

Article type: Opinion

**The use and misuse of regression models in landscape genetic analyses**

Running title: Regression models in landscape genetics

William E. Peterman<sup>1</sup>

1. School of Environment and Natural Resources, The Ohio State University, Columbus, Ohio, United States of America

\*Corresponding author: [Peterman.73@osu.edu](mailto:Peterman.73@osu.edu)

## **Abstract**

The field of landscape genetics has been rapidly evolving, adopting and adapting analytical frameworks to address research questions. As landscape genetic analyses have shifted away from Mantel-based analytical frameworks, studies are increasingly using regression-based frameworks to understand the individual contributions of landscape and habitat variables on genetic differentiation. This paper outlines appropriate and inappropriate uses of multiple regression for these purposes. Of concern is the prevalence of studies seeking to explain genetic differences by fitting regression models with effective distance variables calculated independently on separate landscape resistance surfaces. When moving across the landscape, organisms cannot respond independently and uniquely to habitat and landscape features. Therefore, independent resistance surfaces and their effective distance measures have no mechanistic meaning or relevant statistical interpretation. There are also tremendous challenges to fitting and interpreting regression models that include ‘independent’ effective distance measures as predictors, including statistical suppression. As such, regression analyses seeking to understand how landscape resistance affects gene flow should be univariate models, with the creation of a single resistance surface being a necessary precursor to the regression analysis. There are, however, important statistical advances underway that explicitly model the covariance of allele frequencies or genetic distances as functions of spatial landscape variables. The growth and evolution of landscape genetics as a field has been rapid and exciting. It is the goal of this opinion paper to highlight past missteps and to ensure that future use of regression models will appropriately consider the process being modeled, which will provide clarity to model interpretation.

## **Keywords:**

Landscape genetics, landscape genomics, landscape resistance, maximum likelihood population effects mixed effects model, MLPE, multiple regression on distance matrices

## Background

A fundamental goal of landscape genetic research is to understand how landscape structure and composition contribute to patterns of genetic (dis)similarity between individuals or populations. However, achieving this goal has been exceedingly difficult. Over the relatively short history of the discipline, landscape genetics has incrementally made progress toward statistically defensible models and analytical frameworks. Mantel and partial Mantel tests were among the first models used to correlate genetic (dis)similarity with least cost path distance, cumulative cost distance, or resistance distance (hereafter effective distance) across the landscape. Mantel tests were bolstered through the use of rigorous frameworks for assessing alternative hypotheses of landscape resistance, specifically the causal modeling framework (Cushman *et al.* 2006) and later the reciprocal causal modeling framework (Cushman *et al.* 2013). However, Mantel tests have received extensive scrutiny (e.g., Balkenhol *et al.* 2009; Guillot & Rousset 2013; Legendre *et al.* 2015; Legendre & Fortin 2010; Zeller *et al.* 2016), and are now known to be a relatively low-power method for correctly inferring genetic-distance relationships.

As of 2008, nearly 60% of landscape genetic studies used Mantel tests, while linear regression (~20%) and multiple matrix regression (<5%) were much less common (Storfer *et al.* 2010). However, the removal of the Mantel test from the statistical toolbox left researchers looking for alternative analytical approaches, of which there are many (Balkenhol *et al.* 2009). Two regression-based methods, multiple regression on distance matrices (MRDM; Legendre *et al.* 1994) and novel formulations of mixed effects models (Clarke *et al.* 2002), have received extensive use in more recent landscape genetic analyses. Originally developed as a statistical framework for evaluating the effects of independent traits coded as distance matrices, MRDM had a logical and intuitive extension to modeling effective distances between spatial samples and their corresponding genetic distances. In contrast, the maximum likelihood population-effects parameterization of a mixed effects model (MLPE; Clarke *et al.* 2002) was specifically developed for accommodating the non-independence inherent in pairwise genetic and geographic or environmental distance. MLPE was first used in a landscape genetics context by van Strien *et al.* (2012),

and has since seen widespread application. Reinforcing the value of the MLPE model, Shirk *et al.* (2017) identified it as the best overall regression-based approach when conducting model selection in individual-based analyses.

Unfortunately, the use of regression-based models in landscape genetics is not without its pitfalls, and it is my opinion and concern that researchers are misusing regression models for landscape genetic analyses (e.g., Balkenhol *et al.* 2014; Blair *et al.* 2013; Emel & Storfer 2015; Kozakiewicz *et al.* 2019; Mims *et al.* 2015; Prunier *et al.* 2017; Row *et al.* 2015; Trumbo *et al.* 2019; Trumbo *et al.* 2013). Specifically, effective distance should not be calculated separately across individual resistance surfaces and then used as independent predictor variables in regression models. The purpose of this perspective piece is to highlight the conceptual flaws in this approach when the objective is to understand how landscape resistance effects gene flow.

### *Landscape resistance and effective distance*

It is important to think mechanistically about the underlying process we hope to understand through a regression-based landscape genetics analysis. When we sample individuals or populations in space, the occurrence or frequency of alleles can be used to calculate genetic (dis)similarity. It is typically assumed that space (isolation-by-distance), the intervening landscape (isolation-by-resistance), and/or environmental differences (isolation-by-environment) are factors affecting the movement, survival, and subsequent gene flow of organisms. Understanding how landscape features affect gene flow requires a movement model, such as least cost path or circuit resistance (McRae 2006; McRae & Beier 2007), to calculate effective distances between spatial samples. These models of movement require a single resistance surface as an input. Similarly, individual- and population-based genetic simulations that include a landscape resistance effect require a single distance matrix describing the effective distance between spatial locations on the landscape (e.g., Adamack & Gruber 2014; Landguth *et al.* 2016; Landguth & Cushman 2009). This single effective distance matrix must encapsulate the combined effects of all

landscape and environmental features. The requirements of current genetic simulation software also accurately reflect the movement process of organisms, as all landscape and environmental features are perceived and experienced simultaneously. For example, a salamander dispersing across the landscape experiences the land cover, hill slope, and soil moisture simultaneously, and all these features combine to shape the movement paths of salamanders. It therefore does not make biological sense to calculate effective distances across landscape features independently. It is quite likely that the movement paths or effective distances across individual surfaces will not align with those across the combined resistance surface (Fig. 1). Individual effective distance measures are not independent nor additive components of the true effective distance that results from a resistance surface derived from multiple spatial features. As such, the fitted regression model should be

$$\text{genetic distance} \sim \text{effective distance.} \quad (1)$$

If the movement process is the result of experiencing all facets of the landscape at once, what can we hope to meaningfully ascertain from artificially creating distance variables from each landscape surface independently? The challenge for landscape genetic researchers is determining which variables combine in which way to create a resistance surface and subsequent effective distances between individuals or populations that best reflect the patterns in observed genetic distances. Calculating independent effective distance measures is not a valid approach to surmounting this challenge.

### *Multicollinearity and parameter interpretation*

As with any regression-based approach, it is important to assess and minimize collinearity among variables used as independent variables in the model. Assessing zero-order correlations and variance inflation factors are steps that seem to be widely adopted and correctly implemented in regression-based landscape genetic analyses that include multiple predictor variables. However, it is quite alarming that effective distances between sample locations can be uncorrelated and ‘independent’. This highlights how unique and divergent movement paths between sample locations can be, as demonstrated in Figure 1.

Milanesi *et al.* (2016) found that not only can least cost path distances calculated on different resistance surfaces be uncorrelated, they can be significantly different.

Even in the absence of multicollinearity, effective distances as independent variables are likely to act as suppressors in the regression model (Paulhus *et al.* 2004; Prunier *et al.* 2017; Prunier *et al.* 2015). In short, nuisance variance is explained by the suppressor variable, which can elevate the importance of other parameters and result in a superior fitting model. The presence of a suppressor variable is often manifested as a negative regression parameter estimate. This phenomena has been described in previous regression-based analyses (Row *et al.* 2015; Trumbo *et al.* 2019; Trumbo *et al.* 2013). Negative parameter coefficients have been interpreted as indicating that a particular resistance surface has a negative effect on gene flow (Trumbo *et al.* 2019; Trumbo *et al.* 2013), or have been discounted as indicating a non-true relationship (Kozakiewicz *et al.* 2019). In their simulation study and assessment of multiple regression of distance matrices, Prunier *et al.* (2015) used correspondence analysis to demonstrate how suppression can influence parameter estimates, challenging interpretation of parameters and model selection. The reality is that effective distances, as calculated across a hypothesized resistance surface, should have either a positive or negligible effect. The only way to generate a ‘negative effect’ is to invert the resistance values of the landscape. A positive regression parameter estimate on this inverted resistance surface with confidence intervals that do not include zero is statistical support of this ‘negative effect’.

Although multicollinearity is no doubt hugely problematic in regression-based analyses and cannot be ignored, analyses seeking to relate pairwise effective distances across the landscape to pairwise genetic distances should only contain a single predictor variable consisting of the pairwise effective distances between sample locations (Eq 1). Therefore, multicollinearity should not be an issue of concern.

*Can multiple predictor variables ever be used in a distance-based landscape genetics regression analysis?*

In short, yes, there are ways to approach distance-based landscape genetic analyses that can defensibly accommodate multiple independent landscape predictor variables. These scenarios broadly fall into two

different modeling approaches: (1) corridor- or transect-based analyses and (2) derived distance analyses. In corridor-based analyses, transects or corridors between all sample locations are first created, then landscape features are summarized along transects or within corridors (e.g., maximum slope, average canopy cover, proportion of corridor consisting of developed land). This was the approach used by van Strien *et al.* (2012) in their initial application of MLPE for landscape genetic analyses, which resulted in largely independent predictor variables. Using this framework, regression parameter coefficients can be interpreted as promoting or impeding movement within the corridor, as well as their magnitude of effect. There are still challenges and limitations to this approach. First, one must decide how spatial locations will be connected. Studies have simply used a straight line (e.g., Coster *et al.* 2015; Goldberg & Waits 2010; Murphy *et al.* 2010), or alternatively have used a least cost path between locations (e.g., Emel & Storfer 2015; Spear & Storfer 2008; van Strien *et al.* 2012). The former approach assumes that a summary of the landscape directly between locations is influencing movement, while the latter approach requires that a resistance surface be defined in order to generate least cost paths. Second, the width of the corridor surrounding the connection lines must be defined. Researchers have previously considered the corridor width to be indicative of the scale at which a species responds to a particular landscape or habitat feature (Murphy *et al.* 2010; van Strien *et al.* 2012), however this interpretation differs from definitions of scale in the landscape ecological literature (e.g., McGarigal *et al.* 2016; Zeller *et al.* 2012). Finally, it must be assumed that the statistic used to summarize landscape features within the corridor (e.g., proportion, mean, maximum, minimum) appropriately reflects how movement through the corridor might be affected.

The other approach for generating independent landscape variables is to create distance matrices (e.g., Euclidean distance) of the differences in landscape or habitat features between sample locations. This approach was used by Franckowiak *et al.* (2017) in their assessment of the model selection abilities of MRDM, is in line with the original application of MRDM (Legendre *et al.* 1994), and is an assessment of isolation-by-environment (Wang 2013; Wang & Bradburd 2014). It is important to recognize that this approach does not require or follow any specific movement model across the landscape and is not an explicit evaluation of how landscape configuration or matrix composition affect gene flow.

### *Creating a resistance surface*

If the end goal of an analysis is to identify the best-supported resistance surface and/or to infer movement paths across the landscape, the real challenge facing researchers is how to combine individual landscape features to create a resistance surface. This is a long-standing challenge in landscape genetics (Spear *et al.* 2010). Unfortunately, neither of the frameworks described above will achieve this goal (although full implementation of the multi-step framework of van Strien *et al.* (2012) does include identifying a resistance surface for creating least cost paths). Numerous approaches for creating and optimizing resistance surfaces have been used (e.g., Peterman 2018; Shirk *et al.* 2010; Wang *et al.* 2009). Central to these is an assessment of effective distance calculated on a resistance surface in relation to genetic distance (Eq 1). Essentially, a suite of resistance surfaces are created by transforming and combining individual surfaces. Effective distances are then calculated across each resistance surface and univariate models are then fit with each effective distance measure in turn. Finally, some form of model selection is used identify the best-supported resistance surface. Just as in instances of regression misuse, there is still a desire in these analyses to identify the landscape features promoting or impeding movement, but these approaches recognize that it is not possible to ascertain individual landscape effects in a regression model.

### *Future advances*

What is apparent from the misuse of regression-based models is that an end goal of many analyses is to understand the individual contributions of landscape and habitat features on gene flow. Development of novel, model-based analytical frameworks is ongoing, and show great promise (e.g., Bradburd *et al.* 2013; Hanks 2017; Hanks & Hooten 2013; Peterson *et al.* 2019; Pope 2019). In these model-based approaches, allelic frequencies or genetic distances are directly modeled as a function of landscape predictor variables and covariance relationships. These frameworks allow individual covariates to be explicitly modeled,



provide estimates of parameter uncertainty, and facilitate model selection. Currently proposed model-based approaches can be computationally demanding, and/or relevant for certain types of genetic data (e.g., bi-allelic loci), which perhaps has limited their uptake and broader application in landscape genetic analyses. Flexible and computationally efficient model-based methods are the future of landscape genetic analyses. As a young and still developing field, landscape genetics has a relatively finite lens through which to ask and answer spatial genetic questions (Dyer 2015). Models and methods developed to maximally leverage the amount and types of spatial and genetic data currently available to researchers will undoubtedly lead to tremendous advances and novel questions.

### *Conclusions*

Landscape genetics has been a rapidly evolving field, always striving to improve the power and precision of its quantitative inferences. Unfortunately, the necessary transition from Mantel-based models to more statistically defensible regression-based models has resulted in analyses that likely lack ecological interpretation and are potentially subject to statistical artifacts, such as suppression. Effective distances calculated separately on individual landscape resistance surfaces should not be used as independent predictor variables in regression models. Doing so creates an artificial caricature of the movement process that generated the observed patterns of genetic diversity across the landscape. The desire to understand the individual contributions of landscape features to movement is well-intentioned, but misuse of regression-based methods such as MRDM and MLPE, is likely to lead to inaccurate or incomplete interpretations of model results.

## References

- Adamack AT, Gruber B (2014) PopGenReport: simplifying basic population genetic analyses in R. *Methods in Ecology and Evolution* **5**, 384–387.
- Balkenhol N, Holbrook JD, Onorato D, *et al.* (2014) A multi-method approach for analyzing hierarchical genetic structures: a case study with cougars *Puma concolor*. *Ecography*, no-no.
- Balkenhol N, Waits LP, Dezzani R (2009) Statistical approaches in landscape genetics: an evaluation of methods for linking landscape and genetic data. *Ecography* **32**, 818–830.
- Blair C, Jiménez Arcos VH, Mendez de la Cruz FR, Murphy RW (2013) Landscape genetics of leaf-toed geckos in the tropical dry forest of Northern Mexico. *PLoS ONE* **8**.
- Bradburd GS, Ralph PL, Coop GM (2013) Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution* **67**, 3258–3273.
- Clarke RT, Rothery P, Raybould AF (2002) Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. *Journal of Agricultural, Biological, and Environmental Statistics* **7**, 361–372.
- Coster SS, Babbitt KJ, Cooper A, Kovach AI (2015) Limited influence of local and landscape factors on fine-scale gene flow in two pond-breeding amphibians. *Molecular Ecology* **24**, 742–758.
- Cushman S, Wasserman T, Landguth E, Shirk A (2013) Re-evaluating causal modeling with Mantel tests in landscape genetics. *Diversity* **5**, 51.
- Cushman SA, McKelvey KS, Hayden J, Schwartz MK (2006) Gene flow in complex landscapes: Testing multiple hypotheses with causal modeling. *American Naturalist* **168**, 486–499.
- Dyer RJ (2015) Is there such a thing as landscape genetics? *Molecular Ecology* **24**, 3518–3528.
- Emel SL, Storfer A (2015) Landscape genetics and genetic structure of the southern torrent salamander, *Rhyacotriton variegatus*. *Conservation Genetics* **16**, 209–221.
- Franckowiak RP, Panasci M, Jarvis KJ, *et al.* (2017) Model selection with multiple regression on distance matrices leads to incorrect inferences. *PLoS ONE* **12**, e0175194.
- Goldberg CS, Waits LP (2010) Comparative landscape genetics of two pond-breeding amphibian species in a highly modified agricultural landscape. *Molecular Ecology* **19**, 3650–3663.
- Guillot G, Rousset F (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution* **4**, 336–344.

224 Hanks EM (2017) Modeling spatial covariance using the limiting distribution of spatio-temporal random  
225 walks. *Journal of the American Statistical Association* **112**, 497–507.

226 Hanks EM, Hooten MB (2013) Circuit theory and model-based inference for landscape connectivity.  
227 *Journal of the American Statistical Association* **108**, 22–33.

228 Kozakiewicz CP, Burridge CP, Funk WC, *et al.* (2019) Urbanization reduces genetic connectivity in  
229 bobcats (*Lynx rufus*) at both intra- and inter-population spatial scales. *Molecular Ecology* **28**,  
230 5068–5085.

231 Landguth EL, Bearlin A, Day CC, Dunham J (2016) CDMetaPOP: an individual-based, eco-evolutionary  
232 model for spatially-explicit simulation of landscape demogenetics. *Methods in Ecology and*  
233 *Evolution* **8**, 4–11.

234 Landguth EL, Cushman SA (2009) CDPOP: A spatially explicit cost distance population genetics  
235 program. *Molecular Ecology Resources* **10**, 156–161.

236 Legendre P, Fortin M-J, Borcard D (2015) Should the Mantel test be used in spatial analysis? *Methods in*  
237 *Ecology and Evolution* **6**, 1239–1247.

238 Legendre P, Fortin MJ (2010) Comparison of the Mantel test and alternative approaches for detecting  
239 complex multivariate relationships in the spatial analysis of genetic data. *Molecular Ecology*  
240 *Resources* **10**, 831–844.

241 Legendre P, Lapointe F-J, Casgrain P (1994) Modeling brain evolution from behavior: A permutational  
242 regression approach. *Evolution* **48**, 1487–1499.

243 McGarigal K, Wan HY, Zeller KA, Timm BC, Cushman SA (2016) Multi-scale habitat selection  
244 modeling: a review and outlook. *Landscape Ecology*.

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

246 McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations.  
247 *Proceedings of the National Academy of Sciences of the United States of America* **104**, 19885–  
248 19890.

249 Milanesi P, Holderegger R, Caniglia R, Fabbri E, Randi E (2016) Different habitat suitability models  
250 yield different least-cost path distances for landscape genetic analysis. *Basic and Applied Ecology*  
251 **17**, 61–71.

252 Mims MC, Phillipsen IC, Lytle DA, Hartfield Kirk EE, Olden JD (2015) Ecological strategies predict  
253 associations between aquatic and genetic connectivity for dryland amphibians. *Ecology* **96**, 1371–  
254 1382.

- 255 Murphy MA, Evans JS, Storfer A (2010) Quantifying *Bufo boreas* connectivity in Yellowstone National  
256 Park with landscape genetics. *Ecology* **91**, 252–261.
- 257 Paulhus DL, Robins RW, Trzesniewski KH, Tracy JL (2004) Two Replicable Suppressor Situations in  
258 Personality Research. *Multivariate Behavioral Research* **39**, 303–328.
- 259 Peterman WE (2018) ResistanceGA: An R package for the optimization of resistance surfaces using  
260 genetic algorithms. *Methods in Ecology and Evolution* **9**, 1638–1647.
- 261 Peterson EE, Hanks EM, Hooten MB, Ver Hoef JM, Fortin M-J (2019) Spatially-structured statistical  
262 network models for landscape genetics. *Ecological Monographs* **89**.
- 263 Pope N (2019) *Genetic mark-recapture provides insights into beemovement and plant reproductive*  
264 *success*, University of Texas-Austin.
- 265 Prunier JG, Colyn M, Legendre X, Flamand M-C (2017) Regression commonality analyses on  
266 hierarchical genetic distances. *Ecography* **40**, 1412–1425.
- 267 Prunier JG, Colyn M, Legendre X, Nimon KF, Flamand MC (2015) Multicollinearity in spatial genetics:  
268 separating the wheat from the chaff using commonality analyses. *Molecular Ecology* **24**, 263–  
269 283.
- 270 Row JR, Oyler-McCance SJ, Fike JA, *et al.* (2015) Landscape characteristics influencing the genetic  
271 structure of greater sage-grouse within the stronghold of their range: a holistic modeling  
272 approach. *Ecology and Evolution* **5**, 1955–1969.
- 273 Shirk AJ, Landguth EL, Cushman SA (2017) A comparison of regression methods for model selection in  
274 individual-based landscape genetic analysis. *Molecular Ecology Resources* **18**, 55–67.
- 275 Shirk AJ, Wallin DO, Cushman SA, Rice CG, Warheit KI (2010) Inferring landscape effects on gene  
276 flow: A new model selection framework. *Molecular Ecology* **19**, 3603–3619.
- 277 Spear SF, Balkenhol N, Fortin MJ, McRae BH, Scribner K (2010) Use of resistance surfaces for  
278 landscape genetic studies: considerations for parameterization and analysis. *Molecular Ecology*  
279 **19**, 3576–3591.
- 280 Spear SF, Storfer A (2008) Landscape genetic structure of coastal tailed frogs (*Ascaphus truei*) in  
281 protected vs. managed forests. *Molecular Ecology* **17**, 4642–4656.
- 282 Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we  
283 now? *Molecular Ecology* **19**, 3496–3514.
- 284 Trumbo D, Salerno P, Logan K, *et al.* (2019) Urbanization impacts apex predator gene flow but not  
285 genetic diversity across an urban-rural divide. *Molecular Ecology* **28**, 4926–4940.

286 Trumbo DR, Spear SF, Baumsteiger J, Storfer A (2013) Rangewide landscape genetics of an endemic  
287 Pacific northwestern salamander. *Molecular Ecology* **22**, 1250–1266.

288 van Strien MJ, Keller D, Holderegger R (2012) A new analytical approach to landscape genetic  
289 modelling: least-cost transect analysis and linear mixed models. *Molecular Ecology* **21**, 4010–  
290 4023.

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

294 Wang IJ, Bradburd GS (2014) Isolation by environment. *Molecular Ecology* **23**, 5649–5662.

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

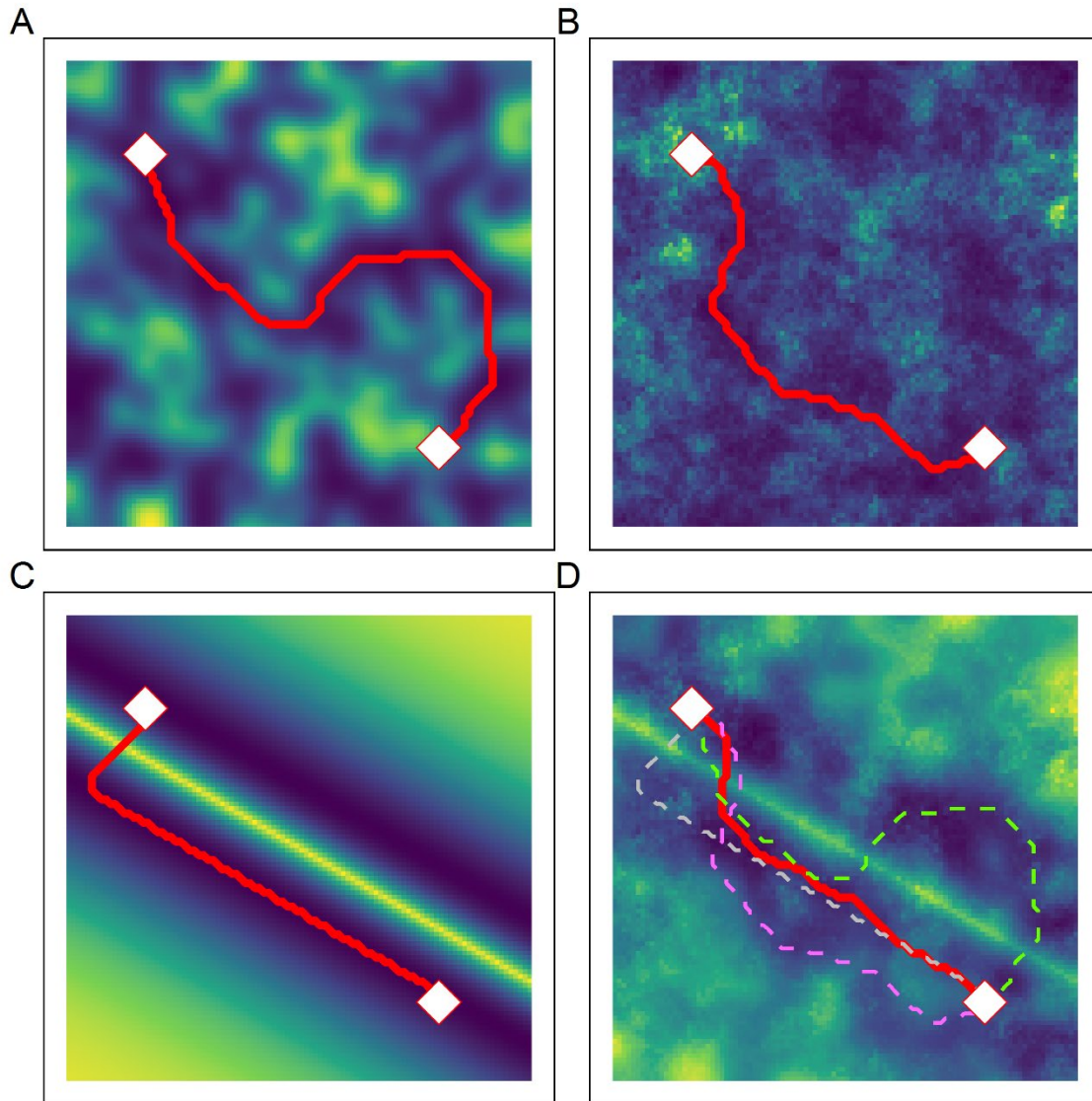
298 Zeller K, McGarigal K, Whiteley A (2012) Estimating landscape resistance to movement: A review.  
299 *Landscape Ecology* **27**, 777–797.

300 Zeller KA, Creech TG, Millette KL, *et al.* (2016) Using simulations to evaluate Mantel-based methods for  
301 assessing landscape resistance to gene flow. *Ecology and Evolution* **6**, 4115–4128.  
302

303    **Data Accessibility**

304    Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## Figures



**Figure 1.** Least cost paths between two locations on a smoothed categorical surface (A), a random Gaussian surface (B), distance from feature (C), and the composite resistance surface (D) that results from summing A–C. In all panels, resistance goes from low to high as the color transitions from blue to green, and green to yellow. The solid red line in each panel indicates the least cost path for that surface. The dashed lines in panel D are the least cost paths for each of the three component resistance surfaces.