PLOS ONE

RESEARCH ARTICLE

Model selection with multiple regression on distance matrices leads to incorrect inferences

Ryan P. Franckowiak[1]*, Michael Panasci[2], Karl J. Jarvis[3], Ian S. Acuña-Rodriguez[4,5], Erin L. Landguth[6], Marie-Josée Fortin[7], Helene H. Wagner[7]

1 Environmental & Life Sciences Graduate Program, Trent University, Peterborough, Ontario, Canada, **2** Department of Natural Resources Management, Texas Tech University, Lubbock, Texas, United States of America, **3** Department of Biology, Southern Utah University, Cedar City, Utah, United States of America, **4** Centro de Ecología Molecular y Aplicaciones Evolutivas en Agroecosistemas (CEM), Instituto de Ciencias Biológicas, Universidad de Talca, Talca, Chile, **5** Departamento de Biología, Facultad de Ciencias, Universidad de La Serena, La Serena, Chile, **6** Division of Biological Sciences, University of Montana, Missoula, Montana, United States of America, **7** Department of Ecology & Evolutionary Biology, University of Toronto, Toronto, Ontario, Canada

* ryanfranckowiak@trentu.ca

Citation: Franckowiak RP, Panasci M, Jarvis KJ, Acuña-Rodriguez IS, Landguth EL, Fortin M-J, et al. (2017) Model selection with multiple regression on distance matrices leads to incorrect inferences. PLoS ONE 12(4): e0175194. https://doi.org/10.1371/journal.pone.0175194

Editor: Duccio Rocchini, Universita degli Studi di Trento, ITALY

Received: October 8, 2016

Accepted: March 22, 2017

Published: April 13, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: Funding for this research comes from Discovery Grants by the National Science and Engineering Research Council of Canada (NSERC) to H. Wagner and to M-J. Fortin.

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

Abstract

In landscape genetics, model selection procedures based on Information Theoretic and Bayesian principles have been used with multiple regression on distance matrices (MRM) to test the relationship between multiple vectors of pairwise genetic, geographic, and environmental distance. Using Monte Carlo simulations, we examined the ability of model selection criteria based on Akaike's information criterion (AIC), its small-sample correction (AICc), and the Bayesian information criterion (BIC) to reliably rank candidate models when applied with MRM while varying the sample size. The results showed a serious problem: all three criteria exhibit a systematic bias toward selecting unnecessarily complex models containing spurious random variables and erroneously suggest a high level of support for the incorrectly ranked best model. These problems effectively increased with increasing sample size. The failure of AIC, AICc, and BIC was likely driven by the inflated sample size and different sum-of-squares partitioned by MRM, and the resulting effect on delta values. Based on these findings, we strongly discourage the continued application of AIC, AICc, and BIC for model selection with MRM.

Introduction

A primary goal of landscape genetics is to determine the relative influence of landscape composition (e.g., amount of habitat), configuration (spatial arrangement of habitat patches), and matrix quality (landscape between habitat patches) on patterns of gene flow, genetic discontinuities and population genetic structure [1–5]. Gene flow may be restricted by geographic distance (isolation-by-distance) and by resistance of land-cover types to movement (isolation-by-resistance). Because gene flow depends on what lies between patches and not

the conditions within patches (sampling locations), hypotheses are expressed in terms of pairwise distances between patches [6]. While the genetic data are collected within patches, genetic differentiation resulting from restricted gene flow is quantified in terms of pairwise genetic distances. Hypotheses concerning the association of pairwise distances between sampling units (i.e., genetic, geographic, environmental, or temporal distances) are often analyzed using Mantel tests [7] or its derivatives, such as partial Mantel test [8] and multiple regression with distance matrices (MRM) ([9–11], for examples see [12–14]). Competing hypotheses are typically defined in either of two ways: (1) each hypothesis is represented by a single distance matrix D_x that integrates hypothesized effects of multiple landscape features, or (2) each factor p is represented by its own distance matrix D_p and each hypothesis is defined by a set of predictor matrices [6,10,15]. Various model selection approaches have been proposed for identifying the model that best explains the observed spatial genetic structure and assessing the level of support for each competing hypothesis [8,16–23], but the accuracy and reliability of these approaches remain a topic of considerable debate in the context of spatial analysis (e.g., [24,25]).

Model selection procedures based on Akaike's information criterion (AIC) [26], its small sample size correction (AICc) [27], and the Bayesian information criterion (BIC) [28] have been suggested as a potential alternative to traditional statistical hypothesis testing for analyzing landscape genetic data [5,29], and these methods have been used increasingly with the Mantel test [30–33] and MRM [23,34–41]. AIC and AICc are information theoretic indices and aim to identify the fitted model with the minimum loss of Kullback-Leibler (K-L) information compared to the full reality, whereas BIC aims to identify the model with the fewest parameters that is nearest to the truth as measured by K-L distance [17,18]. In practice, AIC has a tendency to include too many predictors (overfitting) irrespective of sample size, whereas BIC has a tendency towards underfitting that increases with sample size [42]. AIC, AICc, and BIC values are not directly interpretable due to unknown scaling constants and strong dependence on sample size, but instead rely on delta Δ_i values, which represent the difference in AIC, AICc, or BIC values between candidate model i and the selected best model (i.e., $\Delta_i = AIC_i - AIC_{min}$), and provide a quantitative measure of support for each competing hypothesis [17,18]. In situations where more than one model from the candidate set of models is supported by the data, model averaging procedures may be used based on model weights $w_i = \exp(-1/2\,\Delta_i)/\Sigma_{r=1}^{R}\exp(-1/2\,\Delta_r)$ ([17], p. 75). Because each model i is weighted with respect to all other models r in the entire set of candidate models R, model averaging generally results in more robust parameter estimates and model predictions [17,18].

A linear relationship r_{xy} between two normally distributed variables x and y observed at n sampling locations translates into a linear relationship between two vectors of pairwise distances D_x and D_y, where each element in D_x is the difference $(x_j - x_i)$ between two values of x observed at locations i and j, with a linear (Mantel) correlation between D_x and D_y slightly smaller than r^2_{xy} [43]. Here, we refer to the analysis of the relationship between x and y as *node-based* analysis, and the analysis of the relationship between D_x and D_y as *distance-based* analysis [6]. Mantel tests evaluate the (full or partial) correlation between D_x and D_y, whereas MRM performs regression analysis of D_y on one or more predictors D_x. While distance-based analysis is a round-about and inefficient way for assessing the linear relationship between x and y where node-based analysis can be applied, it is useful in cases where the predictor variable exists only in the form of pairwise differences [44]. In the case of hypotheses about landscape resistance to gene flow, the ecological distance between two sampling locations (predictor variable D_x) depends on the resistance values of all land-cover types between the two locations, not on the values at the sampling locations.

MRM as a distance-based analysis differs from standard, node-based regression analysis in important ways [45], as it tests the relationship between two or more vectors of $N = n(n - 1)/2$ unique distance values derived from n independent observations. Thus, values are not independent, as each of the n original observations will contribute to $n - 1$ of the N values in the distance vector. This leads to several complications. (1) Due to the non-independence of pairwise observations, statistical significance tests must be based on appropriate permutation tests rather than parametric procedures (e.g., [43])–this is now routinely implemented. (2) Spatial autocorrelation may further jeopardize statistical significance testing by inflating type I error rates [24,25]. (3) MRM minimizes a different residual sum-of-squares (RSS) than linear regression of the node-based data from which the pairwise distances were derived [45]. Hence, even if based on the same original data, we should not expect to find the same parameter estimates. Indeed, the Mantel correlation r_M, calculated from the N pairwise distances, is generally much lower than the corresponding Pearson correlation r calculated from the n original values [44,45]. (4) The non-independence of pairwise observations invalidates the use of AIC and similar measures in MRM [22]. This problem cannot be easily fixed by adjusting for inflated sample size, as the true degrees of freedom in distance matrices are unknown [22,46]. Hence, adjusting the sample size in the calculation of AIC, AICc, and BIC is not a recommended strategy. While these issues are known to exist, there is a lack of research that would allow authors and reviewers to judge the severity of the consequences of using AIC, AICc or BIC with MRM to assess the empirical support for competing models.

In this study, we used a simple Monte Carlo simulation approach to evaluate the behavior and performance of AIC, AICc, and BIC when applied with MRM. Rather than mimicking the full complexity, e.g., of landscape genetic data, we present an artificially ideal situation, where pairwise distances are derived from node-based data simulated as multivariate normal variables with known linear correlation structure and without complicating factors, such as spatial autocorrelation or collinearity, among predictor variables. This approach allowed us to use the results from node-based analysis as a benchmark for the results from distance-based analysis. We determined the ability of AIC, AICc, and BIC to (1) identify and provide empirical support (i.e., delta values Δ_i and model weights w_i) for the correct, single-predictor model when confronted with a candidate set of models containing an increasing number of spurious predictors, and (2) identify the correct model with multiple predictors varying in strength of correlation (i.e., tapering effects) with the response variable. This study aims to address, in part, a current research priority in landscape genetics, which is to (1) evaluate how well various analytical approaches perform at identifying the relevant factors controlling gene flow in complex landscapes, (2) determine under what conditions they perform reliably, and (3) assess how they are affected by common violations of assumptions.

Methods

Using a Monte Carlo simulation approach, we evaluate the ability of AIC, AICc, and BIC to identify the correct model when applied to MRM on distance transformed data. Simulations allowed us to compare multiple linear regression and MRM analysis under conditions where the relationship between the response and predictor variables in each simulated dataset were known. Simulated data sets consisted of six random variables sampled from a multivariate normal distribution with zero means and a pre-specified covariance matrix, where all diagonal values (variances) were set to one and all off-diagonal terms (covariances, i.e., expected correlations ρ_{ij}) were set to zero unless specified otherwise below. First, node-based data were generated using the *mvrnorm* function in the MASS package [47] in R [48]. Each data set included a single response variable y and five predictor variables $x_1 - x_5$, with all predictors

being independent of each other (i.e., no collinearity) and without spatial autocorrelation. These conditions present an ideal situation, where the model selection methods we wanted to test would be most likely to perform well. In a second step, node-based data were transformed into distance-based data. Data were simulated under two scenarios: (1) with a single meaningful predictor and four spurious predictors, and (2) with three meaningful predictors with decreasing effects on y and two spurious predictors.

Node-based analysis–In the first series of simulations, we examined the ability of AIC, AICc, and BIC to correctly penalize for spurious predictors, i.e., to select the correct model containing a single meaningful predictor variable and zero independent random variables, $y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$, when confronted with a candidate set of models containing an increasing number of independent random variables. In each simulation run, the single meaningful predictor x_1 was assigned an expected correlation of $\rho_{xy} = 0.60$ with the response y, whereas variables $x_2 - x_5$ were simulated to be independent of y ($\rho_{xy} = 0.0$). Following Peres-Neto et al. [49], we generated four incorrect models by sequentially adding four additional independent random variables $x_2 - x_5$ to the correct model with the meaningful predictor x_1.

In the second series of simulations, we examined the ability of AIC, AICc, and BIC to identify the correct model across different levels of strength of correlation with the response variable. Specifically, we assessed the performance at correctly ranking models containing weak tapering effects incorporated in the data; keeping in line with the variable selection problem which is often the focus of multiple regression analysis [17,18]. We modified the correlation matrix used to generate the simulated data sets assigning expected correlation values of $\rho_{xy} = 0.30, 0.25, 0.20$ to variables $x_1 - x_3$, with x_1 having the highest and x_3 the lowest correlation with y. We simulated the remaining two variables x_4 and x_5 to be independent of y ($\rho_{xy} = 0.0$). The candidate set of models contained again five models, but in this series of simulations the correct model contained three meaningful predictors and zero independent random variables, $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_i$, where the linear effects of variables $x_1 - x_3$ on y were assumed to be additive. We performed regression analyses for both node-based simulations using the *lm* function in R [48]. We calculated AIC and BIC using the basic R functions *AIC* and *BIC*, and we calculated AICc using function *AICc* of R package 'MuMIn' [50].

Distance-based analysis–The six normally distributed random variables within each node-based data set (described above) were transformed into Euclidean distance matrices using the *dist* function in the R package 'stats', from which we extracted the lower-triangle values as a vector of $N = n(n - 1)$ pairwise distances, which we subsequently analyzed with MRM. For pairs of linearly correlated normally distributed variables, we expected the correlation between distance-transformed variables to be less than the square of the node-based values [44]. The transformation into distance values can also introduce non-linearity of the relationship that further reduces linear correlation [45]. To account for the reduction in the strength of correlation caused by the distance transformation, we modified the correlation matrix to generate a second set of node-based data with higher expected correlation values of $\rho_{xy} = 0.8$ instead of 0.6, and {0.58, 0.52, 0.47} instead of {0.30, 0.25, 0.20}. These data were again transformed into a second set of Euclidean distance matrices (in this case, only the distance-based data were retained for analysis). Thus, we were able to evaluate the reliability of AIC, AICc, and BIC when applied with MRM on distance data with a reduced correlation derived directly from the raw data (low correlation set of distance vectors) and independently simulated distance data with an empirical correlation equal to the original raw data (high correlation set of distance vectors). The calculations for fitting an MRM model are no different than those for multiple regression with raw data, and thus we fitted the same five models used in the node-based analysis using the *lm* function in R [48], using the same functions to calculate AIC, AICc and BIC.

For each set of simulations, we determined the reliability with which model selection algorithms based on AIC, AICc, and BIC were able to identify the correct model when applied with MRM by the proportion of 1000 replicate data sets where we identified the correct model as the best model. To further understand how the number of observations n in the original raw data influences the behavior of AIC, AICc, and BIC, we ran each simulation with three different initial sample sizes, n = 30, 100, and 300. R code for generating and analyzing simulated data sets is available as S1 File.

Results

Simulations with an increasing number of spurious predictors

The first set of simulations resulted in empirical correlations between y and x_1 for the node-based data that varied around the expected correlation of $\rho_{xy} = 0.60$, with mean = 0.598 (sd = 0.063) based on 1000 replicate simulations with a sample size of n = 100. Transformation to distance vectors reduced the average correlation to 0.322 and increased the standard deviation of empirical correlation coefficients to 0.078. Increasing the pre-specified covariance for the second set of distance vectors resulted in a mean correlation of distance vectors of 0.604 (sd = 0.062), closely matching the properties of the node-based correlations.

The results from a typical single simulation run with n = 100 (Fig 1) illustrate how the behavior of AIC changed markedly when used with MRM on distance transformed data (patterns for AICc and BIC were similar, not shown). For the node-based analysis (left column), the correct model with a single meaningful predictor and zero independent random variables, $y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$, had the lowest absolute value (top row), and thus the lowest delta Δ_i value (middle row), with values for both increasing monotonically with each additional variable added to the model. The correct model also had the highest model weight w_i (bottom row).

When applied to MRM, the absolute values, and more importantly, the delta values Δ_i, were considerably larger than those observed for the node-based regression analysis, with similar patterns for both sets of distance vectors (center column: low correlation, right column: high correlation). As in this example, the correct model often had the largest absolute value and, contrary to expectations, the values generally decreased with each additional variable added to the model. Moreover, delta values Δ_i not only reversed the rank order of the five candidate models, but also increased more rapidly between successive models than those reported for the node-based analysis, and thus, provided high weight w_i of support (bottom row) for the incorrect model, with the remaining models, including the correct model, receiving little support.

Across 1000 replicate simulations with an increasing number of spurious variables (Fig 2), AIC, AICc, and BIC applied with multiple regression on the original raw data (node-based analysis) were able to select the correct model as the best model in the majority of simulations, and the ability of all three criteria to identify the correct model generally increased with larger sample size n. Under this scenario, BIC performed more reliably than either AIC or AICc, selecting the correct model in more than 76 percent of simulations, whereas, AIC and AICc selected the correct model in less than 57 and 68 percent of simulations, respectively. These results serve as a benchmark for distance-based analysis.

When applied to distance-based analysis MRM, AIC, AICc, and BIC exhibited a strong bias toward selecting models containing spurious effects and, more surprisingly, the severity of this bias increased markedly with larger sample size n. For simulations run with n = 300, AIC and AICc selected the model with all four additional spurious variables in more than 76 percent of simulations (for both low and high correlation), whereas BIC selected this same model in 45 and 47 percent of simulations for distance vector data generated with low and high correlation, respectively.

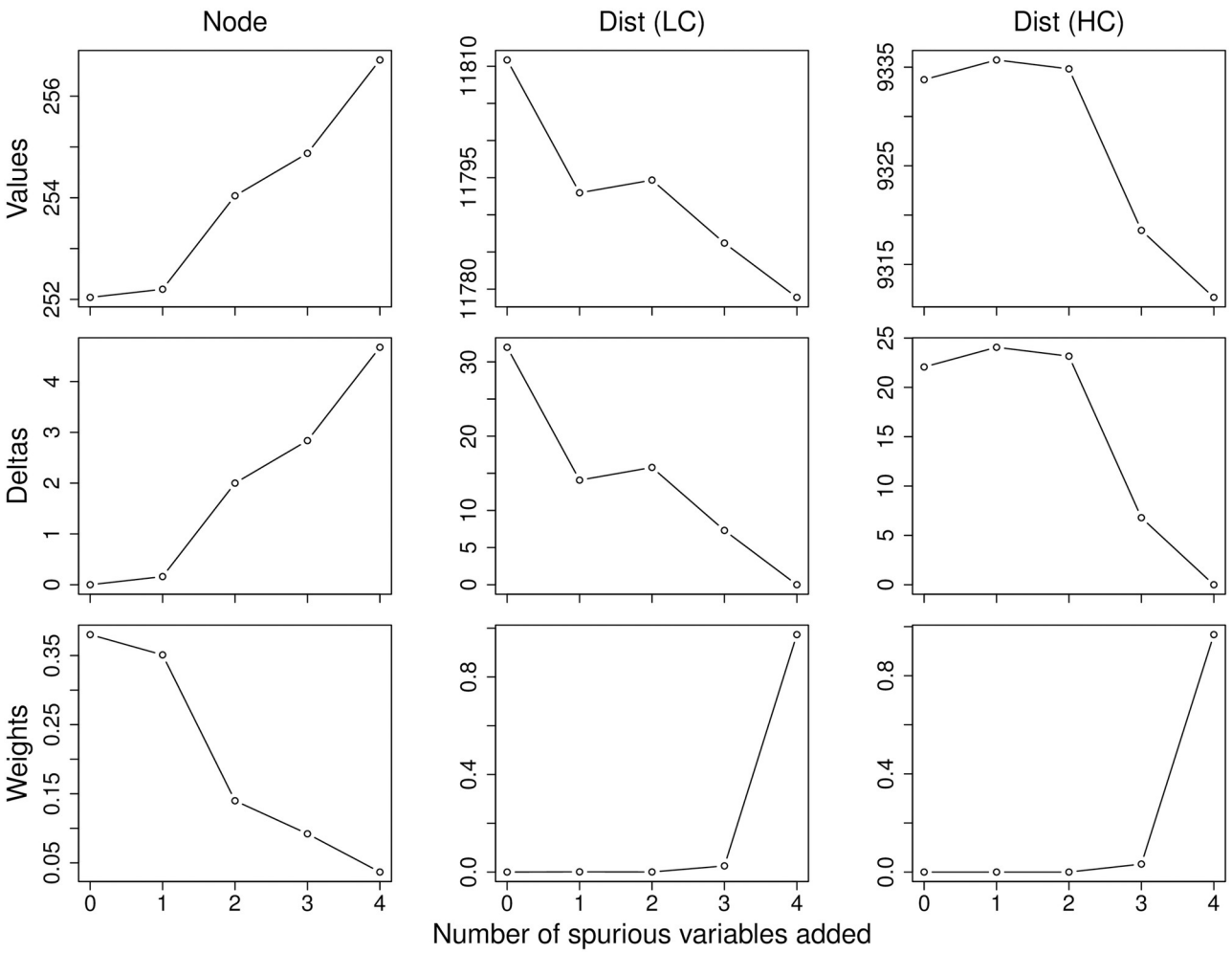

Fig 1. Results from a single simulation run. The absolute values (top row), delta values Δ_i (middle row), and model weights w_i (bottom row) for node-based analysis (Node: left column), distance-based analysis with low correlation (Dist (LC): middle column), and distance-based analysis with high correlation (Dist: (HC): right column) as a function of the number of spurious random variables added sequentially to the correct model with a single meaningful predictor x_1 ($\rho_{xy} = 0.6$) for node-based, based on $n = 100$.

https://doi.org/10.1371/journal.pone.0175194.g001

Simulations with multiple meaningful predictor variables

The second set of simulations resulted in empirical correlations between y and $x_1 - x_3$ for the node-based data that varied around the expected correlation of $\rho_{xy} = 0.30$, 0.25, and 0.20, with mean = 0.299, 0.247, and 0.198 ($sd = 0.092$, 0.094 and 0.096) based on 1000 replicate simulations with a sample size of $n = 100$. Transformation to distance vectors reduced the mean correlation to 0.078, 0.053, and 0.037 ($sd = 0.063$, 0.060, and 0.055) for the low-correlation data set (LC). Increasing the pre-specified covariance for the second set of distance vectors (HC) resulted in a mean correlation of distance vectors of 0.30, 0.24, and 0.19 ($sd = 0.080$, 0.076, and 0.077), closely matching the properties of the node-based correlations.

For the 1000 replicate simulations run with three meaningful predictors with tapering effects (Fig 3), AIC, AICc, and BIC applied with multiple regression on the original raw data (node-based analysis) exhibited considerable uncertainty in selecting the correct model, particularly for small sample size $n = 30$, though the ability of all three criteria to identify the correct model improved with larger sample size n. In simulations run with $n = 300$, AIC and AICc

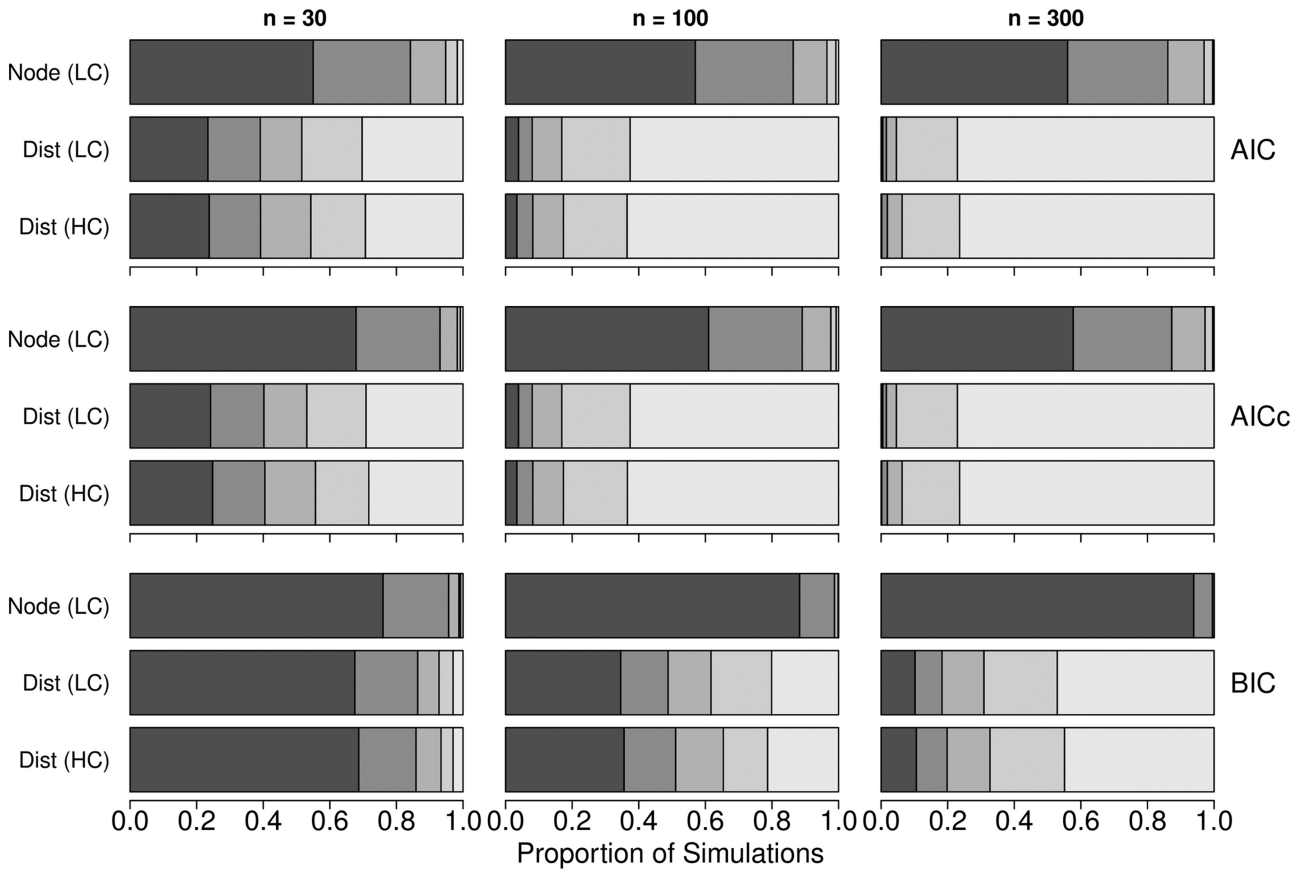

Fig 2. Proportional selection of the correct model by means of MRM among 1000 simulated data sets with a different number of spurious predictors. The proportion of 1000 simulated data sets where each of the five candidate models was selected as the best model using AIC (top row), AICc (middle row), and BIC (bottom row) with three different sample sizes of $n = 30$ (left column), $n = 100$ (middle column), $n = 300$ (right column) for the node-based analysis with low correlation (Node LC), the distance-based analysis with low correlation (Dist LC), and the distance-based analysis with high correlation (Dist HC). The correct model included only the single meaningful predictor x_1 (black), whereas, the four additional models contained the single meaningful predictor x_1 and one (dark grey), two (medium dark grey), three (medium light grey), and four (light grey) spurious variables ($x_2 - x_5$).

https://doi.org/10.1371/journal.pone.0175194.g002

performed less reliably than BIC; both measures identifying the correct model in 70 percent of simulation, whereas, BIC identified the correct model in 100 percent of simulations. Again, these results of node-based analysis provide a benchmark for the results from distance-based analysis.

When applied to distance-based analysis with MRM, AIC, AICc, and BIC were unable to reliably identify the correct model regardless of sample size n or strength of correlation used to generate the raw data. The overall performance of AIC, AICc, and BIC in distance-based analysis decreased markedly with larger sample size n. AIC and AICc exhibited a strong bias toward selecting the model containing all five variables (i.e. full model), but the bias was slightly less severe for BIC data generated with high correlation.

Discussion

When applied with MRM (distance-based analysis), model selection procedures based on AIC, AICc, and BIC were unable to reliably rank candidate models or consistently identify the correct model, and thus often led to incorrect inferences about the relationships between

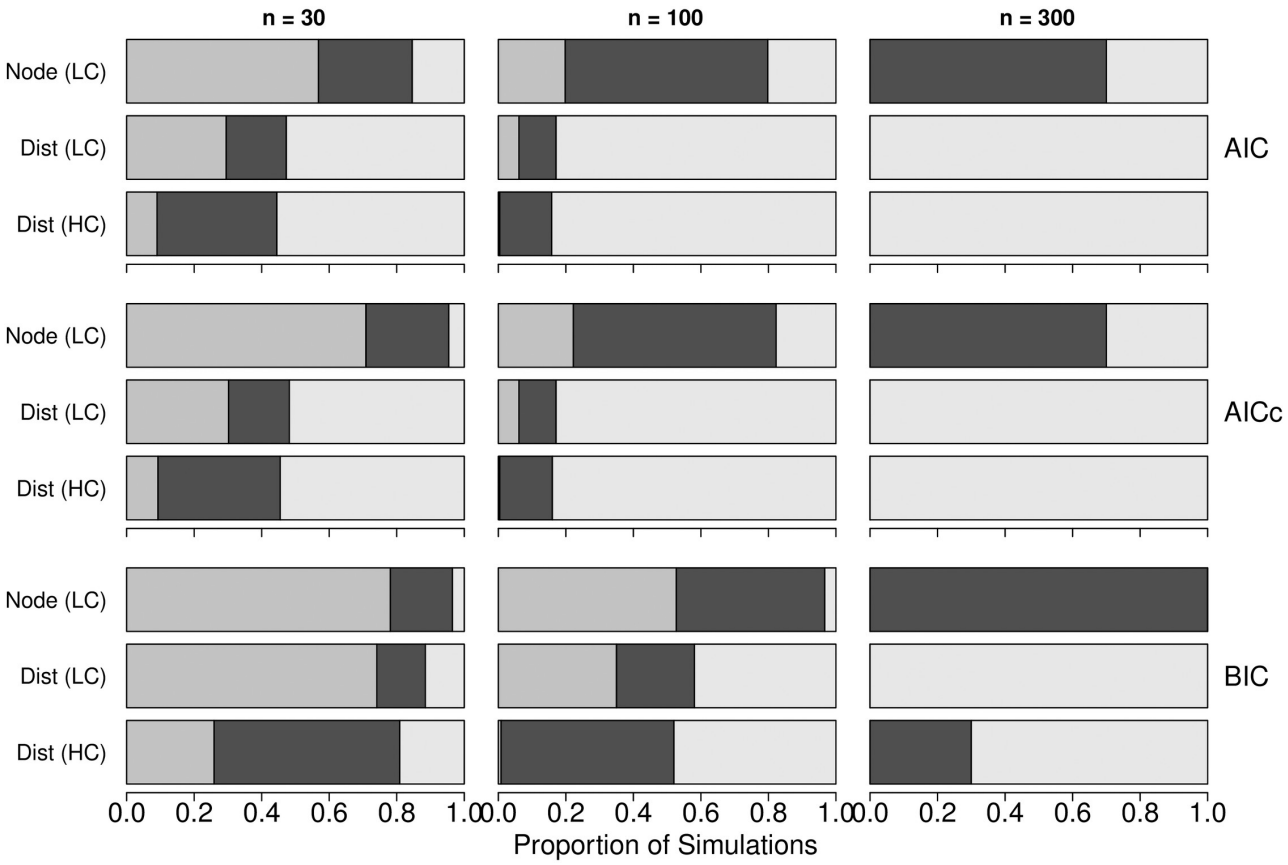

Fig 3. Proportional selection of the correct model by means of MRM among 1000 simulated data sets for different levels of correlated predictors. The proportion of 1000 simulated data sets where each of the five candidate models were selected as the best model using AIC (top row), AICc (middle row), and BIC (bottom row) with three different sample sizes $n = 30$ (left column), $n = 100$ (middle column), $n = 300$ (right column) for the node-based analysis with low correlation (Node LC), the distance-based analysis with low correlation (Dist LC), and the distance-based analysis with high correlation (Dist HC). We were primarily interested in determining whether AIC, AICc, and BIC selected the correct model containing three meaningful variables with tapering effects (black) or selected an underfitted (dark grey) or overfitted (light grey) model.

https://doi.org/10.1371/journal.pone.0175194.g003

response and predictor variables. All three criteria exhibited a systematic bias toward selecting unnecessarily complex models and indeed would often select the full model (with two or four spurious predictors included depending on the scenario) as the best model. The absolute and relative values (i.e., delta values Δ_i) of AIC, AICc, and BIC exhibited very different behavior that interfered with the ability of each criterion to correctly rank candidate models. The large delta values suggested a high level of support for the incorrectly selected best model, and thus provided little support for other models in the candidate set, including the correct model. The observed change in behavior goes beyond the general bias toward overfitting often cited for AIC (and AICc) when comparing models containing a small number of predictors with large effect [17,18]; a pattern that can be seen in the results from the node-based analysis. The observed bias became more pronounced when we increased the sample size n used to generate the original raw data, which is directly related to using $N = n(n - 1)/2$ pairwise distance values to calculate AIC, AICc, and BIC values for MRM on distance matrices. BIC performed slightly better, as its penalty increases with sample size, but this was not sufficient to correct for the problem in distance-based analysis. Manually adjusting for sample size in the calculation of AIC, AICc, and BIC may confer some improvement (S2 File; S1–S3 Figs) but cannot be

recommended as it does not adequately address the problem of unknown degrees of freedom [46]. Thus, our results suggest considerable caution should be taken when interpreting and evaluating studies that have relied on model selection with MRM to assess the relationship between landscape features and patterns of gene flow [23,34–41].

Another substantial barrier to the application of AIC, AICc, and BIC with MRM is the non-independence of pairwise distances; this violates a basic assumption of linear regression analysis [51] and can strongly bias model selection results [17,18]. Clarke et al. [52] developed the maximum-likelihood population effects model with a covariate structure to explicitly model the correlated error structure in MRM, relying on restricted maximum likelihood (REML) to generate unbiased estimates of the variance components of the mixed effects models. However, Van Strien et al. [22] stated that model selection procedures based on AIC, AICc, and BIC should not be used to compare mixed models where parameter estimation is performed using REML with different fixed effects. Autocorrelated residuals resulting from various spatial processes (e.g., isolation-by-distance; IBD), can also severely affect regression results; potentially leading to spurious correlations in genetic analyses [53]. Recent studies demonstrated that including a vector of geographic distance to account for IBD does not sufficiently remove spatial autocorrelation [24,25]. Therefore, additional research is required to develop approaches that explicitly model the correlated error structure within vectors of pairwise distances d_{ij}, as well as spatial autocorrelation among errors, and while still allowing the use of AIC, AICc, and BIC for model selection.

Depending on the nature of the data and question being addressed, a number of alternative analytical approaches are available that are not subject to the statistical issues associated with the analysis of pairwise distances [6]. Neighborhood-level approaches can be used to reduce pairwise distance matrices into node-level data vectors based on connectivity indices calculated for each focal site with all other sites within its local neighborhood; these represent either a single environmental factor or a resistance surface containing multiple factors [6]. Canonical redundancy analysis (RDA) [11], which has been shown to have greater power than Mantel-based approaches [25,45], can then be used to test for relationships between connectivity indices and measures of genetic diversity, genetic differentiation, a matrix of allele frequencies or a set of PCoA scores (distance based RDA) [54]. The functional connectivity (as measured by gene flow) through a network of observations can also be evaluated using gravity models which incorporate both at- and between-site variables, and allowing multiple parameters to be estimated from the sample data [55]. Alternatively, a predictor distance matrix may be used to define the error correlation structure in a node-based framework, using a table of allele frequencies rather than a genetic distance matrix as the response (e.g., [56]), either in a Bayesian context or with generalized linear mixed models (GLMM) [57,58]. There are, however, concerns about the validity of such covariance models [59], and it is unclear how valid model selection with multiple competing hypotheses could be performed in this type of analysis. While these approaches offer considerable promise for incorporating geographic and environmental distance into a single analysis, valid methods for statistical inference and model selection with landscape genetic data urgently require further development and evaluation.

Conclusions

The development of statistically valid methods for comparing alternative hypotheses regarding the effects of landscape features on patterns of gene flow remains an important area of research in landscape genetics. Our results clearly demonstrated that AIC, AICc, and BIC were unable to reliably rank candidate models when applied with MRM, even under artificially ideal conditions, leading to systematically incorrect inferences. While e.g. AIC is known to overfit models

in node-based analysis e.g. by including one predictor more than necessary, application to distance-based analysis typically resulted in AIC reversing the expected ranking of candidate models and preferentially selecting the full model with the maximum number of spurious predictors available. Methods for explicitly modeling the correlated error structure within vectors of pairwise distances d_{ij} resulting from non-independence of observations or spatial autocorrelation within a MRM framework are currently being explored, but additional research is needed to develop and test approaches that permit the use of AIC, AICc and/or BIC in a MRM framework [6]. Until these issues have been adequately addressed, we strongly discourage the continued used of AIC, AICc, and BIC with MRM.

Supporting information

S1 File. Simulation R code. The R code used to generate the node- and distance-based data vectors used in the simulation analysis.
(DOCX)

S2 File. Evaluation of sample-size corrected model selection criteria. Description of a simple sample-size correction for AIC, AIC, and BIC (i.e., AICd, AICcd, and BICd) and its relative performance when used with MRM on distance matrices. Corrected measures were applied to the same data used in the manuscript. This is presented for illustration only and we do not recommend application of such correction in any form. Rather, we call on statisticians to help develop valid alternatives.
(DOCX)

S1 Fig. Results from a single simulation run with sample-size corrected model selection criteria.
(TIF)

S2 Fig. Proportional selection of the correct model with sample-size corrected model selection criteria by means of MRM among 1000 simulated data sets with a different number of spurious predictors.
(TIF)

S3 Fig. Proportional selection of the correct model with sample-size corrected model selection criteria by means of MRM among 1000 simulated data sets for different levels of correlated predictors.
(TIF)

Acknowledgments

We would like to thank all participating members of the Distributed Graduate Course in Landscape Genetics 2012 for their constructive comments throughout this project, particularly Tara Crewe and Lynne Gardner for their contribution during the early development stages of this project.

Author Contributions

Conceptualization: RPF MP KJJ ISAR ELL MJF HHW.

Data curation: RPF MP KJJ ISAR ELL MJF HHW.

Formal analysis: RPF MP KJJ ISAR.

Funding acquisition: MJF HHW.

Investigation: RPF MP KJJ ISAR.

Methodology: RPF MP KJJ ISAR ELL MJF HHW.

Project administration: ELL MJF HHW.

Supervision: ELL MJF HHW.

Validation: RPF MP KJJ ISAR ELL MJF HHW.

Visualization: RPF MP KJJ ISAR.

Writing – original draft: RPF MP KJJ ISAR ELL MJF HHW.

Writing – review & editing: RPF MP KJJ ISAR ELL MJF HHW.

References

1. Manel S, Schwartz MK, Luikart G, Taberlet P. Landscape genetics: combining landscape ecology and population genetics, Trends Ecol Evol. 2003; 18: 189–197.
2. Holderegger R, Wagner HH. A brief guide to landscape genetics. Landsc Ecol. 2006; 21: 793–796.
3. Holderegger R, Wagner HH. Landscape genetics. Bioscience. 2008; 58: 199–207.
4. Storfer A, Murphy M, Evans J, Goldberg C, Robinson S, Spear S, et al. Putting the 'landscape' in landscape genetics. Heredity. 2007; 98: 128–142. https://doi.org/10.1038/sj.hdy.6800917 PMID: 17080024
5. Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP. Landscape genetics where are we now? Mol Ecol. 2010; 19: 3496–3514. https://doi.org/10.1111/j.1365-294X.2010.04691.x PMID: 20723061
6. Wagner HH, Fortin MJ. A conceptual framework for the spatial analysis of landscape genetic data. Conserv Genet. 2013; 14: 253–261.
7. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Res. 1967; 27: 209–220. PMID: 6018555
8. Smouse PE, Long JC, Sokal RR. Multiple regression and correlation extensions of the Mantel test of matrix correspondence. Syst Zool. 1986; 35: 627–632.
9. Manly B. Randomization and regression methods for testing for associations with geographical environmental and biological distances between populations. Res Popul Ecol. 1986; 28: 201–218.
10. Lichstein JW. Multiple regression on distance matrices: A multivariate spatial analysis tool. Plant Ecol. 2007; 188: 117–131.
11. Legendre P, Legendre L. Numerical Ecology. 3rd ed. San Diego: Elsevier Science & Technology Books; 2012.
12. Krackhardt D. Predicting with networks: Nonparametric multiple regression analysis of dyadic data. Soc Networks. 1988; 10: 359–381.
13. Leduc A, Drapeau P, Bergeron Y, Legendre P. Study of spatial components of forest cover using partial mantel tests and path analysis. J. Veg. Sci. 1992; 3: 69–78.
14. Nantel P, Neumann P. Ecology of ectomycorrhizal-basidiomycete communities on a local vegetation gradient. Ecology. 1992; 73: 99–117.
15. Spear SF, Balkenhol N, Fortin MJ, McRae BH, Scribner KT. (2010). Use of resistance surfaces for landscape genetic studies: Considerations for parameterization and analysis. Molecular Ecology 2010; 19: 3576–3591. https://doi.org/10.1111/j.1365-294X.2010.04657.x PMID: 20723064
16. Legendre P, Troussellier M. Aquatic heterotrophic bacteria: modeling in the presence of spatial autocorrelation. Limnol Oceanogr. 1988; 33: 1055–1067.
17. Burnham KP, Anderson DR. Model selection and multimodel inference: A practical information-theoretic approach. Heidelberg and New York: Springer; 2002.
18. Burnham KP, Anderson DR. Multimodel inference: Understanding AIC and BIC in model selection. Sociol Methods Res. 2004; 33: 261–304.
19. Cushman SA, McKelvey KS, Hayden J, Schwartz MK. Gene flow in complex landscapes: testing multiple hypotheses with causal modeling. Am Nat. 2006; 168: 486–499. https://doi.org/10.1086/506976 PMID: 17004220
20. Cushman SA, Wasserman TN, Landguth EL, Shirk AJ. Re-evaluating causal modeling with Mantel tests in landscape genetics. Diversity. 2013; 5: 51–72.

21. Shirk A, Wallin D, Cushman S, Rice C, Warheit K. Inferring landscape effects on gene flow: a new model selection framework. Mol Ecol. 2010; 19: 3603–3619. https://doi.org/10.1111/j.1365-294X.2010.04745.x PMID: 20723066

22. Van Strien MJ, Keller D, Holderegger R. A new analytical approach to landscape genetic modelling: Least-cost transect analysis and linear mixed models. Mol Ecol. 2012; 21: 4010–4023. https://doi.org/10.1111/j.1365-294X.2012.05687.x PMID: 22738667

23. Dudaniec RY, Spear SF, Richardson JS, Storfer A. Current and historical drivers of landscape genetic structure differ in core and peripheral salamander populations. PLoS ONE, 2012; 7: 36769. http://dx.doi.org/10.1371/journal.pone.0036769.

24. Guillot G, Rousset F. Dismantling the Mantel tests. Methods Ecol Evol. 2013; 4: 336–344.

25. Legendre P, Fortin MJ, Borcard D. Should the Mantel test be used in spatial analysis? Methods Ecol Evol. 2015; 6: 1239–1247.

26. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr, 1974; 19: 716–723.

27. Hurvich CM, Tsai C. Regression and time series model selection in small samples. Biometrika, 1989; 76: 297–307.

28. Schwarz GE. Estimating the dimension of a model. Ann Stat. 1978; 6: 461–464.

29. Balkenhol N, Gugerli F, Cushman SA, Waits LP, Coulon A, Arntzen JW, et al. Identifying future research needs in landscape genetics: where to from here? Landsc Ecol. 2009; 24: 455–463.

30. Roach JL, Stapp P, Van Horne B, Antolin MF. Genetic structure of a metapopulation of black-tailed prairie dogs. J. Mammal. 2001; 82: 946–959.

31. Spear SF, Peterson CR, Matocq MD, Storfer A. Landscape genetics of the blotched tiger salamander (*Ambystoma tigrinum melanostictum*). Mol Ecol. 2005; 14: 2553–2564. https://doi.org/10.1111/j.1365-294X.2005.02573.x PMID: 15969734

32. Wang IJ. Fine-scale population structure in a desert amphibian: Landscape genetics of the black toad (*Bufo exsul*). Mol Ecol. 2009; 18: 3847–3856. https://doi.org/10.1111/j.1365-294X.2009.04338.x PMID: 19708887

33. Yang J, Jiang Z, Zeng Y, Turghan M, Fang H, Li C. Effect of anthropogenic landscape features on population genetic differentiation of Przewalski's gazelle: Main role of human settlement. PLoS ONE. 2011; 6: 20144. http://dx.doi.org/10.1371/journal.pone.0020144.

34. Emaresi G, Pellet J, Dubey S, Hirzel A, Fumagalli L. Landscape genetics of the alpine newt (*Mesotriton alpestris*) inferred from a strip-based approach. Conserv Genet. 2011; 12: 41–50.

35. Jaquiéry J, Broquet T, Hirzel AH, Yearsley J, Perrin N. Inferring landscape effects on dispersal from genetic distances: How far can we go? Mol Ecol. 2011; 20: 692–705. https://doi.org/10.1111/j.1365-294X.2010.04966.x PMID: 21175906

36. Igawa T, Oumi S, Katsuren S,Sumida M. Population structure and landscape genetics of two endangered frog species of genus *Odorrana*: different scenarios on two islands. Heredity. 2012; 110: 46–56. https://doi.org/10.1038/hdy.2012.59 PMID: 22990312

37. Engler JO, Balkenhol NN, Filz KJ, Habel JC, Rödder D. Comparative landscape genetics of three closely related sympatric hesperid butterflies with diverging ecological traits. PLoS ONE. 2014; 9: e106526. http://dx.doi.org/10.1371/journal.pone.0106526. PMID: 25184414

38. Fitzpatrick SW, Crocket H, Funk WC. Water availability strongly impacts population genetic patterns of an imperiled Great Plains endemic fish. Conserv Genet. 2014; 15: 771–788.

39. Medley KA, Jenkins DG, Hoffman EA. Human-aided and natural dispersal drive gene flow across the range of an invasive mosquito. Mol Ecol. 2014; 24: 1–12.

40. Paz A, Ibáñez R, Lips KR, Crawford AJ. Testing the role of ecology and life history in structuring genetic variation across a landscape: a trait-based phylogeographic approach. Mol Ecol. 2015; 24: 3723–3737. https://doi.org/10.1111/mec.13275 PMID: 26080899

41. Jenkins DA, Lecomte N, Schaefer JA, Olsen SM, Swingedouw D, Cote SD, et al. Loss of connectivity among island-dwelling Peary caribou following sea ice decline. Biol. Lett. 2016; 12: 20160235. https://doi.org/10.1098/rsbl.2016.0235 PMID: 27651531

42. Dziak JJ, Coffman DL, Lanza ST, Li R. Sensitivity and specificity of information criteria. The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University. 2012; 1–10.

43. Dutilleul P, Jason D, Stockwell-Frigon D, Legendre P. The Mantel test versus Pearson's correlation analysis: Assessment of the differences for biological and environmental studies. J Agric Biol Environ Stat. 2000; 5: 131–150.

44. Legendre P, Fortin MJ. Comparison of the Mantel test and alternative approaches for detecting complex multivariate relationships in the spatial analysis of genetic data. Mol Ecol Res. 2010; 10: 831–844.

45. Dow MM, Cheverud JM, Friedlaender JS. Partial correlation of distance matrices in studies of population structure. Am J Phys Anthropol. 1987; 72: 343–352. https://doi.org/10.1002/ajpa.1330720307 PMID: 3578497

46. Legendre P, Fortin MJ. Spatial pattern and ecological analysis. Vegetatio. 1989; 80:107–138.

47. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth Edition. Springer, New York. 2002.

48. R Core Team. R: A language and environment for statistical computing R Foundation for Statistical Computing, Vienna. 2016. http://www.R-project.org

49. Peres-Neto PR, Legendre P, Dray S, Borcard D. Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology. 2006; 87: 2614–2625. PMID: 17089669

50. Bartoń K. MuMIn: Multi-model inference R-package v.1140. 2013. http://CRANR-projectorg/package=MuMIn

51. Sokal RR, Rohlf FJ. Biometry: the principles and practice of statistics in biological research. 4th ed. New York: WH Freeman and Co.; 2012.

52. Clarke RT, Rothery P, Raybould AF. Confidence limits for regression relationships between distance matrices: Estimating gene flow with distance. J Agric Biol Environ Stat. 2002; 7: 361–372.

53. Meirmans PG, The trouble with isolation by distance. Mol Ecol. 2012; 21: 2839–2846. https://doi.org/10.1111/j.1365-294X.2012.05578.x PMID: 22574758

54. Legendre P, Anderson MJ. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol Monogr. 1999; 69: 1–24.

55. Murphy MA, Dezzani R, Pilliod DS, Storfer A. Landscape genetics of high mountain frog metapopulations. Mol Ecol. 2010; 19: 3634–3649. https://doi.org/10.1111/j.1365-294X.2010.04723.x PMID: 20723055

56. Bradburd G, Ralph P, Coop GM. Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution. 2013; 67: 3258–3273. https://doi.org/10.1111/evo.12193 PMID: 24102455

57. Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith G. Mixed Effects Models and Extensions in Ecology with R, New York: Springer; 2009.

58. Galecki A, Burzykowski T. Linear mixed-effects models using R, New York: Springer-Verlag; 2013.

59. Guillot G, Schilling RL, Porcu E, Bevilacqua M. Validity of covariance models for the analysis of geographical variation. Methods Ecol Evol. 2014; 5: 329–335.