayed. The 'Signif.' column indicates whether the correlation is still significant after p-value adjustment (* : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$). For the cost scenarios, refer to table **??** in the main document. 'DistNN' means 'Distance to the Nearest Neighbour'.

Genetic index	**Habitat metric**				**Correlation**	
	Metric	Cost sc.	MDD	β	$r_{Pearson}$	Signif.
AR	Capacity	W50-R1000	1500	∅	0.426	
AR	Local.Buffer	Euc.	200	∅	0.339	
AR	DistNN	W1000-R1000	∅	∅	0.125	
AR	F	W50-R1000	2000	0	0.480	
AR	F	W50-R1000	2000	1	0.400	
AR	Large.Buffer	W50-R1000	1500	∅	0.440	
AR	BC	W1000-R1000	7500	0	0.324	
AR	BC	W1000-R1000	7500	1	0.325	
Priv. AR	Capacity	W1000-R1000	1500	∅	0.263	
Priv. AR	Local.Buffer	W1000-R50	500	∅	0.284	
Priv. AR	DistNN	Euc.	∅	∅	-0.223	
Priv. AR	F	Euc.	1500	0	-0.507	
Priv. AR	F	Euc.	1500	1	-0.484	
Priv. AR	Large.Buffer	W50-R1000	2500	∅	0.448	
Priv. AR	BC	Euc.	4500	0	-0.380	
Priv. AR	BC	W1000-R50	1500	1	0.397	
$MIW_{comp.DPS}$	Capacity	W50-R1000	1500		0.196	
$MIW_{comp.DPS}$	Local.Buffer	Euc.	200	∅	0.211	
$MIW_{comp.DPS}$	DistNN	W1000-R1000	∅		0.167	
$MIW_{comp.DPS}$	F	Euc.	3500	0	0.466	
$MIW_{comp.DPS}$	F	Euc.	3500	1	0.383	
$MIW_{comp.DPS}$	Large.Buffer	W50-R1000	3500	∅	-0.347	
$MIW_{comp.DPS}$	BC	Euc.	3500	0	0.351	
$MIW_{comp.DPS}$	BC	Euc.	1500	1	0.291	
$MIW_{prun.DPS}$	Capacity	W50-R1000	1500	∅	0.144	
$MIW_{prun.DPS}$	Local.Buffer	Euc.	300	∅	0.190	
$MIW_{prun.DPS}$	DistNN	W1000-R1000	∅	∅	0.156	
$MIW_{prun.DPS}$	F	Euc.	3500	0	0.614	*
$MIW_{prun.DPS}$	F	Euc.	3000	1	0.544	
$MIW_{prun.DPS}$	Large.Buffer	W50-R50	4500	∅	-0.393	
$MIW_{prun.DPS}$	BC	Euc.	3500	0	0.467	
$MIW_{prun.DPS}$	BC	Euc.	1500	1	0.408	
$MIW_{comp.FST}$	Capacity	W50-R50	1500	∅	0.224	
$MIW_{comp.FST}$	Local.Buffer	W50-R1000	500	∅	0.189	
$MIW_{comp.FST}$	DistNN	Euc.	∅	∅	-0.198	
$MIW_{comp.FST}$	F	W50-R50	2500	0	0.508	
$MIW_{comp.FST}$	F	W50-R50	2000	1	0.527	
$MIW_{comp.FST}$	Large.Buffer	W50-R50	1500	∅	0.264	
$MIW_{comp.FST}$	BC	W1000-R1000	1500	0	0.475	
$MIW_{comp.FST}$	BC	W1000-R1000	1500	1	0.622	*
$MIW_{prun.FST}$	Capacity	W50-R50	1500	∅	0.224	
$MIW_{prun.FST}$	Local.Buffer	W50-R1000	500	∅	0.187	
$MIW_{prun.FST}$	DistNN	Euc.	∅	∅	-0.199	
$MIW_{prun.FST}$	F	W50-R50	2500	0	0.509	
$MIW_{prun.FST}$	F	W50-R50	2000	1	0.529	
$MIW_{prun.FST}$	Large.Buffer	W50-R50	1500	∅	0.264	
$MIW_{prun.FST}$	BC	W1000-R1000	1500	0	0.475	
$MIW_{prun.FST}$	BC	W1000-R1000	1500	1	0.623	*

C - Rationale behind the use of the Q2 to analyse PLS regression results

When performing a PLS-R1 regression, we express the response variable y as a linear combination of H components $t_1, t_2, \ldots, t_H$ such that :

$$y = c_1 t_1 + c_2 t_2 + \ldots + c_H t_H$$

where $c_1, c_2, \ldots, c_H$ are regression coefficients and $t_1, t_2, \ldots, t_H$ are components obtained such that :

$$t_1 = w_{11} x_1 + w_{12} x_2 + \ldots + w_{1p} x_p$$

and

$$w_{1j} = \frac{cov(x_j, y)}{\sqrt{\sum_{j=1}^{p} cov^2(x_j, y)}}$$

Therefore, $t_1, t_2, \ldots, t_H$ components are also linear combinations of the predictor variables $x_1, x_2, \ldots, x_p$. From that, we can express y as a function of $x_1, x_2, \ldots, x_p$:

$$\begin{aligned} y = & c_1 w_{11} x_1 + c_1 w_{12} x_2 + \ldots + c_1 w_{1p} x_p + \\ & c_2 w_{21} x_1 + c_2 w_{22} x_2 + \ldots + c_2 w_{2p} x_p + \\ & \ldots \\ & c_H w_{H1} x_1 + c_H w_{H2} x_2 + \ldots + c_H w_{Hp} x_p \end{aligned} \tag{5.1}$$

The number H of components to compute is determined through a cross-validation. For each value of h, a model with h components is computed, either from all the observations or leaving one (Leave One Out cross Validation, LOOV) or a block of observations (k-fold cross validation) out. From these models, predicted values of y are computed, either $\hat{y_{hi}}$, the prediction of y_i from the model with h components calibrated from all the observations, or $\hat{y_{h(-i)}}$, the prediction of y_i from the model with h components calibrated from a subset of the observations in which observation i is absent. Two criteria are then computed to assess the goodness of fit of these models :

$$RSS_h = \sum (y_i - \hat{y_{hi}})^2$$

and

$$PRESS_h = \sum (y_i - \hat{y_{h(-i)}})^2$$

which are respectively referred to as the Residual Sum of Squares (RSS) and PRediction Error Sum of Squares (PRESS). Adding a component is relevant if :

$$\sqrt{PRESS_h} \leq 0.95 \sqrt{RSS_{h-1}}$$

which means that when adding another component the prediction error is lower than 90.25 % of the residual sum of squares without adding this component :

$$PRESS_h \leq 0.9025 RSS_{h-1}$$

Then,

$$\frac{PRESS_h}{RSS_{h-1}} \leq 0.9025$$

and

$$1 - \frac{PRESS_h}{RSS_{h-1}} \geq 0.0975$$

Accordingly, the criterion Q^2 is equal to :

$$Q^2 = 1 - \frac{PRESS_h}{RSS_{h-1}}$$

The value of Q^2 is computed for every component h of the models. A component is considered as having a significant effect in the model if it improves the prediction of y, and therefore if $Q^2 > 0.0975$ (Tenenhaus, 1998).

Annexe A6

Validating graph-based connectivity models for a forest tropical bird species using independent presence and genetic datasets

Abstract

Habitat functional connectivity is commonly modeled by landscape graphs, i.e. sets of habitat patches (nodes) connected by potential dispersal paths (links), because graph-based distances and connectivity metrics are used for conservation or knowledge-driven approaches. They are often built from expert opinion or species distribution models (SDM) and therefore lack of validation from data more closely reflecting functional connectivity. Accordingly, we aimed to answer the following question : are landscape graphs validated by independent genetic data ? To that purpose, we modeled the habitat network of a forest bird species (Plumbeous warbler, *Setophaga plumbea*) in the Guadeloupe island with graphs built from either expert opinion, specialization indices or a SDM. In parallel, we used genetic data (712 birds from 27 populations) for computing local genetic indices and pairwise genetic distances. We finally studied the relationships between (i) genetic distances or indices and (ii) cost-distances or connectivity metrics, using MLPE distance models and Spearman correlations between metrics. Overall, the landscape graphs reliably reflected the influence of connectivity on population genetic structure, with validation R^2 above 0.25 and correlation coefficients up to 0.72. Yet, the relationship between graph ecological relevance, data-requirements and construction and analysis methods was not straightforward as the graph based upon the most complex construction method (SDM) had sometimes a lower ecological relevance than the others. Cross-validation methods and sensitivity analyses allowed us to make the advantage and limitation of every construction method spatially-explicit. In sum, we confirmed the relevance of landscape graphs for conservation modeling but we call for a case-specific consideration of the cost effectiveness of their construction methods.

Keywords : conservation modeling, habitat connectivity, landscape graphs, species distribution models, landscape genetics

Cet article est en préparation pour une soumission dans la revue *Conservation Biology* en 2021 :

Savary, P.[*], Daniel, A.[*], Foltête, J. C., Khimoun, A., Faivre, B., Ollivier, A., Moal, H., Éraud, C., Vuidel, G. & Garnier, S. Validating graph-based connectivity models for a forest tropical bird species using independent presence and genetic datasets. In prep. for *Conservation Biology*. ([*] : contributions égales des auteurs)

1 Introduction

The functional connectivity of species habitat patches depends on both the area and spatial configuration of these patches and on the degree to which the landscape matrix resists movements between them (Taylor *et al.*, 1993, 2006). It determines dispersal events and resulting ecological processes such as genetic and demographic rescue effects (Den Boer, 1968 ; Ingvarsson, 2001 ; Levins, 1969), which ensure population fitness and persistence (Frankham, 2015 ; Nieminen *et al.*, 2001 ; Saccheri *et al.*, 1998). Connectivity is therefore crucial for biodiversity conservation (Bennett, 1999 ; Correa Ayram *et al.*, 2016 ; Crooks et Sanjayan, 2006) and a key objective of current conservation policies (Hilty *et al.*, 2020). Consequently, several connectivity modeling approaches have been developed either for supporting decision making (Carroll *et al.*, 2012 ; Clevenger *et al.*, 2002) or for investigating biological responses to connectivity (Fletcher *et al.*, 2016 ; Lindenmayer *et al.*, 2020).

Connectivity modeling approaches have benefited from the increase of computational capacities and spatial and biological data availability, which opened the way for complex models closely reflecting the ecological reality (e.g. Pe'er *et al.* (2011)). Yet, the data and resources needed to implement these approaches are not available on a regular basis, making more "pragmatic" modeling approaches highly valuable (Fagan et Calabrese, 2006). Among them, landscape graphs have been argued to be an optimal compromise for balancing data requirements, model complexity and ecological relevance (Calabrese et Fagan, 2004). These tools make it possible to model an habitat network as a graph whose nodes are habitat patches and links potential dispersal paths between these nodes (Bunn *et al.*, 2000 ; Urban et Keitt, 2001). Graph-theoretical metrics quantify the role of every node and link for the connectivity of the whole network (Galpern *et al.*, 2011 ; Rayfield *et al.*, 2011) and can be predictor variables of biological responses in subsequent analyzes (Mony *et al.*, 2018 ; Ribeiro *et al.*, 2011).

Despite their strengths, landscape graphs suffer from several limitations (Moilanen, 2011). Indeed, in order to estimate functional connectivity, most graph-based connectivity models imply the computation of dispersal paths from cost surfaces. This requires the formulation of assumptions regarding the costs endured by the study species when moving across the landscape matrix (Beier *et al.*, 2008 ; Zeller *et al.*, 2012). Most of these assumptions are based upon expert opinion and whether they are close to the ecological reality is hardly ever tested for (Foltête *et al.*, 2020 ; Sawyer *et al.*, 2011). Besides, graph nodes are defined from spatial data and supposed to reflect the spatial distribution of the species habitat. Yet, it is undoubtedly a difficult task to locate such habitat areas reliably from spatial data alone, which questions the validity of the habitat delineation (Moilanen, 2011). Finally, a large range of connectivity metrics can be derived from landscape graphs (Baranyi *et al.*, 2011 ; Laita *et al.*, 2011 ; Rayfield *et al.*, 2011) but the relationship between the connectivity pattern they quantify and the biological processes they are supposed to explain has rarely been tested (Moilanen, 2011).

To overcome these limitations, several types of biological data have been used in landscape graph modeling (Foltête *et al.*, 2020). Frequently, presence data were used in a first step for deriving a species distribution model (SDM hereafter), which was then used as the basis for defining habitat patches based upon suitability thresholds and/or for creating a cost surface by converting suitability scores into cost values (Clauzel et Godet, 2020 ; Duflot *et al.*, 2018 ; Tarabon *et al.*, 2019). Modeling connectivity using presence data is not without potential limits given that i) habitat suitability is often a poor predictor of resistance to movement (Keeley *et al.*, 2017) and ii) observing the presence of an

individual is not an evidence for a successful dispersal event. Such an approach can therefore be questionable despite offering obvious advantages. Bourdouxhe *et al.* (2020) and Godet et Clauzel (2021) created landscape graphs using expert opinion, SDM or a combination of both and then compared the outputs. Although these authors precisely described the sensitivity of landscape graph modeling to the input data and prior assumptions, they lacked an independent source of biological data closely reflecting the response to connectivity for identifying which method provides the most realistic output.

Genetic data well reflect landscape connectivity (Zeller *et al.*, 2018), as they provide direct insights into dispersal-driven gene flow. According to Beier *et al.* (2008), using genetic data, animal-movement data or interpatch movement measurements is the best option to estimate cost values. These same authors consider that the second best option to that purpose is to use animal occurrence data whereas relying upon expert opinion or literature review should be reserved to cases where empirical data are not available. Yet, these assumptions have rarely been explicitly checked. To that purpose, provided that genetic data are not used in the graph modeling step, they could be good candidate for assessing the ecological validity of graphs constructed using either presence data or expert opinion (Foltête *et al.*, 2020). In studies using both genetic data and landscape graphs, these data were most often used for calibrating cost values prior to graph construction, making it impossible to independently assess the reliability of the graph a posteriori. Such an independent validation would be highly needed given the frequency of use of landscape graphs constructed without biological data closely reflecting functional connectivity. Besides, this would ensure that connectivity models reflect reliably landscape effects on genetic diversity and gene flow, two target parameters of biodiversity conservation programs (Hoban *et al.*, 2020).

Accordingly, we used two independent presence/absence and genetic datasets in order to answer the following question : are connectivity models based upon landscape graphs built from expert opinion or presence/absence data validated by empirical genetic data ? To that purpose, we modeled the habitat network of a forest bird species (Plumbeous warbler, *Setophaga plumbea*) in the Guadeloupe island because connectivity has been shown to be important for bird species whose forest habitats have been reduced (Callens *et al.*, 2011 ; Nevil Amos *et al.*, 2014) and particularly for this species (Khimoun *et al.*, 2016a, 2017). Besides, this species is endemic of a biodiversity hotspot (Myers *et al.*, 2000) facing habitat destruction and fragmentation, making the reliable modeling of its dispersal constraints a conservation issue.

2 Material & Methods

The methodology we implemented is summarized on Figure 32. We built landscape graphs with three construction methods using either expert-based information, Jacobs' specialization indices or a SDM and computed cost-distances along the links of these graphs and connectivity metrics at the level of their nodes. In parallel, we acquired genetic data from birds of 27 populations and computed population-level genetic indices and pairwise genetic distances. We then studied the statistical relationships between these genetic responses and the different cost-distance matrices and connectivity metrics deriving from each graph in order to compare the ecological validity of their outputs. We

hypothesized that distances and connectivity metrics computed from the SDM-derived graph would provide the best predictions of the genetic responses.

2.1 Study area

The French Guadeloupe island (1,713 km^2) is located in the Lesser Antilles (Figure 33). Forests cover 44 % of the island but are mainly located on Basse-Terre (BT), its western part, because volcanic soils, a more humid climate and a high relief make it less suitable for human settlement and agriculture. It is connected by a narrow isthmus to Grande-Terre (GT), the eastern part of the island in which forests have for long undergone more substantial destruction and fragmentation (Éraud *et al.*, 2009). The Plumbeous warbler (*Setophaga plumbea*) is an endemic bird species of the Caribbean known to be a forest specialist in which gene flow is dependent upon forest cover and influenced by landscape resistance (Curson, 2014 ; Khimoun *et al.*, 2016b, 2017 ; Leblond, 2008 ; Lovette *et al.*, 1998). The corollary is that studying the genetic structure of this species could give hints about the influence of connectivity on genetic processes, thereby making it useful for validating connectivity models. Besides, although this species is endemic and threatened by forest degradation, its abundance levels still make it possible to sample and contact sufficient individuals for describing its genetic structure and modelling its distribution (Éraud *et al.*, 2012).

2.2 Connectivity modeling

2.2.1 Spatial and climatic data

We used land cover and climatic spatial data for creating the SDM and the landscape graphs. CORINE Land Cover data of 2012 were complemented by the IGN BD TOPO database for locating roads and built-up elements. Besides, we used the 2010 map of 17 woody vegetation formations created by the Conseil Départemental of Guadeloupe, the IGN and the ONF using manually interpreted aerial photographs (Supporting information 14). Mean precipitation and temperature raster layers with a resolution of 20 m were obtained by interpolation of punctual data of 61 Meteo-France stations for the period 2012-2014, following Joly *et al.* (2012) and Castel *et al.* (2017).

Initial land cover types were classified into nine types (Supporting information 13 and 37). A categorical raster cost surface (resolution : 20 m) was created from these data. Besides, continuous raster layers indicating the distance from each pixel to the closest road, forest, agricultural or built-up elements and the proportion of these landscape features in circular landscapes of 500 m radius around each pixel were created for species distribution modeling.

2.2.2 Expert based information

We obtained expert-based information by asking a local ornithologist expert to :

1. Identify the vegetation formations that constitute the Plumbeous warbler habitat areas. We used them for creating the habitat patches (graph nodes) of the expert-based landscape graph (Table 10). These areas then represented a tenth class of the cost surface.
2. Assign each of the ten land cover types a cost value for computing the least-cost paths between the habitat patches (graph links) of the expert-based landscape graph. Cost values could be chosen in a free range of values starting from 1.

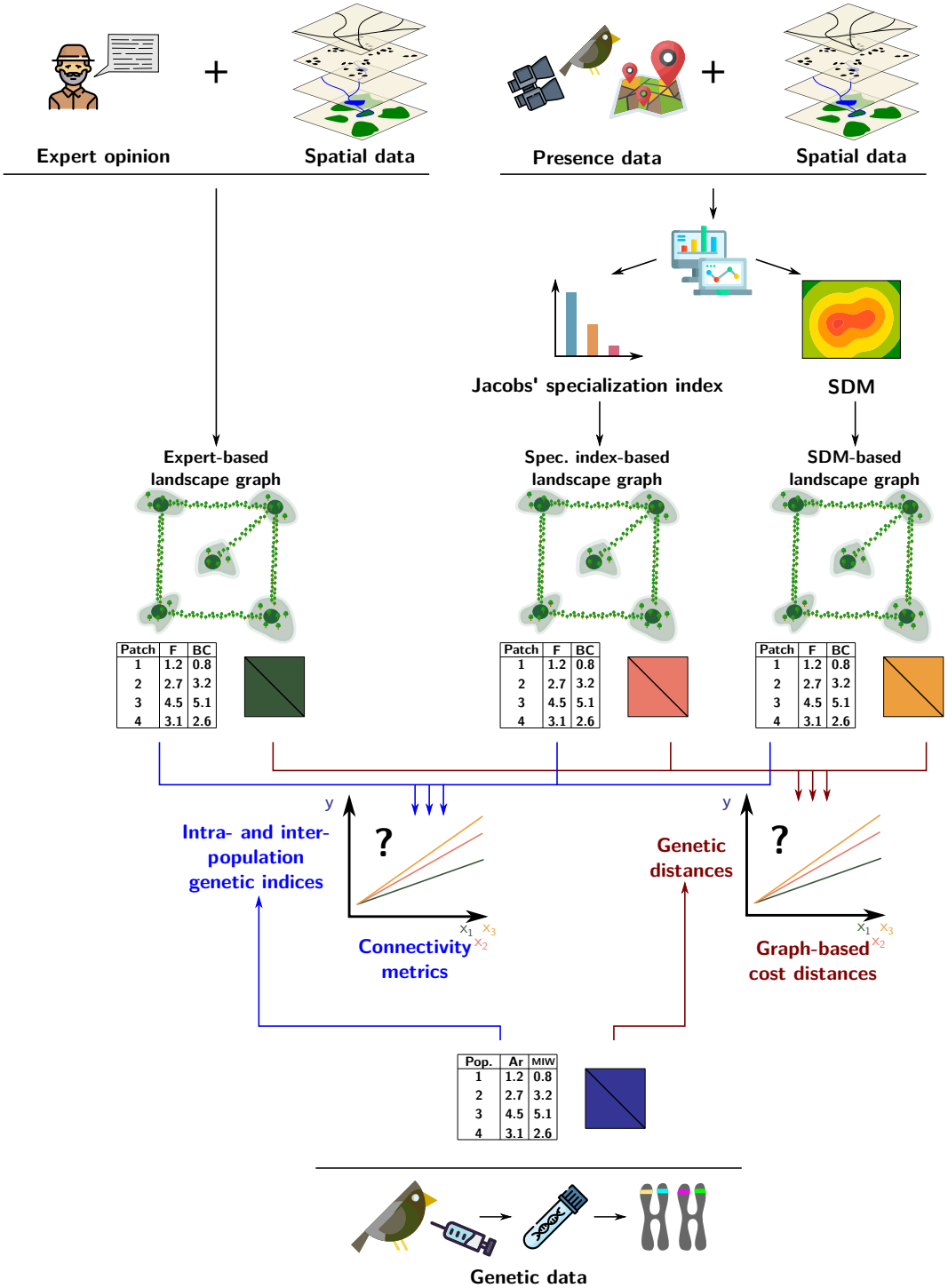


FIGURE 32 – Schematic illustration of the overall methodology

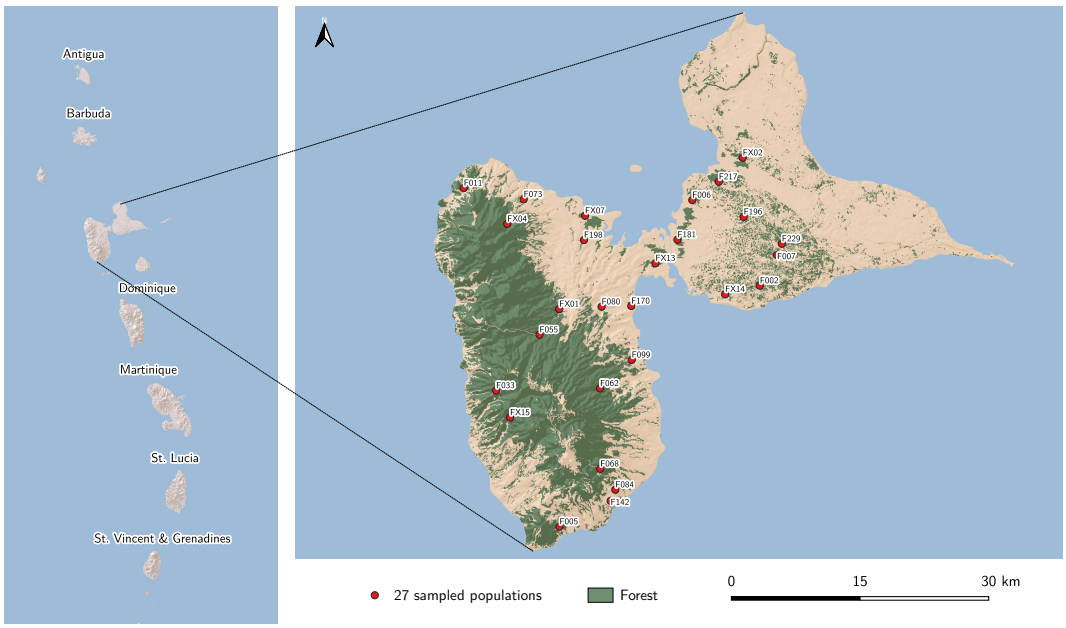


FIGURE 33 – The Guadeloupe island within the Caribbean region.

2.2.3 Specialization indices

We used 991 bird count point data acquired by the ONCFS, the Guadeloupe National Park and several NGOs between 2009 and 2011. They used a stratified sampling protocol covering the different vegetation units, modified in areas of most difficult access. Each point was surveyed twice a year from mid-april to mid-july. The observers stayed at each point for five minutes and recorded the number and species name of seen or heard individuals (Éraud *et al.*, 2012).

From the presence/absence data relative to the study species, we computed a specialization index for each of the ten land cover types following the formula of Jacobs (1974)(Supporting information B). This index varies from -1 (the species avoids the land cover type) to 1 (the species is a specialist of the land cover type). If the index is equal to 0, then the species is rather neutral relative to the land cover type. The land cover type with the maximum Jacobs' index was considered as the species habitat and used for delineating the habitat patches (graph nodes) of the specialization index-derived landscape graph (Table 10). Land cover types were assigned a cost value using a formula adapted from Bourdouxhe *et al.* (2020) :

$$Cost_i = e^{\frac{-\ln(Cost_{max})\times(Jacobs_i+1)}{Jacobs_{habitat}+1}} \times Cost_{max}$$

where $Cost_i$ and $Jacobs_i$ are respectively the cost-value and Jacobs' index associated with land cover type i and $Cost_{max}$ is the maximum cost value. $Jacobs_{habitat}$ is the maximum index value, associated with the species habitat. This formula ensured that the cost value of this preferred land cover type was equal to 1. The negative exponential distribution of the cost values reflects the fact that habitat choices are poor predictors of species dispersal capacities in a given area (Keeley *et al.*, 2016, 2017). The maximum cost value was set to 1000, reflecting previous empirical results (Gurrutxaga *et al.*, 2010).

2.2.4 Species distribution model

We created a distribution model of the Plumbeous warbler for delineating its habitat patches and deriving cost values from empirical data (Table 10). We preferred to use logistic regressions rather than using the Maxent algorithm because presence-absence methods should be used when both presence and absence data are available (Guillera-Arroita *et al.*, 2014). Besides, sampling biases are known to be better accounted for when using both presence and absence data (Fletcher et Fortin, 2018). We considered predictor variables describing the climatic and land cover contexts (see above). We controlled for the collinearity of predictor variables and conserved a set of variables that were not overly correlated ($r < 0.7$) and were potential factors explaining the distribution of the species. After preliminary analyses, we retained the following variables in the model : percentage of agricultural and artificial areas at a radius of 500 m around the data point, distance from the point to the closest road and forest and mean annual precipitation. The forest areas considered for computing the distance to the forests were those identified as being potential habitats by the experts. It reflects the fact that in practice, such information is available for the modeling.

To correct for the fact that bird count points in BT mostly followed linear and accessible tracks and were somehow aggregated, we iteratively subsampled these points to conserve only two points per track. Using this approach, we obtained consistent results and therefore randomly selected one of these samplings for computing the final model. From the confusion matrix obtained, we computed the Matthews correlation coefficient (Matthews, 1975) to assess the prediction quality of the model at different suitability thresholds. The relevance of this index comes from its consideration of every value of the confusion matrix (Baldi *et al.*, 2000).

The threshold suitability value maximizing the MCC was selected as the probability threshold above which we considered a pixel to be a habitat pixel. We converted suitability scores into cost values using the same formula as for converting specialization indices and the same maximum cost value $Cost_{max}$, such that the cost value $Cost_i$ associated with every pixel was :

$$Cost_i = e^{\frac{-\ln(Cost_{max}) \times S_i}{S_{thr.}}} \times Cost_{max}$$

where $S_{thr.}$ is the threshold probability used for delineating habitat patches and S_i is the suitability score of pixel i.

2.2.5 Landscape graphs

Graph construction

We created four landscape graphs whose nodes and links were defined in three different ways according to the information source (Table 10), using Graphab software (Foltête *et al.*, 2012a). The minimum patch size was set to 1 ha because this is the minimum area needed by a breeding pair (Éraud *et al.*, 2012). The capacity of the nodes was equal to their area. The size of some habitat patches in BT exceeded the scale at which habitat amount is supposed to influence population dynamics, which has been shown to decrease the explanation power of graph-based connectivity metrics (Laroche *et al.*, 2020). We thus controlled for the maximum size of habitat patches by fixing their maximum extent to 2,500 m side length.

Landscape graph	Node delineation	Cost-values for link weights
Expert-based (1 and 2)	Preferred vegetation formation according to expert opinion	Cost values following expert opinion
Specialization index-derived	Land cover type with the highest specialization index value	Cost values converted from specialization index values following Bourdouxhe *et al.* (2020) and assuming a cost of 1 in the preferred habitat
SDM-derived	Areas with suitability scores above the suitability threshold maximizing the prediction performance of the SDM	Cost values converted from suitability scores following Bourdouxhe *et al.* (2020) and assuming a cost of 1 in the habitat

TABLE 10 – Landscape graph elements according to the different graph construction methods. Two cost scenarios were derived from the expert responses (cf. section 3.1.1).

The topology of the landscape graphs was planar as it is a good approximation of the complete graph (Fall *et al.*, 2007) and reduces computation times. The links were weighted with the cost-distances associated with the least-cost paths between habitat patches.

Connectivity metrics

From these graphs, we computed three complementary metrics reflecting the ways each patch contributes to the connectivity of the whole habitat network (Baranyi *et al.*, 2011 ; Rayfield *et al.*, 2011). The capacity of every patch was one of these metrics as it reflects intra-patch connectivity (Pascual-Hortal et Saura, 2006) and is a proxy for the carrying capacity of the patches, an important parameter for the demographic component of the connectivity (Drake *et al.*, 2021). We then computed the Flux index (F) in order to measure the contribution of each habitat patch to immigration and emigration flows. We used the following formula :

$$F_i = \sum_{j=1 j \neq i}^{n} Capa_j \times e^{-\alpha \times d_{ij}}$$

with i the index of the focal patch, j the index of all the other n patches and d_{ij} the cost-distance between patches i and j. $Capa_j$ is the capacity of patch j. α was computed according to different dispersal kernels in order to test for the influence of the scale at which between-patch connections are assigned significant weights for computing the metrics. Exponential functions assuming that landscape effects on biological responses progressively decay with distance have already been shown to outperform weighting functions based on fixed distance thresholds (Miguet *et al.*, 2017). To that purpose, we set α values such that $p = e^{-\alpha d_{ij}} = 0.05$ for distances d_{ij} ranging from 500 to 15,000 m (with steps of 500 m), thereby considering the amount of reachable habitat (*sensu* Saura et de la Fuente (2017)) beyond the immediate neighbourhood of a population. The Euclidean dispersal distances considered for computing α were converted into cost distances using a log-log linear regression (Tournant *et al.*, 2013).

We finally computed the Betweenness Centrality (BC) index in order to quantify the role of each patch for the traversability (*sensu* Urban et Keitt (2001)) of the whole habitat network (Baranyi *et al.*,

2011 ; Bodin et Norberg, 2007) :

$$BC_i = \sum_j \sum_k Capa_j \times Capa_k e^{-\alpha \times d_{jk}}$$

$$j, k \in \{1, \ldots, n\}, k < j, i \in P_{jk}$$

P_{jk} represents the set of crossed habitat patches along the least-cost path between patches j and k. α took the same range of values as for the F calculation. This index makes it possible to identify patches that play the role of stepping stones in the habitat network. It has therefore been hypothesized that the BC index would identify habitat patches maintaining high genetic diversity levels (Zetterberg *et al.*, 2010). We computed these three metrics for every landscape graph.

2.3 Genetic data analysis

2.3.1 Field sampling and genotyping

712 birds were mist-netted in 27 forest patches (9 and 18 from GT and BT, respectively, Figure 33). 20 sites were sampled in 2015 and the other seven in 2020, all following the protocol outlined by Khimoun *et al.* (2017). To determine whether these two data sets could be pooled, four sites were sampled twice and genetic differentiation from 2015 to 2020 was tested using GENEPOP (Raymond et Rousset, 1995). This temporal genetic differentiation was not significant for three of the four populations. The relative level of genetic differentiation between the only differentiated population and the other ones remained stable from 2015 to 2020. We thus pooled the data.

The total DNA extraction was performed following either a standard phenol-chloroform protocol (Khimoun *et al.*, 2017) or from blood samples stored in Queen's lysis buffer using Blood Genomic DNA Mini-Preps Kits (BIO BASIC INC., Markham, Canada), for the 2015 and 2020 datasets respectively. 12 microsatellite loci were genotyped following PCR conditions reported in Khimoun *et al.* (2016a). Loci were amplified in simplex in a Dyad thermal cycler (Bio-Rad, Hercules, CA, USA), PCR products were multi-loaded for analyses in an automated sequencer (ABI3730), and allele scoring was performed using GENEIOUS R.8 (Kearse *et al.*, 2012).

2.3.2 Genetic structure indices

We checked for Hardy-Weinberg Equilibrium in each population and for the absence of linkage disequilibrium between pairs of loci using GENEPOP (Raymond et Rousset, 1995). The genetic diversity of each population was assessed by its allelic richness (noted Ar hereafter), computed using a rarefaction method implemented in ADZE 1.0 (Szpiech *et al.*, 2008) to account for differing numbers of individuals in the populations. The relative genetic differentiation level of each population was assessed by the Mean of Inverse Weight metric (noted MIW hereafter), using the `graph4lg` R package (Savary *et al.*, 2021b). The MIW is the mean of the inverse pairwise genetic differentiation values between each population and the 26 others. Highest values indicate populations that are the least different from the others from a genetic point of view. Koen *et al.* (2016) have shown that this index correlates well with the number of dispersing individuals in a population. Pairwise genetic differentiation values (referred to as genetic distances hereafter) were estimated by the F_{ST} (Weir et Cockerham, 1984).

2.4 Validation of landscape graph modeling

2.4.1 Cost scenario and link validation

We wanted to identify the cost scenario reflecting best how land cover types influence the dispersal movements of the study species. Genetic data were assumed to reflect such movements, if followed by gene flow. Therefore, the cost-distance values explaining best pairwise genetic distances were supposed to be the most reliable estimates of landscape connectivity among pairs of populations. For each graph, we thus computed the matrices of cost-distances between populations along the links of the landscape graphs.

We then modeled the relationship between cost-distances and pairwise genetic distances using MLPE models (Clarke *et al.*, 2002). These models explain genetic distance as a function of cost-distance and include a population random effect to control for the non-independence of the observations inherent to models based upon distance matrices. They have been shown to perform well in landscape genetic studies (Shirk *et al.*, 2017b).

The spatial scale at which genetic distances have reached an equilibrium reflecting the influence of both migration and drift has been shown to influence the assessment of the relationship between landscape distance and genetic distance (Savary *et al.*, 2021a ; Van Strien *et al.*, 2015). Accordingly, we assessed the relationship between landscape distances and genetic distances by considering the links between all population pairs or alternatively only the links between population pairs located on the same part of the island (intra-island pruning conserving only BT-BT or GT-GT links) or all population pairs separated by a cost-distance lower than several iterative threshold values (threshold pruning). The first approach reflects what is commonly done in landscape genetics, while the other two reflect the fact that a study area is sometimes divided in several parts for carrying out the analysis (Angelone et Holderegger, 2009 ; Reed *et al.*, 2017 ; Wang *et al.*, 2008), or that iterative thresholding can help understanding the scale of landscape effects on genetic structure (Angelone *et al.*, 2011 ; Emaresi *et al.*, 2011). Adopting these different approaches was also a way to assess the interpretation errors potentially resulting from the wrong choice of the spatial scale of the analysis.

We controlled for potential model overfitting by adapting the Leave One Out Cross Validation (LOOCV) method to our specific objectives and genetic distance data. When creating the MLPE models, we removed iteratively one population and all the links including it from the training data, created the model and then predicted the genetic distance values involving this population using the calibrated model. We therefore predicted 26 genetic distances at each iteration ; the genetic distance associated with a given population pair being predicted twice overall. The mean predicted genetic distances (from the two predicted values) were compared with observed genetic distances to assess the performance of each cost scenario and link set in reliably modeling the relationship between landscape distance and genetic distance. We computed a validation R-squared to quantify the prediction error (Supporting information C). We also computed the root mean square of the errors (RMSE) associated with each population pair from the two corresponding predicted values. We finally computed the mean of these RMSE at the population level by averaging the values corresponding to the 26 population pairs involving a given population and we mapped the results for locating the areas where the model

performed worst.

2.4.2 Metric validation

We aimed to identify the connectivity metrics reflecting best the genetic diversity and relative genetic differentiation indices computed at the population level. To that purpose, we extracted the values of each habitat connectivity metric corresponding to the habitat patches occupied by the sampled populations. We then computed Spearman correlations between the two genetic indices (Ar, MIW) and the three connectivity metrics (Capacity, F, BC), for all the parameters used for computing the connectivity metrics. Finally, in order to illustrate the potential interpretation errors resulting from inadequate calculation parameters, we compared the rank of the occupied habitat patches in terms of connectivity according to (i) the F metric most strongly correlated to the genetic indices or alternatively, (ii) according to this same metric computed using different computation parameters, and mapped the rank differences.

3 Results

3.1 Landscape graphs

3.1.1 Graph construction parameters

Expert opinion

According to the expert, the Plumbeous warbler occupies all the woody vegetation formations but two : coastal thickets and fallow lands with low-growth woody vegetation (Supporting information 14). The assigned cost values ranged from 1 to 10, the smallest values being assigned to forests, followed by agricultural areas, semi-open areas and wetlands. On the opposite, the ocean and artificialized areas were the most resistant land cover types. For allowing comparisons with the other cost scenarios, we rescaled these cost values to set the maximum at 1000 by either (i) multiplying the values by 100, thereby conserving the relative contrasts (expert-based 1 scenario), or by (ii) rescaling them between 1 and 1000 to conserve the same range as in the other cost scenarios (expert-based 2 scenario)(Table 11, Supporting information 39).

Specialization indices

The maximum Jacobs' specialization index was obtained for the habitat areas delineated by the expert (0.73) whereas the minimum were obtained for open areas (-0.83), agricultural areas (-0.68) and artificial areas (-0.26)(Table 11). Given the absence of bird count points in water, wet areas or in the ocean, we set the corresponding indices at -1. Although there was not any point in semi-natural areas, these areas were mostly urban green spaces and we assigned them the same index as for artificial areas. The cost values converted from these indices ranged from 1 to 1000, with sharp contrasts as illustrated by the costs of 52 and 241 respectively assigned to artificial and agricultural areas (Supporting information 39).

Species Distribution Model

The bird count point dataset included 206 presence points and 785 absence points of the study species. The SDM we used for constructing the graph had a good prediction accuracy, with an AUC

Land cover type	Initial values		Cost values			
	Exp.	Spec.index	Exp.1	Exp.2	Spec.index-derived	SDM-derived
Habitat	1	0.73	100	1	1	9 ± 84
Forest	2	0.46	200	112	3	127 ± 308
Semi-natural	4	-0.26	400	334	52	314 ± 367
Semi-open	3	0.29	300	223	6	577 ± 450
Open areas	4	-0.83	400	334	509	544 ± 428
Agricultural	3	-0.68	300	223	241	598 ± 420
Wet areas	3	-1	300	223	1000	109 ± 264
Water	4	-1	400	334	1000	459 ± 441
Ocean	10	-1	1000	1000	1000	993 ± 76
Artificial	10	-0.26	1000	1000	52	510 ± 405

TABLE 11 – Specialization indices and raw expert-based cost values associated with every land cover type and corresponding cost values according to each cost scenario. Exp.1 and Exp.2 are the two cost scenarios deriving from expert opinion. Spec.index refers to the cost values obtained from the Jacobs' specialization indices and SDM-derived refers to the cost values obtained from the Species Distribution Model. In this latter case, the range of cost values is continuous rather than discrete and mean (± SD) cost values have been computed within each land cover type for comparison purposes.

equal to 0.918 (Supporting information 38). The suitability threshold maximizing the Matthews' Correlation Coefficient and above which we considered a pixel to be a habitat pixel was equal to 0.328 (MCC : 0.547). The distance to the closest forest had the strongest effect. It negatively influenced the suitability scores, as did the proportion of agricultural and artificial pixels in the surrounding of a pixel. In contrast, an increase of the distance to the closest road and of the mean precipitation tended to increase the suitability.

In the SDM-derived cost scenario, the values were very contrasted, the expert-based habitat areas being the only land cover type with a mean cost value lower than 100 (Table 11, Supporting information 37 and 39). Agricultural areas (25.4 % of the terrestrial area) took a substantially larger mean cost value as compared with their value in the expert-based (1 and 2) or specialization index-derived cost scenarios (598 *vs* 300, 223 and 241, respectively). Artificial areas (20.8 %) took in average a lower cost value (510) than in the expert-based scenarios (1000) but a larger value than in the specialization index-derived scenario (52, Table 11).

3.1.2 Graph element properties

The landscape graphs built using the different patch delineations and cost scenarios differed by both their node and link properties. The expert-based and specialization index-derived graphs shared the same nodes. They had more nodes with a smaller average area (1653, mean area : 39 ha) than the SDM-derived graph (621, mean area : 102 ha). Indeed, the inclusion of the distance to the closest forest in the SDM tended to incorporate non-forest pixels located on the margin of forest pixels in the habitat patches delineated from this model, sometimes enlarging them. Besides, the effect of the other predictor variables tended to decrease suitability scores in less forested areas, thereby decreasing the number of patches (Supporting information 40 and 41).

Because of the lower number of patches and of the high cost values assigned to both agricultural and artificial areas when building the graph from the SDM, the distribution of the cost-distances between populations along the links of this graph had a different shape than those obtained from the other graphs (Supporting information 40 and 41). In particular, the cost-distances between populations

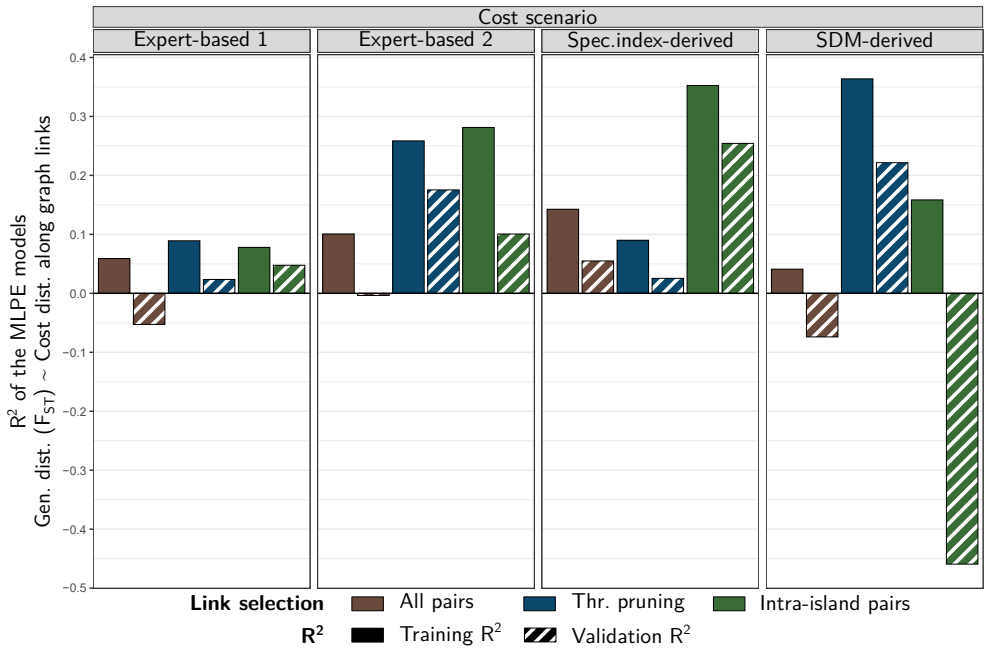


FIGURE 34 – R^2 of the MLPE models explaining pairwise genetic distances (F_{ST}) as a function of the cost-distances between populations along the links of the landscape graphs. Vertically separated boxes distinguish the different cost scenarios used for computing cost-distances when creating landscape graphs. 'Expert-based 1 and 2' : expert-based cost scenarios, 'Spec.index-derived' : cost values converted from Jacobs' specialization indices, 'SDM-derived' : cost values converted from suitability scores of the Species Distribution Model. Bar colors indicate the set of links (population pairs) considered in the models. 'Complete' : all population pairs, 'Thr. pruning' : population pairs separated by a cost-distance lower than a given cost-distance threshold (only the results of the best models obtained using a distance threshold are reported for each cost scenario), 'Intra-island' : only pairs of populations located on the same part of the island are considered (BT-BT or GT-GT). Plain and striped bars respectively indicate the R^2 obtained with the training data or with the validation data (Leave-One-Population-Out Cross Validation).

separated by the isthmus (BT-GT) were much larger when using the SDM-derived scenario rather than the other scenarios. Overall, the latter cost-distances were larger than the cost-distances between populations pairs from BT or GT (BT-BT or GT-GT).

3.2 Genetic structure

Populations from BT and GT had similar levels of genetic diversity in average although the allelic richness was more variable in BT (BT : 4.63 ± 0.34, GT : 4.82 ± 0.13, Supporting information 15). In BT, the main differences were observed between populations located inside the largest forest patches and those located on their margins. Similarly, the indices of relative genetic differentiation (MIW) took values larger than 50 in seven populations of the largest forest patches in BT, intermediate values (26-50) in all GT populations and values lower than 26 in nine BT populations, mostly located on the margins of the main forest patches (Supporting information 15).

3.3 Cost scenario and graph link validation

The MLPE models explaining genetic distances (F_{ST}) in function of the cost-distances between populations along the graph links provided contrasted results according to the cost scenario and the set of links considered (Figure 34). Whatever the cost scenario, the models considering all population pairs had low calibration R^2 and validation R^2. The highest calibration R^2 were obtained with the cost scenarios deriving either from the SDM (0.36) or the specialization indices (0.35), followed by the expert-based 2 scenario (0.28)(Figure 34).

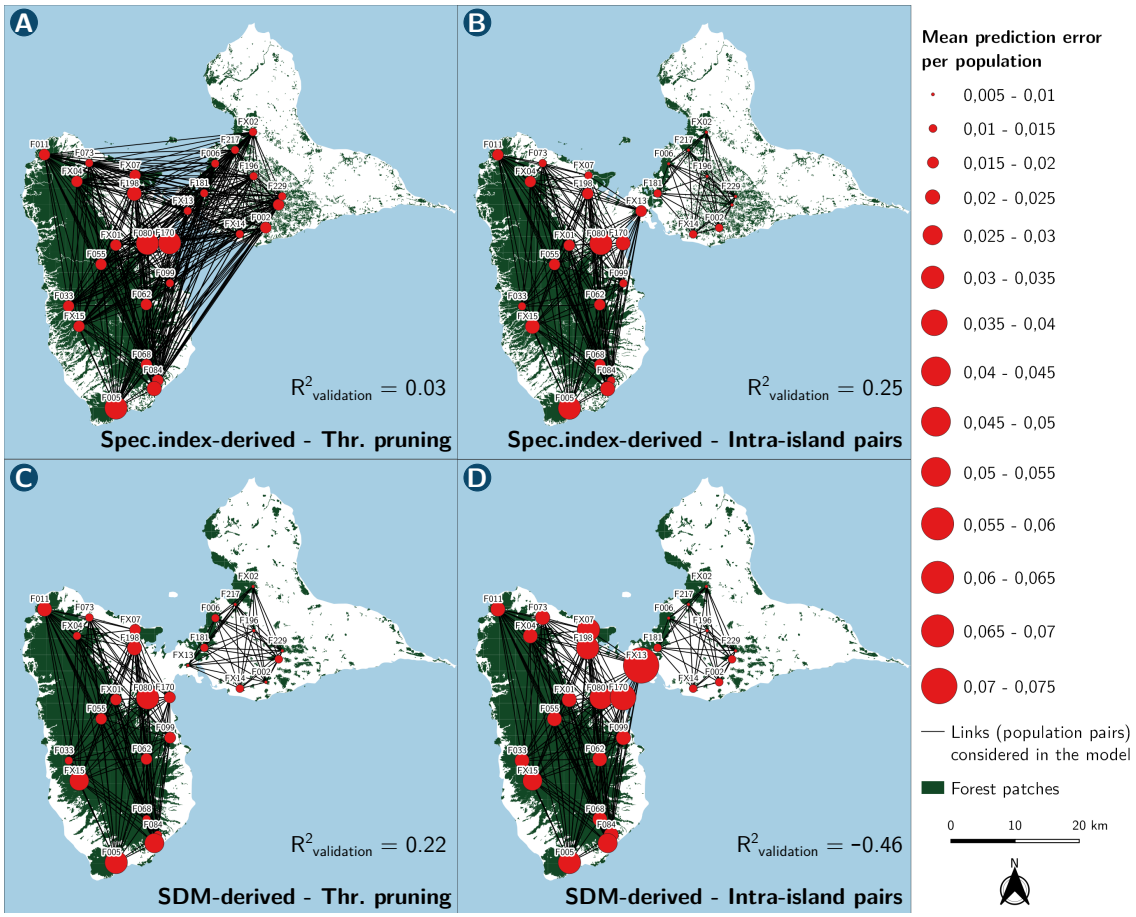


FIGURE 35 – Genetic distance prediction errors of the MLPE models. The mean of the prediction errors of the genetic distances involving every population and all the 26 others has been computed for each population. The prediction error considered is the square root of the mean of the squared differences between the observed genetic distance and the genetic distance predicted by the MLPE models in which one of the two populations of a pair has been excluded from the training dataset (two values for each pair). The corresponding validation R^2 of the MLPE model is reported on the figure. Results are displayed for models including cost-distances computed from cost scenarios deriving either from the Jacobs' specialization indices (A, B) or from the Species Distribution Model (SDM)(C, D), and considering populations pairs either selected based on a threshold pruning (A, C) or corresponding to populations located on the same subpart of the island (BT-BT or GT-GT).

The set of links to consider for maximizing the model goodness-of-fit was not the same according to the cost-distances used (Figure 35). When using the SDM-derived scenario, genetic distances were best explained by a subset of cost-distances lower than a given threshold. Using this threshold separated the populations in two subsets : (i) the populations from GT and one BT population located close to the isthmus and to GT forest patches (FX13) and (ii) all the other BT populations (Figure 35C). This means that large cost-distance values computed from suitability scores across the large area without forest patches at the west of the isthmus did not allow for modeling genetic distances correctly. In contrast, when using the specialization index-derived scenario, the best model explained the genetic distances between pairs of populations located on the same side of the Guadeloupe island (BT-BT or GT-GT pairs, Figure 35B). This included population pairs separated by large unfavorable areas (e.g. between FX13 and the other BT populations).

The validation R^2 confirmed that cost-distances computed from landscape graphs were often reliable predictors of genetic distances (Figure 34). The different performances depending on the link set considered partly stemmed from the fact that the genetic distances between FX13 and the other BT populations were not predicted correctly by the SDM-derived cost-distances, leading to a negative

validation R^2. In contrast, the genetic distances between this population and the GT populations were well predicted using these same cost-distances. Besides, the prediction errors of the best models mainly concerned population pairs involving BT populations, and especially those located on the margin of the largest forest patches. When using a cost-distance pruning threshold, the expert-based 2 scenario led to the third highest validation R^2 (0.18) across all cost scenario and link set combinations. This means that, comparatively, an expert-based scenario could still provide reliable genetic distance predictions.

3.4 Metric validation

The two genetic indices (Ar, MIW) were each significantly correlated to all the three habitat connectivity metrics (Capacity, F, BC). For each cost scenario, we only report (i) the highest correlation coefficients obtained when using the different distances d at which we considered that $p(d) = 0.05$ for fixing α and (ii) the corresponding correlation obtained when considering that $d = 5000$ m (Table 12). These coefficients took large values (up to 0.72) and tended to be slightly higher between the connectivity metrics and the MIW rather than the allelic richness. Besides, the optimal distance d used for computing the metric was always equal or lower than 3500 m.

The correlations obtained with metrics computed using the SDM-derived landscape graph were overall different and lower than those obtained with the three other graphs (Table 12). Indeed, in all but one case (MIW $\propto$ Capacity), the metrics computed from the expert-based 1 landscape graph led to stronger correlations than metrics computed using the other graphs.

When considering that $d = 5000$ m, the correlations were still significant in most cases, but much lower than when using optimal d values. This was due to differences in the ranking of the habitat patches in terms of connectivity, which less closely reflected in this case the genetic responses observed in the populations occupying the patches. Interestingly, the largest rank differences were observed for populations located either in the most forested area of GT or in forest patches disconnected from the largest forest patches of BT. In the former case, using a high d value under-estimated the actual connectivity level of GT patches surrounded by forested areas at a small spatial scale. In contrast, this over-estimated the connectivity of the BT patches which are disconnected from large forest patches when considering connections at a restricted scale (Figure 36).

Genetic index	Connect. metric	Habitat delineation/cost scenario	$r_{Sp.}$ - Optim.		$d_{Optim.}$	$r_{Sp.}$ - 5000 m	
MIW	Capacity	Exp. 1/ Exp.2 / Spec.ind.	0.50	*			
MIW	Capacity	SDM-derived	**0.57**	**			
All.rich.	Capacity	Exp. 1/ Exp.2 / Spec.ind.	**0.55**	*			
All.rich.	Capacity	SDM-derived	0.45				
MIW	F	Expert-based 1	**0.72**	***	1000	0.55	**
MIW	F	Expert-based 2	0.68	***	1000	0.55	**
MIW	F	Spec.ind.-derived	0.71	***	500	0.53	**
MIW	F	SDM-derived	0.58	**	3500	0.54	**
All.rich.	F	Expert-based 1	**0.53**	**	500	0.22	
All.rich.	F	Expert-based 2	0.42	*	1000	0.20	
All.rich.	F	Spec.ind.-derived	0.48	*	500	0.26	
All.rich.	F	SDM-derived	0.37		3500	0.31	
MIW	BC	Expert-based 1	**0.68**	***	2500	0.58	**
MIW	BC	Expert-based 2	0.63	***	1000	0.55	**
MIW	BC	Spec.ind.-derived	**0.68**	***	500	0.49	*
MIW	BC	SDM-derived	0.63	***	3500	0.62	***
All.rich.	BC	Expert-based 1	0.50	**	2500	0.44	*
All.rich.	BC	Expert-based 2	**0.51**	**	500	0.39	*
All.rich.	BC	Spec.ind.-derived	0.50	**	500	0.41	*
All.rich.	BC	SDM-derived	0.42	*	1500	0.35	

TABLE 12 – Spearman correlation coefficients between genetic indices and connectivity metrics. Values are reported according to the genetic index (MIW : Mean Inverse Weight, All.rich. : allelic richness), the connectivity metric (Capacity, F, BC), the scenario considered for delineating habitat patches and fixing cost values ('Expert-based 1 and 2' : expert-based scenarios, 'Spec.index-derived' : habitat delineation and cost values deriving from Jacobs' specialization indices, 'SDM-derived' : habitat delineation and cost values deriving from suitability scores of the Species Distribution Model) and the distance at which dispersal probability is set to 0.05 for computing α values. For comparison purposes, for each combination of a genetic index, a connectivity metric and a construction scenario, we report the correlation coefficients obtained when using the optimal distance and corresponding α value for computing the connectivity metric, and also the value obtained when considering that $p(d = 5000m) = 0.05$ for fixing α. The highest correlation coefficient for each combination of a genetic index and a connectivity metric is displayed in bold. Stars refer to correlation significance : * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

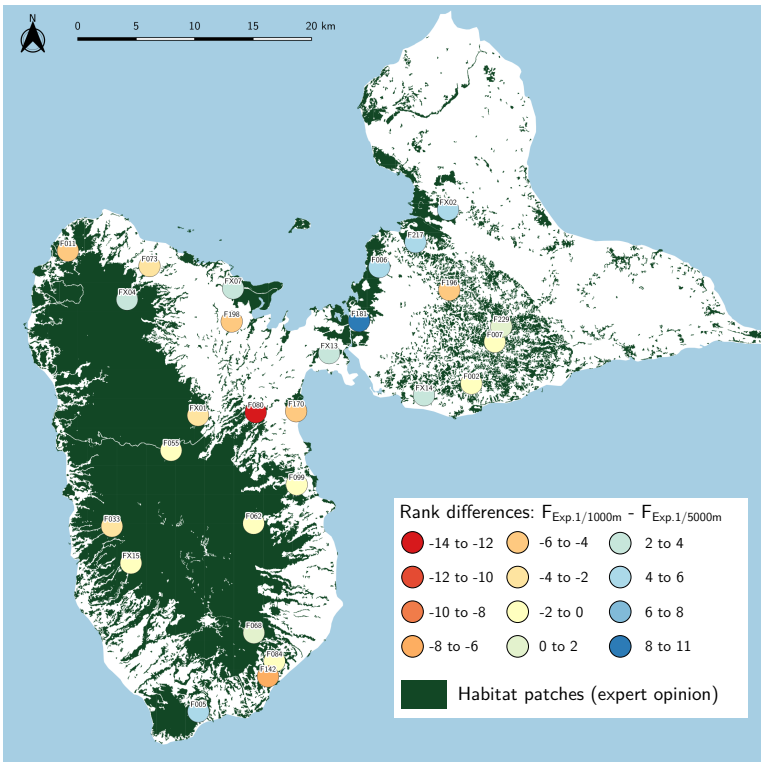


FIGURE 36 – Differences between the ranks of the habitat patches occupied by the sampled populations in terms of connectivity. For each occupied patch, we computed the difference between the rank of the F values computed using the expert-based cost scenario 1 and considering that dispersal probability is equal to 0.05 at a distance d equal to either 1000 m or 5000 m. The former d value is optimal according to Table 12. Negative rank differences indicate that the connectivity of the patch is overestimated when using $d = 5000$ m as compared with $d = 1000$ m.

4 Discussion

Most connectivity models based on landscape graphs are lacking of *a posteriori* empirical validation (Foltête *et al.*, 2020 ; Godet et Clauzel, 2021 ; Kadoya, 2009). Using genetic data, we here demonstrated that landscape graphs constructed using either expert-opinion or presence/absence data were reliable models of the influence of habitat patch connectivity on population genetic structure. We thus confirmed their interest for conservation modeling. Yet, the relationship between graph ecological relevance, data-requirements and construction and analysis methods was not straightforward. Indeed, the graph based upon the most complex construction method (SDM) had a similar or even lower ecological relevance than the others. In the following sections, we discuss these points while mentioning their implications for biodiversity conservation.

4.1 Landscape graphs are empirically validated by genetic data

Genetic distances are well explained and predicted by the cost-distances associated with graph links

We first evidenced that whatever the landscape graph construction method, the cost-distances associated with graph links explained a substantial share of the variance of the genetic distances between populations sampled in graph nodes (> 25 %). Using linear mixed models (MLPE) and an *ad hoc* cross-validation approach, we additionally showed that these cost-distances had a predictive accuracy. We thus validated in a proper way the relationship between genetic distances and cost-distances computed under different cost scenarios. Indeed, this relationship had most often been evidenced using correlative approaches (Balbi *et al.*, 2018 ; Creech *et al.*, 2014 ; Wang *et al.*, 2008). Our method also illustrates that landscape graphs could be useful for implementing predictive approaches in landscape genetics. Such approaches would strengthen the interest of this field for conservation modeling (Keller *et al.*, 2015 ; Richardson *et al.*, 2016) but have rarely been implemented so far (but see Van Strien *et al.* (2014)).

Graph-based connectivity metrics were significantly correlated with genetic indices

Moilanen (2011) raised concern about the uncertain ecological relevance of the multitude of metrics deriving from landscape graphs. Yet, it has been shown that a restricted set of these metrics would be sufficient to reflect how habitat connectivity influences ecological processes such as recruitment, migration flows and long-distance rescue (Baranyi *et al.*, 2011 ; Rayfield *et al.*, 2011 ; Urban et Keitt, 2001). Accordingly, we computed three complementary connectivity metrics (Capacity, F, BC) and showed that they were each significantly correlated with the local genetic diversity of each population and with their level of genetic differentiation with other populations. In contrast with previous studies evidencing similar landscape genetic relationships (Castillo *et al.*, 2016 ; Creech *et al.*, 2014), our landscape graphs were independent from any information deriving from genetic data, thereby strengthening the significance of the relationship between genetic responses and connectivity metrics. Besides, the relationships between our connectivity metrics and both genetic diversity and differentiation confirms previous results obtained in separate studies of the relationships between similar graph-based connectivity metrics and either genetic diversity (Bertin *et al.*, 2017 ; Capurucho *et al.*, 2013) or population-specific genetic differentiation indices (Peterman *et al.*, 2015). Therefore, the metrics associated with the nodes and links of landscape graphs are reliably reflecting the drivers of both genetic drift and gene flow, as expected from theory (DiLeo et Wagner, 2016). This means that ranking habitat patches (nodes) according to different connectivity metrics provides insight into the

respective genetic responses of the populations occupying them and would result helpful for designing conservation measures.

4.2 Relationship between graph ecological relevance, data requirements and construction and analysis methods

Apart from validating the landscape graph approach, we aimed at comparing the ecological relevance of graphs built using either expert-based information or empirical data. We hypothesized that the graph deriving from a SDM would provide the most realistic outputs because this approach is supposed to prevent from relying upon subjective expert opinion (Duflot *et al.*, 2018). However, this hypothesis was only partly supported.

Using empirical data and complex construction methods does not guarantee graph ecological relevance

We used 991 presence/absence observations and both climatic and landscape context variables for fitting the SDM while the same presence/absence data and only land cover data were needed for computing specialization indices. Only land cover data and expert opinion were needed for the simplest expert-based approach. The methods on which the construction of the graphs depended thus differed in terms of both data requirement and complexity. Despite these differences, the connectivity metrics deriving from expert-based graphs were often the most correlated to the genetic indices, although those deriving from the other graphs also exhibited significant correlations. Besides, even if the best model explaining genetic distances included cost-distances computed using the SDM-derived cost scenario, the cost-distances computed from cost values deriving from specialization indices provided very similar goodness-of-fit. One of the expert-based scenario also provided relatively accurate predictions. This indicates that, compared with the SDM approach, simpler empirical approaches or expert-based approaches could be reliable and more cost-effective in many instances. Similarly, Poor *et al.* (2012) showed that expert-based models of migration corridors were the most cost-effective given the slight improvement provided by models based on SDM. Therefore, although we confirm that integrating empirical data in landscape graph modeling leads to reliable models (Kadoya, 2009), we question the cost-effectiveness of such an approach when it is based on a SDM.

Construction methods appear to be equally important as data and information source

The quality of the SDM cannot explain why the SDM-derived graph did not always appear as the most realistic. Indeed, the SDM had a good accuracy (AUC > 0.9), reflecting the quality of these models for specialist species such as the Plumbeous warbler (Hernandez *et al.*, 2006 ; McPherson et Jetz, 2007). The availability of both presence and absence data, and our consideration of a potential sampling bias, may be other reasons for the quality of this model (Fletcher et Fortin, 2018). Interestingly, the use of presence/absence data does not seem to explain either the main differences between the graphs. Indeed, we used these same data for computing specialization indices and the metrics deriving from the graph built from these indices were more similar to the metrics deriving from the expert-based graphs as compared with those deriving from the SDM-derived graph. Therefore, the choices made for delineating habitat patches and setting cost values seem to be equally important, if not more, than the data used for modeling landscape graphs. This result is somehow in contradiction with the common belief that the use of empirical data prevents from making arbitrary choices.

Using suitability scores for delineating habitat patches produced less patches of larger sizes, partly because of the smooth contrasts between suitability scores in areas dominated by suitable pixels. This has already been observed (Bourdouxhe *et al.*, 2020 ; Godet et Clauzel, 2021), although an opposite result can also be obtained depending on the variables included in the SDM (Stevenson-Holt *et al.*, 2014) or when landscape features such as transport infrastructures are added to the cost surface deriving from the SDM (Ziółkowska *et al.*, 2012). The different correlations between genetic indices and connectivity metrics deriving from the graphs could be due to these node differences.

In addition, the methods used for setting the cost values substantially affected our results. First, the expert-based cost scenarios 1 and 2 differed by the range and the contrast of the cost values. This led to different correlation levels and model fits when using either one or the other scenario for computing metrics and cost-distances. The differences were larger in the case of cost-distances, the scenario with the widest range (1-1000) being the best for explaining genetic distances. Besides, the negative exponential function introduced by Keeley *et al.* (2016) for converting suitability scores into cost values led to high cost values in widespread agricultural and artificial areas. Accordingly, we obtained large cost-distance values between populations separated by large tracts of such land cover types, which did not allow for predicting well the genetic distances between these populations. Yet, the cost scenario based on specialization indices led to better predictions of these same genetic distances, illustrating that two scenarios can be equally relevant for predicting genetic distances in the same species, but in different areas. For this same reason, Reed *et al.* (2017) concluded that cost values deriving from either telemetry data or expert opinion were complementary for connectivity analysis as they were each equally validated when confronted to genetic data in different areas. Besides, methods for converting suitability scores into cost values based on thresholds could have resulted in stronger correlations between cost-distances and genetic distances, as in Wang *et al.* (2008) study.

The spatial scale of the analyses strongly influences the results

Finally, the scale at which the connections between populations or patches were considered largely affected the results. Computing connectivity metrics by assigning connections between patches a large weight at a larger scale (5000 m) than the optimal scale (< 3500 m) decreased the correlation between these metrics and genetic responses. Interestingly, it tended to either under- or over-estimate patch connectivity depending on the patch location within the island. Similarly, the set of links considered for predicting genetic distances from cost-distances largely influenced the quality of the predictions, and in differing ways depending on the cost scenario. Although these complex results partly stem from the specific topography and spatial distribution of the Guadeloupe island forests, they recall the importance of the "scale of effect" in landscape ecology (Jackson et Fahrig, 2012 ; Stuber et Gruber, 2020) and the importance of the topology of the population networks for the reliable assessment of landscape genetic relationships (Savary *et al.*, 2021a ; Van Strien, 2017). The sensitivity of landscape graph modeling outputs to the spatial scale of the analysis should therefore be seen as equally important as their construction parameters. It also confirms that graph-based approaches are perfectly suited for the consideration of scale effects, as outlined in seminal works (Keitt *et al.*, 1997).

4.3 Implications for biodiversity conservation

Our results confirm the relevance of the wide range of graph-based methods specifically developed for conservation purposes (Foltête *et al.*, 2014 ; Tarabon *et al.*, 2019 ; Zetterberg *et al.*, 2010). Besides,

we further illustrate how landscape graphs can be used for addressing conservation issues whether genetic data are available or not.

Genetic data allowed us to assess the ecological validity of each graph element by implementing a new cross-validation approach and by studying the rank differences of the patches according to metrics computed differently. This evidenced that the Plumbeous warbler is positively affected by the availability of forest areas at a restricted scale and negatively affected by the breaking apart of forest patches even when large forest areas are located nearby (as in BT). We also distinguished populations whose genetic differentiation mostly originated from the habitat pattern and matrix resistance, from populations where other factors potentially explained it as the prediction errors associated with these populations were consistently high. In the latter case, complementary field surveys should be carried out. This illustrates that making the model limitations spatially-explicit using genetic data could expand the usefulness of the landscape graph approach for conservation modelling.

In contrast, when genetic data are not available, the optimal construction and analysis parameters cannot be optimized but we showed that most parameters led to cost-distances and metrics significantly explaining the genetic responses in our case study. Thus, although landscape graph modeling should preferably be used for studying specialist species occupying discrete habitat patches (Urban et Keitt, 2001), it would continue to be a useful tool for addressing conservation issues in a wide range of species for which genetic data acquisition is difficult or expensive. The availability and the quality of either (i) information from literature or expert opinion or (ii) empirical data should determine the graph construction method, as we here showed that empirical approaches were not necessarily the most cost-effective ones.

4.4 Limits and perspectives

The connectivity models we validated using genetic data were reflecting the influence of habitat connectivity on multi-generation drift and gene flow processes. However, other data such as GPS tracks or interpatch movement data would better reflect daily or single-generation dispersal movements and could be used for validating the spatial location of the graph links. Similar approaches have been implemented for validating least cost path models (Driezen *et al.*, 2007 ; Laliberté et St-Laurent, 2020) and graph-based metrics (Poli *et al.*, 2020).

Our approach could also be replicated for species with different ecological traits or habitat specialisation levels for assessing the usefulness of the landscape graph approach in a wider range of situations. Indeed, habitat connectivity has been shown to affect species in different ways, including among tropical forest birds (Radford *et al.*, 2021). Finally, although we here used genetic data for the *a posteriori* validation of landscape graphs built using empirical data or expert opinion, another valid option would consist in using genetic data *a priori*, e.g. for optimizing cost values prior to graph construction (Beier *et al.*, 2008 ; Zeller *et al.*, 2018).

A - Supplementary tables

CLC initial category	New category
Deciduous forest	Forest
Mangrove	Forest
Non-irrigated arable land	Agricultural areas
Sugar cane	Agricultural areas
Orchards	Agricultural areas
Banana plantations	Agricultural areas
Olive groves	Agricultural areas
Complex cultivation patterns	Agricultural areas
Land principally occupied by agriculture, with significant areas of natural vegetation	Agricultural areas
Grasslands	Open areas
Other natural grasslands and pastures	Open areas
Sparse vegetation	Open areas
Shrubland	Semi-open areas
Sclerophillous vegetation	Semi-open areas
Transitional forest and shrub vegetation stage	Semi-open areas
Urban green areas	Semi-natural areas
Bare rocks	Semi-natural areas
Rivers and streams	Rivers and lakes
Lakes and ponds	Rivers and lakes
Coastal lagoon	Rivers and lakes
Wetlands	Wetlands
Ocean	Ocean
Continuous urban fabric	Artificial areas
Discontinuous urban fabric	Artificial areas
Industrial or commercial units	Artificial areas
Road and rail networks and associated land	Artificial areas
Port areas	Artificial areas
Airports	Artificial areas
Mineral extraction sites	Artificial areas
Dump sites	Artificial areas
Construction sites	Artificial areas
Sport and leisure facilities	Artificial areas

Table 13 – Classification of the initial land cover categories from the CORINE Land Cover (CLC) database into nine new land cover categories

Vegetation formation	**Habitat (Y/N)**
Altitude thickets	Y
Low-growth altitude thickets	Y
Low-growth pioneer formations	Y
Diversified sylvicultural areas	Y
High-altitude formations	Y
Limestone lowland forests	Y
Coastal forests	Y
Rain forests	Y
Valley floor forests	Y
Fallow lands with low-growth woody vegetation	**N**
Coastal thickets	**N**
Mangroves	Y
Swamp forests	Y
Mahogany forests	Y
Semi-deciduous forests	Y
Seasonal evergreen forests	Y
Forests of agricultural areas	Y

TABLE 14 – Vegetation formations of the 2010 map created by the Conseil Departemental of Guadeloupe, the IGN and the ONF using manually interpreted aerial photographs. The "Habitat" column indicates whether the formation was considered as being part of the Plumbeous warbler habitat according to the expert.

Pop.ID	**Island part**	**Nb.ind.**	**Ar**	**MIW**
F005	BT	8	4.14	15.86
F011	BT	6	4.33	25.22
F033	BT	8	5.26	39.38
F055	BT	28	4.89	83.58
F062	BT	27	4.89	388.55
F068	BT	30	4.88	250.20
F073	BT	52	4.67	36.01
F080	BT	52	4.52	16.16
F084	BT	15	4.83	67.19
F099	BT	29	4.57	134.50
F142	BT	23	4.28	20.04
F170	BT	26	3.99	15.49
F198	BT	12	4.23	15.88
FX01	BT	27	4.98	146.66
FX04	BT	19	4.89	88.90
FX07	BT	36	4.51	16.79
FX13	BT	21	4.88	25.55
FX15	BT	12	4.73	20.87
F002	GT	48	4.58	26.26
F006	GT	33	4.95	30.04
F007	GT	28	4.94	34.39
F181	GT	36	4.97	40.64
F196	GT	22	4.75	28.04
F217	GT	47	4.79	34.76
F229	GT	30	4.78	33.76
FX02	GT	19	4.90	31.06
FX14	GT	18	4.76	27.15

TABLE 15 – Genetic indices of each population

B - Jacobs' specialization index

The Jacobs' specialization index was computed using the following formula (Jacobs, 1974) :

$$I_{Jacobs_{ik}} = \frac{r_{ik} - p_i}{r_{ik} + p_i - 2r_{ik}p_i}$$

where r_{ik} is the ratio between the number of presence points n_{ik} of species k in land cover i and the total number of presence points n_k for species k ($r_{ik} = \frac{n_{ik}}{n_k}$) and p_i is the ratio between the total number of sampling points m_i in land cover i and the total number m of sampling points.

C - Validation R-squared calculation

Training R^2

For assessing the goodness-of-fit of a model explaining a response variable y with one or p predictor variables $x_1, x_2, \ldots, x_p$, the coefficient of determination (R^2) is computed by comparing the residual and explained sum of squares ($SS_{res.}$, $SS_{exp.}$) to the total sum of squares ($SS_{tot.}$) of the explained variable y, such that :

$$R^2 = 1 - \frac{SS_{res.}}{SS_{tot.}} = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2}$$

with n the number of observations, $\hat{y}_i$ the predicted value of y_i according to the model (e.g. $\hat{y}_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$) and $\bar{y}$ the grand mean of y. By construction, we usually have :

$$SS_{tot.} = SS_{exp.} + SS_{res.}$$

such that the R^2 can be interpreted as the share of the variance of y that is explained by the model. It thus ranges from 0 to 1. We refer to this value as the training R^2 because it is computed by considering a given set of n observations for calibrating the model and assessing its goodness-of-fit.

Validation R^2

Alternatively, the dataset can be splitted into several parts : a training dataset and a validation dataset. In such a case, the training dataset is used for calibrating a model of the following form :

$$\hat{y_{i_t}} = \beta_1 x_{1i_t} + \beta_2 x_{2i_t} + \ldots + \beta_p x_{pi_t}$$

with $\hat{y_{i_t}}$ and x_{1i_t} referring respectively to the predicted value of the response variable and to the value of the predictor variable x_1 for the observation i_t included in the training dataset.

This same model can be used to predict the values of the response variable for the observations of the validation dataset :

$$\hat{y_{i_v}} = \beta_1 x_{1i_v} + \beta_2 x_{2i_v} + \ldots + \beta_p x_{pi_v}$$

One way of assessing the goodness-of-fit of the model while taking into account its potential over-fitting is then to compute a validation R^2 from this independent predicted values. That is what we did in this study. Because we used distance data and a particularly Leave-One-Out Cross Validation method (cf. section 2.4.1), each observation corresponded to a population pair and the corresponding

genetic distance was predicted twice (once every time one of the two populations was excluded from the training dataset). We therefore computed the mean predicted values for each population pair to obtain $\hat{y_{i_v}}$ values.

We then computed the validation R^2 using the following formula :

$$R^2_{validation} = 1 - \frac{SS_{res.v}}{SS_{tot.v}} = 1 - \frac{\sum_i^n (y_{i_v} - \hat{y_{i_v}})^2}{\sum_i^n (y_{i_v} - \bar{y_v})^2}$$

Because the model is calibrated with the independent training dataset, there is no reason that $SS_{tot.v} = SS_{exp.v} + SS_{res.v}$. In some cases, $SS_{res.v}$ can even be higher than $SS_{tot.v}$, such that the validation R^2 can be negative, in contrast with the training R^2, which is always between 0 and 1. However, negative validation R^2 values mean that a model has a very low predictive accuracy.

D - Supplementary figures

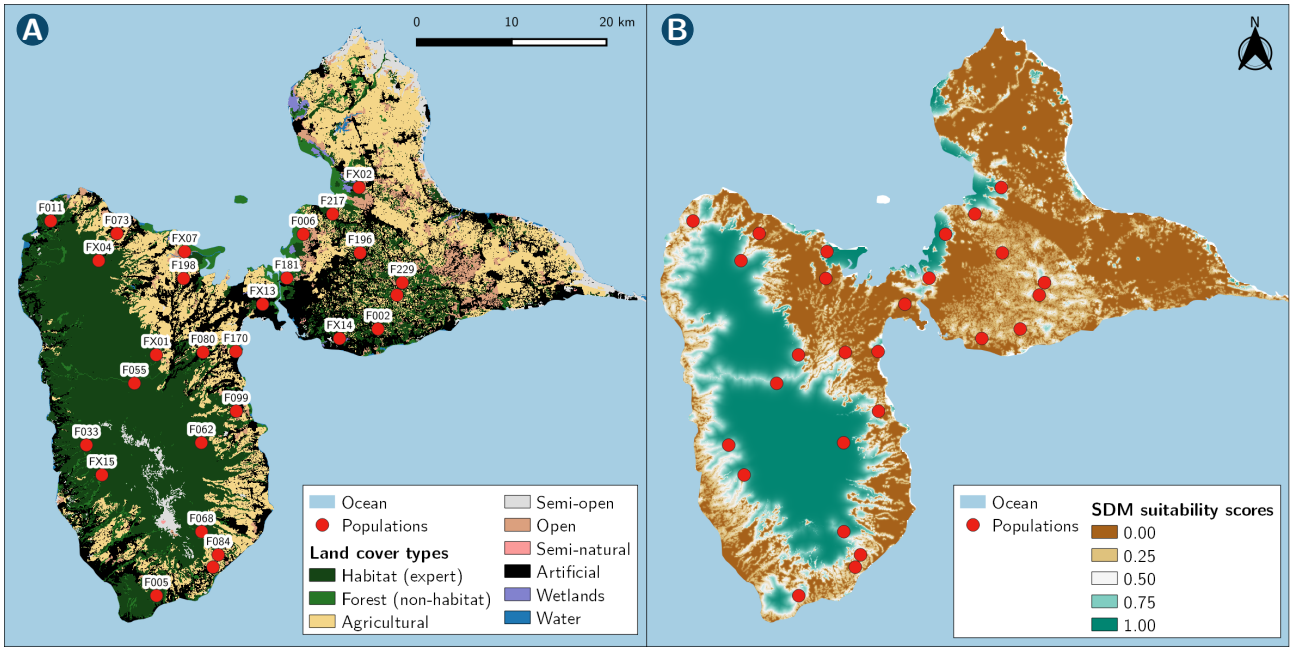


FIGURE 37 – (A) Land cover map used for building the expert-based and specialization index-derived landscape graphs and (B) SDM raw suitability scores used for building the SDM-derived landscape graph.

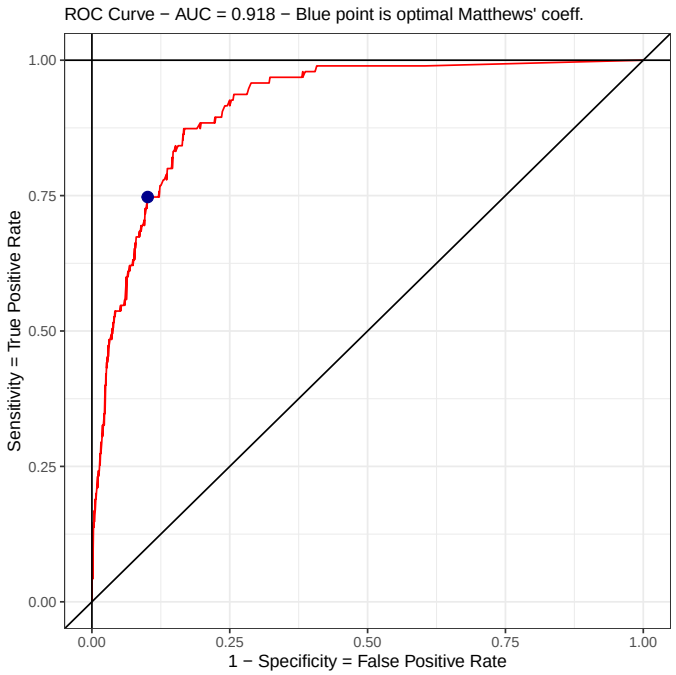


FIGURE 38 – ROC curve obtained with the SDM, leading to an AUC of 0.918.

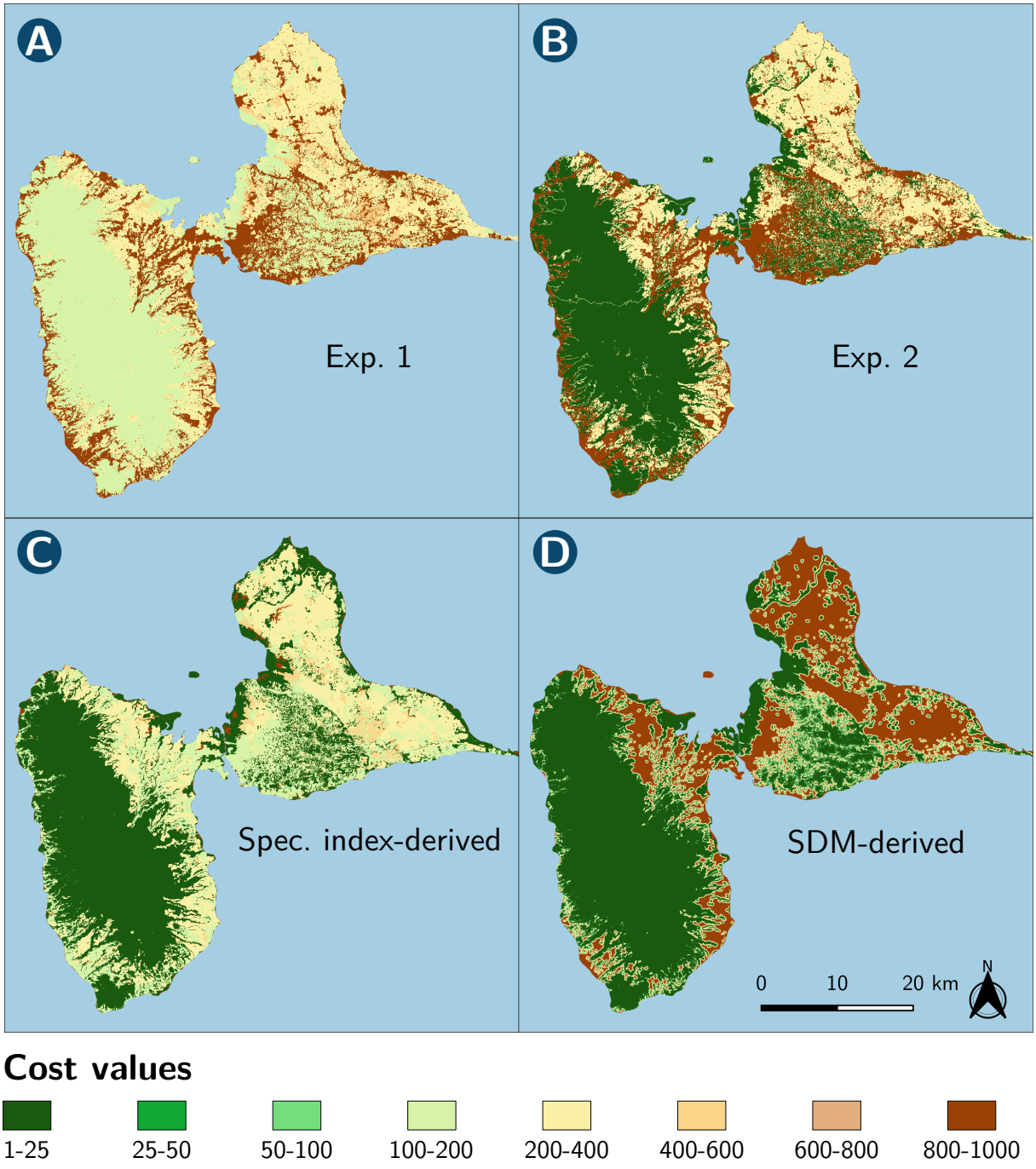


FIGURE 39 – Cost surfaces obtained when applying the four cost scenarios : (A) expert-based 1, (B) expert-based 2, (C) specialization index-derived, (D) SDM-derived. The same palette of colors is used for the four maps.

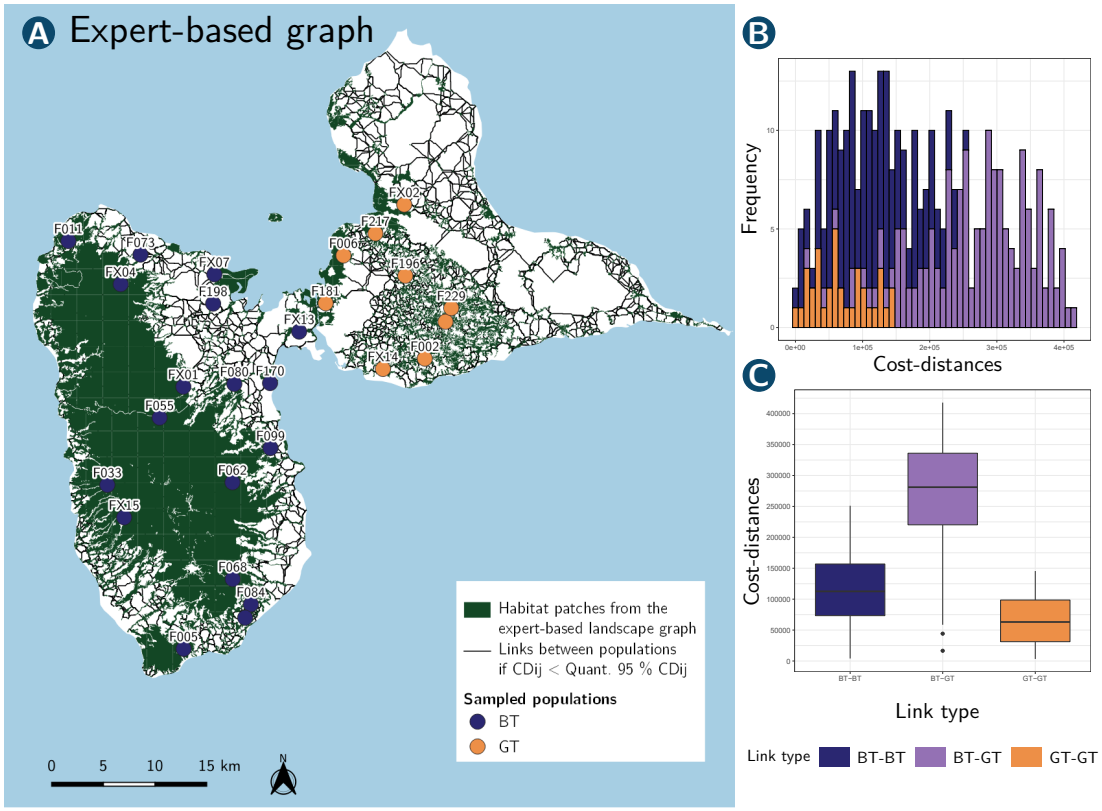


FIGURE 40 – Expert-based landscape graph (cost scenario 1). (A) Map of the patches and links of the landscape graph, and of the sampled bird populations. Only the links whose cost-distance values are not in the 95-100 % quantile of values are displayed. Distribution of the cost-distance values associated with the landscape graph links displayed as an histogram (B) or as boxplots (C). On these figures, the colors refer to the type of links considered (BT-BT, BT-GT, GT-GT).

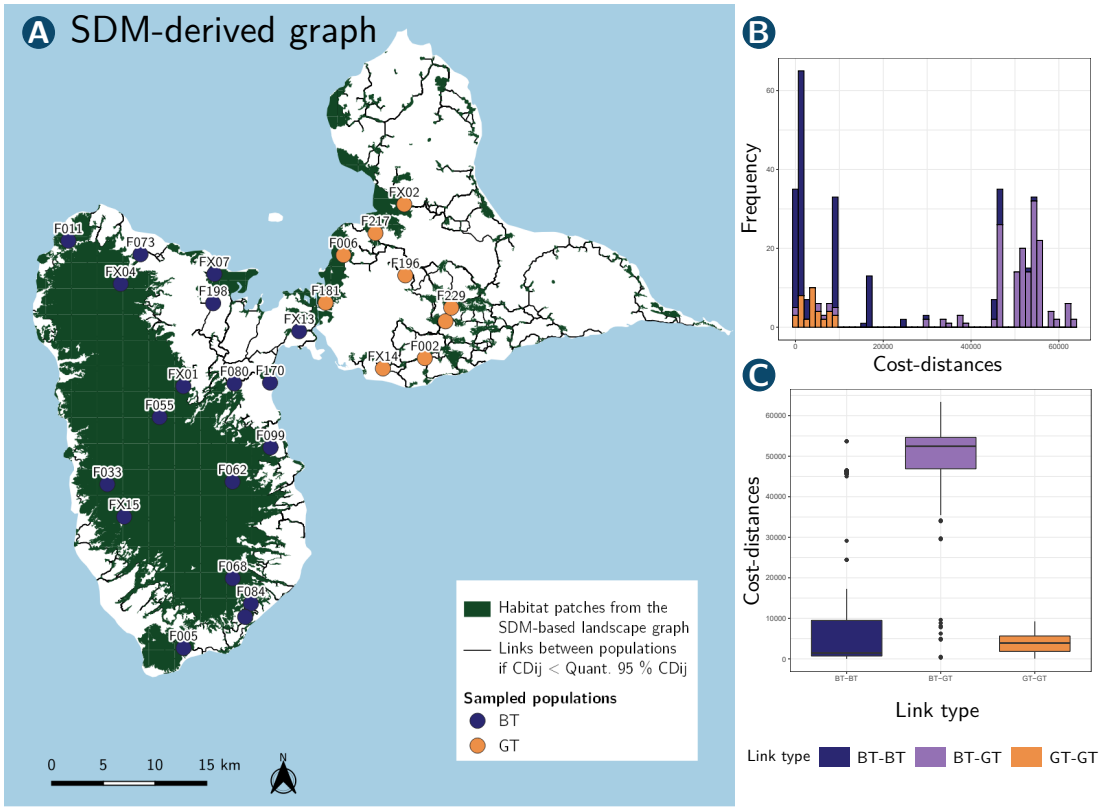


FIGURE 41 – SDM-derived landscape graph. (A) Map of the patches and links of the landscape graph, and of the sampled bird populations. Only the links whose cost-distance values are not in the 95-100 % quantile of values are displayed. Distribution of the cost-distance values associated with the landscape graph links displayed as an histogram (B) or as boxplots (C). On these figures, the colors refer to the type of links considered (BT-BT, BT-GT, GT-GT).

Annexe A7

Cost distances and least cost paths respond differently to cost scenario variations - A sensitivity analysis of ecological connectivity modeling

Abstract

Biodiversity conservation measures designed to ensure ecological connectivity depend on the reliable modeling of species movements. Least cost path modeling makes it possible to identify the most likely dispersal paths within a landscape and provide two items of ecological relevance : (i) the spatial location of these least cost paths (LCPs) and (ii) the accumulated cost along them ('cost distance', CD). This spatial analysis requires that cost values be assigned to every type of land cover. The sensitivity of both LCPs and CDs to the cost scenarios has not been comprehensively assessed across realistic landscapes and diverging cost scenarios. We therefore assessed it in diverse landscapes sampled over metropolitan France and with widely diverging cost scenarios. The spatial overlap of the LCPs was more sensitive to the cost scenario than the CD values were. Besides, highly correlated CD matrices could derive from very different cost scenarios. Although the range of the cost values and the properties of each cost scenario significantly influenced the outputs of LCP modeling, landscape composition and configuration variables also explained their variations. Accordingly we provide guidelines for the use of LCP modeling in ecological studies and conservation planning.

Keywords : least cost path modeling, sensitivity analysis, ecological connectivity, spatial ecology, landscape ecology

Cet article a été re-soumis après des modifications mineures à l'*International Journal of Geographical Information Science* en septembre 2021 :

Savary, P., Foltête, J. C. & Garnier, S. Cost distances and least cost paths respond differently to cost scenario variations - A sensitivity analysis of ecological connectivity modeling. Submitted to *International Journal of Geographical Information Science*

1 Introduction

In the last decades, a variety of spatial models have been put forward for mapping and conserving ecological connectivity, largely benefiting from the development of GIS tools and landscape ecology theories (Rayfield *et al.*, 2011 ; Zeller *et al.*, 2012). Since the landscape matrix was shown to exert a heterogeneous effect on species' movement depending on its composition and configuration (Ricketts, 2001), landscapes have been represented as cost surfaces, i.e. raster grids on which every pixel value is supposed to reflect resistance to movement. Modeling ecological connectivity on these surfaces then consists in computing the paths followed by individuals for bridging pairs of habitat patches while minimizing movement costs (Adriaensen *et al.*, 2003). These 'least cost paths' (LCP hereafter) provide two kinds of ecological information : (i) their spatial location, i.e. the LCP itself, and (ii) the accumulated cost summed along the LCP, also known as the 'cost distance' (CD hereafter). Although similar connectivity modeling approaches have been developed (Marrotte et Bowman, 2017 ; McRae, 2006 ; Panzacchi *et al.*, 2016), the relevance of LCP modeling continues to be reflected by their frequent application in spatial analyses with decision-oriented aims such as wildlife linkage planning (Beier *et al.*, 2008 ; Carroll *et al.*, 2012 ; Sawyer *et al.*, 2011), invasive species control (Etherington et Perry, 2016) or in statistical analyses directed at hypothesis testing in ecological studies (Balbi *et al.*, 2019 ; Mony *et al.*, 2018).

In this approach, cost values are key modeling inputs that most often depend upon the types of land cover in which the pixels lie (Zeller *et al.*, 2012). Their choice is frequently driven by knowledge from literature surveys, field experience, or expert opinion regarding the movement behavior of the study species (Clevenger *et al.*, 2002 ; Pullinger et Johnson, 2010). Therefore an element of arbitrariness remains at this stage, which could influence the outputs. On the one hand, different cost scenarios can produce substantially different outputs if they differ widely in terms of the absolute values assigned to land cover types and the contrast between them (Gonzales et Gergel, 2007 ; Murekatete et Shirabe, 2018 ; Rayfield *et al.*, 2010). On the other hand, the order of the cost values assigned to land cover types is known to influence CDs and LCPs (Beier *et al.*, 2009). This order directly reflects the preference of individuals for land cover types against others. Yet we may wonder whether two highly correlated CD matrices or two sets of spatially equivalent LCPs could derive from cost scenarios ranking land cover types in different orders. Although Murekatete et Shirabe (2018) have assessed the influence of the distribution of cost values on LCPs, their study was based on simulated landscapes and continuous cost ranges. Therefore an assessment of the sensitivity of LCP modeling in real landscapes to changes in the distribution of discrete cost values commonly used by practitioners in conservation modeling should be carried out.

In addition to cost values, the landscape structure in itself has been shown to influence the sensitivity of CD values to cost scenarios in analyses relying on simplified simulated landscapes (Bowman *et al.*, 2020 ; Cushman *et al.*, 2013b ; Marrotte et Bowman, 2017 ; Murekatete et Shirabe, 2018 ; Rayfield *et al.*, 2010 ; Simpkins *et al.*, 2017). For example, CD values were less affected by variations of scenarios when habitat areas were highly connected, according to the study of Cushman *et al.* (2013b). However, a similar assessment over a wide range of existing landscapes is still lacking although it could identify the range of realistic landscape contexts in which modeling results are the most dependent on cost scenarios.

Similarly few studies have assessed the sensitivity of the LCP spatial locations to variations of cost scenarios. Beier *et al.* (2009) modeled a corridor providing a linkage between two Californian protected areas using several expert-based cost scenarios. Although they concluded that this corridor overlapped most of the alternative corridors modeled following alternative scenarios, these scenarios were all somehow similar to a presumably realistic scenario and the cost values ranged from 1 to 10. The relative resistance of land cover types is sometimes known by the experts but the plausibility of the range of cost values cannot be known beforehand. Besides, when this range derives from empirical observations, it is usually much wider (Khimoun *et al.*, 2017 ; Ruiz-González *et al.*, 2014). In contrast, Pullinger et Johnson (2010) compared the paths followed by woodland caribou according to GPS tracks to LCPs modeled under several cost scenarios. They reached more pessimistic conclusions suggesting that the spatial location of LCPs is highly sensitive to cost scenarios. Given that a limited number of scenarios were used in these rare studies with inconsistent findings (see Murekatete et Shirabe (2018) for another example), the spatial overlap of LCPs remains to be investigated under more variable scenarios.

In order to bring landscape resistance assumptions closer to ecological reality, a first step of connectivity analysis often consists in inferring cost values from biological data and environmental variables (Kadoya, 2009 ; Pressey, 2004). This can be done by converting presence or movement probabilities deriving from species distribution models (SDM) or step selection functions based on telemetry data into cost values (de la Torre *et al.*, 2019 ; Duflot *et al.*, 2018 ; Zeller *et al.*, 2018). Alternatively, statistical approaches have been developed for inferring cost values from the relationship between pairwise CDs and pairwise genetic distances between populations occupying habitat patches (Peterman *et al.*, 2019 ; Peterman et Pope, 2020 ; Zeller *et al.*, 2016). In this case, the cost value scenario leading to the strongest statistical relationship between CD and pairwise genetic distances is supposed to reflect the CD perceived by individuals during their dispersal movements leading to gene flow. The set of inferred cost values associated with land cover types is then the input of LCP modeling. Their absolute values, rank, and contrasts inherently contain information of ecological relevance, making it possible to determine that one land cover type is more resistant than another, or to ascertain how many times more resistant it is, among other interpretations (see Khimoun *et al.* (2017) or Ruiz-González *et al.* (2014) for an illustration).

However, the latter statistical approaches may be unable to identify the cost scenario closest to the real ecological situation when competing scenarios lead to highly correlated CD values because they may be equally correlated to the pairwise biological response used for the inference (Zeller *et al.*, 2016). This reinforces the need to understand how cost value distributions and landscape structure influence the sensitivity of LCP modeling to cost scenarios. Similarly we do not know precisely whether optimization approaches maximizing the strength of a statistical relationship between pairwise biological responses and CDs under several scenarios can lead to reliable predictions of the spatial location of LCPs. Two highly correlated CD matrices computed from the same locations but using different scenarios could lead to spatially distinct LCPs. Using either one or the other set of CD values in a statistical analysis would not significantly affect the output and both cost scenarios could be assigned the same likelihood from an inference using empirical data. In contrast, the choice of one of them would largely influence the LCP, potentially leading to the implementation of spatially inadequate conservation measures. In sum, although it has already been shown that LCPs and CDs

are sensitive to both the cost scenarios and the landscape contexts in which they are computed, the relative sensitivities of these outputs and their main drivers need to be investigated simultaneously and in a realistic context if the reliability of connectivity analysis is to be improved.

In this study, we assessed the sensitivity of both LCP spatial locations and corresponding CD values deriving from LCP modeling to variations in cost scenarios. For that purpose, we randomly sampled 77 existing landscapes, geolocated points within them, and computed LCPs and the corresponding CDs under 100 widely diverging cost scenarios. We fixed an arbitrary but plausible scenario that we considered as the ecological 'truth'. We then assessed (i) the correlation between alternative CD matrices and the true CD matrix and (ii) the spatial overlap between the alternative LCP and the true ones. Finally we performed statistical analyses to identify the drivers of the sensitivity of CD values and LCPs to the cost scenario. This novel approach allowed us to identify (i) landscape contexts and (ii) cost scenario characteristics influencing the sensitivity of LCP modeling to cost scenarios.

2 Methods

2.1 Landscape sampling

With the aim of providing guidelines for LCP modeling in realistic conditions, we randomly sampled 250 landscapes of 30 km × 30 km with a spatial resolution of 10 m across metropolitan France from the OSO land cover raster map (Inglada *et al.*, 2017). This map is based on remote sensing imagery and initially included 23 land cover types. As this thematic resolution did not reflect the simplified land cover maps commonly used for connectivity modeling and would not have allowed a fine assessment of the influence of the cost assigned to each land cover type, we reclassified it into five land cover types : (1) forests, (2) grasslands and woody perennial crops (grasslands hereafter), (3) annual crops, (4) artificial areas (built-up land, roads and transport infrastructures), and (5) others (water and other land cover types). The spatial and thematic resolutions of this raster layer allowed us to correctly account for the barrier effects of linear landscape features such as transport infrastructures, which can largely influence LCP modeling (Hoover *et al.*, 2020).

We considered the landscape constraints on movement faced by an arbitrary forest species and we therefore only conserved the sampled landscapes with a proportion of forest above 15 %. This proportion of habitat is close to the threshold below which a specialist forest species becomes extinct (Balkenhol *et al.*, 2013 ; Hanski *et al.*, 1996). To prevent the results from being influenced by the absence of one of the land cover types while allowing the land cover type proportions to vary substantially, we ensured that the proportion of grasslands, crops and artificial areas were above 5%, 5%, and 2% respectively. Finally, we removed coastal landscapes including large maritime areas, which led us to retain 77 landscapes for the analyses (cf. section 3.1).

2.2 Cost scenario creation

For comparative purposes, we chose a reference cost scenario ('true scenario' hereafter), in which the cost values associated with forests, grasslands, crops, artificial areas and water and other land cover types were respectively 1, 10, 100, 1000 and 100. They reflected the movement behavior of a forest specialist species. Note that similar cost values have already been used for modeling connectivity for

forest species (Gurrutxaga *et al.*, 2010 ; Schadt *et al.*, 2002) and a similar range (1-1000) has already been inferred from field data (Khimoun *et al.*, 2017 ; Pérez-Espona *et al.*, 2008 ; Ruiz-González *et al.*, 2014 ; Wang *et al.*, 2008).

In order to test for the sensitivity of LCP modeling to cost scenarios, we randomly created 100 widely different alternative cost scenarios. They differed by both the order of the land cover types and the contrast between cost values. We used Shirk *et al.* (2010) approach to set 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 $C_{max} = 1000$ associated with the i-th land cover type. $Rank_i$ is the rank of the land cover type i between 1 and $Rank_{max} = 4$. We used x values equal to 1, 2, 4, 8 or 16. We therefore obtained five series of values : [1, 1, 11, 1000], [1, 4, 11, 1000], [4, 63, 317, 1000], [63, 250, 563, 1000], [250, 500, 750, 1000](Supporting information, figure 49). Using each of them, we randomly assigned cost values to forests, grasslands, crops and artificial areas before randomly selecting 100 alternative cost scenarios among these combinations. The cost value associated with water and other land cover types was set to 100 in each cost scenario in order to limit the number of combinations to test and because this land cover type was absent from several landscapes.

2.3 Least cost path modeling

In every landscape, we randomly selected 50 point locations within forest patches, separated by a distance of more than 500 m. We then computed LCPs between every pair of points in every landscape and under every cost scenario (Figure 42A). We thus obtained in each case a set of LCPs and the corresponding 50×50 pairwise CD matrix. We created buffer zones of 200 m on each side of every spatial line and merged them. We call these polygons of equal width around least cost paths least cost corridors hereafter.

2.4 Spatial and distance-based comparisons of LCPs

We first measured the proportion of the area of every true least cost corridor between a pair of locations that was overlapped by the corresponding alternative least cost corridor. We averaged the 1225 values obtained when considering every pair of locations, thereby obtaining a spatial overlap measure for each combination of a landscape and an alternative cost scenario. Besides, we assessed the statistical relationship between each alternative CD matrix and the true one by computing their Mantel r correlation coefficient (Mantel, 1967). This coefficient is commonly used for assessing the relationship between distance matrices.

2.5 Landscape structure and cost scenario descriptors

We first aimed to explain the sensitivity of LCP modeling to the cost scenarios according to landscape structure. To that end, we computed the proportion of forests, grasslands, crops, and artificial areas in every landscape (landscape composition variables). We also computed the Shannon index as a landscape composition diversity variable. It was divided by $log(n)$ where n is the number of land cover types so that it ranges from 0 to 1. In order to assess the influence of landscape configuration, we computed several FRAGSTATS configuration metrics (McGarigal, 1995). At the landscape level, we

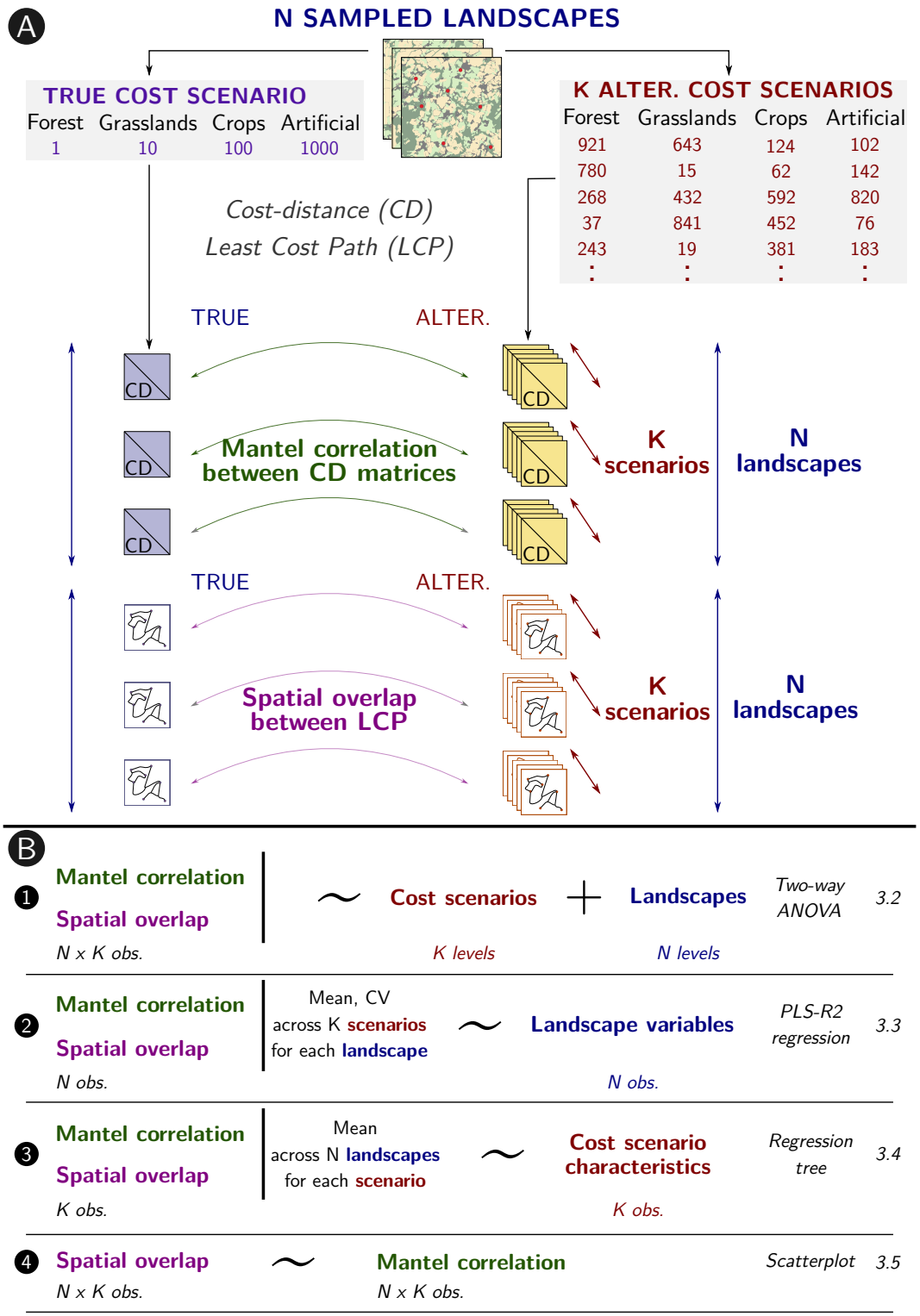


FIGURE 42 – Schematic representation of (A) the spatial and distance-based comparisons of the LCP computed under true and $K = 100$ alternative cost scenarios within $N = 77$ landscapes, and of (B) the statistical analyses performed for assessing the sensitivity of LCP modeling to cost scenarios. We performed separate two-way ANOVA for assessing the respective influences of the landscape and cost scenario on (i) the spatial overlap and (ii) the Mantel correlations. Then we assessed the influence of landscape composition and configuration on the sensitivity of LCP modeling to cost scenarios in every landscape (mean and coefficient of variation across scenarios) by carrying out a PLS-R2 regression. We identified the characteristics of the cost scenarios explaining the values of spatial overlap and Mantel correlations averaged across landscapes for every cost scenario with a regression tree. Finally, we assessed the relationship between the spatial overlap and the Mantel correlations for every combination of the landscape and cost scenario by displaying it on a scatterplot. The numbers in italics at the bottom-right of the figure refer to the results section where the corresponding results are described.

computed the contagion index which reflects the degree of aggregation of the cells of the same land cover type. For the two land cover types that were assigned extreme cost values in the true scenario, i.e. forest and artificial areas, we computed the number of patches, the shape complexity, and the 'clumpy' index of fragmentation. Finally, as a way to assess the global permeability of the landscapes, we computed the sum of the pixel costs on each landscape according to the true scenario.

In parallel, we aimed at explaining the sensitivity of LCP modeling in the different landscapes according to the cost value distribution of each scenario. For that purpose, we first computed binary variables indicating whether each alternative scenario ordered land cover types in the same way as the true scenario does, or whether the contrast of the cost values was similar. In the latter case, we considered that the cost values [1, 4, 101, 1000] were the closest to the true contrast. Finally we identified the pairs of land cover types that were not ordered in the same way as in the true scenario.

2.6 Statistical analyses of the drivers of LCP modeling sensitivity to cost scenarios

The values of (i) the Mantel r correlation coefficients between true and alternative CD matrices (Mantel correlation hereafter) and of (ii) the proportions of spatial overlap of the alternative least cost corridors with the true ones (spatial overlap hereafter) were supposed to reflect the sensitivity of LCP modeling to cost values. The greater the variability of these metrics in a given landscape for the different scenarios, the greater the sensitivity of LCP modeling in this landscape. Besides, for a given cost scenario, these metrics took values across the landscapes reflecting the overall similarity of this scenario to the true scenario. By computing these two metrics in every landscape and for every cost scenario ($2 \times 77 \times 100$ values), we could compare their sensitivity, assess whether and to what extent the landscape structure explained their sensitivity and identify which characteristics of the cost scenarios make them similar to the true scenario. Accordingly, we first performed separate two-way ANOVA of these two metrics by considering the cost scenario and landscape as the factor variables explaining their variations (Figure 42 B1). This allowed us to quantify the contribution of each factor to the variations of both Mantel correlations and spatial overlaps.

Then we studied whether landscape composition and configuration variables could explain the sensitivity to cost scenarios (Figure 42 B2). For that purpose, we computed the mean and coefficient of variation of the Mantel correlations and of the spatial overlaps for each landscape across the different cost scenarios. Large mean values indicate that, independently of the cost scenario, the Mantel correlations or spatial overlaps tend to be high for a given landscape, whereas large coefficients of variation indicate that these metrics are highly variable depending on the cost scenario for a given landscape. We then modeled these numeric indicators as a function of the landscape variables. We carried out two separate Partial Least Squares (PLS) regressions with the mean and the coefficient of variation of these two metrics as the response variables (one model for each metric) and the landscape variables as the predictor variables. PLS regressions are an alternative to multiple linear regression and principal component regression (Carrascal *et al.*, 2009 ; Roy *et al.*, 2015 ; Wold *et al.*, 2001), particularly suitable for cases in which predictor variables are collinear. This type of regression identifies the factorial space components that simultaneously maximize the explained variance of the response variables and of the predictor variables. This makes it possible to model a set of response variables (PLS-R2). Following Tenenhaus (1998), we computed the Q^2 index to assess the role of every component for improving the prediction of the response variables when performing leave-one-out cross-validation. We described

only the results regarding the effects of the components which significantly improved the prediction of the response variables, i.e. when the Q^2 associated with these components were larger than 0.0975.

Finally, we aimed at identifying the characteristics of the cost scenarios driving the sensitivity of LCP modeling to cost scenarios (Figure 42 B3). We computed the mean values of the Mantel correlations and of the spatial overlaps for each cost scenario across the different landscapes. High values indicate that, independently of the landscape, a given cost scenario leads to LCPs and CDs that are very similar to those derived from the true scenario. We expected this similarity to be explained by the raw cost values of each cost scenario, by their orders and contrasts and by their differences with those of the true scenario (cf. previous section). In order to obtain a decision tree showing the cost scenario characteristics leading to the similarity of LCP modeling outputs with the true ones, we built regression trees (Breiman *et al.*, 1984) to explain either the mean Mantel correlation or mean spatial overlap as a function of the cost scenario characteristics. 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). These trees can take both continuous and categorical predictor variables and have been shown to perform better than linear models in the presence of non-linear relationships. They were pruned according to a cost-complexity criterion to prevent overfitting, using `rpart` package (Therneau *et al.*, 2010) in R.

3 Results

3.1 Structure of the sampled landscapes

After applying our selection criteria to the sampled landscapes, we ended up with 77 landscapes (Figure 43), all very different in terms of both landscape composition (land cover type proportions and diversity) and configuration (fragmentation, number of patches, and contagion)(Table 16). This sample included fine-grained and coarse-grained agricultural landscapes (Figures 43A and 43F, respectively) and widely forested landscapes in both lowlands and highlands (Figures 43G and 43C, respectively).

Variable	Minimum	Median	Maximum
% forest (15 - 100)	15.61	28.57	79.16
% grasslands (5 - 100)	5.38	26.43	69.66
% crops (5 - 100)	5.08	21.97	59.98
% artificial areas (2 - 100)	2.04	8.42	24.65
Shannon div. index (0 - 1)	0.46	0.75	0.93
Frag. forest patches ('clumpy') (-1 - 1)	0.81	0.93	0.97
Frag. artif. patches ('clumpy') (-1 - 1)	0.67	0.79	0.91
Nb. forest patches (0 - 9×10^6)	2467	6563	31918
Nb. artif. patches (0 - 9×10^6)	7602	18610	48569
Shape complex. forest patches (>1)	1.24	1.39	1.49
Shape complex. artif. patches (>1)	1.29	1.36	1.43
Contagion (0 - 100)	46.24	58.29	74.60
Total cost ($\times 10^9$)(0.009 - 9)	0.41	1.09	2.34

TABLE 16 – Landscape characteristic distributions observed among the 77 sampled landscapes. The possible range of variation of the variables is shown in brackets after the variable name.

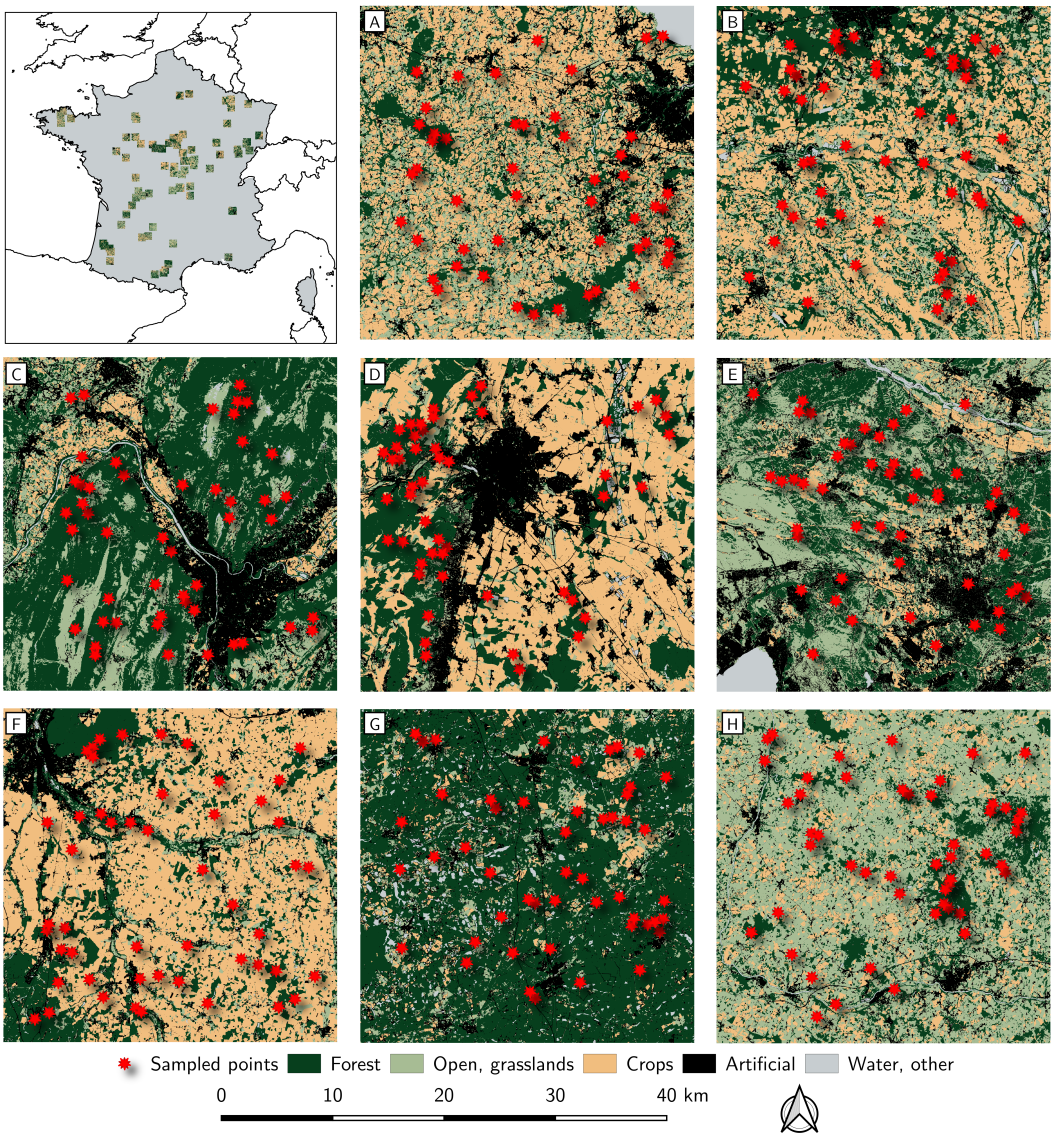


FIGURE 43 – Results of the landscape sampling : 77 landscapes of 30 km × 30 km with a spatial resolution of 10 m were randomly sampled across metropolitan France. They were filtered applying the following criteria : > 15 % forests, > 5 % grasslands, > 5 % crops, > 2 % artificial areas. Fifty points were randomly sampled in the forest areas, all more than 500 m apart. Eight contrasted examples (A to H) are shown on the map.

3.2 Relative influences of cost values and landscape structure on the sensitivity of LCP modeling outputs

The Mantel correlations ranged from -0.40 to 0.99 and the spatial overlaps exhibited similar variations (from 7% to 94%) but never reached 100%, their theoretical maximum (Figure 44, see figure 50 for a similar variation using the Spearman Mantel correlation coefficient). This indicates that the wide range of cost scenarios we considered was sufficient for creating contrasted outputs and studying their variability and its drivers. The variance of the spatial overlap for every combination of a landscape and a cost scenario was accounted for much more by the cost scenario than by the landscape considered (F values from the two-way ANOVA : 498.17 vs 33.71, respectively), although both influences were highly significant. Indeed the cost scenario and the landscape together explained 87% of the variance of this metric but applying the same cost scenario to the different landscapes led to lower variations in terms of spatial overlap than applying the different scenarios to the same landscape did (Figures 44A and 44B).

The cost scenario and the landscape together explained a slightly lower yet significant share of variance of the Mantel correlations (79 %). Similarly the magnitude of variation of the Mantel correlations was much lower for a given cost scenario across the landscapes than for a given landscape across the cost scenarios (F values from the two-way ANOVA : 235.39 vs 67.34, respectively, both highly significant, Figures 44C and 44D).

The spatial overlap was more sensitive to the cost scenario than the Mantel correlation, as shown by the rapid decrease of the median spatial overlaps computed with each cost scenario (Figure 44A) compared with the slower decrease of the median Mantel correlations (Figure 44B). In contrast, when considering the distribution of these two metrics for each landscape across the different cost scenarios, we observed less variation for the spatial overlap than for the Mantel correlations (Figures 44B and 44D). Indeed the spatial overlaps across the cost scenarios were overall small for a given landscape (Figure 44B). In contrast, the Mantel correlations were much more variable (Figure 44D) and consistently took large values in some landscapes, whereas they took smaller values with larger variations in others.

3.3 Landscape structure influence on the sensitivity of LCP modeling outputs

The PLS regressions identified the landscape variables responsible for the sensitivity of LCP modeling to the cost scenarios. Only the first component of the PLS regression explaining the mean and the coefficient of variation of the spatial overlaps across the cost scenarios had a significant effect ($Q^2 = 0.34$, Figure 45A). The mean spatial overlap was highly and positively correlated to the first component of the PLS whereas the coefficient of variation of this variable was only slightly correlated and not significantly explained by this component. The mean spatial overlap was positively influenced by the contagion variable and by the proportion of forests in the landscapes, and negatively influenced by the total cost of the landscape, the Shannon diversity index, the number of patches of artificial area and of forest, and by the proportion of artificial areas (Figure 45A). This means that the spatial overlap was higher in landscapes relatively favorable to species movements, with little diversified land cover types dominated by forests and containing large and aggregated patches.

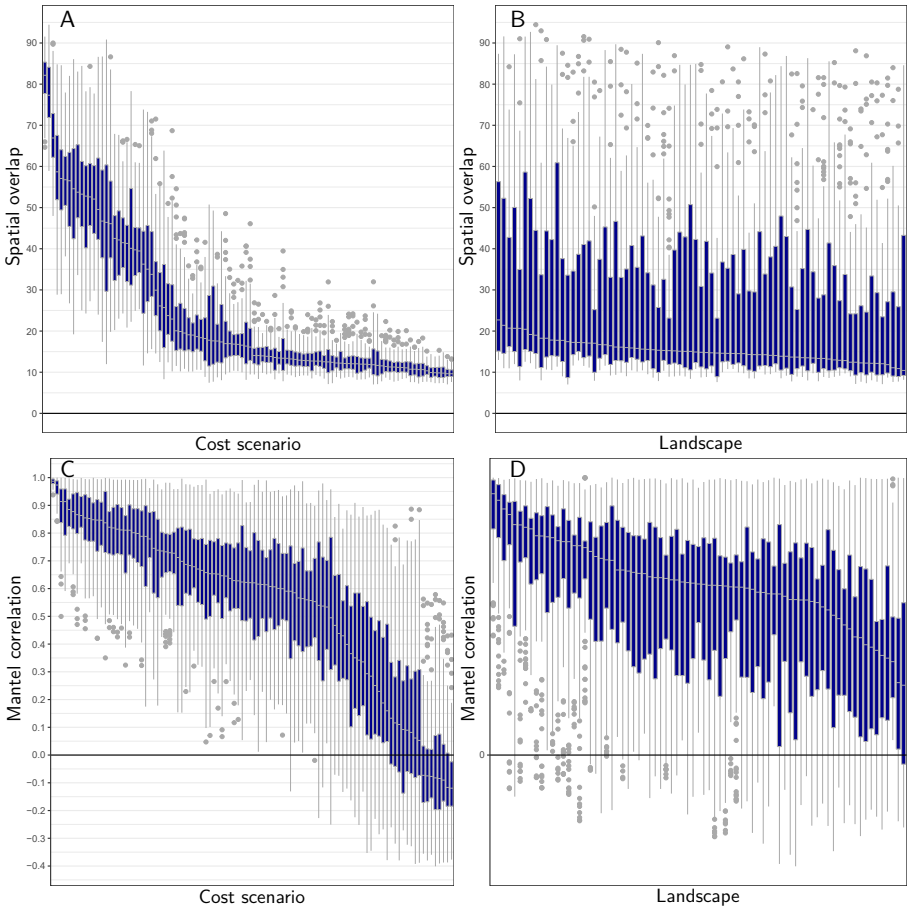


FIGURE 44 – Distribution of the spatial overlap (A, B) or the Mantel correlations (C, D) according to the cost scenario (A, C) or the landscape (B, D). Cost scenarios (A, C) and landscapes (B, D) are placed in decreasing order of median spatial overlap (A, B) or median Mantel correlations (C, D) along the x axis. When the distribution is displayed for each cost scenario (A, C), the 100 boxes are made up of 77 values each corresponding to a landscape, whereas when it is displayed for each landscape (B, D), the 77 boxes are made up of 100 values each corresponding to a cost scenario.

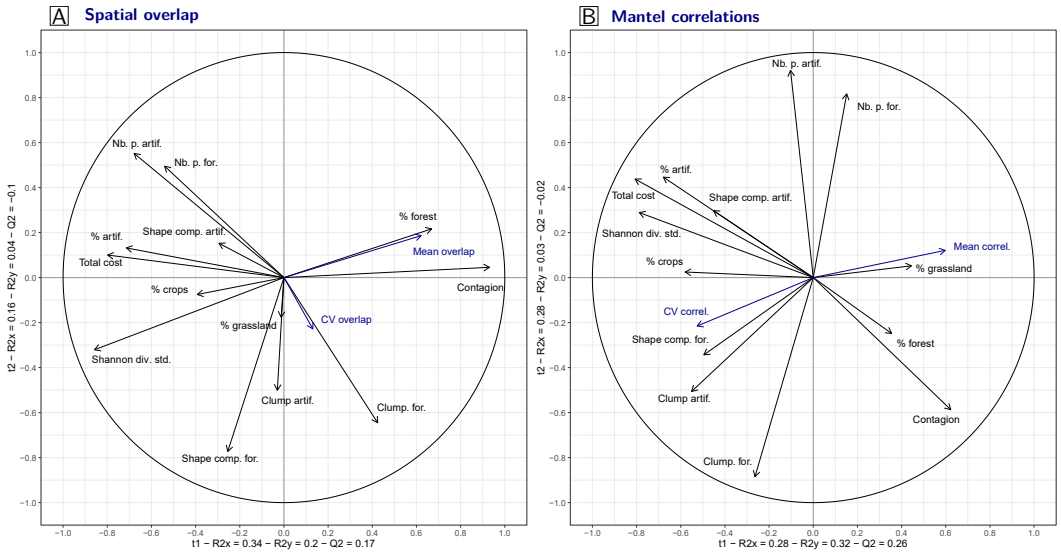


FIGURE 45 – Projection of the response (dark blue labels) and predictor variables (black labels) of the PLS-R2 regression in the factorial space derived from the two first components (t1, t2). The left panel (A) shows the results obtained when modeling the mean and coefficient of variation of the spatial overlap in a given landscape across all the cost scenarios whereas the right panel (B) shows results obtained when modeling the mean and coefficient of variation of the Mantel correlation in a given landscape across all the cost scenarios. The axis labels indicate the percentage of variance of the predictor variables table (R2x) or of the response variables table (R2y) explained by each component (t1 or t2), as well as the percentage of variance of the response variables table explained by these components when performing a cross-validation (Q2).

Similarly only the first component of the PLS regression explaining the mean and the coefficient of variation of the Mantel correlations for a given landscape had a significant effect ($Q^2 = 0.26$, Figure 45B). This component was positively correlated with the mean Mantel correlation whereas it was negatively correlated with its coefficient of variation. Mean correlation coefficients were positively influenced by the contagion index. Conversely the coefficients of variation of these coefficients were positively influenced by the total cost of the landscape, the Shannon diversity index, the proportion of crops and artificial areas, the shape complexity of the forest patches, and the Clumpy index of forest and artificial area patches. This means that CD matrices consistently exhibited high correlations with the true CD matrix in landscapes with large contiguous patches. In contrast, alternative CD matrices tended to be less strongly and more variably correlated with the true CD in diverse landscapes with complex patch shapes and large areas of the least favorable land cover types.

3.4 Cost scenario characteristics influence on the sensitivity of LCP modeling outputs

The regression trees identified the cost scenario characteristics explaining their spatial overlaps and Mantel correlations with the true scenario across the different landscapes (Figure 46). The first split of the two regression trees created were conditions regarding the cost value assigned to forest. Indeed, cost scenarios with forest cost values lower than 82 led to mean spatial overlaps averaging 40% (Figure 46A). Conversely the cost scenarios assigning forests a cost value lower than 875 (i.e. different from 1000) led to mean Mantel correlations averaging 0.63 across landscapes (Figure 46B). Interestingly, when the forest cost value was equal to 1000, if the cost values were drawn from either the [63, 250, 563, 1000] or [250, 500, 750, 1000] gradients, the Mantel correlations still averaged 0.55 although such cost scenarios differed largely from the true one. In this case, assigning much lower cost values to grasslands, crops, and artificial areas ([1, 1, 11], [1, 4, 101], or [4, 63, 317]) led to negligible Mantel correlations (Figure 46B). Accordingly, the gradient of values of the cost scenarios was the

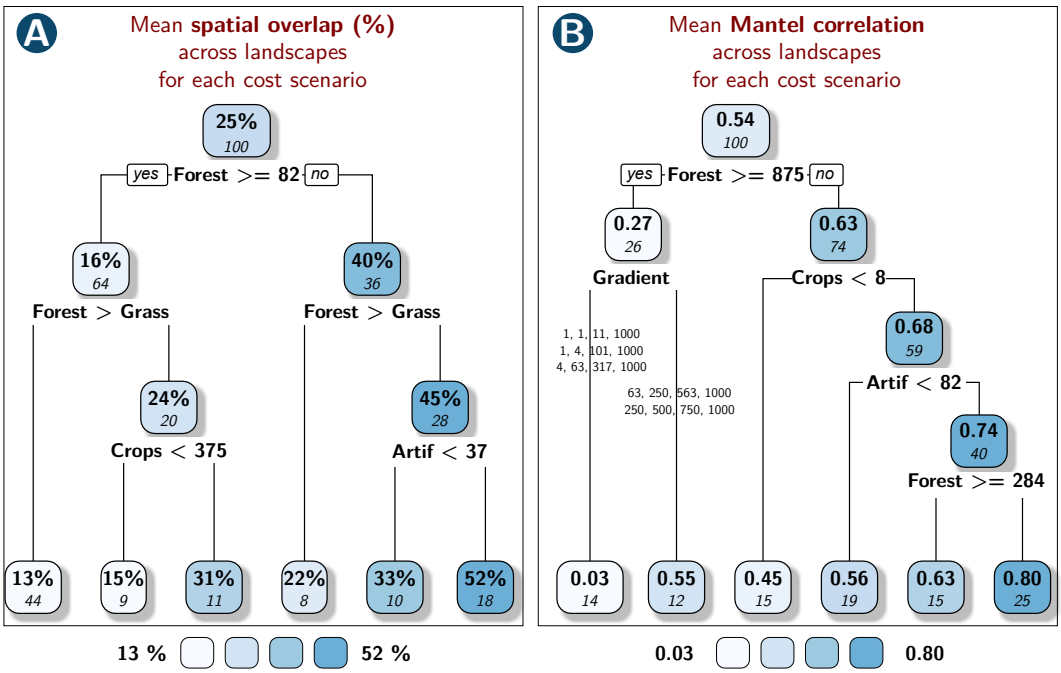


FIGURE 46 – Regression trees explaining the mean spatial overlap (A) or the mean Mantel correlation across landscapes for each cost scenario as a function of the characteristics of the distribution of the cost values. At each node (branch split), the criterion displayed is verified in all the leaves stemming from branches located on the left side of the split, whereas its opposite is verified in all the leaves stemming from branches located on the right side of the split. The colored boxes indicate the mean value of the response variable across all the leaves stemming from a given node (in bold) and the number of observations, i.e. cost scenarios, included in these leaves (in italics).

second most important criterion explaining the Mantel correlations obtained across the landscapes for a given cost scenario.

In contrast, the second most important criterion explaining the spatial overlap was the difference between cost values assigned to forest and grassland, which are the two least resistant land cover types in the true scenario (Figure 46A). Making forests more resistant than grasslands invariably reduced the spatial overlap with the true LCP. Finally, in both trees, the other splitting criteria concerned the costs associated with crops and artificial areas. For example, when forest cost value is both lower than 82 and lower than the grassland cost value, assigning artificial areas a cost value lower than 37 (true cost value : 1000) led to spatial overlaps averaging 33 %, which is a rather large value in light of the range of variation. Similarly, provided the cost value was lower than 284 for forests, greater than 8 for the crops and greater than 82 for artificial areas, the mean Mantel correlation across landscapes averaged over the corresponding scenarios reached 0.8, independently of the order of cost values and the contrast between them. The binary variables comparing each scenario to the true one in terms of order and contrasts were not retained in the best trees computed for both metrics.

3.5 Relationship between the spatial overlap of LCP and the correlation between CD matrices

From the application of every alternative cost scenario in every landscape (7700 combinations), we observed that the spatial overlaps and the Mantel correlations were somehow related (Spearman's $r = 0.66$) but their relationship was highly non-linear (Figure 47). Spatial overlaps above 65% were only obtained with LCPs whose associated CDs were moderately to highly correlated with the true CD ($r > 0.5$, Figure 47). Yet the degree of correlation between CD matrices was a poor proxy for the

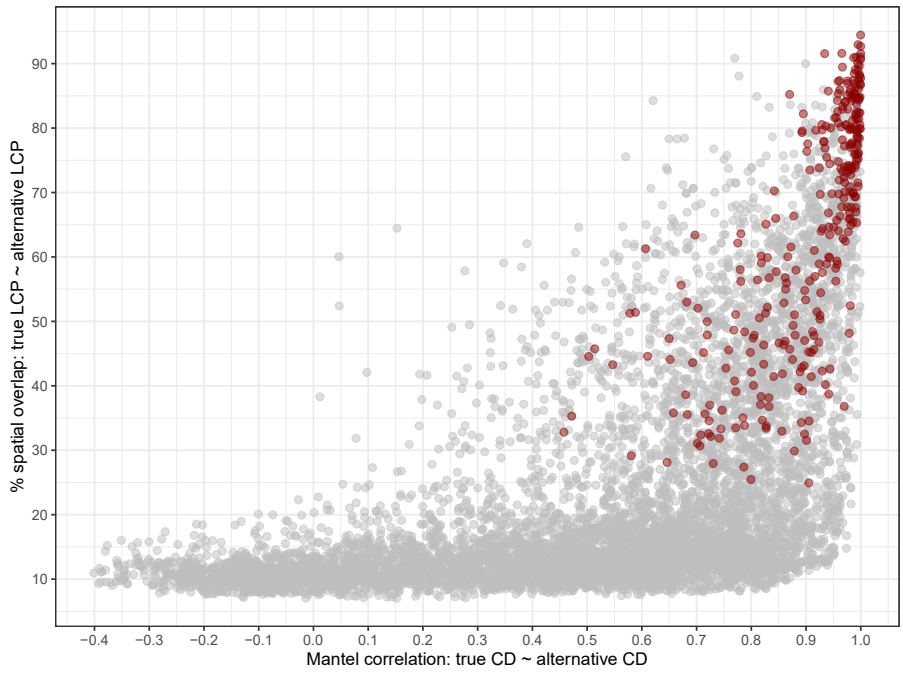


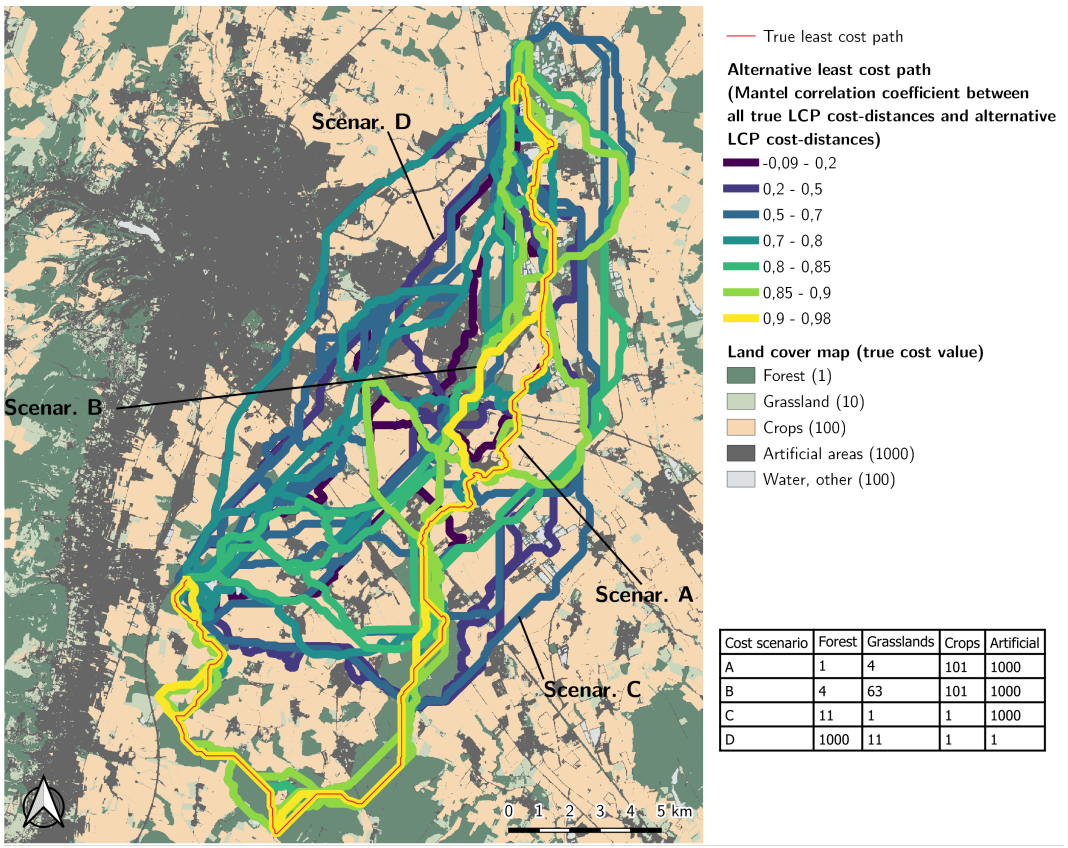
FIGURE 47 – Relationship between the Mantel correlations (x axis) and the spatial overlap (y axis). The spatial overlap is expressed as the proportion of the area of the 200 m wide buffer around the true LCP that overlaps the same buffer created from each alternative LCP. $n = 7700$, corresponding to the combination of 100 cost scenarios and 77 landscapes. Red dots correspond to alternative cost scenarios ordering the four land cover types in the same way as in the true cost scenario.

spatial overlap of LCPs. Indeed, spatial overlaps below 20% were frequently obtained while corresponding CD matrices were highly correlated with the true CD matrix ($r > 0.9$, Figure 47 and Table 17).

In addition, spatial overlaps above 80% were mostly observed when land cover types were arranged in the same order of resistance as in the true cost scenario (Figure 47). However, some scenarios incorrectly ordering these land cover types still reached spatial overlaps up to 91%. The cost scenario leading to the largest Mantel correlations was [1, 4, 101, 1000](for forest, grasslands, crops, and artificial areas, respectively ; Mantel r : mean across landscapes = 0.98, Table 17). This scenario was apparently the most similar to the true one (i.e. [1, 10, 100, 1000]) in terms of order and contrast of the cost values but surprisingly the mean spatial overlap of the corresponding LCP across the 77 landscapes was not the maximum (77.81% vs 81.00%, Table 17). Indeed, the best scenario in terms of spatial overlap was [4, 63, 317, 1000] and also led to CD values highly correlated with the true ones (mean Mantel correlation across landscapes = 0.96).

Although the best cost scenarios in terms of spatial overlap always assigned a larger cost value to grasslands than they did to forests, they did not systematically assign a larger cost value to artificial areas than they did to crops (e.g. scenario [1, 4, 1000, 101] : mean spatial overlap : 68.04%, Table 17). In contrast, in the ten cost scenarios with the strongest Mantel correlations, two cost scenarios assigned a lower cost to grasslands than they did to forests. If the Mantel correlations obtained in these two cases were above 0.85, the corresponding spatial overlaps were nevertheless below 40%.

Finally, when projected into a spatially-explicit layout, we observed large differences between LCPs resulting from different cost scenarios (Supporting information, figure 48). Interestingly, even highly correlated CD matrices could be derived from LCPs diverging rather markedly from each other.



Cost scenario	Forest	Grasslands	Crops	Artificial
A	1	4	101	1000
B	4	63	101	1000
C	11	1	1	1000
D	1000	11	1	1

FIGURE 48 – Spatial representation of the influence of cost scenarios on least cost path locations. The least cost corridor between two locations computed applying the true cost values to land cover types is displayed in red on the map. Alternative least cost paths computed using other cost values are displayed with colors reflecting the mean Mantel correlation between the corresponding alternative cost distance matrices and the true cost distance matrix (including all the paths computed on this landscape). Scenarios A, B, C, and D are examples of cost scenarios that diverge to varying degrees from the true one. Their corresponding costs are in the table included in this figure.

Forest	Grasslands	Crops	Artif	For. <Grass.	Order	% overlap	Mantel r
1	4	101	1000	Yes	Same	**77.81**	**0.98**
4	63	317	1000	Yes	Same	**81.00**	**0.96**
1	4	1000	101	Yes	Diff.	**68.04**	**0.88**
4	1	101	1000	No	Diff.	25.78	**0.88**
4	317	63	1000	Yes	Diff.	**58.83**	**0.87**
1	1	11	1000	No	Diff.	36.86	**0.87**
63	250	563	1000	Yes	Same	**56.08**	**0.86**
63	563	1000	250	Yes	Diff.	**55.52**	**0.84**
63	563	250	1000	Yes	Diff.	**54.59**	**0.84**
1	101	4	1000	Yes	Diff.	**54.44**	**0.83**
63	1000	563	250	Yes	Diff.	52.25	0.83
63	1000	250	563	Yes	Diff.	51.60	0.83
1	11	1	1000	Yes	Diff.	45.86	0.81
4	317	1000	63	Yes	Diff.	**57.19**	0.81
250	500	1000	750	Yes	Diff.	40.09	0.79
250	750	1000	500	Yes	Diff.	41.67	0.79
4	1000	63	317	Yes	Diff.	**53.70**	0.79
250	500	750	1000	Yes	Same	39.80	0.79
250	1000	750	500	Yes	Diff.	41.89	0.79
250	750	500	1000	Yes	Diff.	40.13	0.78
4	1000	317	63	Yes	Diff.	52.70	0.76
1	11	1000	1	Yes	Diff.	43.99	0.74
1	1000	4	101	Yes	Diff.	48.63	0.67
1	1000	101	4	Yes	Diff.	47.03	0.62

TABLE 17 – Differences between the alternative cost scenarios and the true cost scenario. We included the cost scenarios with the 20 highest mean proportions of spatial overlaps or mean Mantel correlation coefficients (24 scenarios in total). They are ordered in descending order of Mantel correlation coefficient (Mantel r). Their spatial overlap with the LCPs derived from the true cost scenario are displayed (% overlap). The 10 largest values of the columns 'Mantel r' and '% overlap' are displayed in bold. Values obtained for all LCPs, CD matrices and landscapes have been averaged for each cost scenario. Cost values associated with each of the four land cover types are included in the 'Forest', 'Grasslands', 'Crops', and 'Artif' columns. The 'For. < Grass.' column indicates whether the cost value associated with forest is lower than that associated with grasslands. The 'Order' column indicates whether cost values associated with land cover types in each alternative cost scenario follow the same order as those in the true cost scenario.

4 Discussion

Using a wide range of cost scenarios in real diversified landscapes, we analyzed the relative sensitivities of both LCP spatial locations and CD values to the choice of cost values and identified their drivers. As expected, these outputs of LCP modeling were sensitive to the cost scenarios but their sensitivities differed and were not influenced by the same characteristics of the cost scenarios, nor to the same extent according to landscape structure.

4.1 Sensitivity of LCP and CD to cost scenarios

The spatial overlap of LCP is very sensitive to the cost scenario

An analysis of the spatial sensitivity of LCP modeling to widely diverging scenarios was lacking. Using the spatial overlap between true and alternative LCPs obtained from different cost scenarios as a measure of sensitivity, we showed that LCPs are highly dependent upon the cost scenarios given that just a few scenarios allowed us to reach large proportions of overlap with the true LCPs. Previous studies regarding the spatial overlaps of LCPs provided inconsistent findings. Pullinger et Johnson (2010) showed that LCPs did not follow caribou GPS tracks, whereas Beier *et al.* (2009) obtained large spatial overlaps of alternative corridors between two protected areas. Our results can explain these opposing conclusions. Although the cost scenario was the main driver of the spatial overlap, in landscapes with

large proportions of favorable land cover types, large contiguous patches and a low diversity of land cover types, the spatial overlap between true and alternative LCPs tended to be larger. In this case, using similar scenarios in terms of cost value order and range would lead to large spatial overlaps across the scenarios. This could potentially explain the results of Beier *et al.* (2009) because these authors used a set of biologically plausible cost scenarios. In contrast, in fine-grained landscapes with the highest diversity of land cover types and the largest proportions of adverse land cover types, the spatial overlap was consistently low. This situation could reflect the study by Pullinger et Johnson (2010), performed in a landscape in which very contrasted altitude classes were intersected with 10 different land cover types to define a fine-grained permeability map. This also indicates that the thematic resolution of land cover maps could influence the outputs of LCP modeling by influencing their grain and diversity, although spatial resolution is often given more consideration (Lechner et Rhodes, 2016).

Highly correlated CD matrices can derive from very different cost scenarios

We studied the sensitivity of the statistical properties of CD to cost scenarios by comparing the Mantel correlations between the true CD matrix and every alternative CD matrix. Our results showed that CD matrices highly correlated to the true one can be obtained using several cost scenarios differing widely from the true cost scenario. Besides, although the sensitivity of this correlation depended more on the characteristics of the cost scenarios, the landscape context was also responsible for the variable sensitivity to cost scenarios observed in the different landscapes. In landscapes with large amounts of favorable land cover types and large patches, whatever the alternative cost scenario, the CD matrices tended to be highly correlated with the true CD matrix. Conversely, in diverse landscapes with patches of complex shapes, correlations with the true CD were lower and much more variable. This result recalls those of Cushman *et al.* (2013b) showing that Euclidean distances and CDs were equivalent for explaining genetic distances when the proportion of habitat in the landscape is high and the contrast between cost values is low.

Highly correlated CD matrices can derive from spatially distinct LCPs

Spatial overlaps and Mantel correlations exhibited different sensitivity to cost scenarios when considered separately. The main contribution and novelty of our analyses is to provide insights into the relationship between the spatial locations of LCPs and their corresponding CD values, and into the drivers of the mismatches between them. We first showed that two highly correlated CD matrices can derive from paths whose spatial overlap is very low (as low as 15% with correlation coefficients above 0.9). Nevertheless, the reverse does not hold true because large spatial overlaps between paths invariably involve high correlations between CD matrices. This result is explained by the greater sensitivity to cost scenarios of the LCPs than the corresponding CDs.

These contrasted sensitivities partly stem from the fact that LCPs and CDs are not influenced by the same characteristics of cost scenarios. While the relative order of the cost values associated with the least resistant land cover types (forests and grasslands) is a key factor explaining the spatial overlap of a given cost scenario with the true scenario, the correlation of alternative CDs with the true CDs depended more heavily on the gradient of cost values. This seems logical given that the order of the cost values determines whether the path should better cross some land cover types than others, whereas the gradient of cost values determines the CD statistical distribution independently of the

spatial location of the corresponding LCP. Interestingly, when the least resistant land cover type was assigned a large cost value (forest : 1000), limiting the contrast with the costs of other land cover types still led to CDs highly correlated with the true CDs. Such strong correlations between CDs deriving from the most homogeneous cost scenarios and the reference CD recall the strong correlations often observed between CDs and Euclidean distances (Marrotte et Bowman, 2017). This strong correlation has been a reason for preferring the accumulated cost along the LCP (CD) over the length of the LCP as a measure of connectivity (Etherington et Holland, 2013 ; Simpkins *et al.*, 2018). However, the dependence of CDs upon Euclidean distances is still a limitation of this measure and can make it difficult to distinguish several CD matrices.

Furthermore, although the cost scenario being at first sight the most similar to the true scenario ([1, 4, 101, 1000] vs [1, 10, 100, 1000]) led to the CDs most strongly correlated with the true CDs, it did not lead to the highest spatial overlap with the true LCPs, obtained with the scenario [4, 63, 317, 1000]. This could be explained by the sensitivity of LCPs to the contrasts of cost values between the least resistant land cover types, the ratios 4/63 and 63/317 being both closer to 1/10 and 10/100 than 1/4 and 4/100.

4.2 Implications for cost value inference and LCP modeling

Ecological interpretations and use of inferred cost values must be subject to caution

The statistical distribution of two CD matrices can be almost identical although they correspond to spatially distinct LCPs that derive from cost scenarios implying different ecological interpretations. Assuming that cost value inference from biological data depends essentially upon the statistical properties of CD matrices, care has to be taken when interpreting inferred cost values and using them for mapping LCPs. Similarly the cost scenarios leading to the largest spatial overlap are not necessarily the scenarios whose values are most like the true cost values. Indeed, given that inferred cost values may be closely related to the statistical properties of the CDs, these inferred cost values should better reflect the gradient of cost values and the difference between the lowest and largest cost values than their relative order. Note that this limitation does not concern the cost values inferred from presence or telemetry data. Yet, in the latter case, the method used for converting SDM or step selection function outputs into cost values could significantly affect CD statistical distribution by determining the range and contrasts between these values.

Outline solutions for the use of inferred cost values in LCP modeling

When LCP modeling supports decision-making in conservation, the spatial location of the LCP can be used to design restoration measures such as wildlife crossings for example (Clevenger *et al.*, 2002 ; Mimet *et al.*, 2016). Such a location optimization based upon LCPs can be suboptimal due to the sensitivity of these paths to the cost scenario. Although it may be problematic when LCP modeling is based upon cost values inferred from the relationship between CDs and biological data, we provide outline solutions to this problem. First, the scenario leading to the highest spatial overlap with the true LCPs was always within the scenarios leading to the CD matrices most closely correlated with the true one. This means that the set of cost scenarios closely reflecting the true landscape constraints on movements share similar statistical properties and could be retained as the 'best ones' in cost value inference. Our results thus call into question the common practice of optimization of a single best cost scenario. Rather than retaining the single 'best' cost scenario from the inference, retaining a set made

of several best scenarios could ensure that the 'ecological truth' is part of the inference results. This strategy is not unlike the use of Circuitscape software (McRae et Beier, 2007), which models several alternative paths between locations in a landscape considered as an electric circuit. Similarly, Pinto et Keitt (2009) developed methods for modeling multiple shortest paths between habitat patches and Rayfield *et al.* (2010) suggested identifying such multiple low-cost routes for coping with the sensitivity of LCP modeling. The Linkage Mapper software (McRae et Kavanagh, 2011) makes this possible by creating least cost corridors of varying width according to the cost surface, which could provide insight into the existence of alternative and equally probable paths around the least cost path (see also Shirabe (2016)).

Yet, instead of modeling several alternative paths under one cost scenario, we here suggest to model several LCPs under a set of highly likely scenarios because this strategy could maximize the likelihood of taking into account the 'true' LCP. It would mirror the growing interest for multi-model inference (Burnham et Anderson, 2004) in ecological science where considering a single best model is often a poor approximation of the stochastic ecological reality. Similarly, the set of highly likely scenarios could be selected on the basis of a model fit criterion, e.g. the AIC (Burnham et Anderson, 2004). We acknowledge that the alternative LCPs thereby identified may occupy very different spatial locations. In such a case, their intersections may be the only information that can be used for conservation purposes.

Another strategy would be to limit the number of cost scenarios to maximize the contrasts between them and their corresponding CD matrices, because statistical inference cannot distinguish them if they are too strongly correlated (Zeller *et al.*, 2016). Besides, although cost values inferred from such an approach should be used carefully for locating LCPs, the CD matrices derived could be used for estimating the importance of the locations linked by LCPs for the connectivity of a whole network of patches using graph-theoretical connectivity metrics (Foltête *et al.*, 2014). This could represent a reliable alternative to the spatial application of the results of this type of inference. However, note that when CD thresholds are used to define the connections between patches, we can expect the statistical distribution of CD values and consequently the range of the cost values to affect metric calculations.

Methodological perspectives for LCP modeling

Although the competing cost scenarios can be controlled until the very end of a study, the studied landscape is determined in the early stages. We here showed that in landscapes with high proportions of favorable land cover types, reduced land cover diversity and large contiguous patches, the correlation coefficients between CD matrices deriving from very different cost scenarios consistently reached high values. In such a case, it can be determined beforehand that the reliability of the cost value inference will be reduced, as already shown by Cushman *et al.* (2013b). The sensitivity of LCPs to cost scenarios should therefore be tested prior to any study if the main objective is to infer the resistance to movements. For that purpose, we included the function `link_compar()` which computes the spatial overlap between several sets of LCPs within the `graph4lg` package in R (Savary *et al.*, 2021b). This function makes it possible to specify the width of the least cost corridors. Indeed we used here a constant total width of 400 m, which reflects the scale at which conservation measures can be implemented following connectivity modeling (Ford *et al.*, 2020 ; Spackman et Hughes, 1995) and prevents overestimating the spatial overlap for short LCPs.

The lower sensitivity of LCP modeling outputs to the cost scenarios in landscapes with large proportions of favorable land cover types and large patches may be due to the sampling of points within forests. Although this reflects the fact that connectivity analyses aim at identifying favorable paths between similar habitat patches, it also means that whatever the forest cost value, these areas had to be crossed by LCPs and over larger distances in such landscapes. This could have reduced the differences between LCPs and CD matrices computed under different scenarios. Considering resistance distances using the circuit theory (McRae, 2006) could have decreased the correlations between distance matrices obtained with cost scenarios assigning different cost values to forests. In contrast, using current maps of connectivity obtained from the circuit theory would probably have increased the overall spatial overlap between the most similar cost scenarios due to the consideration of alternative LCPs which are potentially shared across similar scenarios.

Finally we raised concerns about the risk of identifying cost scenarios in data based inference leading to incorrect qualitative and spatial output, while being highly correlated with biological responses. Previous landscape genetic studies investigating the promises and pitfalls of cost surface parametrization from genetic data (Cushman *et al.*, 2013b ; Graves *et al.*, 2012, 2013 ; Koen *et al.*, 2012 ; Spear *et al.*, 2010) should be completed by considering our results.

Acknowledgements

This study is 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. Part of the analyses were carried out on the calculation "Mésocentre" facilities of the University of Bourgogne-Franche-Comté. We thank Christopher Sutcliffe for revising the English manuscript

Data and Codes Availability Statement

The functions `sample_raster()`, `graphab_link()`, `mat_cost_dist()` and `link_compar()` respectively used for sampling points within landscapes, computing LCPs between these points, computing cost distances and comparing LCPs spatially have been included into the R package `graph4lg` and are directly available here : https ://cran.r-project.org/web/packages/graph4lg/index.html. The OSO land cover raster data are available here : http ://osr-cesbio.ups-tlse.fr/ oso/posts/2018-04-09-carte-s2-2017/.

A - Supplementary figures

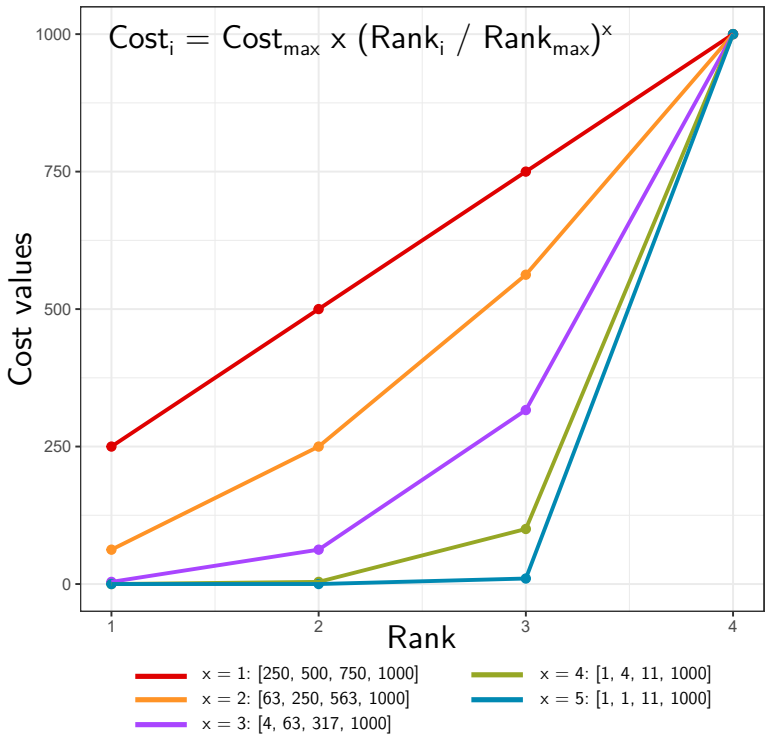


FIGURE 49 – Gradient of cost values according to the alternative cost scenarios created using the method of Shirk *et al.* (2010).

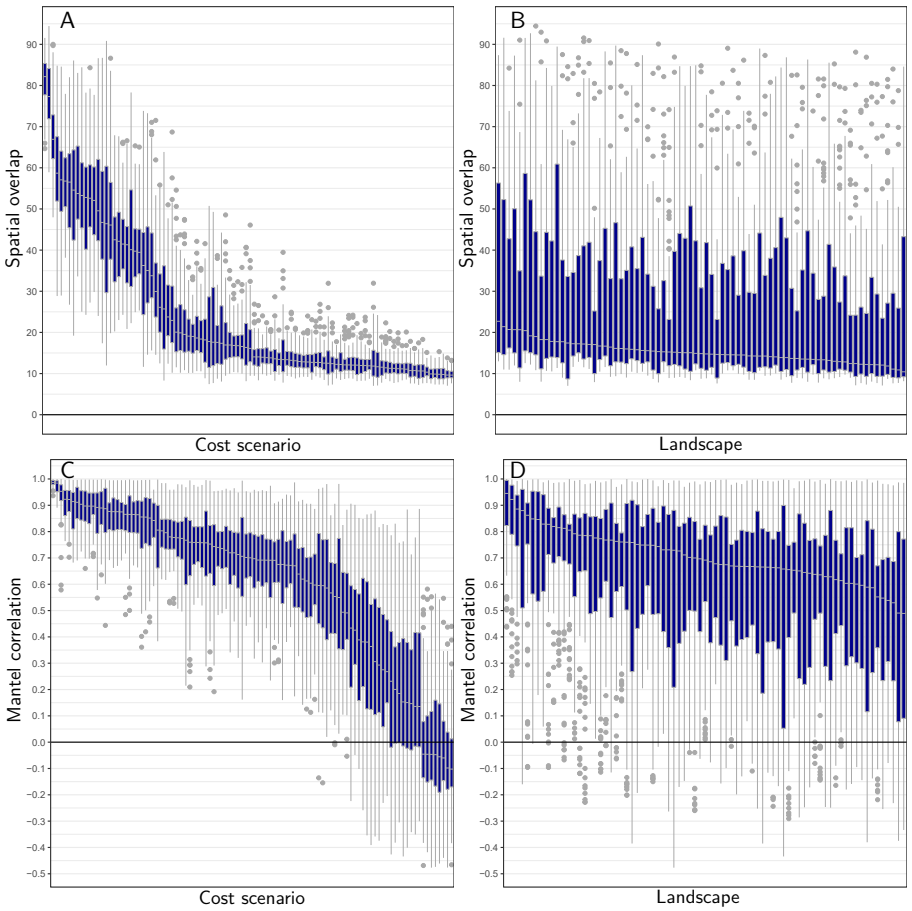


FIGURE 50 – Distribution of the proportions of spatial overlap between the alternative and the true least cost paths (A, B) or the Spearman Mantel r correlation coefficients between the alternative cost distance matrices and the true one (C, D) according to the cost scenario (A, C) or the landscape (B, D). Cost scenarios (A, C) and landscapes (B, D) are placed in decreasing order of mean proportion of spatial overlap (A, B) or mean correlation coefficient (C, D) along the x axis. When the distribution is displayed for each cost scenario (A, C), the 100 boxes are made up of 77 values each corresponding to a landscape, whereas when it is displayed for each landscape (B, D), the 77 boxes are made up of 100 values each corresponding to a cost scenario.

Annexe A8

Inferring landscape resistance to gene flow using gravity models

Abstract

Connectivity modelling requires to associate land cover types with cost values characterising their resistance to species movements. Landscape genetic methods allow for the inference of these values from the relationship between genetic differentiation and cost-distances. In this inference, the spatial heterogeneity of population sizes is usually not included while it is known to influence genetic differentiation. Similarly, migration rates and population spatial distribution patterns are potentially influencing this inference. Our objective was therefore to assess the reliability of cost value inference under several migration rates, population spatial patterns and degrees of population size heterogeneity. We also wanted to assess whether the inclusion of intra-population variables in gravity models improved this inference. To that purpose, we simulated several intensities of gene flow between sets of populations of different sizes with various spatial distribution patterns. We then computed gravity models explaining simulated genetic distances as a function of the 'true' 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 aimed at determining conditions making the identification of the 'true' cost-distances and of their corresponding cost value scenario possible and at assessing the contribution of intra-population variables to this objective. Cost value inference was reliable in most cases but was hampered when migration was very restricted, population sizes were most heterogeneous and some populations were spatially aggregated. We further demonstrate the interest that intra-population variables and gravity models represent for the inference of cost values from genetic data.

Keywords : gravity models, landscape genetics, population genetics

Cet article est en préparation pour une soumission dans la revue *X* en 2021 :

Savary, P., Foltête, J. C., Moal, H., Vuidel, G. & Garnier, S. 2021. Inferring landscape resistance to gene flow using gravity models. In prep.

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. To that purpose, 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 last landscape changes for the genetic differentiation pattern to reach an equilibrium, we can expect a positive linear relationship between genetic differentiation and effective distances between populations (Hutchison et 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 population size heterogeneity is high, pairwise population size heterogeneity contributes to genetic differentiation more significantly than the distance between populations does. Furthermore, these authors developed metrics measuring SHNe 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 SHNe 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 SHNe into account in the analyses, but is undoubtedly a difficult task (Wang, 2005). Yet, 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.*, 2010a ; Wang, 2013 ; Wang *et al.*, 2013), though seldom considered in landscape genetic analyses (Pflüger et Balkenhol, 2014), and could positively contribute 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 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, population sizes are spatially heterogeneous and some populations are spatially aggregated. In addition, 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 et O'Kelly, 1989) have already been used in landscape genetics and allow for the test of these hypotheses because these models enable to assess the influence of intra- and inter-population variables on measures of genetic differentiation (Murphy *et al.*, 2010a ; Robertson *et al.*, 2018b ; 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 SHNe.

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 computed 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 aimed at determining the conditions making cost value inference possible and at identifying cases 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

We adopted a 'virtual ecologist' approach (Zurell *et al.*, 2010) in order to assess a commonly used approach in landscape genetics for inferring landscape resistance from the relationship between landscape distances and genetic distances. To that purpose, we simulated the genetic differentiation pattern emerging through gene flow over several generations in a species with limited dispersal capacities (Figure 51). We knew the 'true' cost values associated with land cover types and the resulting cost-distances (CD) driving dispersal in the simulated landscapes. Our objective was to assess 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. In particular, 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.

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.*, 2013b ; Rayfield *et al.*, 2010). This variability ensures that alternative scenarios lead to different cost-distance matrices, thereby making the inference possible. To that purpose, we simulated 200 landscapes with four land cover types using spatially correlated Gaussian random fields models (Schlather *et al.*, 2015) and a level of land cover auto-correlation leading to variable cost-distances across the cost scenarios (`autocor=30` in `nlm_gaussianfield()` function).

We simulated the movement of a forest specialist species with limited dispersal capacities avoiding anthropogenic land cover types when dispersing, corresponding to a usual focal species of connectivity studies. 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 an 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

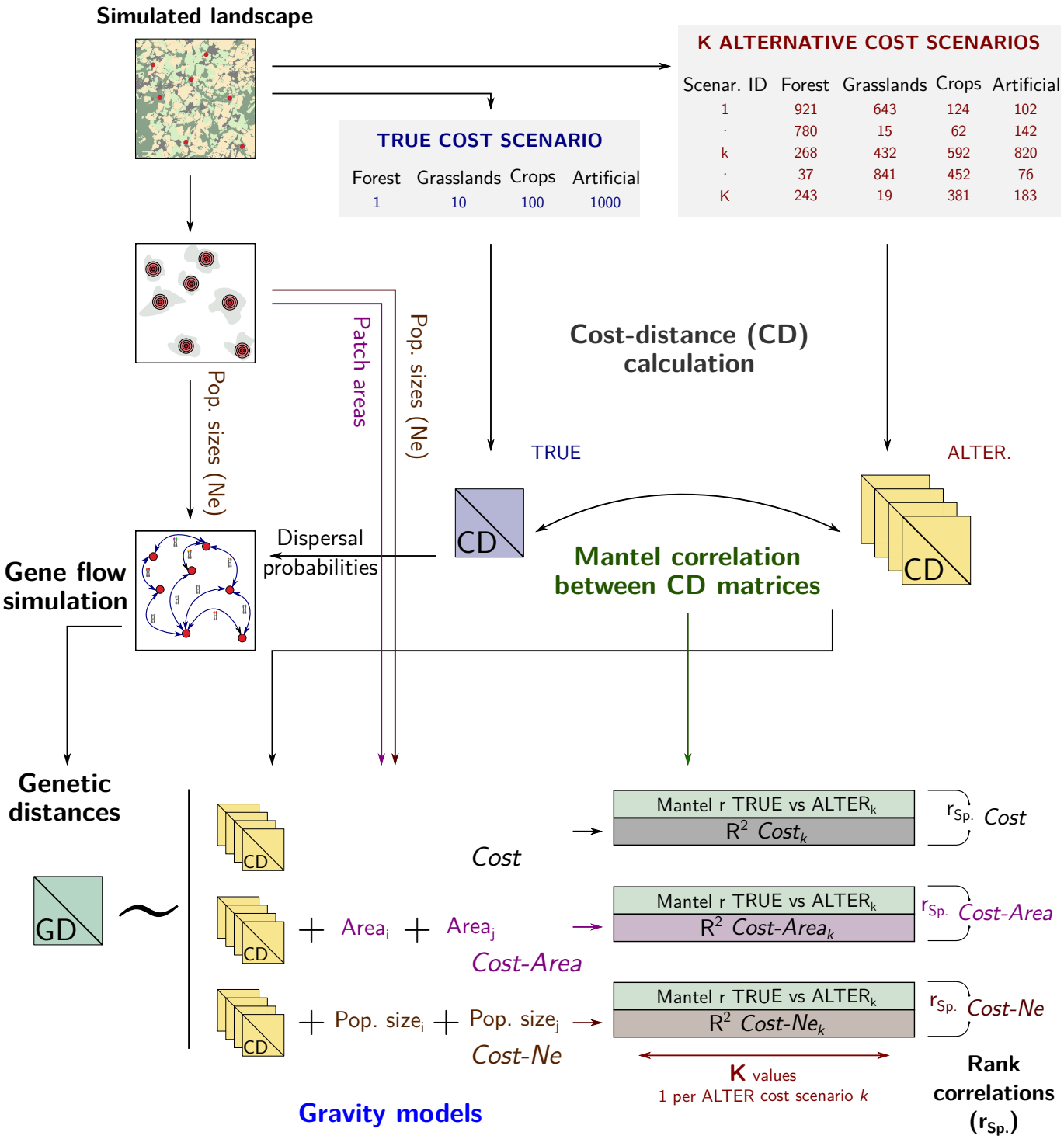


FIGURE 51 – Overall methodological approach

studies (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 aimed at simulating a large range of patterns of population sizes and habitat patch areas. We randomly sampled 60 populations within 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 point at a distance lower than 500 m (Supporting information, Figure 54). This allowed us to vary the area of the habitat patches according to the landscape structure. Then, we set population sizes between 10 and 100 individuals with a total of 3300 individuals but we distinguished two spatial distributions of population sizes :

1. Equal population sizes (Equal) : all 60 populations contained 55 individuals.
2. Area-dependent population sizes (Area) : population size was heterogeneous and dependent upon the area of the habitat patch occupied by the population.

The 'Equal' setting constituted the reference baseline making it possible to assess the effect of SHNe on the inference by comparison with the 'Area' setting. For the 'Area' setting, we aimed at covering a wide gradient of SHNe while mimicking realistic conditions in which population sizes are driven by patch areas. To 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 55). In parallel, we computed the Gini index for 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. In every case, the maximum population density was equal to 2 individuals per hectare of forest. 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.*, 2013b). 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 Shirk *et al.* (2010) approach 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 distribution and the 'true' one (Figure 51). 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 CDPOP software (Landguth et 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 loci 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 assigned at generation 0. 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, dispersal probability p_{ij} between populations i and j followed a negative exponential distribution (Urban et Keitt, 2001) and was a function of CD_{ij}, such that $p_{ij} = e^{-\beta CD_{ij}}$. 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 SHNe 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 SHNe. Preliminary analyses showed that migration rates above 0.005 led to situations in which gene flow was too important for heterogeneous drift effects to influence the inference whereas migration rates below 0.0005 led to situations in which drift effects were too strong for inference to be possible, whatever the landscape and population parameters. In total, 1000 simulations were performed.

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.*, 2021a).

2.3 Gravity models

Gravity models have been initially used in geography and economics (Anderson, 1979 ; Fotheringham et 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.*, 2010a ; Robertson *et al.*, 2018b ; Watts *et al.*, 2015 ; Zero *et al.*, 2017) is more recent. They model spatial interaction or fluxes as a function of both the variables characterizing the objects involved in the interaction and of the distance between them (masses and distance in Newton's gravity theory, respectively). Here, we used them to model the genetic distance G_{ij} between populations i and j (response variable, link-level) as a function of several predictors computed at the population-level (nodes) or between populations (link-level) : cost-distance CD_{ijk} between populations i and j in the cost scenario k, patch areas (a_i, a_j) and population sizes (N_i, N_j). We computed three types of models of the following form in order to test for our hypotheses :

Cost : $G_{ij} \sim c \times CD_{ijk}^{m}$

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

Cost-Ne : $G_{ij} \sim c \times CD_{ijk}^{m} \times N_i^{p} \times N_j^{q}$

c was a constant. We computed these three models using the CD values obtained with every 'true' or alternative cost scenario. A natural log was applied to these formula to obtain the classical formula of a multiple regression model whose parameters (c, m, n, o, p and q in our case) can be estimated. To account for the non-independence inherent to distance matrices, we performed constrained models by adding a random effect corresponding to the identity of the population i (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 51). From these results, we aimed at deriving general guidelines for cost value inference in landscape genetics.

We first used Edward's R^2_β (Edwards *et al.*, 2008) to assess the goodness of fit because other model selection criteria are not relevant when fitting mixed models with residual maximum likelihood estima-

tion (Van Strien *et al.*, 2012). Under our settings, if a given model ('Cost', 'Cost-Area' or 'Cost-Ne') performs well in distinguishing among cost scenarios, the largest R^2_β values should be obtained when the CD values included in this model are the most correlated with the 'true' cost-distance values driving the simulation. In contrast, when a model behaves badly, the ranks of the models obtained with the different cost scenarios according either to the R^2_β values or to their correlation with the 'true' cost-scenario should be independent. To quantify the performance of every model in every case, we therefore 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 comparing 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-Ne' 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 51).

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 (Breiman *et al.*, 1984). 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 (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. To 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 `NLMR` package (Sciaini *et al.*, 2018) to simulate landscapes, `graph4lg` package (Savary *et al.*, 2021b) 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. The variation of the Gini indices, ranging from 0.170 to 0.290 (median : 0.232), outlines the simulation of contrasted population size distributions in the 'Area' settings. 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 coefficients of correlation 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. Larger migration rates made long distance dispersal events more frequent, such that after 50 generations, the number of different dispersal paths followed by individuals averaged 84 (± 15), 145 (± 16), 240 (± 26) and 423 (± 68) with migration rates of 0.0005, 0.001, 0.002 and 0.005 respectively, although there were important variations among landscapes. This led to contrasted influences of genetic drift relative to gene flow in these situations and contrasted influences of large distance dispersal events on genetic differentiation.

3.2 Gravity models

R^2_β values obtained for a given simulation and model with different CD values exhibited large variations meaning that the models were able to distinguish cost scenarios among them (Table 18). 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). Although the median R^2_β values 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 18). This means that in every case, the model with the best goodness of fit was computed with CD values not deriving 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.

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 19). 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 18 and 19).

When population sizes were heterogeneous and depended on patch area, especially when the migration rate was low (0.0005 or 0.001), r_{Sp} values were more variable (Table 19). They took slightly lower values than when using similar migration rates and models ('Cost', 'Cost-Area') but with equal population sizes, meaning that SHNe 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-Ne' models in the 'Area' case were larger with the lowest migration rates (Table 19). In particular, although maximum r_{Sp} values obtained using either the 'Cost' model and the 'Cost-Area' or 'Cost-Ne' models were relatively similar, their respective median and mean values were more different ; those obtained with the latter models being larger than those obtained with the former (Table 19). This means that although in some landscapes, including intra-population variables provided a very slight advantage, there were landscapes in which it improved the quality of the inference more significantly. In the next section we therefore focus on the results obtained with

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-Ne	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-Ne	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-Ne	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-Ne	0.554	0.373	0.733	0.737

TABLE 18 – 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 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. Reported values were averaged over the different landscapes, simulation runs and cost scenarios (ALTER case only).

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-Ne') averaged 0.050 overall when the migration rate was either 0.0005 or 0.001 (Figure 52) and ranged from -0.330 to 0.590 (see figures S56 and S57 for similar results when considering all the migration 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-Ne' models (Table 19). This regression tree explained 86 % of the variations of D and was made of six leaves corresponding to different regions of the predictor space (Figures 52 and 53). 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 53).

According to the splitting rules of the regression tree (Figure 52), 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

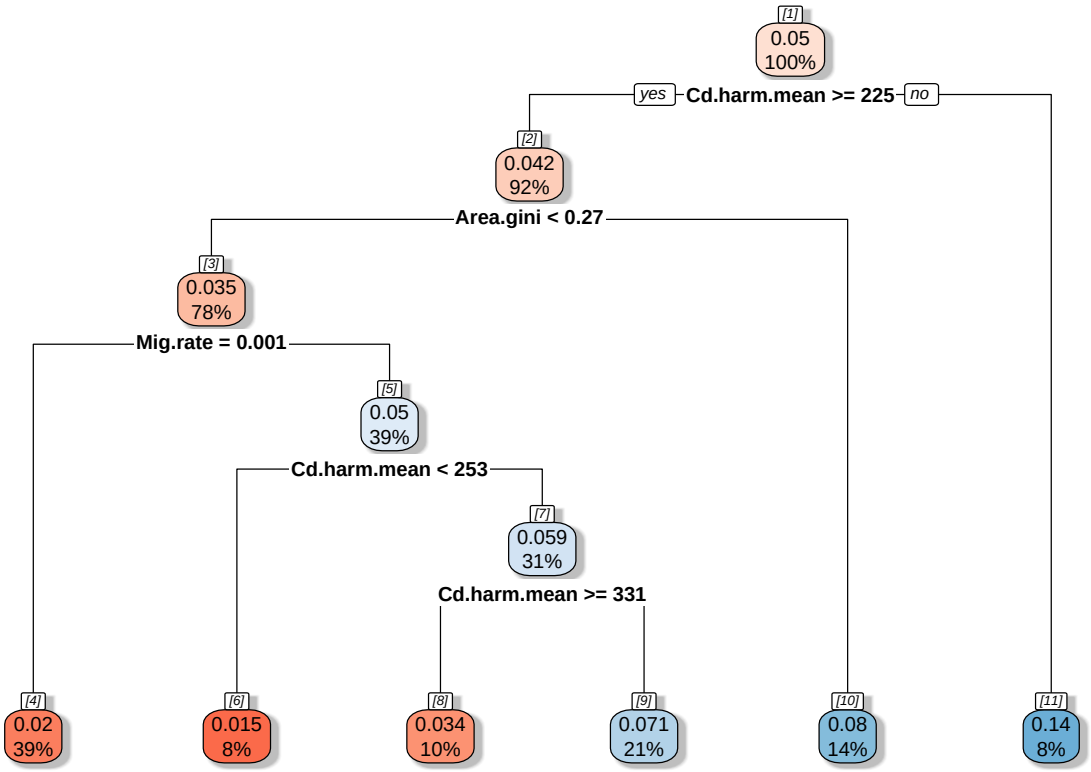


FIGURE 52 – Regression tree obtained when considering five predictors (Model, Mig.rate, CD.harm.mean, Area.gini) to explain the response variable D. Only the cases corresponding to the scenario in which population size and habitat patch areas are rank-correlated, and migration rates are equal to 0.0005 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.

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-Ne	-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-Ne	-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-Ne	-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-Ne	0.666	0.960	0.944	0.988

TABLE 19 – 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. Large values indicate that the models are able to identify the cost scenarios most similar to the 'true' one.

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 52), 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 52). Besides, when this harmonic mean was larger 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 S56 and S57).

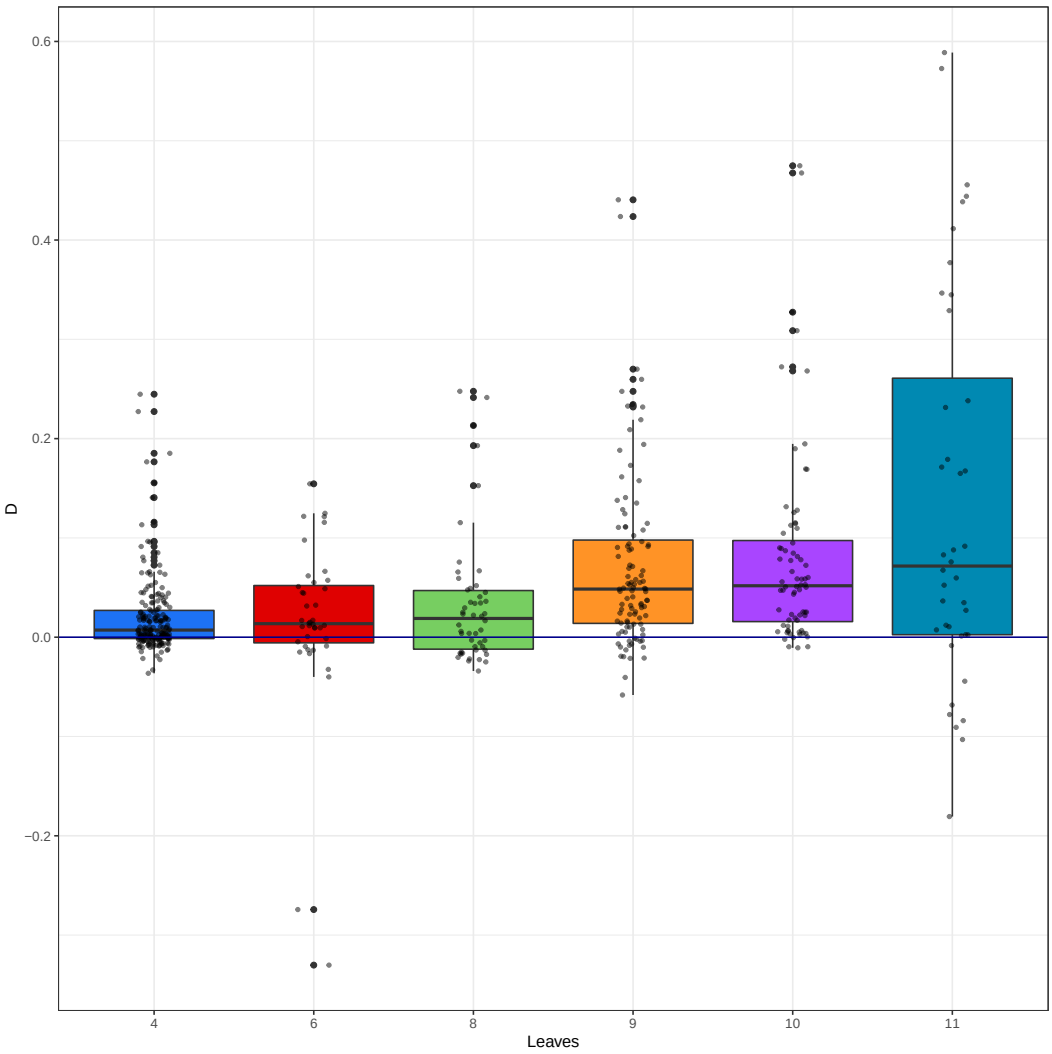


FIGURE 53 – Distribution of D in each leaf of the regression tree displayed on figure 52(refer to this figure for the leaf numbers). Only the cases corresponding to the scenario in which population size and habitat patch areas are rank-correlated, and migration rates are equal to 0.0005 or 0.001, are considered.

4 Discussion

4.1 Is cost value inference from genetic data reliable ?

Overall, the models performed well in inferring cost values, which confirms the interest of genetic data in this type of analysis, as suggested by Beier *et al.* (2008) and empirically validated by Zeller *et al.* (2018). Models with the lowest goodness of fit and capacity to identify the most realistic cost scenarios were obtained when migration rates were the lowest. This may stem from the stronger influence of drift relative to gene flow on genetic differentiation in these cases (Hutchison et 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. This means that even when signal to noise ratios are low, inferring landscape resistance to gene flow is possible with genetic data.

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. These scenarios were different from the 'true' scenario in both their absolute cost values and the relative ranking of these values. Such an erroneous output could lead to wrong conservation measures if used for the spatial modelling of potential dispersal paths. Indeed, closely correlated cost-distance values can correspond to least-cost paths that are spatially distinct (Savary et al., in prep.). Retaining the set of 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 this study, we ensured that there was some variability among the cost scenarios when simulating both these cost scenarios and the landscapes. Yet, in many landscapes and/or study designs, these conditions are likely not to be met which could compromise the reliability of the inference, even in situations here identified as being optimal. Similarly, Cushman *et al.* (2013b) 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 useful to identify landscape properties responsible for the similarity of cost scenarios, prior to cost value inference.

4.2 Does SHNe influence cost value inference ?

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 (5×10^{-4}, 10×10^{-4}). This inference relies upon the assumption that genetic differentiation reflects landscape influence on gene flow (Savary *et al.*, 2021a). It is therefore 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 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.

4.3 Can we improve cost value inference by taking SHNe into account in gravity models ?

In order to take SHNe into account in the cost value inference, we computed gravity models in which both CD values and intra-population variables such as patch areas ('Cost-Area') and population sizes ('Cost-Ne') were predictor variables explaining pairwise genetic distances between populations. In cases where SHNe 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. Our results therefore extend to the specific context of cost value inference the recommendation of Prunier *et al.* (2017) to account for SHNe in landscape genetic analyses.

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

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) although these authors quantified overall population size heterogeneity using the coefficient of variation of these sizes.

On the other hand, in accordance with our hypothesis, the spatial aggregation of the populations tended to decrease the quality of cost value inference. We used the harmonic mean of CD values for distinguishing landscapes in which gene flow events frequently occurred at a restricted scale because populations tended to form spatial aggregates. Thus, this result could potentially stem from the fact that 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 gene flow 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, making 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.* (2013b).

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 SHNe 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.

4.5 Limits and perspectives

The migration rates for which we observed a significant influence of SHNe 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 study had different objectives and although our results show that intra-population variables help inferring cost values when gene flow is very reduced, they do not mean that SHNe is not substantially affecting genetic differentiation for larger migration rates.

Finally, in our simulations, we considered that we knew and sampled the exhaustive set of populations. 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 et 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 SHNe effects are at play.

5 Conclusion

Landscape genetic studies have soon considered matrix heterogeneity when inferring landscape resistance to gene flow. 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 are appropriate for the inclusion of these variables in the inference of cost values associated with land cover types.

A - Supplementary figures

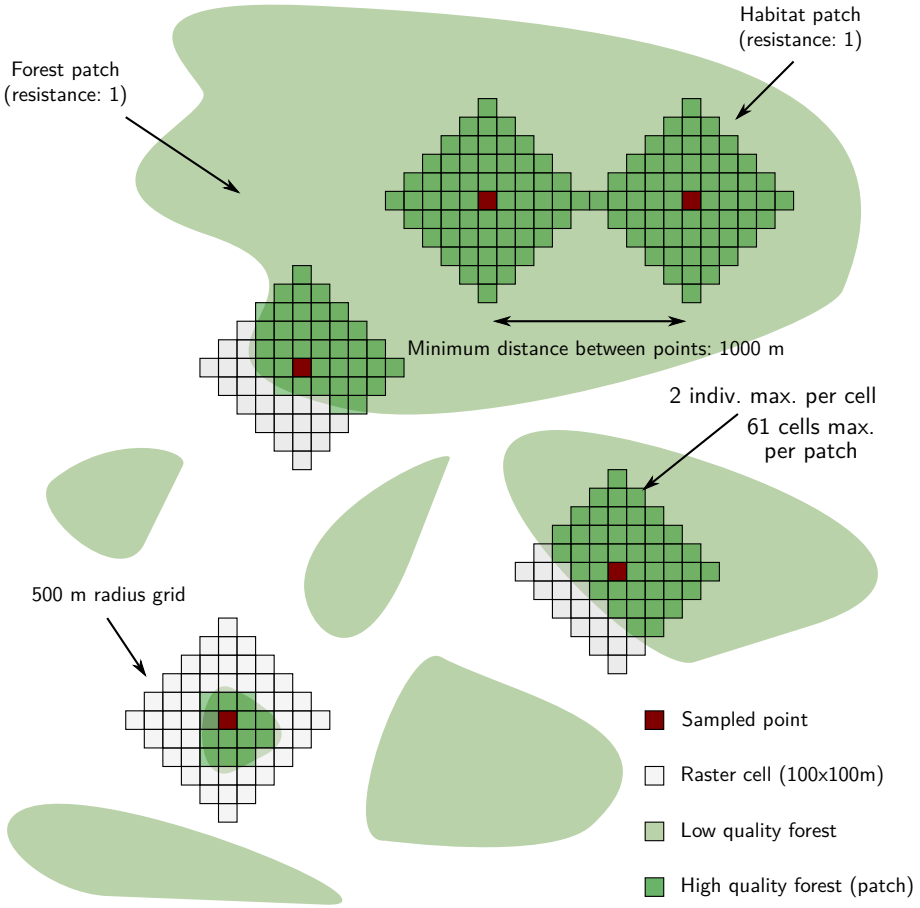


FIGURE 54 – Method used for calculating patch capacities from the simulated landscapes and the sampled points

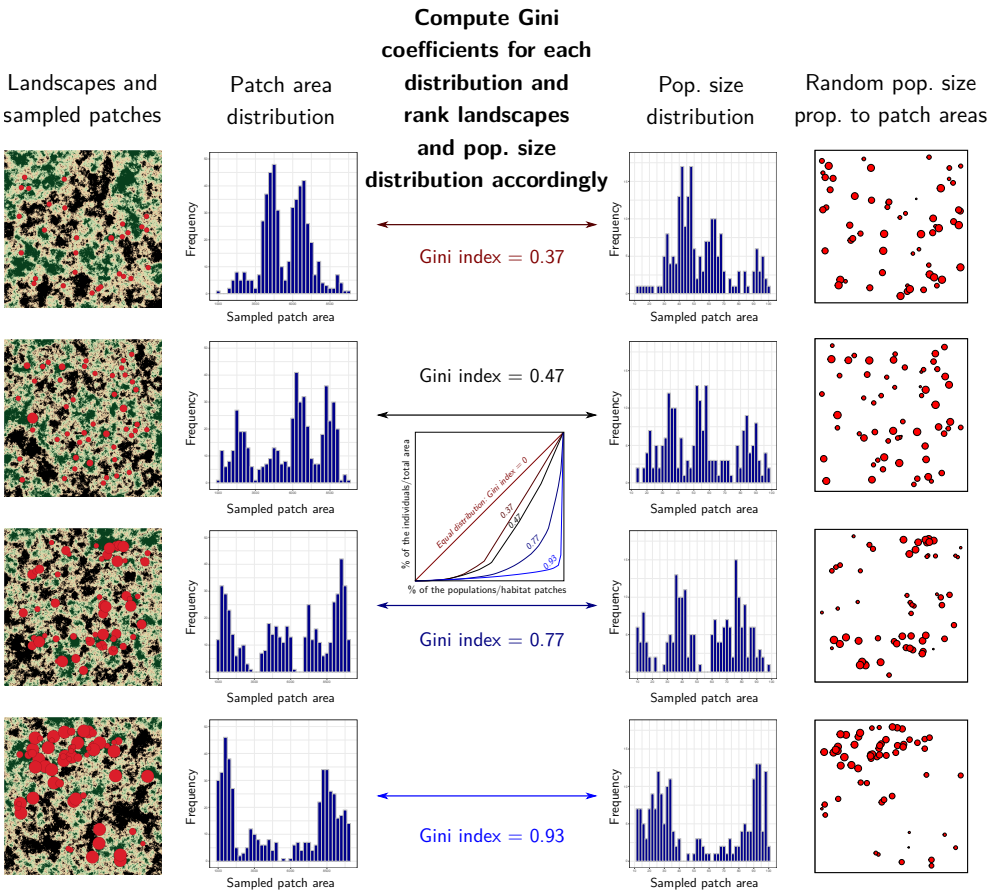


FIGURE 55 – Method used for assigning population sizes to the sampled populations according to the patch area heterogeneity

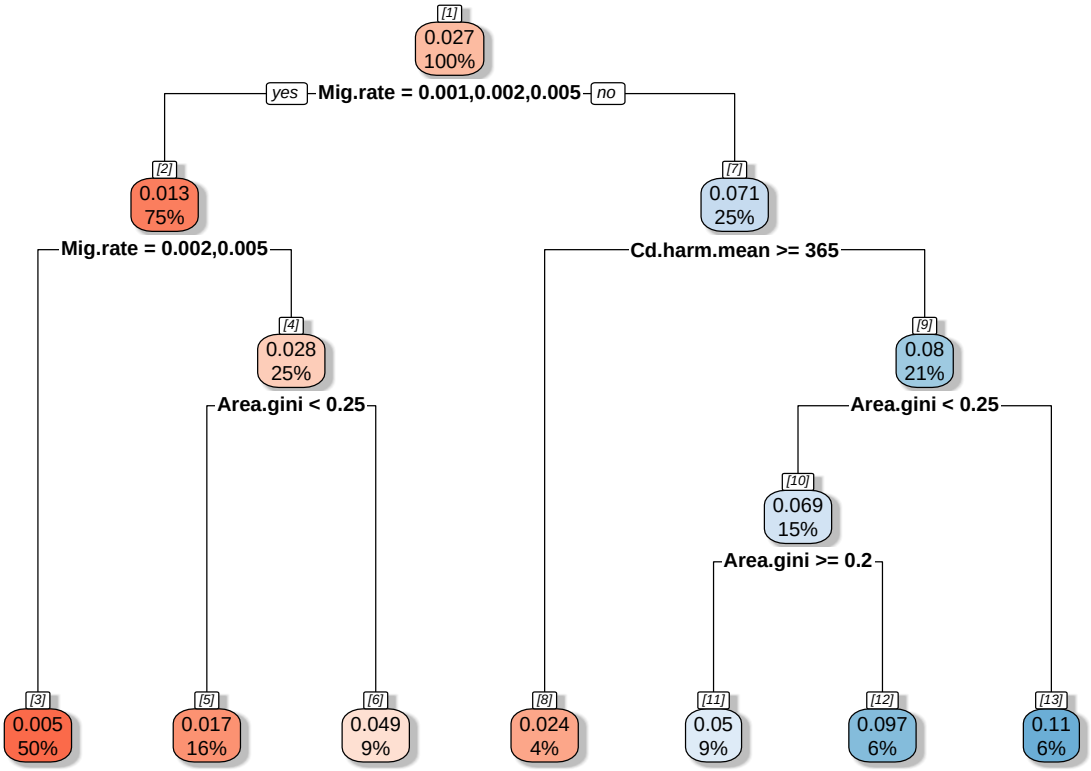


FIGURE 56 – Regression tree obtained when considering seven predictors (Pop.size, Model, Mig.rate, CD.harm.mean, Area.gini) to explain the response variable D. The tree was pruned in order to have at least 40 observations in every leaf. Only the cases corresponding to the scenario in which population size and habitat patch areas are rank-correlated are considered. The total number of observations is 1000. 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 1000 observations included in each leaf.

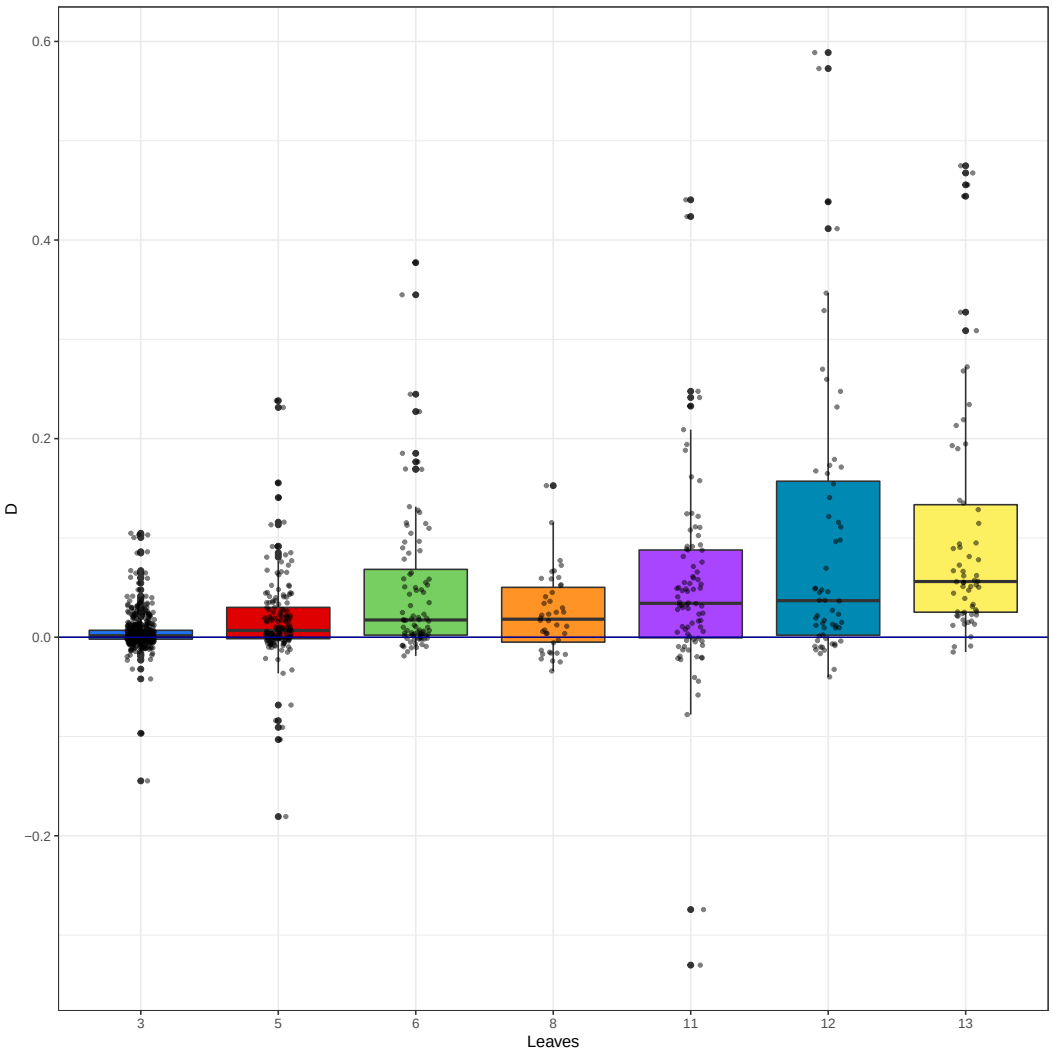


FIGURE 57 – Distribution of D in each leaf of the regression tree displayed on figure 56(refer to this figure for the leaf numbers). Only the cases corresponding to the scenario in which population size and habitat patch areas are rank-correlated are considered.

Bibliographie

Addicott, J. F., Aho, J. M., Antolin, M. F., Padilla, D. K., Richardson, J. S. et Soluk, D. A. (1987). Ecological neighborhoods : scaling environmental patterns. *Oikos*, 49(3):340–346.

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

Al-Asadi, H., Petkova, D., Stephens, M. et Novembre, J. (2019). Estimating recent migration and population-size surfaces. *PLoS Genetics*, 15(1):1–21.

Albert, E. M., Fortuna, M. A., Godoy, J. A. et Bascompte, J. (2013). Assessing the robustness of networks of spatial genetic variation. *Ecology Letters*, 16(1):86–93.

Allendorf, F. W., Luikart, G. et Aitken, S. N. (2007). Conservation and the genetics of populations. *mammalia*, 2007(2007):189–197.

Allendorf, F. W., Luikart, G. H. et Aitken, S. N. (2012). *Conservation and the genetics of populations*. Wiley Blackwell, New York, 2 édition.

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

Andersson, E. et Bodin, Ö. (2009). Practical tool for landscape planning ? an empirical investigation of network based models of habitat fragmentation. *Ecography*, 32(1):123–132.

Angelone, S. et Holderegger, R. (2009). Population genetics suggests effectiveness of habitat connectivity measures for the European tree frog in Switzerland. *Journal of Applied Ecology*, 46(4):879–887.

Angelone, S., Kienast, F. et Holderegger, R. (2011). Where movement happens - scale-dependent landscape effects on genetic differentiation in the European tree frog. *Ecography*, 34(5):714–722.

Arnaud, J.-F. (2003). Metapopulation genetic structure and migration pathways in the land snail *Helix aspersa* : influence of landscape heterogeneity. *Landscape Ecology*, 18(3):333–346.

Avon, C. et Bergès, L. (2016). Prioritization of habitat patches for landscape connectivity conservation differs between least-cost and resistance distances. *Landscape Ecology*, 31(7):1551–1565.

Awade, M., Boscolo, D. et Metzger, J. P. (2012). Using binary and probabilistic habitat availability indices derived from graph theory to model bird occurrence in fragmented forests. *Landscape Ecology*, 27(2):185–198.

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

Balbi, M., Ernoult, A., Poli, P., Madec, L., Guiller, A., Martin, M.-C., Nabucet, J., Beaujouan, V. et Petit, E. J. (2018). Functional connectivity in replicated urban landscapes in the land snail (*Cornu aspersum*). *Molecular Ecology*, 27(6):1357–1370.

Balbi, M., Petit, E. J., Croci, S., Nabucet, J., Georges, R., Madec, L. et Ernoult, A. (2019). Ecological relevance of least cost path analysis : An easy implementation method for landscape urban planning. *Journal of Environmental Management*, 244:61–68.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. et Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification : an overview. *Bioinformatics*, 16(5):412–424.

Balkenhol, N., Cushman, S., Storfer, A. et Waits, L. (2016). *Landscape genetics : concepts, methods, applications*. John Wiley & Sons.

Balkenhol, N., Gugerli, F., Cushman, S. A., Waits, L. P., Coulon, A., Arntzen, J., Holderegger, R., Wagner, H. H. *et al.* (2009a). Identifying future research needs in landscape genetics : where to from here ? *Landscape Ecology*, 24(4):455–463.

Balkenhol, N., Pardini, R., Cornelius, C., Fernandes, F. et Sommer, S. (2013). Landscape-level comparison of genetic diversity and differentiation in a small mammal inhabiting different fragmented landscapes of the Brazilian Atlantic Forest. *Conservation Genetics*, 14(2):355–367.

Balkenhol, N., Waits, L. P. et Dezzani, R. J. (2009b). Statistical approaches in landscape genetics : an evaluation of methods for linking landscape and genetic data. *Ecography*, 32(5):818–830.

Baranyi, G., Saura, S., Podani, J. et Jordán, F. (2011). Contribution of habitat patches to network connectivity : redundancy and uniqueness of topological indices. *Ecological Indicators*, 11(5):1301–1310.

Barr, K. R., Kus, B. E., Preston, K. L., Howell, S., Perkins, E. et Vandergast, A. G. (2015). Habitat fragmentation in coastal southern California disrupts genetic connectivity in the Cactus Wren (*Campylorhynchus brunneicapillus*). *Molecular Ecology*, 24(10):2349–2363.

Barton, P. S., Lentini, P. E., Alacs, E., Bau, S., Buckley, Y. M., Burns, E. L., Driscoll, D. A., Guja, L. K., Kujala, H., Lahoz-Monfort, J. J. *et al.* (2015). Guidelines for using movement science to inform biodiversity policy. *Environmental management*, 56(4):791–801.

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

Beier, P., Majka, D. R. et Newell, S. L. (2009). Uncertainty analysis of least-cost modeling for designing wildlife linkages. *Ecological Applications*, 19(8):2067–2077.

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

Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 57(1):289–300.

Bennett, A., Crooks, K. et Sanjayan, M. (2006). The future of connectivity conservation. *In* Crooks, K. et Sanjayan, M., éditeurs : *Connectivity conservation*, pages 676–694. Cambridge University Press.

Bennett, A. F. (1999). *Linkages in the landscape : the role of corridors and connectivity in wildlife conservation*. IUCN.

Bergerot, B., Tournant, P., Moussus, J.-P., Stevens, V.-M., Julliard, R., Baguette, M. et Foltête, J.-C. (2013). Coupling inter-patch movement models and landscape graph to assess functional connectivity. *Population Ecology*, 55(1):193–203.

Bergès, L., Avon, C., Bezombes, L., Clauzel, C., Duflot, R., Foltête, J.-C., Gaucherand, S., Girardet, X. et Spiegelberger, T. (2020). Environmental mitigation hierarchy and biodiversity offsets revisited through habitat connectivity modelling. *Journal of Environmental Management*, 256:1–10.

Bergsten, A. et Zetterberg, A. (2013). To model the landscape as a network : A practitioner's perspective. *Landscape and Urban Planning*, 119:35–43.

Bertin, A., Gouin, N., Baumel, A., Gianoli, E., Serratosa, J., Osorio, R. et Manel, S. (2017). Genetic variation of loci potentially under selection confounds species–genetic diversity correlations in a fragmented habitat. *Molecular Ecology*, 26(2):431–443.

Betbeder, J., Laslier, M., Hubert-Moy, L., Burel, F. et Baudry, J. (2017). Synthetic Aperture Radar (SAR) images improve habitat suitability models. *Landscape Ecology*, 32(9):1867–1879.

Blazquez-Cabrera, S., Bodin, Ö. et Saura, S. (2014). Indicators of the impacts of habitat loss on connectivity and related conservation priorities : Do they change when habitat patches are defined at different scales ? *Ecological Indicators*, 45:704–716.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. et Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics - Theory and Experiment*, 10:1–13.

Bodin, Ö. et Norberg, J. (2007). A network approach for analyzing spatially structured populations in fragmented landscape. *Landscape Ecology*, 22(1):31–44.

Bönsel, A. B. et Sonneck, A.-G. (2011). Habitat use and dispersal characteristic by *Stethophyma grossum* : the role of habitat isolation and stable habitat conditions towards low dispersal. *Journal of Insect Conservation*, 15(3):455–463.

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

Boulanger, E., Dalongeville, A., Andrello, M., Mouillot, D. et Manel, S. (2020). Spatial graphs highlight how multi-generational dispersal shapes landscape genetic patterns. *Ecography*, 15(1):1–13.

Bourdouxhe, A., Duflot, R., Radoux, J. et Dufrêne, M. (2020). Comparison of methods to model species habitat networks for decision-making in nature conservation : The case of the Wildcat in southern Belgium. *Journal for Nature Conservation*, 58:125901.

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

Bowman, J., Adey, E., Angoh, S. Y., Baici, J. E., Brown, M. G., Cordes, C., Dupuis, A. E., Newar, S. L., Scott, L. M. et Solmundson, K. (2020). Effects of cost surface uncertainty on current density estimates from circuit theory. *PeerJ*, 8:e9617.

Bowne, D. R. et Bowers, M. A. (2004). Interpatch movements in spatially structured populations : a literature review. *Landscape Ecology*, 19(1):1–20.

Bradbury, I. R. et Bentzen, P. (2007). Non-linear genetic isolation by distance : implications for dispersal estimation in anadromous and marine fish populations. *Marine Ecology Progress Series*, 340:245–257.

Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z. et Wagner, D. (2008). On modularity clustering. *IEEE transactions on knowledge and data engineering*, 20(2):172–188.

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

Brodie, J. F., Mohd-Azlan, J. et Schnell, J. K. (2016). How individual links affect network stability in a large-scale, heterogeneous metacommunity. *Ecology*, 97(7):1658–1667.

Brooks, C. (2003). A scalar analysis of landscape connectivity. *Oikos*, 102(2):433–439.

Brooks, C. P. (2006). Quantifying population substructure : extending the graph-theoretic approach. *Ecology*, 87(4):864–872.

Bruggeman, D. J., Wiegand, T. et Fernandez, N. (2010). The relative effects of habitat loss and fragmentation on population genetic variation in the red-cockaded woodpecker (*Picoides borealis*). *Molecular Ecology*, 19(17):3679–3691.

Bunn, A., Urban, D. et Keitt, T. (2000). Landscape connectivity : a conservation application of graph theory. *Journal of Environmental Management*, 59(4):265–278.

Burnham, K. P. et Anderson, D. R. (2004). Multimodel inference : understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304.

Calabrese, J. M. et Fagan, W. F. (2004). A comparison-shopper's guide to connectivity metrics. *Frontiers in Ecology and the Environment*, 2(10):529–536.

Callens, T., Galbusera, P., Matthysen, E., Durand, E. Y., Githiru, M., Huyghe, J. R. et Lens, L. (2011). Genetic signature of population fragmentation varies with mobility in seven bird species of a fragmented Kenyan cloud forest. *Molecular Ecology*, 20(9):1829–1844.

Capurucho, J. M. G., Cornelius, C., Borges, S. H., Cohn-Haft, M., Aleixo, A., Metzger, J. P. et Ribas, C. C. (2013). Combining phylogeography and landscape genetics of *Xenopipo atronitens* (Aves : Pipridae), a white sand campina specialist, to understand Pleistocene landscape evolution in Amazonia. *Biological Journal of the Linnean Society*, 110(1):60–76.

Carrascal, L. M., Galván, I. et Gordo, O. (2009). Partial Least Squares regression as an alternative to current regression methods used in ecology. *Oikos*, 118(5):681–690.

Carroll, C., McRae, B. et Brookes, A. (2012). Use of linkage mapping and centrality analysis across habitat gradients to conserve connectivity of gray wolf populations in western North America. *Conservation Biology*, 26(1):78–87.

Castel, T., Lecomte, C., Richard, Y., Lejeune-Hénaut, I. et Larmure, A. (2017). Frost stress evolution and winter pea ideotype in the context of climate warming at a regional scale. *OCL*, 24(1):D106.

Castillo, J. A., Epps, C. W., Jeffress, M. R., Ray, C., Rodhouse, T. J. et Schwalm, D. (2016). Replicated landscape genetic and network analyses reveal wide variation in functional connectivity for American pikas. *Ecological Applications*, 26(6):1660–1676.

Cavalli-Sforza, L. L. et Edwards, A. W. (1967). Phylogenetic analysis : models and estimation procedures. *Evolution*, 21(3):550–570.

Ciofi, C., Beaumont, M. A., Swingland, I. R. et Bruford, M. W. (1999). Genetic divergence and units for conservation in the Komodo dragon *Varanus komodoensis*. *Proceedings of the Royal Society B*, 266(1435):2269–2274.

Clarke, R. T., Rothery, P. et 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.

Clauset, A., Newman, M. E. et Moore, C. (2004). Finding community structure in very large networks. *Physical review E*, 70(6):1–6.

Clauzel, C., Girardet, X. et Foltête, J.-C. (2013). Impact assessment of a high-speed railway line on species distribution : Application to the European tree frog (*Hyla arborea*) in Franche-Comté. *Journal of Environmental Management*, 127:125–134.

Clauzel, C. et Godet, C. (2020). Combining spatial modeling tools and biological data for improved multispecies assessment in restoration areas. *Biological Conservation*, 250:1–15.

Clegg, S. M. et Phillimore, A. B. (2010). The influence of gene flow and drift on genetic and phenotypic divergence in two species of Zosterops in Vanuatu. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 365(1543):1077–1092.

Clevenger, A. P., Wierzchowski, J., Chruszcz, B. et Gunson, K. (2002). GIS-generated, expert-based models for identifying wildlife habitat linkages and planning mitigation passages. *Conservation Biology*, 16(2):503–514.

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

Correa Ayram, C. A., Mendoza, M. E., Etter, A. et Salicrup, D. R. P. (2016). Habitat connectivity in biodiversity conservation : a review of recent studies and applications. *Progress in Physical Geography*, 40(1):7–37.

Coster, S. S., Babbitt, K. J., Cooper, A. et Kovach, A. I. (2015). Limited influence of local and landscape factors on finescale gene flow in two pond-breeding amphibians. *Molecular Ecology*, 24(4):742–758.

Coulon, A., Cosson, J., Angibault, J., Cargnelutti, B., Galan, M., Morellet, N., Petit, E., Aulagnier, S. et Hewison, A. (2004). Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape : an individual–based approach. *Molecular Ecology*, 13(9):2841–2850.

Creech, T. G., Epps, C. W., Monello, R. J. et Wehausen, J. D. (2014). Using network theory to prioritize management in a desert bighorn sheep metapopulation. *Landscape Ecology*, 29(4):605–619.

Crispo, E., Moore, J.-S., Lee-Yaw, J. A., Gray, S. M. et Haller, B. C. (2011). Broken barriers : human-induced changes to gene flow and introgression in animals : an examination of the ways in which humans increase genetic exchange among populations and species and the consequences for biodiversity. *BioEssays*, 33(7):508–518.

Crooks, K. R. et Sanjayan, M. (2006). *Connectivity conservation*, volume 14. Cambridge University Press.

Cross, T. B., Schwartz, M. K., Naugle, D. E., Fedy, B. C., Row, J. R. et Oyler-McCance, S. J. (2018). The genetic network of greater sage-grouse : Range-wide identification of keystone hubs of connectivity. *Ecology and Evolution*, 8(11):1–19.

Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. *International Journal of Complex Systems*, 1695(5):1–9.

Curson, J. (2014). *Plumbeous warbler* (*Dendroica plumbea*). *In* del Hoyo, J., Elliott, A., Sargatal, J., Christie, D. et de Juana, E., éditeurs : *Handbook of the Birds of the World alive*. Barcelona - Lynx Edicions.

Cushman, S. A., McKelvey, K. S., Hayden, J. et 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., McRae, B., Adriaensen, F., Beier, P., Shirley, M. et Zeller, K. (2013a). Biological corridors and connectivity. *In* Macdonald, D. et Willis, K., éditeurs : *Key Topics in Conservation Biology*, volume 2, chapitre 21, pages 384–404. Wiley Online Library.

Cushman, S. A., Shirk, A., Howe, G. T., Dyer, R. J., Murphy, M. A. et Joost, S. (2018). The least cost path from landscape genetics to landscape genomics : challenges and opportunities to explore NGS data in a spatially explicit context. *Frontiers in Genetics*, 9:215.

Cushman, S. A., Shirk, A. et Landguth, E. L. (2012). Separating the effects of habitat area, fragmentation and matrix resistance on genetic differentiation in complex landscapes. *Landscape Ecology*, 27(3):369–380.

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

Dale, M. et Fortin, M.-J. (2010). From graphs to spatial graphs. *Annual Review of Ecology, Evolution, and Systematics*, 41:21–38.

Dale, M. R. (2017). *Applying Graph Theory in Ecological Research*. Cambridge University Press.

Danon, L., Diaz-Guilera, A., Duch, J. et Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics : Theory and Experiment*, 2005(09):P09008.

de la Torre, J. A., Lechner, A. M., Wong, E. P., Magintan, D., Saaban, S. et Campos-Arceiz, A. (2019). Using elephant movements to assess landscape connectivity under peninsular malaysia's central forest spine land use policy. *Conservation Science and Practice*, 1(12):e133.

Den Boer, P. J. (1968). Spreading of risk and stabilization of animal numbers. *Acta biotheoretica*, 18(1-4):165–194.

Díaz, S. M., Settele, J., Brondízio, E., Ngo, H., Guèze, M., Agard, J., Arneth, A., Balvanera, P., Brauman, K., Butchart, S. *et al.* (2019). The global assessment report on biodiversity and ecosystem services : Summary for policy makers. Rapport technique, Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services.

Didham, R. K., Kapos, V. et Ewers, R. M. (2012). Rethinking the conceptual foundations of habitat fragmentation research. *Oikos*, 121(2):161–170.

Dieringer, D. et Schlötterer, C. (2003). Microsatellite analyser (MSA) : a platform independent analysis tool for large microsatellite data sets. *Molecular Ecology Notes*, 3(1):167–169.

DiLeo, M. F., Siu, J. C., Rhodes, M. K., López-Villalobos, A., Redwine, A., Ksiazek, K. et 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.

DiLeo, M. F. et Wagner, H. H. (2016). A landscape ecologist's agenda for landscape genetics. *Current Landscape Ecology Reports*, 1(3):115–126.

Diniz, M. F., Cushman, S. A., Machado, R. B. et Júnior, P. D. M. (2020). Landscape connectivity modeling from the perspective of animal dispersal. *Landscape Ecology*, 35(1):41–58.

Dondina, O., Saura, S., Bani, L. et Mateo-Sánchez, M. C. (2018). Enhancing connectivity in agroecosystems : focus on the best existing corridors or on new pathways ? *Landscape Ecology*, 33(10):1741–1756.

Draheim, H. M., Moore, J. A., Etter, D., Winterstein, S. R. et Scribner, K. T. (2016). Detecting black bear source–sink dynamics using individual-based genetic graphs. *Proceedings of the Royal Society B*, 283(1835):1–9.

Drake, J., Lambin, X. et Sutherland, C. (2021). The value of considering demographic contributions to connectivity : a review. *Ecography*.

Driezen, K., Adriaensen, F., Rondinini, C., Doncaster, C. P. et 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(2-4):314–322.

Duflot, R., Avon, C., Roche, P. et Bergès, L. (2018). Combining habitat suitability models and spatial graphs for more effective landscape conservation planning : An applied methodological framework and a species case study. *Journal for Nature Conservation*, 46:38–47.

Dyer, R. J. (2014). R package gstudio : analyses and functions related to the spatial analysis of genetic marker data. *R package version*, 1.

Dyer, R. J. (2015a). Is there such a thing as landscape genetics ? *Molecular Ecology*, 24(14):3518–3528.

Dyer, R. J. (2015b). Population graphs and landscape genetics. *Annual Review of Ecology, Evolution, and Systematics*, 46:327–342.

Dyer, R. J. et Nason, J. D. (2004). Population graphs : the graph theoretic shape of genetic structure. *Molecular Ecology*, 13(7):1713–1727.

Dyer, R. J., Nason, J. D. et Garrick, R. C. (2010). Landscape modelling of gene flow : improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology*, 19(17):3746–3759.

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

Emaresi, G., Pellet, J., Dubey, S., Hirzel, A. H. et Fumagalli, L. (2011). Landscape genetics of the Alpine newt (*Mesotriton alpestris*) inferred from a strip-based approach. *Conservation Genetics*, 12(1):41–50.

Éraud, C., Arnoux, É., Levesque, A., Van Laere, G. et Magnin, H. (2012). Biologie des populations et statut de conservation des oiseaux endémiques des Antilles en Guadeloupe. Rapport d'étude, ONCFS-Parc National Guadeloupe.

Éraud, C., Magnin, H., Redaud, L., Tartaglino, O. et Levesque, A. (2009). Oiseaux endémiques des Petites Antilles - enjeux et orientations de recherche en Guadeloupe. *Faune sauvage*, 284:13–16.

Estoup, A., Jarne, P. et 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. et 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.

Estrada-Peña, A. (2005). Effects of habitat suitability and landscape patterns on tick (Acarina) metapopulation processes. *Landscape Ecology*, 20(5):529–541.

Etherington, T. R. (2012). Least-cost modelling on irregular landscape graphs. *Landscape Ecology*, 27(7):957–968.

Etherington, T. R. et Holland, E. P. (2013). Least-cost path length versus accumulated-cost as connectivity measures. *Landscape Ecology*, 28(7):1223–1229.

Etherington, T. R. et Perry, G. L. (2016). Visualising continuous intra-landscape isolation with uncertainty using least-cost modelling based catchment areas : common brushtail possums in the Auckland isthmus. *International Journal of Geographical Information Science*, 30(1):36–50.

Everitt, B. et Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. Springer Science & Business Media.

Excoffier, L., Smouse, P. E. et Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes : application to human mitochondrial DNA restriction data. *Genetics*, 131(2):479–491.

Ezard, T. et Travis, J. M. J. (2006). The impact of habitat loss and fragmentation on genetic drift and fixation time. *Oikos*, 114(2):367–375.

Fagan, W. et Calabrese, J. (2006). Quantifying connectivity : balancing metric performance with data requirements. *In* Crooks, K. et Sanjayan, M., éditeurs : *Connectivity conservation*, pages 297–317. Cambridge University Press.

Fahrig, L. (2003). Effects of habitat fragmentation on biodiversity. *Annual Review of Ecology, Evolution, and Systematics*, 34(1):487–515.

Fahrig, L. (2013). Rethinking patch size and isolation effects : the habitat amount hypothesis. *Journal of Biogeography*, 40(9):1649–1663.

Fahrig, L. (2017). Ecological responses to habitat fragmentation per se. *Annual Review of Ecology, Evolution, and Systematics*, 48(1):1–23.

Fall, A., Fortin, M.-J., Manseau, M. et O'Brien, D. (2007). Spatial graphs : principles and applications for habitat connectivity. *Ecosystems*, 10(3):448–461.

Farine, D. R. et Whitehead, H. (2015). Constructing, conducting and interpreting animal social network analysis. *Journal of Animal Ecology*, 84(5):1144–1163.

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

Flavenot, T., Fellous, S., Abdelkrim, J., Baguette, M. et Coulon, A. (2015). Impact of quarrying on genetic diversity : an approach across landscapes and over time. *Conservation Genetics*, 16(1):181–194.

Fletcher, R. et Fortin, M.-J. (2018). *Spatial ecology and conservation modeling*. Springer.

Fletcher, R. J., Acevedo, M. A., Reichert, B. E., Pias, K. E. et Kitchens, W. M. (2011). Social network models predict movement and connectivity in ecological landscapes. *Proceedings of the National Academy of Sciences*, 108(48): 19282–19287.

Fletcher, R. J., Burrell, N. S., Reichert, B. E., Vasudev, D. et Austin, J. D. (2016). Divergent perspectives on landscape connectivity reveal consistent effects from genes to communities. *Current Landscape Ecology Reports*, 1(2):67–79.

Fletcher, R. J., Revell, A., Reichert, B. E., Kitchens, W. M., Dixon, J. D. et Austin, J. D. (2013). Network modularity reveals critical scales for connectivity in ecology and evolution. *Nature Communications*, 4(1):1–7.

Fletcher Jr, R. J., Didham, R. K., Banks-Leite, C., Barlow, J., Ewers, R. M., Rosindell, J., Holt, R. D., Gonzalez, A., Pardini, R., Damschen, E. I. *et al.* (2018). Is habitat fragmentation good for biodiversity ? *Biological Conservation*, 226:9–15.

Foll, M. et Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers : a Bayesian perspective. *Genetics*, 180(2):977–993.

Foltête, J.-C., Clauzel, C. et Vuidel, G. (2012a). A software tool dedicated to the modelling of landscape networks. *Environmental Modelling & Software*, 38:316–327.

Foltête, J.-C., Clauzel, C., Vuidel, G. et Tournant, P. (2012b). Integrating graph-based connectivity metrics into species distribution models. *Landscape Ecology*, 27(4):557–569.

Foltête, J.-C., Girardet, X. et Clauzel, C. (2014). A methodological framework for the use of landscape graphs in land-use planning. *Landscape and Urban Planning*, 124:140–150.

Foltête, J.-C. et Giraudoux, P. (2012). A graph-based approach to investigating the influence of the landscape on population spread processes. *Ecological Indicators*, 18:684–692.

Foltête, J.-C., Savary, P., Clauzel, C., Bourgeois, M., Girardet, X., Sahraoui, Y., Vuidel, G. et Garnier, S. (2020). Coupling landscape graph modeling and biological data : a review. *Landscape Ecology*, 35(5):1035–1052.

Foltête, J.-C. et Vuidel, G. (2017). Using landscape graphs to delineate ecologically functional areas. *Landscape Ecology*, 32(2):249–263.

Ford, A. T., Sunter, E. J., Fauvelle, C., Bradshaw, J. L., Ford, B., Hutchen, J., Phillipow, N. et Teichman, K. J. (2020). Effective corridor width : linking the spatial ecology of wildlife with land use policy. *European Journal of Wildlife Research*, 66(4):1–10.

Fortin, M.-J., James, P. M., MacKenzie, A., Melles, S. J. et Rayfield, B. (2012). Spatial statistics - spatial regression, and graph theory in ecology. *Spatial Statistics*, 1:100–109.

Fortuna, M. A., Albaladejo, R. G., Fernández, L., Aparicio, A. et Bascompte, J. (2009). Networks of spatial genetic variation across species. *Proceedings of the National Academy of Sciences*, 106(45):19044–19049.

Fotheringham, A. S. et 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. (2015). Genetic rescue of small inbred populations : Meta-analysis reveals large and consistent benefits of gene flow. *Molecular Ecology*, 24(11):2610–2618.

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

Fruchterman, T. M. et Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software : Practice and experience*, 21(11):1129–1164.

Gaggiotti, O. E. et Foll, M. (2010). Quantifying population structure using the F-model. *Molecular Ecology Resources*, 10(5):821–830.

Galpern, P., Manseau, M. et Fall, A. (2011). Patch-based graphs of landscape connectivity : a guide to construction, analysis and application for conservation. *Biological Conservation*, 144(1):44–55.

Galpern, P., Manseau, M. et Wilson, P. (2012). Grains of connectivity : analysis at multiple spatial scales in landscape genetics. *Molecular Ecology*, 21(16):3996–4009.

Garroway, C. J., Bowman, J., Carr, D. et Wilson, P. J. (2008). Applications of graph theory to landscape genetics. *Evolutionary Applications*, 1(4):620–630.

Garroway, C. J., Bowman, J. et Wilson, P. J. (2011). Using a genetic network to parameterize a landscape resistance surface for fishers, *Martes pennanti*. *Molecular Ecology*, 20(19):3978–3988.

Gil-Tena, A., Nabucet, J., Mony, C., Abadie, J., Saura, S., Butet, A., Burel, F. et Ernoult, A. (2014). Woodland bird response to landscape connectivity in an agriculture-dominated landscape : a functional community approach. *Community Ecology*, 15(2):256–268.

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

Gippoliti, S. et Battisti, C. (2017). More cool than tool : Equivoques, conceptual traps and weaknesses of ecological networks in environmental planning and conservation. *Land Use Policy*, 68:686–691.

Godet, C. et Clauzel, C. (2021). Comparison of landscape graph modelling methods for analysing pond network connectivity. *Landscape Ecology*, 36(3):735–748.

Gonzales, E. K. et Gergel, S. E. (2007). Testing assumptions of cost surface analysis - a tool for invasive species management. *Landscape Ecology*, 22(8):1155–1168.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338.

Graves, T. A., Beier, P. et 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. et 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.

Greenbaum, G. et Fefferman, N. H. (2017). Application of network methods for understanding evolutionary dynamics in discrete habitats. *Molecular Ecology*, 26(11):2850–2863.

Greenbaum, G., Templeton, A. R. et Bar-David, S. (2016). Inference and analysis of population structure using genetic data and network theory. *Genetics*, 202(4):1299–1312.

Griffioen, R. (1996). Over het dispersievermogen van de moerassprinkhaan. *Nieuwsbrief Saltabel*, 15(1):39–41.

Guillera-Arroita, G., Lahoz-Monfort, J. J. et Elith, J. (2014). Maxent is not a presence–absence method : a comment on Thibaud et al. *Methods in Ecology and Evolution*, 5(11):1192–1197.

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

Gurrutxaga, M., Lozano, P. J. et 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.

Haddad, N. M., Gonzalez, A., Brudvig, L. A., Burt, M. A., Levey, D. J. et Damschen, E. I. (2017). Experimental evidence does not support the Habitat Amount Hypothesis. *Ecography*, 40(1):48–55.

Hahn, T., Kettle, C. J., Ghazoul, J., Hennig, E. I. et Pluess, A. R. (2013). Landscape composition has limited impact on local genetic structure in mountain clover, *Trifolium montanum* L. *Journal of Heredity*, 104(6):842–852.

Hall, L. A. et Beissinger, S. R. (2014). A practical toolbox for design and analysis of landscape genetics studies. *Landscape Ecology*, 29(9):1487–1504.

Hänfling, B. et Weetman, D. (2006). Concordant genetic estimators of migration reveal anthropogenically enhanced source-sink population structure in the river sculpin *Cottus gobio*. *Genetics*, 173(3):1487–1501.

Hanski, I. (1998). Metapopulation dynamics. *Nature*, 396(6706):41–49.

Hanski, I. (2015). Habitat fragmentation and species richness. *Journal of Biogeography*, 42(5):989–993.

Hanski, I., Alho, J. et Moilanen, A. (2000). Estimating the parameters of survival and migration of individuals in metapopulations. *Ecology*, 81(1):239–251.

Hanski, I., Moilanen, A. et Gyllenberg, M. (1996). Minimum viable metapopulation size. *The American Naturalist*, 147(4):527–541.

Hardy, O. J. et Vekemans, X. (1999). Isolation by distance in a continuous population : reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity*, 83(2):145.

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

Hedrick, P. (2011). *Genetics of populations*. Jones & Bartlett Learning.

Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution*, 59(8):1633–1638.

Hernandez, P. A., Graham, C. H., Master, L. L. et Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, 29(5):773–785.

Hilty, J., Worboys, G. L., Keeley, A., Woodley, S., Lausche, B., Locke, H., Carr, M., Pulsford, I., Pittock, J., White, J. W. *et al.* (2020). Guidelines for conserving connectivity through ecological networks and corridors. Best practice protected area guidelines series, The World Conservation Union (IUCN).

Hoban, S., Bruford, M., Jackson, J. D., Lopes-Fernandes, M., Heuertz, M., Hohenlohe, P. A., Paz-Vinas, I., Sjögren-Gulve, P., Segelbacher, G., Vernesi, C. *et al.* (2020). Genetic diversity targets and indicators in the CBD post-2020 Global Biodiversity Framework must be improved. *Biological Conservation*, 248:108654.

Holderegger, R. et Gugerli, F. (2012). Where do you come from, where do you go ? directional migration rates in landscape genetics. *Molecular Ecology*, 21(23):5640–5642.

Holderegger, R., Kamm, U. et Gugerli, F. (2006). Adaptive vs. neutral genetic diversity : implications for landscape genetics. *Landscape Ecology*, 21(6):797–807.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 6(2):65–70.

Holzhauer, S. I., Ekschmitt, K., Sander, A.-C., Dauber, J. et Wolters, V. (2006). Effect of historic landscape change on the genetic structure of the bush-cricket *Metrioptera roeseli*. *Landscape Ecology*, 21(6):891–899.

Hoover, B., Yaw, S. et Middleton, R. (2020). CostMAP : an open-source software package for developing cost surfaces using a multi-scale search kernel. *International Journal of Geographical Information Science*, 34(3):520–538.

Horskins, K., Mather, P. B. et Wilson, J. C. (2006). Corridors and connectivity : when use and function do not equate. *Landscape Ecology*, 21(5):641–655.

Hubert, L. et Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hutchison, D. W. et 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.

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D. et Rodes, I. (2017). Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1):95.

Ingvarsson, P. K. (2001). Restoration of genetic variation lost–the genetic rescue hypothesis. *Trends in Ecology & Evolution*, 16(2):62–63.

Jackson, H. B. et Fahrig, L. (2012). What size is a biologically relevant landscape ? *Landscape Ecology*, 27(7):929–941.

Jackson, N. D. et Fahrig, L. (2015). Habitat amount - not habitat configuration - best predicts population genetic structure in fragmented landscapes. *Landscape Ecology*, 31(5):951–968.

Jacobs, J. (1974). Quantitative measurement of food selection. *Oecologia*, 14(4):413–417.

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

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

Jaquiéry, J., Broquet, T., Hirzel, A. H., Yearsley, J. et Perrin, N. (2011). Inferring landscape effects on dispersal from genetic distances : how far can we go ? *Molecular Ecology*, 20(4):692–705.

Jeltsch, F., Bonte, D., Pe'er, G., Reineking, B., Leimgruber, P., Balkenhol, N., Schröder, B., Buchmann, C. M., Mueller, T., Blaum, N. *et al.* (2013). Integrating movement ecology with biodiversity research-exploring new avenues to address spatiotemporal biodiversity dynamics. *Movement Ecology*, 1(1):1–13.

Joly, C. A., Metzger, J. P. et Tabarelli, M. (2014). Experiences from the brazilian atlantic forest : ecological findings and conservation initiatives. *New Phytologist*, 204(3):459–473.

Joly, D., Bois, B. et Zakšek, K. (2012). Rank-ordering of topographic variables correlated with temperature. *Atmospheric and Climate Sciences*, 2(2):139–147.

Jombart, T. (2008). adegenet : a R package for the multivariate analysis of genetic markers. *Bioinformatics*, 24(11):1403–1405.

Jordán, F., Magura, T., Tóthmérész, B., Vasas, V. et Ködöböcz, V. (2007). Carabids (Coleoptera : Carabidae) in a forest patchwork : a connectivity analysis of the Bereg Plain landscape graph. *Landscape Ecology*, 22(10):1527–1539.

Jost, L. (2008). GST and its relatives do not measure differentiation. *Molecular Ecology*, 17(18):4015–4026.

Kadoya, T. (2009). Assessing functional connectivity using empirical data. *Population ecology*, 51(1):5–15.

Kalinowski, S. T. (2004). Counting alleles with rarefaction : private alleles and hierarchical sampling designs. *Conservation Genetics*, 5(4):539–543.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C. *et al.* (2012). Geneious basic : an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28(12):1647–1649.

Keeley, A. T., Beier, P. et Gagnon, J. W. (2016). Estimating landscape resistance from habitat suitability : effects of data source and nonlinearities. *Landscape Ecology*, 31(9):2151–2162.

Keeley, A. T., Beier, P., Keeley, B. W. et Fagan, M. E. (2017). Habitat suitability is a poor proxy for landscape connectivity during dispersal and mating movements. *Landscape and Urban Planning*, 161:90–102.

Keitt, T., Urban, D. et Milne, B. (1997). Detecting critical scales in fragmented landscapes. *Conservation Ecology*, 1(1):1–17.

Keller, D., Holderegger, R. et Strien, M. J. (2013). Spatial scale affects landscape genetic analysis of a wetland grasshopper. *Molecular Ecology*, 22(9):2467–2482.

Keller, D., Holderegger, R., van Strien, M. J. et Bolliger, J. (2015). How to make landscape genetics beneficial for conservation management ? *Conservation Genetics*, 16(3):503–512.

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

Keyghobadi, N., Roland, J., Matter, S. F. et Strobeck, C. (2005). Among- and within-patch components of genetic diversity respond at different rates to habitat fragmentation : an empirical demonstration. *Proceedings of the Royal Society B*, 272(1562):553–560.

Khimoun, A., Arnoux, E., Martel, G., Pot, A., Eraud, C., Condé, B., Loubon, M., Théron, F., Covas, R., Faivre, B. *et al.* (2016a). Contrasted patterns of genetic differentiation across eight bird species in the Lesser Antilles. *Genetica*, 144(1):125–138.

Khimoun, A., Eraud, C., Ollivier, A., Arnoux, E., Rocheteau, V., Bely, M., Lefol, E., Delpuech, M., Carpentier, M.-L., Leblond, G. *et al.* (2016b). Habitat specialization predicts genetic response to fragmentation in tropical birds. *Molecular Ecology*, 25(16):3831–3844.

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

Kierepka, E. M., Anderson, S. J., Swihart, R. K. et Rhodes, O. E. (2020). Differing, multiscale landscape effects on genetic diversity and differentiation in eastern chipmunks. *Heredity*, 124(3):457–468.

Kimura, M. et Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, 49(4):561–576.

Kininmonth, S., van Oppen, M. J. et Possingham, H. P. (2010). Determining the community structure of the coral *Seriatopora hystrix* from hydrodynamic and genetic networks. *Ecological Modelling*, 221(24):2870–2880.

Koen, E. L., Bowman, J., Garroway, C. J. et Wilson, P. J. (2013). The sensitivity of genetic connectivity measures to unsampled and under-sampled sites. *PLoS ONE*, 8(2):e56204.

Koen, E. L., Bowman, J. et Walpole, A. A. (2012). The effect of cost surface parameterization on landscape resistance estimates. *Molecular Ecology Resources*, 12(4):686–696.

Koen, E. L., Bowman, J. et Wilson, P. J. (2016). Node-based measures of connectivity in genetic networks. *Molecular Ecology Resources*, 16(1):69–79.

Koenig, W. D., Van Vuren, D. et Hooge, P. N. (1996). Detectability, philopatry, and the distribution of dispersal distances in vertebrates. *Trends in Ecology & Evolution*, 11(12):514–517.

Koh, I., Rowe, H. I. et Holland, J. D. (2013). Graph and circuit theory connectivity models of conservation biological control agents. *Ecological Applications*, 23(7):1554–1573.

Kong, F., Yin, H., Nakagoshi, N. et 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.

Kool, J. T., Moilanen, A. et Treml, E. A. (2013). Population connectivity : recent advances and new perspectives. *Landscape Ecology*, 28(2):165–185.

Koschuh, A. (2004). Verbreitung, lebensräume und gefährdung der sumpfschrecke (stethophyma grossum, l., 1758)(saltatoria) in der steiermark. *Joannea, Zool*, 6:223–246.

Krause, S. (1996). Populationsstruktur, Habitatbindung und Mobilität der Larven von *Stethophyma grossum* (Linné, 1758). *Articulata*, 11(2):77–89.

Krzanowski, W. et Marriott, F. (1995). *Multivariate Analysis vol. 2 : Classification, Covariance Structures, and Repeated Measurements*. London : Arnold.

Kuehn, R., Schroeder, W., Pirchner, F. et Rottmann, O. (2003). Genetic diversity - gene flow and drift in Bavarian red deer populations (*Cervus elaphus*). *Conservation Genetics*, 4(2):157–166.

Kuismin, M., Saatoglu, D., Niskanen, A. K., Jensen, H. et Sillanpää, M. J. (2020). Genetic assignment of individuals to source populations using network estimation tools. *Methods in Ecology and Evolution*, 11(2):333–344.

Kuismin, M. O., Ahlinder, J. et Sillanpää, M. J. (2017). CONE : community oriented network estimation is a versatile framework for inferring population structure in large-scale sequencing data. *G3 : Genes, Genomes, Genetics*, 7(10):3359–3377.

Laita, A., Kotiaho, J. et Mönkkönen, M. (2011). Graph-theoretic connectivity measures : what do they tell us about connectivity ? *Landscape Ecology*, 26(7):951–967.

Laliberté, J. et St-Laurent, M.-H. (2020). Validation of functional connectivity modeling : The Achilles' heel of landscape connectivity mapping. *Landscape and Urban Planning*, 202:103878.

Landguth, E., Cushman, S., Schwartz, M., McKelvey, K., Murphy, M. et Luikart, G. (2010). Quantifying the lag time to detect barriers in landscape genetics. *Molecular Ecology*, 19(19):4179–4191.

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

Laroche, F., Balbi, M., Grébert, T., Jabot, F. et Archaux, F. (2020). Three points of consideration before testing the effect of patch connectivity on local species richness : patch delineation, scaling and variability of metrics. *bioRxiv, 640995, ver. 5 peer-reviewed and recommended by PCI Ecology*, 1:1–21.

Latta, R. G. (2006). Integrating patterns across multiple genetic markers to infer spatial processes. *Landscape Ecology*, 21(6):809–820.

Leblond, G. (2008). Étude sur les structures de peuplement de l'avifaune du massif forestier du Parc national de Guadeloupe. Rapport technique, Parc National de Guadeloupe.

Lechner, A. M. et Rhodes, J. R. (2016). Recent progress on spatial and thematic resolution in landscape ecology. *Current Landscape Ecology Reports*, 1(2):98–105.

Legendre, P. et Fortin, M.-J. (2010). Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology Resources*, 10(5):831–844.

Lehnen, L., Jan, P.-L., Besnard, A.-L., Fourcy, D., Kerth, G., Biedermann, M., Nyssen, P., Schorcht, W., Petit, E. et Puechmaille, S. (2021). Genetic diversity in a long-lived mammal is explained by the past's demographic shadow and current connectivity. *Molecular Ecology*, 00(1).

Lenormand, T. (2002). Gene flow and the limits to natural selection. *Trends in Ecology & Evolution*, 17(4):183–189.

Leroux, S. J., Albert, C. H., Lafuite, A.-S., Rayfield, B., Wang, S. et Gravel, D. (2017). Structural uncertainty in models projecting the consequences of habitat loss and fragmentation on biodiversity. *Ecography*, 40(1):36–47.

Levins, R. (1969). Some demographic and genetic consequences of environmental heterogeneity for biological control. *American Entomologist*, 15(3):237–240.

Lindenmayer, D. B., Blanchard, W., Foster, C. N., Scheele, B. C., Westgate, M. J., Stein, J., Crane, M. et Florance, D. (2020). Habitat amount versus connectivity : An empirical study of bird responses. *Biological Conservation*, 241:108377.

Lindenmayer, D. B. et Fischer, J. (2006). *Habitat fragmentation and landscape change : an ecological and conservation synthesis*. Island Press, Washington.

Long, F. H. (2013). Multivariate analysis for metabolomics and proteomics data. *In Proteomic and Metabolomic Approaches to Biomarker Discovery*, pages 299–311. Elsevier.

Lovette, I. J., Bermingham, E., Seutin, G. et Ricklefs, R. E. (1998). Evolutionary differentiation in three endemic West Indian warblers. *The Auk*, 115(4):890–903.

Luque, S., Saura, S. et Fortin, M.-J. (2012). Landscape connectivity analysis for conservation : insights from combining new methods with ecological and genetic data. *Landscape Ecology*, 27(2):153–157.

MacArthur, R. et Wilson, E. (1967). *The theory of island biogeography*. Princeton University Press, Princeton, NJ.

Magwene, P. M. (2001). New tools for studying integration and modularity. *Evolution*, 55(9):1734–1745.

Malkus, J. (1997). Habitatpräferenzen und mobilität der sumpfschrecke (stethophyma grossum l. 1758) unter besonderer berücksichtigung der mahd. *Articulata*, 12(1):1–18.

Manel, S. et Holderegger, R. (2013). Ten years of landscape genetics. *Trends in Ecology & Evolution*, 28(10):614–621.

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

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209–220.

Marrotte, R. R. et Bowman, J. (2017). The relationship between least-cost and resistance distance. *PLoS ONE*, 12(3):e0174212.

Martensen, A. C., Saura, S. et Fortin, M.-J. (2017). Spatio-temporal connectivity : assessing the amount of reachable habitat in dynamic landscapes. *Methods in Ecology and Evolution*, 8(10):1253–1264.

Martín-Queller, E., Albert, C. H., Dumas, P.-J. et Saatkamp, A. (2017). Islands, mainland, and terrestrial fragments : How isolation shapes plant diversity. *Ecology and Evolution*, 7(17):6904–6917.

Martín-Queller, E. et Saura, S. (2013). Landscape species pools and connectivity patterns influence tree species richness in both managed and unmanaged stands. *Forest Ecology and Management*, 289:123–132.

Marzelli, M. (1994). Ausbreitung von mecostethus grossus auf einer ausgleichs-und renaturierungsfläche. *Articulata*, 9(1):25–32.

Mateo-Sánchez, M. C., Balkenhol, N., Cushman, S., Pérez, T., Domínguez, A. et Saura, S. (2015). Estimating effective landscape distances and movement corridors : comparison of habitat and genetic data. *Ecosphere*, 6(4):1–16.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

McGarigal, K. (1995). *FRAGSTATS : spatial pattern analysis program for quantifying landscape structure*, volume 351. US Department of Agriculture, Forest Service, Pacific Northwest Research Station.

McPherson, J. M. et Jetz, W. (2007). Effects of species' ecology on the accuracy of distribution models. *Ecography*, 30(1):135–151.

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

McRae, B. H. et Beier, P. (2007). Circuit theory predicts gene flow in plant and animal populations. *Proceedings of the National Academy of Sciences*, 104(50):19885–19890.

McRae, B. H. et Kavanagh, D. M. (2011). *Linkage mapper connectivity analysis software*. The Nature Conservancy, Seattle WA.

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

Melles, S., Fortin, M.-J., Badzinski, D. et Lindsay, K. (2012). Relative importance of nesting habitat and measures of connectivity in predicting the occurrence of a forest songbird in fragmented landscapes. *Avian Conservation and Ecology*, 7(2).

Méndez, M., Tella, J. L. et Godoy, J. A. (2011). Restricted gene flow and genetic drift in recently fragmented populations of an endangered steppe bird. *Biological Conservation*, 144(11):2615–2622.

Miele, V., Matias, C., Robin, S. et Dray, S. (2019). Nine quick tips for analyzing network data. *PLoS Computational Biology*, 15(12):e1007434.

Miguet, P., Fahrig, L. et Lavigne, C. (2017). How to quantify a distance-dependent landscape effect on a biological response. *Methods in Ecology and Evolution*, 8(12):1717–1724.

Miguet, P., Jackson, H. B., Jackson, N. D., Martin, A. E. et Fahrig, L. (2016). What determines the spatial extent of landscape effects on species ? *Landscape Ecology*, 31(6):1177–1194.

Millette, K. L. et Keyghobadi, N. (2015). The relative influence of habitat amount and configuration on genetic structure across multiple spatial scales. *Ecology and Evolution*, 5(1):73–86.

Milligan, B. G. (in prep). Probabilistic graph models for landscape genetics. *PeerJ Preprints*.

Milligan, B. G., Archer, F. I., Ferchaud, A.-L., Hand, B. K., Kierepka, E. M. et Waples, R. S. (2018). Disentangling genetic structure for genetic monitoring of complex populations. *Evolutionary Applications*, 11(7):1149–1161.

Mimet, A., Clauzel, C. et Foltête, J.-C. (2016). Locating wildlife crossings for multispecies connectivity across linear infrastructures. *Landscape Ecology*, 31(9):1955–1973.

Minor, E. S. et Urban, D. L. (2008). A graph-theory framework for evaluating landscape connectivity and conservation planning. *Conservation Biology*, 22(2):297–307.

Miraldo, A., Li, S., Borregaard, M. K., Flórez-Rodríguez, A., Gopalakrishnan, S., Rizvanovic, M., Wang, Z., Rahbek, C., Marske, K. A. et Nogués-Bravo, D. (2016). An anthropocene map of genetic diversity. *Science*, 353(6307):1532–1535.

Moilanen, A. (2011). On the limitations of graph-theoretic connectivity in spatial ecology and conservation. *Journal of Applied Ecology*, 48(6):1543–1547.

Moilanen, A. et Nieminen, M. (2002). Simple connectivity measures in spatial ecology. *Ecology*, 83(4):1131–1145.

Mony, C., Abadie, J., Gil-Tena, A., Burel, F. et Ernoult, A. (2018). Effects of connectivity on animal-dispersed forest plant communities in agriculture-dominated landscapes. *Journal of Vegetation Science*, 29(2):167–178.

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

Munwes, I., Geffen, E., Roll, U., Friedmann, A., Daya, A., Tikochinski, Y. et Gafny, S. (2010). The change in genetic diversity down the core-edge gradient in the eastern spadefoot toad (*Pelobates syriacus*). *Molecular Ecology*, 19(13):2675–2689.

Muratet, A., Lorrilliere, R., Clergeau, P. et Fontaine, C. (2013). Evaluation of landscape connectivity at community level using satellite-derived NDVI. *Landscape Ecology*, 28(1):95–105.

Murekatete, R. M. et Shirabe, T. (2018). A spatial and statistical analysis of the impact of transformation of raster cost surfaces on the variation of least-cost paths. *International Journal of Geographical Information Science*, 32(11):2169–2188.

Murphy, M., Dyer, R. et Cushman, S. A. (2016). Graph theory and network models in landscape genetics. *In* Balkenhol, N., Cushman, S., Storfer, A. et Waits, L., éditeurs : *Landscape genetics : Concepts, methods, applications*, pages 165–180. John Wiley & Sons, 1 édition.

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

Murphy, M. A., Evans, J. S. et Storfer, A. (2010b). Quantifying Bufo boreas connectivity in Yellowstone National Park with landscape genetics. *Ecology*, 91(1):252–261.

Myers, N., Mittermeier, R. A., Mittermeier, C. G., Da Fonseca, G. A. et Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403(6772):853.

Nathan, R., Perry, G., Cronin, J. T., Strand, A. E. et Cain, M. L. (2003). Methods for estimating long-distance dispersal. *Oikos*, 103(2):261–273.

Naujokaitis-Lewis, I. R., Rico, Y., Lovell, J., Fortin, M.-J. et Murphy, M. A. (2013). Implications of incomplete networks on estimation of landscape genetic connectivity. *Conservation Genetics*, 14(2):287–298.

Neel, M. C. (2008). Patch connectivity and genetic diversity conservation in the federally endangered and narrowly endemic plant species *Astragalus albens* (Fabaceae). *Biological Conservation*, 141(4):938–955.

Neigel, J. E. (2002). Is FST Obsolete ? *Conservation Genetics*, 3(2):167–173.

Neuditschko, M., Khatkar, M. S. et Raadsma, H. W. (2012). NETVIEW : a high-definition network-visualization approach to detect fine-scale population structures from genome-wide patterns of variation. *PLoS ONE*, 7(10):e48375.

Nevil Amos, J., Harrisson, K. A., Radford, J. Q., White, M., Newell, G., Nally, R. M., Sunnucks, P. et Pavlova, A. (2014). Species-and sex-specific connectivity effects of habitat fragmentation in a suite of woodland birds. *Ecology*, 95(6):1556–1568.

Nieminen, M., Singer, M. C., Fortelius, W., Schöps, K. et Hanski, I. (2001). Experimental confirmation that inbreeding depression increases extinction risk in butterfly populations. *The American Naturalist*, 157(2):237–244.

O'Brien, D., Manseau, M., Fall, A. et Fortin, M.-J. (2006). Testing the importance of spatial configuration of winter habitat for woodland caribou : an application of graph theory. *Biological Conservation*, 130(1):70–83.

Ortiz-Rodríguez, D. O., Guisan, A., Holderegger, R. et van Strien, M. J. (2019). Predicting species occurrences with habitat network models. *Ecology and evolution*, 9(18):10457–10471.

Panzacchi, M., Van Moorter, B., Strand, O., Saerens, M., Kivimäki, I., St Clair, C. C., Herfindal, I. et Boitani, L. (2016). Predicting the continuum between corridors and barriers to animal movements using step selection functions and randomized shortest paths. *Journal of Animal Ecology*, 85(1):32–42.

Paradis, E. (2010). pegas : an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3):419–420.

Paschou, P., Drineas, P., Yannaki, E., Razou, A., Kanaki, K., Tsetsos, F., Padmanabhuni, S. S., Michalodimitrakis, M., Renda, M. C., Pavlovic, S. *et al.* (2014). Maritime route of colonization of Europe. *Proceedings of the National Academy of Sciences*, 111(25):9211–9216.

Pascual-Hortal, L. et Saura, S. (2006). Comparison and development of new graph-based landscape connectivity indices : towards the priorization of habitat patches and corridors for conservation. *Landscape Ecology*, 21(7):959–967.

Pasinelli, G., Meichtry-Stier, K., Birrer, S., Baur, B. et Duss, M. (2013). Habitat quality and geometry affect patch occupancy of two Orthopteran species. *PLoS ONE*, 8(5):e65850.

Pe'er, G., Henle, K., Dislich, C. et Frank, K. (2011). Breaking functional connectivity into components : A novel approach using an individual-based model, and first outcomes. *PLoS ONE*, 6(8):1–18.

Pe'er, G., Saltz, D. et Frank, K. (2005). Virtual corridors for conservation management. *Conservation Biology*, 19(6):1997–2003.

Pereira, M., Segurado, P. et Neves, N. (2011). Using spatial network structure in landscape management and planning : a case study with pond turtles. *Landscape and Urban Planning*, 100(1):67–76.

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

Pérez-Rodríguez, A., Khimoun, A., Ollivier, A., Eraud, C., Faivre, B. et Garnier, S. (2018). Habitat fragmentation, not habitat loss, drives the prevalence of blood parasites in a Caribbean passerine. *Ecography*, 41(11):1835–1849.

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.

Peterman, W. E., Anderson, T. L., Ousterhout, B. H., Drake, D. L., Semlitsch, R. D. et Eggert, L. S. (2015). Differential dispersal shapes population structure and patterns of genetic differentiation in two sympatric pond breeding salamanders. *Conservation Genetics*, 16(1):59–69.

Peterman, W. E., Connette, G. M., Semlitsch, R. D. et Eggert, L. S. (2014). Ecological resistance surfaces predict fine-scale genetic differentiation in a terrestrial woodland salamander. *Molecular Ecology*, 23(10):2402–2413.

Peterman, W. E. et Pope, N. S. (2020). The use and misuse of regression models in landscape genetic analyses. *Molecular Ecology*, 30(1):37–47.

Peterman, W. E., Winiarski, K. J., Moore, C. E., da Silva Carvalho, C., Gilbert, A. L. et Spear, S. F. (2019). A comparison of popular approaches to optimize landscape resistance surfaces. *Landscape Ecology*, 34(9):2197–2208.

Peterson, E. E., Hanks, E. M., Hooten, M. B., Ver Hoef, J. M. et Fortin, M.-J. (2019). Spatially structured statistical network models for landscape genetics. *Ecological Monographs*, 89(2):e01355.

Pflüger, F. J. et 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.

Pinto, N. et Keitt, T. H. (2009). Beyond the least-cost path : evaluating corridor redundancy using a graph-theoretic approach. *Landscape Ecology*, 24(2):253–266.

Poli, C., Hightower, J. et Fletcher Jr, R. J. (2020). Validating network connectivity with observed movement in experimental landscapes undergoing habitat destruction. *Journal of Applied Ecology*, 57(7):1426–1437.

Pons, P. et Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2):191–218.

Poor, E. E., Loucks, C., Jakes, A. et Urban, D. L. (2012). Comparing habitat suitability and connectivity modeling methods for conserving pronghorn migrations. *PLoS ONE*, 7(11):e49390.

Poor, E. E., Shao, Y. et Kelly, M. J. (2019). Mapping and predicting forest loss in a Sumatran tiger landscape from 2002 to 2050. *Journal of Environmental Management*, 231:397–404.

Pressey, R. (2004). Conservation planning and biodiversity : assembling the best data for the job. *Conservation Biology*, 18(6):1677–1681.

Pritchard, J. K., Stephens, M. et Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959.

Prunier, J. G., Dubut, V., Chikhi, L. et 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.

Pullinger, M. G. et Johnson, C. J. (2010). Maintaining or restoring connectivity of modified landscapes : evaluating the least-cost path model with multiple sources of ecological information. *Landscape Ecology*, 25(10):1547–1560.

Radford, J. Q., Amos, N., Harrisson, K., Sunnucks, P. et Pavlova, A. (2021). Functional connectivity and population persistence in woodland birds : insights for management from a multi-species conservation genetics study. *Emu-Austral Ornithology*, 121(1):1–13.

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

Rayfield, B., Fortin, M.-J. et Fall, A. (2011). Connectivity for conservation : a framework to classify network measures. *Ecology*, 92(4):847–858.

Raymond, M. et Rousset, F. (1995). GENEPOP : Population genetics software for exact tests and ecumenism. Vers. 1.2. *Journal of Heredity*, 86:248–249.

Reed, G., Litvaitis, J., Callahan, C., Carroll, R., Litvaitis, M. et Broman, D. (2017). Modeling landscape connectivity for bobcats using expert-opinion and empirically derived models : how well do they work ? *Animal Conservation*, 20(4):308–320.

Reichert, B. E., Fletcher Jr, R. J., Cattau, C. E. et Kitchens, W. M. (2016). Consistent scaling of population structure across landscapes despite intraspecific variation in movement and connectivity. *Journal of Animal Ecology*, 85(6):1563–1573.

Reinhardt, K., Köhler, G., Maas, S. et Detzel, P. (2005). Low dispersal ability and habitat specificity promote extinctions in rare but not in widespread species : the Orthoptera of Germany. *Ecography*, 28(5):593–602.

Ribeiro, R., Carretero, M. A., Sillero, N., Alarcos, G., Ortiz-Santaliestra, M., Lizana, M. et Llorente, G. A. (2011). The pond network : can structural connectivity reflect on (amphibian) biodiversity patterns ? *Landscape Ecology*, 26(5):673–682.

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

Ricketts, T. H. (2001). The matrix matters : effective isolation in fragmented landscapes. *The American Naturalist*, 158(1):87–99.

Riley, S. P., Pollinger, J. P., Sauvajot, R. M., York, E. C., Bromley, C., Fuller, T. K. et 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, E. P., Fletcher, R. J., Cattau, C. E., Udell, B. J., Reichert, B. E., Austin, J. D. et Valle, D. (2018a). Isolating the roles of movement and reproduction on effective connectivity alters conservation priorities for an endangered bird. *Proceedings of the National Academy of Sciences*, 115(34):8591–8596.

Robertson, E. P., Fletcher Jr, R. J. et Austin, J. D. (2019). The number of breeders explains genetic connectivity in an endangered bird. *Molecular Ecology*, 28(11):2746–2756.

Robertson, J. M., Murphy, M. A., Pearl, C. A., Adams, M. J., Páez-Vacas, M. I., Haig, S. M., Pilliod, D. S., Storfer, A. et Funk, W. C. (2018b). 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.

Robin, V., Gupta, P., Thatte, P. et Ramakrishnan, U. (2015). Islands within islands : two montane palaeo-endemic birds impacted by recent anthropogenic fragmentation. *Molecular Ecology*, 24(14):3572–3584.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, 145(4):1219–1228.

Roy, K., Kar, S. et Das, R. N. (2015). Statistical methods in QSAR/QSPR. *In A primer on QSAR/QSPR modeling*, pages 37–59. Springer.

Rozenfeld, A. F., Arnaud-Haond, S., Hernández-García, E., Eguíluz, V. M., Serrão, E. A. et Duarte, C. M. (2008). Network analysis identifies weak and strong links in a metapopulation system. *Proceedings of the National Academy of Sciences*, 105(48):18824–18829.

Ruiz-Gonzalez, A., Cushman, S. A., Madeira, M. J., Randi, E. et Gómez-Moliner, B. J. (2015). Isolation by distance, resistance and/or clusters ? lessons learned from a forest-dwelling carnivore inhabiting a heterogeneous landscape. *Molecular Ecology*, 24(20):5110–5129.

Ruiz-González, A., Gurrutxaga, M., Cushman, S. A., Madeira, M. J., Randi, E. et 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.

Saccheri, I., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W. et Hanski, I. (1998). Inbreeding and extinction in a butterfly metapopulation. *Nature*, 392(6675):491–494.

Saura, S. (2018). The amount of reachable habitat - jointly measuring habitat amount and connectivity in space and time. *In International Conference of Ecological Sciences of the French Society for Ecology and Evolution.*

Saura, S. (2021). The Habitat Amount Hypothesis implies negative effects of habitat fragmentation on species richness. *Journal of Biogeography*, 48(1):11–22.

Saura, S., Bodin, Ö. et Fortin, M.-J. (2014). Stepping stones are crucial for species' long-distance dispersal and range expansion through habitat networks. *Journal of Applied Ecology*, 51(1):171–182.

Saura, S. et de la Fuente, B. (2017). Connectivity as the amount of reachable habitat : conservation priorities and the roles of habitat patches in landscape networks. *In* Gergel, S. E. et Turner, M. G., éditeurs : *Learning landscape ecology : a practical guide to concepts and techniques*, pages 229–254. Springer.

Saura, S. et Pascual-Hortal, L. (2007). A new habitat availability index to integrate connectivity in landscape conservation planning : comparison with existing indices and application to a case study. *Landscape and Urban Planning*, 83(2):91–103.

Saura, S. et Rubio, L. (2010). A common currency for the different ways in which patches and links can contribute to habitat availability and connectivity in the landscape. *Ecography*, 33(3):523–537.

Saura, S. et Torne, J. (2009). Conefor Sensinode 2.2 : a software package for quantifying the importance of habitat patches for landscape connectivity. *Environmental Modelling & Software*, 24(1):135–139.

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

Sawyer, S. C., Epps, C. W. et Brashares, J. S. (2011). Placing linkages among fragmented habitats : do least-cost models reflect how animals use landscapes ? *Journal of Applied Ecology*, 48(3):668–678.

Schadt, S., Knauer, F., Kaczensky, P., Revilla, E., Wiegand, T. et 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, C. (2003). The influence of spatial scale on quantifying insect dispersal : an analysis of butterfly data. *Ecological Entomology*, 28(2):252–256.

Schneider, D. W., Ellis, C. D. et 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.

Schoville, S. D., Dalongeville, A., Viennois, G., Gugerli, F., Taberlet, P., Lequette, B., Alvarez, N. et Manel, S. (2018). Preserving genetic connectivity in the European Alps protected area network. *Biological Conservation*, 218:99–109.

Sciaini, M., Fritsch, M., Scherer, C. et 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.

Segelbacher, G., Cushman, S. A., Epperson, B. K., Fortin, M.-J., Francois, O., Hardy, O. J., Holderegger, R., Taberlet, P., Waits, L. P. et Manel, S. (2010). Applications of landscape genetics in conservation biology : concepts and challenges. *Conservation Genetics*, 11(2):375–385.

Serrano, M. Á., Boguná, M. et Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488.

Serrouya, R., Paetkau, D., McLELLAN, B. N., Boutin, S., Campbell, M. et 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.

Shirabe, T. (2016). A method for finding a least-cost wide path in raster space. *International Journal of Geographical Information Science*, 30(8):1469–1485.

Shirk, A. et Cushman, S. (2011). sGD : software for estimating spatially explicit indices of genetic diversity. *Molecular Ecology Resources*, 11(5):922–934.

Shirk, A., Landguth, E. et Cushman, S. (2017a). A comparison of indiviudal-based genetic distance metrics for landscape genetics. *Molecular Ecology*, 17(6):1308–1317.

Shirk, A., Wallin, D., Cushman, S., Rice, C. et 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. et Cushman, S. A. (2017b). A comparison of regression methods for model selection in individual-based landscape genetic analysis. *Molecular Ecology Resources*, 18(1):55–67.

Simpkins, C. E., Dennis, T. E., Etherington, T. R. et Perry, G. L. (2017). Effects of uncertain cost-surface specification on landscape connectivity measures. *Ecological Informatics*, 38:1–11.

Simpkins, C. E., Dennis, T. E., Etherington, T. R. et Perry, G. L. (2018). Assessing the performance of common landscape connectivity metrics using a virtual ecologist approach. *Ecological Modelling*, 367:13–23.

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

Smouse, P. E. et Peakall, R. (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, 82(5):561–573.

Song, W. et Kim, E. (2016). Landscape factors affecting the distribution of the great tit in fragmented urban forests of Seoul, South Korea. *Landscape and Ecological Engineering*, 12(1):73–83.

Sonneck, A.-G., Bönsel, A. et Matthes, J. (2008). Der einfluss von landnutzung auf die habitate von stethophyma grossum (linnaeus, 1758) an beispielen aus mecklenburg-vorpommern. *Articulata*, 23:15–30.

Spackman, S. C. et Hughes, J. W. (1995). Assessment of minimum stream corridor width for biological conservation : species richness and distribution along mid-order streams in Vermont, USA. *Biological Conservation*, 71(3):325–332.

Spear, S. F., Balkenhol, N., Fortin, M.-J., McRae, B. H. et Scribner, K. (2010). Use of resistance surfaces for landscape genetic studies : considerations for parameterization and analysis. *Molecular Ecology*, 19(17):3576–3591.

Spielman, D., Brook, B. W. et 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.

Stevenson-Holt, C. D., Watts, K., Bellamy, C. C., Nevin, O. T. et Ramsey, A. D. (2014). Defining landscape resistance values in least-cost connectivity models for the invasive grey squirrel : a comparison of approaches using expert-opinion and habitat suitability modelling. *PLoS ONE*, 9(11):e112119.

Storfer, A., Murphy, M., Evans, J., Goldberg, C., Robinson, S., Spear, S., Dezzani, R., Delmelle, E., Vierling, L. et Waits, L. (2007). Putting the "landscape" in landscape genetics. *Heredity*, 98(3):128–142.

Storfer, A., Murphy, M. A., Spear, S. F., Holderegger, R. et Waits, L. P. (2010). Landscape genetics : where are we now ? *Molecular Ecology*, 19(17):3496–3514.

Stuber, E. F. et Gruber, L. F. (2020). Recent methodological solutions to identifying scales of effect in multi-scale modeling. *Current Landscape Ecology Reports*, 5(1):127–139.

Szpiech, Z. A., Jakobsson, M. et Rosenberg, N. A. (2008). ADZE : a rarefaction approach for counting alleles private to combinations of populations. *Bioinformatics*, 24(21):2498–2504.

Taberlet, P., Zimmermann, N. E., Englisch, T., Tribsch, A., Holderegger, R., Alvarez, N., Niklfeld, H., Coldea, G., Mirek, Z., Moilanen, A. *et al.* (2012). Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. *Ecology Letters*, 15(12):1439–1448.

Tannier, C., Bourgeois, M., Houot, H. et Foltête, J.-C. (2016). Impact of urban developments on the functional connectivity of forested habitats : a joint contribution of advanced urban models and landscape graphs. *Land Use Policy*, 52:76–91.

Tarabon, S., Bergès, L., Dutoit, T. et Isselin-Nondedeu, F. (2019). Environmental impact assessment of development projects improved by merging species distribution and habitat connectivity modelling. *Journal of Environmental Management*, 241:439–449.

Taylor, P. D., Fahrig, L., Henein, K. et Merriam, G. (1993). Connectivity is a vital element of landscape structure. *Oikos*, 68(3):571–573.

Taylor, P. D., Fahrig, L. et With, K. A. (2006). Landscape connectivity : a return to the basics. *In* Crooks, K. R. et Sanjayan, M., éditeurs : *Connectivity conservation*, chapitre 2, pages 29–43. Cambridge University Press.

Taylor, Z. et Hoffman, S. (2014). Landscape models for nuclear genetic diversity and genetic structure in white-footed mice (*Peromyscus leucopus*). *Heredity*, 112(6):588–595.

Tenenhaus, M. (1998). *La régression PLS : théorie et pratique*. Editions TECHNIP.

Tenenhaus, M. et Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50(1):91–119.

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

Toma, Y., Imanishi, J., Yokogawa, M., Hashimoto, H., Imanishi, A., Morimoto, Y., Hatanaka, Y., Isagi, Y. et Shibata, S. (2015). Factors affecting the genetic diversity of a perennial herb *Viola grypoceras* A. Gray var. *grypoceras* in urban fragmented forests. *Landscape Ecology*, 30(8):1435–1447.

Tournant, P., Afonso, E., Roué, S., Giraudoux, P. et Foltête, J.-C. (2013). Evaluating the effect of habitat connectivity on the distribution of lesser horseshoe bat maternity roosts using landscape graphs. *Biological Conservation*, 164:39–49.

Trautner, J. et Hermann, G. (2008). Die Sumpfschrecke (*Stethophyma grossum* L., 1758) im Aufwind-Erkenntnisse aus dem zentralen Baden-Württemberg. *Articulata*, 23(2):37–52.

Ueno, S., Tomaru, N., Yoshimaru, H., Manabe, T. et 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. et Keitt, T. (2001). Landscape connectivity : a graph-theoretic perspective. *Ecology*, 82(5):1205–1218.

Urban, D. L., Minor, E. S., Treml, E. A. et Schick, R. S. (2009). Graph models of habitat mosaics. *Ecology Letters*, 12(3):260–273.

Van Dyck, H. et Baguette, M. (2005). Dispersal behaviour in fragmented landscapes : routine or special movements ? *Basic and Applied Ecology*, 6(6):535–545.

Van Etten, J. (2012). R package gdistance : distances and routes on geographical grids (version 1.1-4). *Journal of Statistical Software*, 76(1):1–13.

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., Holderegger, R. et Van Heck, H. J. (2015). Isolation-by-distance in landscapes : considerations for landscape genetics. *Heredity*, 114(1):27–37.

Van Strien, M. J., Keller, D. et 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.

Van Strien, M. J., Keller, D., Holderegger, R., Ghazoul, J., Kienast, F. et Bolliger, J. (2014). Landscape genetics as a tool for conservation planning : predicting the effects of landscape change on gene flow. *Ecological Applications*, 24(2):327–339.

Varvio, S.-L., Chakraborty, R. et Nei, M. (1986). Genetic variation in subdivided populations and conservation genetics. *Heredity*, 57(2):189–198.

Vellend, M. et Geber, M. A. (2005). Connections between species diversity and genetic diversity. *Ecology Letters*, 8(7):767–781.

Villard, M.-A. et Metzger, J. P. (2014). Beyond the fragmentation debate : a conceptual model to predict when habitat configuration really matters. *Journal of Applied Ecology*, 51(2):309–318.

Wagner, H. H. et Fortin, M.-J. (2013). A conceptual framework for the spatial analysis of landscape genetic data. *Conservation Genetics*, 14(2):253–261.

Waits, L. P. et Storfer, A. (2015). Basics of population genetics : quantifying neutral and adaptive genetic variation for landscape genetic studies. *In* Balkenhol, N., Cushman, S., Storfer, A. et Waits, L., éditeurs : *Landscape genetics : Concepts, methods, applications*, pages 35–57. John Wiley & Sons.

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. et 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. et Lin, L.-K. (2008). Habitat suitability modelling to correlate gene flow with landscape connectivity. *Landscape Ecology*, 23(8):989–1000.

Wasserman, T. N., Cushman, S. A., Schwartz, M. K. et Wallin, D. O. (2010). Spatial scaling and multi-model inference in landscape genetics : *Martes americana* in northern Idaho. *Landscape Ecology*, 25(10):1601–1612.

Watts, A. G., Schlichting, P. E., Billerman, S. M., Jesmer, B. R., Micheletti, S., Fortin, M.-J., Funk, W. C., Hapeman, P., Muths, E. et 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. et 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.

Weir, B. S. et Cockerham, C. C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution*, 38(6):1358–1370.

Whitlock, M. C. et Mccauley, D. E. (1999). Indirect measures of gene flow and migration : Fst≠ 1/(4nm+ 1). *Heredity*, 82(2):117–125.

Whittaker, J. (2009). *Graphical models in applied multivariate statistics*. Wiley Publishing.

Wold, S., Sjöström, M. et Eriksson, L. (2001). PLS-regression : a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2):97–159.

Wright, S. (1943). Isolation by distance. *Genetics*, 28(2):114–138.

Xia, Y., Bjørnstad, O. N. et 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., Creech, T. G., Millette, K. L., Crowhurst, R. S., Long, R. A., Wagner, H. H., Balkenhol, N. et Landguth, E. L. (2016). Using simulations to evaluate Mantel-based methods for assessing landscape resistance to gene flow. *Ecology and Evolution*, 6(12):4115–4128.

Zeller, K. A., Jennings, M. K., Vickers, T. W., Ernest, H. B., Cushman, S. A. et 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. et 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.

Zetterberg, A., Mörtberg, U. M. et Balfors, B. (2010). Making graph theory operational for landscape ecological assessments, planning, and design. *Landscape and Urban Planning*, 95(4):181–191.

Ziółkowska, E., Ostapowicz, K., Kuemmerle, T., Perzanowski, K., Radeloff, V. C. et Kozak, J. (2012). Potential habitat connectivity of European bison (*Bison bonasus*) in the Carpathians. *Biological Conservation*, 146(1):188–196.

Zurell, D., Berger, U., Cabral, J. S., Jeltsch, F., Meynard, C. N., Münkemüller, T., Nehrbass, N., Pagel, J., Reineking, B., Schröder, B. *et al.* (2010). The virtual ecologist approach : simulating data and observers. *Oikos*, 119(4):622–635.