MOLECULAR ECOLOGY

Molecular Ecology (2015) **24**, 263–283 doi: 10.1111/mec.13029

INVITED REVIEWS AND SYNTHESES

Multicollinearity in spatial genetics: separating the wheat from the chaff using commonality analyses

J. G. PRUNIER,* M. COLYN,† X. LEGENDRE,‡ K. F. NIMON§ and M. C. FLAMAND*

Institut des Sciences de la Vie, Université catholique de Louvain, Croix du Sud 4, L7.07.14, 1348, Louvain-la-Neuve, Belgium, †CNRS-UMR 6553, Université de Rennes 1, Station Biologique 35380, Paimpont, France, ‡Muséum National d'Histoire Naturelle (MNHN), DPBZ, Réserve de la Haute Touche 36290, Obterre, France, §Department of Human Resource Development & Technology, University of Texas at Tyler 3900 University Blvd., HPR 226, Tyler, TX 75799, USA

Abstract

Direct gradient analyses in spatial genetics provide unique opportunities to describe the inherent complexity of genetic variation in wildlife species and are the object of many methodological developments. However, multicollinearity among explanatory variables is a systemic issue in multivariate regression analyses and is likely to cause serious difficulties in properly interpreting results of direct gradient analyses, with the risk of erroneous conclusions, misdirected research and inefficient or counterproductive conservation measures. Using simulated data sets along with linear and logistic regressions on distance matrices, we illustrate how commonality analysis (CA), a detailed variance-partitioning procedure that was recently introduced in the field of ecology, can be used to deal with nonindependence among spatial predictors. By decomposing model fit indices into unique and common (or shared) variance components, CA allows identifying the location and magnitude of multicollinearity, revealing spurious correlations and thus thoroughly improving the interpretation of multivariate regressions. Despite a few inherent limitations, especially in the case of resistance model optimization, this review highlights the great potential of CA to account for complex multicollinearity patterns in spatial genetics and identifies future applications and lines of research. We strongly urge spatial geneticists to systematically investigate commonalities when performing direct gradient analyses.

Keywords: CDPOP, commonality analysis, logistic regressions, multiple regressions on distance matrices, spurious correlations

Received 29 September 2014; revision received 24 November 2014; accepted 28 November 2014

Direct gradient analyses in spatial genetics

Spatial genetics, including both landscape and seascape genetics, is an ebullient scientific field that aims at investigating the influence of spatial heterogeneity on the spatial distribution of genetic variation (Manel *et al.* 2003; Holderegger & Wagner 2008; Guillot *et al.* 2009; Storfer *et al.* 2010). In the context of accelerating landscape fragmentation worldwide, spatial genetics allows a thorough assessment of landscape functional connectivity (With 1997) and has now emerged as a valuable way of assisting both landscape management and wildlife conservation (Segelbacher *et al.* 2010).

Correspondence: Jérôme G. Prunier, Fax: (32) 10 47 38 72; E-mail: jerome.prunier@gmail.com

Two main approaches can be used to investigate the influence of landscape (respectively, seascape) features on spatial genetic patterns (Balkenhol *et al.* 2009; Guillot *et al.* 2009): overlay methods and direct gradient analyses (*sensu* ter Braak & Prentice 1988). In overlay methods, genetic structures inferred from clustering algorithms (Pritchard *et al.* 2000; Chen *et al.* 2007; Jombart *et al.* 2008) or edge-detection methods (Monmonier 1973; Barbujani *et al.* 1989) are visually confronted to landscape patterns (e.g. Frantz *et al.* 2012; Prunier *et al.* 2014). In direct gradient analyses, regression procedures such as Mantel tests (Cushman *et al.* 2006; Shirk *et al.* 2010; Castillo *et al.* 2014), multiple regressions on distance matrices (MRDM; Legendre *et al.* 1994; Holzhauer *et al.* 2006; Lichstein 2007; Wang 2013; Balkenhol *et al.* 2014), mixed-effect models (Selkoe *et al.* 2010; Van

Strien *et al.* 2012; Peterman *et al.* 2014) or constrained ordination techniques (e.g. distance-based redundancy analyses or canonical correspondence analyses, which imply multivariate regressions; Angers *et al.* 1999; Legendre & Anderson 1999; Balkenhol *et al.* 2009; Vangestel *et al.* 2012; Orsini *et al.* 2013 are used to investigate the relative contribution of various independent variables (predictors) to the variance of a dependent (response) variable. In spatial genetics, the dependent variable is most often expressed as genetic distances between pairwise sampled individuals and populations (Prunier *et al.* 2013; Luximon *et al.* 2014). In the specific case of constrained ordination techniques, it may also be expressed as the ordination solution of the pairwise distance matrix using principal coordinate analysis (PCoA; Legendre & Anderson 1999; Orsini *et al.* 2013). Explanatory variables are most often derived from categorical (e.g. land cover) or continuous spatial data. When focusing on the effects of landscape on gene flow, explanatory variables can be either expressed as relative proportions of categorical landscape features between pairwise sampled units (transect-based approaches; Angelone *et al.* 2011; Emaresi *et al.* 2011; Van Strien *et al.* 2012; Keller *et al.* 2013), as pairwise effective distances computed from parameterized resistance surfaces with least-cost or isolation-by-resistance modelling (Adriaensen *et al.* 2003; McRae 2006; Peterman *et al.* 2014) or, in constrained ordination techniques, as site-specific environmental measures (e.g. Pilot *et al.* 2006). When focusing on the effects of local environmental conditions on site-specific attraction or productivity, explanatory variables may also be expressed as environmental dissimilarities between sampled points (Wang 2013; Pfluger & Balkenhol 2014).

When the objective of the study is to maximize the predictive power of a regression model and eventually to compile several univariate resistance surfaces into a unique weighted multivariate one (Spear *et al.* 2010; Zeller *et al.* 2012), a model fit index, quantifying the proportion of variance in the dependent variable that is explained by a given model (e.g. zero-order correlation coefficients in univariate procedures, R^2 in MRDM, R^2_β in mixed-effect models or AIC in constrained ordination techniques), is used as a criterion for stepwise regression (e.g. Legendre & Legendre 1998; Graves *et al.* 2012), hierarchical model selection (e.g. Selkoe *et al.* 2010; Emaresi *et al.* 2011; Dudaniec *et al.* 2012; Van Strien *et al.* 2012; Blair *et al.* 2013 or resistance model optimization (Shirk *et al.* 2010; Perez-Espona *et al.* 2012; Dudaniec *et al.* 2013; Castillo *et al.* 2014; Peterman *et al.* 2014). In an explanatory approach, that is, when the goal is simply to gain insight into the relative influence of explanatory variables on an observed biological response, model fit index is often only considered as a way to support the reliability of a unique (full) regression model (e.g. Wang 2013; Balkenhol *et al.* 2014; Guarnizo & Cannatella 2014; Nanninga *et al.* 2014). In both cases though, standardized regression weights (hereafter called beta weights β), or canonical coefficients in the case of constrained ordination techniques, are used to rank predictors according to their contributions in a multivariate regression equation. As pairwise genetic distances are nonindependent, significance levels of model fit and predictors are usually computed through matrix permutations (Legendre *et al.* 1994) or pseudobootstrap procedures (Worthington Wilmer *et al.* 2008).

Multicollinearity issues

There is a growing but unresolved concern about the reliability of regression procedures in correctly identifying the spatial determinants of observed genetic patterns and ruling out noninformative spatial features (Balkenhol *et al.* 2009). Several simulation studies showed that partial Mantel tests may yield poor results in both ways (Cushman & Landguth 2010; Legendre & Fortin 2010; Cushman *et al.* 2013; Graves *et al.* 2013; Guillot & Rousset 2013), and researchers are rather encouraged to use multivariate approaches (Bolliger *et al.* 2014; Guarnizo & Cannatella 2014). Whatever the approach though, beta weights or canonical coefficients, their standard errors and thus marginal statistics used to test their significance as well as model fit indices may be heavily impacted by even weak levels of multicollinearity (nonindependence) among predictors (Angers *et al.* 1999; Graham 2003; Smith *et al.* 2009; Nimon & Reio 2011; Kraha *et al.* 2012). Multicollinearity is thus likely to cause serious difficulties in properly interpreting results of direct gradient analyses (Mac Nally 2000; Nimon *et al.* 2010; Dormann *et al.* 2013). For instance, it may be difficult to identify the likely causal variables among collinear predictors showing significant correlation with the response variable (Graham 2003; Spear *et al.* 2010; Dormann *et al.* 2013). Spurious correlations are also likely to occur in the presence of suppression, that is, when a variable confounds the variance explained by another (Courville & Thompson 2001; Nimon & Reio 2011; Beckstead 2012; Ray-Mukherjee *et al.* 2014). For instance, beta weights may be negative even when the predictor and the dependent variable are positively correlated. In case of multicollinearity, a thorough understanding of the correlation structure among predictors is thus primordial (Dormann *et al.* 2013).

Multicollinearity among explanatory variables is a regular feature in ecology (Graham 2003; Dormann *et al.* 2013) and is probably unavoidable in spatial genetic

studies as predictors are derived from landscape characteristics that cannot be experimentally controlled (Graham 2003; King *et al.* 2005; Whittingham *et al.* 2006; Smith *et al.* 2009). Multicollinearity is multifaceted in origins. It is largely influenced by specific local landscape configuration patterns resulting from climate, geological events, past disturbances and anthropogenic pressures. Some trends can be observed though. Some spatial features usually come together (resulting in positive correlations among predictors), while others are mutually exclusive (resulting in negative correlations). Among features that usually come together: rivers and valleys or altitude and snow cover, obviously, but also rivers and urban areas, because of historical facilities offered by waterway transport, or motorways and agricultural surfaces, because of the regrouping of cultivated plots due to the increase in farmland value (Drescher *et al.* 2001; Prunier *et al.* 2014). On the contrary, categorical land-cover features are in essence mutually exclusive, as a pixel cannot be classified as 'forest' and as 'bare land' at the same time. As a consequence, when the landscape matrix is composed of a few predominant land-cover classes, negative correlations will often be observed between main predictors (King *et al.* 2005; Rioux Paquette *et al.* 2014). The same observation can be made in the context of transect-based approaches, when explanatory variables are expressed as relative quantities summing to 100% (e.g. habitat proportions in a delimited area; Angelone *et al.* 2011; Emaresi *et al.* 2011; Van Strien *et al.* 2012; Dormann *et al.* 2013).

Multicollinearity also typically arises in the specific case of resistance model optimization (Spear *et al.* 2010), when predictors are derived from a large set of alternative but closely related models (e.g. Cushman *et al.* 2013; Dudaniec *et al.* 2013). The use of various functions to reclassify or combine univariate resistance surfaces (Shirk *et al.* 2010; Spear *et al.* 2010; Peterman *et al.* 2014) may also influence patterns of nonindependence among predictors. To further complicate matters, multicollinearity patterns can vary depending on the spatial configuration of sampled points (Graves *et al.* 2013), change through time and differ across spatial scales (Dormann *et al.* 2013), making it an outstanding challenge in spatial genetics (Anderson *et al.* 2010).

Dealing with multicollinearity

There is a growing awareness of multicollinearity issues in spatial genetics (Garroway *et al.* 2011; Wedding *et al.* 2011; Dudaniec *et al.* 2012; Blair *et al.* 2013), and several approaches have been proposed to deal with multicollinearity issues. The simplest one is variable exclusion. The idea is to discard any variable showing correlations with other predictors higher than a certain threshold. A Pearson's correlation coefficient $|r| > 0.7$ is commonly used as a threshold (Dormann *et al.* 2013), although the exact value is left to the discretion of investigators (e.g. Angelone *et al.* 2011; Graves *et al.* 2012; Keller *et al.* 2013; Balkenhol *et al.* 2014). The estimation of variance inflation factors (VIF) can also be used as a way to identify nonindependence among predictors (Dyer *et al.* 2010; Blair *et al.* 2013). VIF is a positive value representing the overall correlation of each predictor with all others in a model. For a predictor X_i, VIF_i is computed as the inverse of the coefficient of nondetermination $(1/(1 - R_i^2))$, where R_i^2 is the model fit of the multiple regression of X_i over all other predictors (Neter *et al.* 1990; Stine 1995). VIF values >10 are usually considered as evidence for substantial multicollinearity and often justify the removal of certain predictors (but see O'Brien 2007 for a discussion on this subject). Variable exclusion may also be based on the investigation of principal component analyses (PCA) to identify and select a few biologically relevant predictors among a set of collinear variables (Manel *et al.* 2010). Nevertheless, variable exclusion, ignoring the unique contribution of discarded predictors, may result in a loss of explanatory power (Graham 2003).

Multicollinearity may also be addressed through the computation of orthogonal predictors using unconstrained ordination techniques. For instance, linear combinations of collinear variables (principal components) can be used as synthetic independent predictors in principal component regressions (Vigneau *et al.* 1997). However, this kind of approach may show serious statistical pitfalls (Hadi & Ling 1998), while the new independent variables will often be difficult to interpret (see Dormann *et al.* 2013 for details and a review of other available methods).

In 2001, Courville & Thompson advocated the simultaneous interpretation of beta weights β and structure coefficients r_s (or squared structure coefficients r_s^2) to improve the interpretation of multivariate regressions (Box 1). A structure coefficient is the bivariate (or zero-order) Pearson's correlation between a predictor X and predicted values $\hat{Y}$ of the dependent variable Y (Pedhazur 1997; Nathans *et al.* 2012). A squared structure coefficient thus represents the amount of variance in model fit that is accounted for by a single predictor (Nimon *et al.* 2008). By construction, structure coefficients may also be considered as rescaled zero-order correlations (Box 1; Courville & Thompson 2001). Structure coefficients are thus independent of collinearity among explanatory variables and allow ranking independent variables based on their direct contribution to model fit (Kraha *et al.* 2012; Ray-Mukherjee *et al.* 2014). For instance, a situation where a predictor X shows low β

but high r_s^2 may indicate that, because of collinearity among predictors, a proportion of the variance in X was assigned to another predictor in the process of computing beta weights. However, structure coefficients are limited. They do not precisely indicate which predictors synergistically or antagonistically contribute to predicting the dependent variable nor do they allow quantifying the amount of shared variance between collinear predictors (Nathans *et al.* 2012). Other indices can be used to dissect the complexity of predictors' relative contribution to model fit (Box 1). Among these indices, unique and common effects computed in commonality analysis are of great value (Box 2; Campbell & Tucker 1992; Nimon & Reio 2011).

Commonality analysis

Commonality analysis (CA) is a detailed variance-partitioning procedure that was developed in the 1960s (Newton & Spurrel 1967). From the field of human sciences, it was very recently brought to the attention of ecologists (Ray-Mukherjee *et al.* 2014). CA can provide substantial guidance for the interpretation of generalized linear models, including linear and logistic regressions, hierarchical mixed-effect models and canonical correlation analyses (not to be confused with canonical correspondence analyses; Legendre & Anderson 1999), by decomposing the overall model fit into its unique and common effects (Campbell & Tucker 1992; Nimon & Oswald 2013; Nimon *et al.* 2013a). Unique and common effects are nonoverlapping components of variance that ensue from formulae involving the regression of the dependent variable over all possible subsets of predictors (Box 3; Nimon & Reio 2011; Nathans *et al.* 2012; Ray-Mukherjee *et al.* 2014). By predictor, we mean any additive term in a regression equation: interaction effects are thus considered as predictors (Ray-Mukherjee *et al.* 2014). For a number k of predictors, CA returns a table of (2^k-1) commonality coefficients (or commonalities) including both unique and common effects. Unique effects U, or first-order effects, quantify the amount of variance in the dependent variable Y that is uniquely accounted for by a single explanatory variable. A negligible value of U indicates that the regression model only improves slightly with the addition of the predictor, when entered last in the model (Nathans *et al.* 2012; Roberts & Nimon 2012). Common effects represent the proportion of variance in the dependent variable that can be jointly explained by two or more predictors together, making CA particularly well suited in the case of multicollinearity (Campbell & Tucker 1992; Ray-Mukherjee *et al.* 2014). The common components of two or three (or k) variables are, respectively, called second- or third-order (or k^{th}-order) commonalities. The sum C of all commonalities involving a specific predictor indicates the amount of variance explained by this predictor that is shared with other explanatory variables. When they are divided by model fit index, U and C, respectively, represent the unique and common contributions of a predictor to the *explained* variance (that is, unique and common contributions to model fit) rather than to the total variance in the dependent variable. The sum $T = (U + C)$ represents the total contribution of a predictor to the dependent variable irrespective of collinearity with other variables, that is, in the case of linear regression, the squared zero-order (Pearson's) correlation r^2 between the predictor and the dependent variable (Nimon & Reio 2011; Kraha *et al.* 2012). When T is divided by the model fit index, it is equivalent to the squared structure coefficient r_s^2 (Kraha *et al.* 2012; Nimon & Oswald 2013; Ray-Mukherjee *et al.* 2014) and indicates how much each predictor contributes to the explanation of the entire model (model fit index) irrespective of other predictors (Box 1).

When predictors are independent, for instance in the case of principal component regressions (Dormann *et al.* 2013), common effects are null and the sum of unique effects equals the total variance of the dependent variable that is explained by the model (model fit index). However, when predictors show even low levels of multicollinearity, common effects are usually non-null and can show either positive or negative values. Positive common effects occur in the case of synergistic association among variables. Situations where C is substantially larger than U indicate that a predictor, showing significant contribution to the regression equation (high beta weight), only contributes indirectly to the dependent variable (or to model fit) because of its high positive correlation with other predictors. On the opposite, negative common effects are generally indicative of suppression (Capraro & Capraro 2001; Kraha *et al.* 2012).

The notion of suppression is an important contribution of CA over other variance-partitioning procedures (see Box 2): indeed, negative common effects are usually considered as embarrassing variation terms and thus interpreted as zero (Legendre & Legendre 1998; Peres-Neto *et al.* 2006). Suppression occurs in a variety of situations, depending on multicollinearity patterns among predictors (Lewis & Escobar 1986; Beckstead 2012). For instance, suppression may occur when correlations among predictors are of opposite sign. In all suppression situations, a predictor X_1 (the suppressor), sharing no or little variance with the response Y, purifies the relationship between a predictor X_2 and Y by removing (or suppressing) the irrelevant variance of X_2 on Y. As a result, the contribution of X_2 to the model fit is higher than would have been observed if X_1 had not

Box 1. Assessing variable importance in multivariate regression models

Here, we provide a brief description of the complementary indices that can be used to investigate the relative contribution of each predictor to the multivariate regression effect (see Kraha *et al*. 2012; Nathans *et al*. 2012 for details).

Beta weights

Beta weights correspond to classical regression weights when variables are z-transformed (by subtracting the mean and dividing by the standard deviation of the variable). Beta weights are thus comparable across various predictors. In logistic regressions, we speak about semi-standardized beta weights as the dichotomous dependent variable cannot be *z*-transformed. Beta weights quantify the change in the dependent variable (in standard deviation units) with a one standard deviation change in a predictor, all other predictors being held constant. They are thus measures of the total effect of a predictor on the dependent variable, accounting for the contribution of other predictors.

Zero-order (or bivariate) correlation coefficients

Zero-order correlation coefficients (also known as validity coefficients) range from −1 to 1 and represent the positive or negative linear (parametric Pearson's *r*) or monotonous (rank-based Spearman's ρ or Kendall's τ) relationship between two variables. In multivariate procedures, zero-order correlation coefficients are measures of the direct effect of a predictor on the dependent variable, without accounting for the contributions of other variables in the model; when *z*-transformed predictors are independent, they are equivalent to beta weights β. A discrepancy between zero-order correlation coefficients and beta weights is indicative of suppression.

Structure coefficients (see main text for details)

A structure coefficient r_s is the zero-order Pearson's correlation between a predictor and predicted values $\hat{Y}$ of the dependent variable Y (that is, $r_s = r_{X \cdot \hat{Y}}$). Structure coefficients are measures of the direct effect of a predictor on the dependent variable, irrespective of the influence of other predictors in the model. When squared, structure coefficients represent the amount of variance in model fit that is accounted for by a single predictor. Note that, in linear regressions, squared structure coefficients r_s^2 may also be computed by dividing the squared zero-order Pearson's correlation between a predictor and the dependent variable by the model fit index R^2 (that is, $r_s^2 = r_{X \cdot Y}^2 / R^2$). Structure coefficients may thus also be considered as rescaled validity coefficients (Courville & Thompson 2001). As previously, a discrepancy between structure coefficients and beta weights is indicative of suppression.

Product measures (or Pratt measures; Pratt 1987)

For a given predictor, product measure is the zero-order correlation coefficient multiplied by the corresponding beta weight, thus reflecting in a single metric both direct and total effects of a predictor on the dependent variable. The computation of product measures is a variance-partitioning procedure, the sum of the *k* product measures (for *k* predictors) being equal to the model fit index. When compared to zero-order correlation coefficients and beta weights, negative product measures may help identify suppression situations.

Relative weights (or relative importance weights)

A relative weights analysis is another variance-partitioning technique, minimizing (but not fully addressing) the problem of multicollinearity among *k* predictors through the use of orthogonal principal components. Relative weights are the proportionate contribution of each predictor to the overall model fit after (partially) correcting for multicollinearity. Suppression situations may be suspected when the sum of the *k* relative weights is larger than model fit index.

General dominance weights

General dominance analysis uses the results from an all-possible-subsets regression (with $2^k - 1$ subset models) to compute a set of *k* general dominance weights for a regression model containing *k* predictors. General dominance weights can be used to rank predictors according to a dominance hierarchy: they indicate the average difference in fit between all subset models of equal size that include a predictor and those that do not include it. As other

variance-partitioning procedures, the sum of general dominance weights equals the model fit index. Other kinds of dominance indices exist (complete and conditional dominance weights) that can also be used to identify suppression situations (Azen & Budescu 2003; Nathans *et al.* 2012).

Commonalities (see main text for details)

As general dominance weights, commonalities result from an all-possible-subsets regression (with 2^k-1 subset models; see Box 3), but there are 2^k-1 commonality coefficients (rather than k), each one indicating the amount of variance that a predictor set uniquely shares with the dependent variable. By decomposing model fit indices into unique U and common (or shared) C variance components, CA helps identify the location and magnitude of multicollinearity as well as suppression situations. Note that CA encompasses several other indices, notably zero-order Pearson's correlation coefficients r and structure coefficients r_s. Indeed, for a given predictor, $T = U + C = r^2$ and $T/R^2 = r^2/R^2 = r_s^2$.

Box 2. Commonality coefficients in comparison with other regression-type metrics

Commonality analysis is similar to other variance-partitioning techniques in that it partitions the regression effect into orthogonal nonoverlapping parts. Unlike product measures, relative weights or general dominance weights that partition the regression effect into k parts (Box 1), commonality analysis partitions the regression effect into 2^k-1 parts, where k equals the number of predictors in the regression equation (Box 3). Despite different analytical processes, general dominance weights and relative weights usually produce similar results. Some researchers therefore prefer to compute relative weights as they are less computationally burdensome than dominance analysis which demands an all-possible-subsets regression. Product measures are similarly easy to compute; however, they are criticized in the literature as they can produce negative coefficients, which some may find counterintuitive as a measure of variance. Among the variance-partitioning techniques reviewed, commonality analysis produces a more specific partitioning of the regression effect than product measures, relative weights or general dominance weights. Like product measures, commonality analysis may produce negative coefficients. While some researchers suggest these coefficients be interpreted as zero, Nimon & Oswald (2013) and others have suggested that negative commonality coefficients can be used to identify the loci and magnitude of suppression. This is a unique advantage of commonality analysis over other variance-partitioning techniques. For example, while a negative product measure signals a variable as a potential suppressor, product measures tell the researchers nothing about what variables are being suppressed. Similarly, by analysing the pattern of conditional dominance weights, researchers may be able to identify a variable as a suppressor. However, as documented by Beckstead (2012), dominance analysis may not always reveal complex suppression effects. Nor does dominance analysis provide information regarding the loci and magnitude of the suppression effect.

Note that commonality coefficients may also be used to derive squared validity and structure coefficients. Summing all the commonality coefficients that involve a predictor X_1 (e.g. U_{X1}, $C_{X_1X_2}$, $C_{X_1X_3}$, $C_{X_1X_2X_3}$) yields the per cent of variance that the predictor shares in common with the criterion (i.e. squared validity coefficient). Similarly, summing the commonality coefficients that have been divided by the magnitude of the regression effect (e.g. multiple R^2) for a given predictor yields the amount of variance that the predictor has in common with $\hat{Y}$ (i.e. squared structure coefficient). Like regression weights, commonality coefficients are considered measures of the total effect of an independent variable (see LeBreton *et al.* 2004), as they both take into account all independent variables in their computation. While regression weights indicate the amount of change in the criterion variable for each unit change in the independent variable holding all other independent variables constant, commonality coefficients indicate the amount of variance each variable set uniquely contributes to the regression effect. Note that in the case of perfect uncorrelated predictors, beta weights and commonality analysis will produce identical results; the unique effects from commonality analysis will be identical to the standardized regression weights, squared validity coefficients and squared structure coefficients.

been considered in the regression, but at the cost of a spurious correlation between X_1 and Y (Nimon & Reio 2011; Ray-Mukherjee *et al.* 2014). Suppression situations are sometimes so complex that no predictor can be identified as a specific suppressor variable (Lewis & Escobar 1986): in that case, predictors act as partial suppressors (Nimon 2010). Negative commonalities are the amount of predictive power that would be lost by

Box 3. Commonality coefficient formulae

For k predictors, there are 2^k-1 commonality coefficients, each with a unique formula. In addition, the formulae for calculating commonality coefficients are based on the number of predictors. Therefore, the commonality coefficients formulae for two predictors will be different than the formulae, for example, for 3, 4, 5 or more predictors. While formulae for calculating commonality coefficients have been published for 2, 3 and 4 predictors (see for example Thompson 2006), Mood (1971) developed a general procedure that allows for developing commonality coefficient formulae for any number of predictors.

In Mood's procedure, $(1-x)$ was used to represent variables in the common variance subset and (x) was used to represent variables not in the common variance subset. By negating the product of the variables in the subset and the variables not in the subset, deleting the -1 that may result from the expansion of the product and replacing x with R^2, Mood noted that the formula for computing any commonality coefficient can be derived (Nimon *et al.* 2008).

Take, for example, a regression model with five predictors. In such a model, there will be 31 (2^5-1) commonality coefficients. To calculate the amount of variance that is uniquely common to variables f_0, f_1, and f_3, the corresponding commonality coefficient ($C_{f_0f_1f_3}$) can be derived using Mood's procedure, expanding the results, and substituting x for R^2 as follows:

$$
\begin{aligned}
&-(1-f_0)(1-f_1)(1-f_3)f_2f_4 = \\
&-(1-f_0-f_1+f_0f_1)(1-f_3)f_2f_4 = \\
&-(1-f_0-f_1-f_3+f_0f_1+f_0f_3+f_1f_3-f_0f_1f_3)f_2f_4 = \\
&-(f_2f_4-f_0f_2f_4-f_1f_2f_4-f_2f_3f_4+f_0f_1f_2f_4+f_0f_2f_3f_4+f_1f_2f_3f_4-f_0f_1f_2f_3f_4) = \\
&-f_2f_4+f_0f_2f_4+f_1f_2f_4+f_2f_3f_4-f_0f_1f_2f_4-f_0f_2f_3f_4-f_1f_2f_3f_4+f_0f_1f_2f_3f_4 = \\
&C_{f_0f_1f_3} = \\
&-R^2_{f_2f_4}+R^2_{f_0f_2f_4}+R^2_{f_1f_2f_4}+R^2_{f_2f_3f_4}-R^2_{f_0f_1f_2f_4}-R^2_{f_0f_2f_3f_4}-R^2_{f_1f_2f_3f_4}+R^2_{f_0f_1f_2f_3f_4}
\end{aligned}
$$

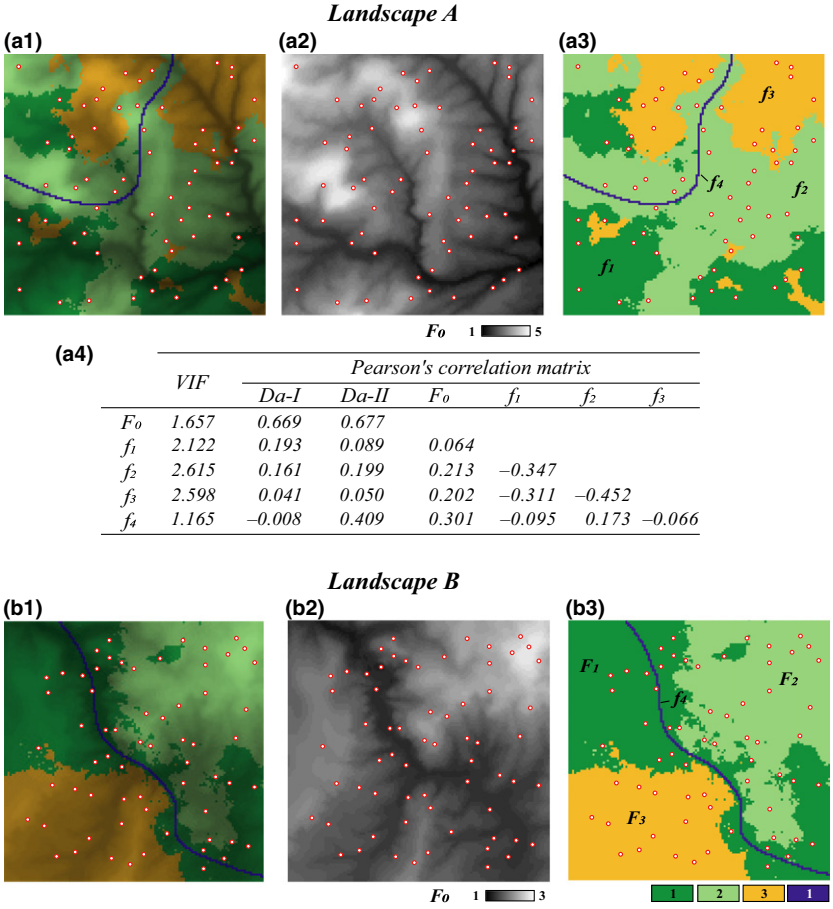
other predictors if the (partial) suppressor variable was not considered in the regression model. Suppressor variables may also be detected by comparing beta weights β to structure coefficients r_s. Situations where β is far larger than r_s or where both indices are of opposite signs are indicative of suppression (Ray-Mukherjee *et al.* 2014).

The following section illustrates how CA can assist the interpretation of multivariate regressions and avoid spurious biological conclusions. We used three simulated genetic data sets so that there was no ambiguity as to the drivers of the observed genetic patterns (Cushman & Landguth 2010; Epperson *et al.* 2010). In each example, we investigated multicollinearity among predictors and performed multivariate regression along with CA. Output parameters (zero-order correlations, VIF, beta weights, structure and commonality coefficients) are reported and discussed in detail.

Interpreting commonality analyses

Simulated data

We first created two distinct artificial landscapes A and B (Fig. 1, panels a_1 and b_1) of 128 × 128 pixels each. The resolution (size of pixels) was arbitrarily set to 10 m. Both landscapes had distinct configurations but the same composition: a continuous feature f_0 (e.g. topography; Fig. 1, panels a_2 and b_2), three categorical landcover features f_1 to f_3 (e.g. woods, meadows and crops) and one categorical linear feature f_4 (e.g. a road; panels a_3 and b_3). We then used CDPOP 1.2.11 (Landguth & Cushman 2010) to simulate gene flow between 64 randomly located populations of 30 individuals each, according to the relative resistance of landscape features to be crossed. We did not consider all features as being resistant to dispersal: the true drivers of gene flow were feature f_0 in landscape A and features f_0 to f_3 in landscape B (Fig. 1; Table 1). In landscape A, feature f_0 was rescaled to range from 1 to 5 (Fig. 1, panel a_2): the resulting resistance surface was used to compute pairwise effective distances based on least-cost paths. In CDPOP, travelled distances were drawn from a probability distribution inversely proportional to a linear function, with the maximal dispersal cost distance that may be travelled onto this resistance surface (associated with a null probability) set to 1500 m in data set I and 500 m in data set II (Table 1). In landscape B, a first continuous resistance surface was created by rescaling feature f_0 to range from 1 to 3 (Fig. 1, panel b_2), while a second categorical



(a4)

	VIF	Pearson's correlation matrix					
		Da-I	*Da*-II	F_0	f_1	f_2	f_3
F_0	1.657	0.669	0.677				
f_1	2.122	0.193	0.089	0.064			
f_2	2.615	0.161	0.199	0.213	−0.347		
f_3	2.598	0.041	0.050	0.202	−0.311	−0.452	
f_4	1.165	−0.008	0.409	0.301	−0.095	0.173	−0.066

(b4)

	VIF	Pearson's correlation matrix				
		Da-III	F_0	F_1	F_2	F_3
F_0	6.625	0.665				
F_1	5.504	−0.293	−0.232			
F_2	14.573	0.476	0.766	−0.469		
F_3	5.666	0.292	−0.042	−0.389	−0.369	
f_4	1.072	−0.045	−0.059	0.242	−0.142	−0.089

Fig. 1 Characteristics of landscapes *A* and *B*. Panels a_1 and b_1: overview of each landscape, combining both continuous and categorical features. Populations are indicated by red circles. Panels a_2 and b_2: overview of continuous features f_0. Panels a_3 and b_3: overview of land-cover features f_1, f_2, f_3 and f_4. Resistance values assigned to these features when simulating genetic data are provided below respective panels. Panels a_4 and b_4: VIF for each predictor and Pearson's correlation matrices among variables (with *Da*-I, *Da*-II and *Da*-III, respectively, indicating dependent variables in data sets I, II and III). For each landscape, the true drivers of gene flow (features with non-null resistance) are in capital letters (e.g. f_1 in landscape *A* but F_1 in landscape *B*).

Table 1 Characteristics of the three illustrative data sets

Data set	Landscape	True drivers of gene flow					Maximal dispersal distance (m)
		f_0	f_1	f_2	f_3	f_4	
I	A	✓					1500
II	A	✓					500
III	B	✓	✓	✓	✓		1500

layer representing the three features f_1, f_2 and f_3 was created by assigning pixels with resistance values of 1, 2 or 3, respectively (Fig. 1, panel b_3). Pixels associated with the 'blank' feature f_4 were assigned a value of 1, as this linear feature was totally included within feature f_1. Both continuous and categorical layers were then summed, and the resulting resistance surface rescaled to range from 1 to 5, as in landscape *A*. This final layer was finally used to compute pairwise effective distances based on least-cost paths, and simulations were performed with a maximal dispersal cost distance set to 1500 m (Table 1). In each case, CDPOP was run for 100 generations with 20 neutral loci of 20 alleles each (see Appendix S1, Supporting information for details).

Genetic distances and spatial predictors

Pairwise genetic distances were computed between populations using the Nei's version of Cavalli-Sforza's chord distance *Da* (Nei *et al.* 1983), as it is not contingent on any theoretical assumption. To compute spatial predictors in landscapes *A* and *B*, a specific layer was created for each landscape feature. Layers

associated with categorical features (f_1 to f_4) were binary, with pixels of value 1 for the considered feature and 0 otherwise. We then overlaid a 32 × 32 pixel grid on these layers and calculated the mean value of continuous feature f_0 and the percentage of categorical features f_1 to f_4 per blocks of 4 × 4 pixels (Balkenhol *et al.* 2014). These layers were finally rescaled to range from 1 to 100 and used in CIRCUITSCAPE 3.5.8 (McRae & Shah 2009) to compute pairwise effective distances between populations. The reasoning for the systematic use of feature density (or mean feature value) per square surface as a proxy for matrix resistance was the following: areas with high densities of unsuitable feature are assumed to hinder dispersal because of higher physiological cost and predatory risk (positive relationship between feature density and genetic distances), while areas with high densities of neutral or suitable feature are not (null or negative relationship). Pairwise distances based on circuit theory being expressed in terms of random walk probabilities (McRae 2006; Spear *et al.* 2010), we did not include Euclidean distances as an additional explanatory variable (Garroway *et al.* 2011; Peterman *et al.* 2014). All variables were *z*-transformed (by subtracting the mean and dividing by the standard deviation) for output parameter estimates to be comparable (Schielzeth 2010).

MRDM and LRDM

When the maximal cost distance was set to 1500 m (data sets I and III), genetic distances were approximately normally distributed (Fig. 2a–c), allowing the use of linear regression such as MRDM (e.g. Braunisch *et al.* 2010; Blair *et al.* 2013; Nanninga *et al.* 2014). MRDM are similar to classical multiple ordinary least-square (OLS) regressions, except that the significance of model fit (multivariate R^2) as well as the significance of beta weights β is assessed through permutations of the dependent matrix (Legendre *et al.* 1994). Linear MRDM and associated CA were, respectively, conducted using packages *ecodist* (Goslee & Urban 2007) and *yhat* (Nimon *et al.* 2013b) in R 3.1.0 (R Development Core Team 2014) with the following full model: $D_a = \sum(\beta_i \cdot f_i)$. All significance levels were assessed with 10 000 permutations after sequential Bonferroni correction (Holm 1979). Ninety-five per cent confidence intervals around beta weights, structure coefficients and commonalities were computed using a bootstrap procedure, with 1000 replicates based on a random selection of 58 out of 64 populations without replacement (Peterman *et al.* 2014). This approach allows assessing the robustness of observed parameters to the random removal of a few sampling points: asymmetrical confidence intervals (e.g. Fig. 3) may result from multimodal distributions of

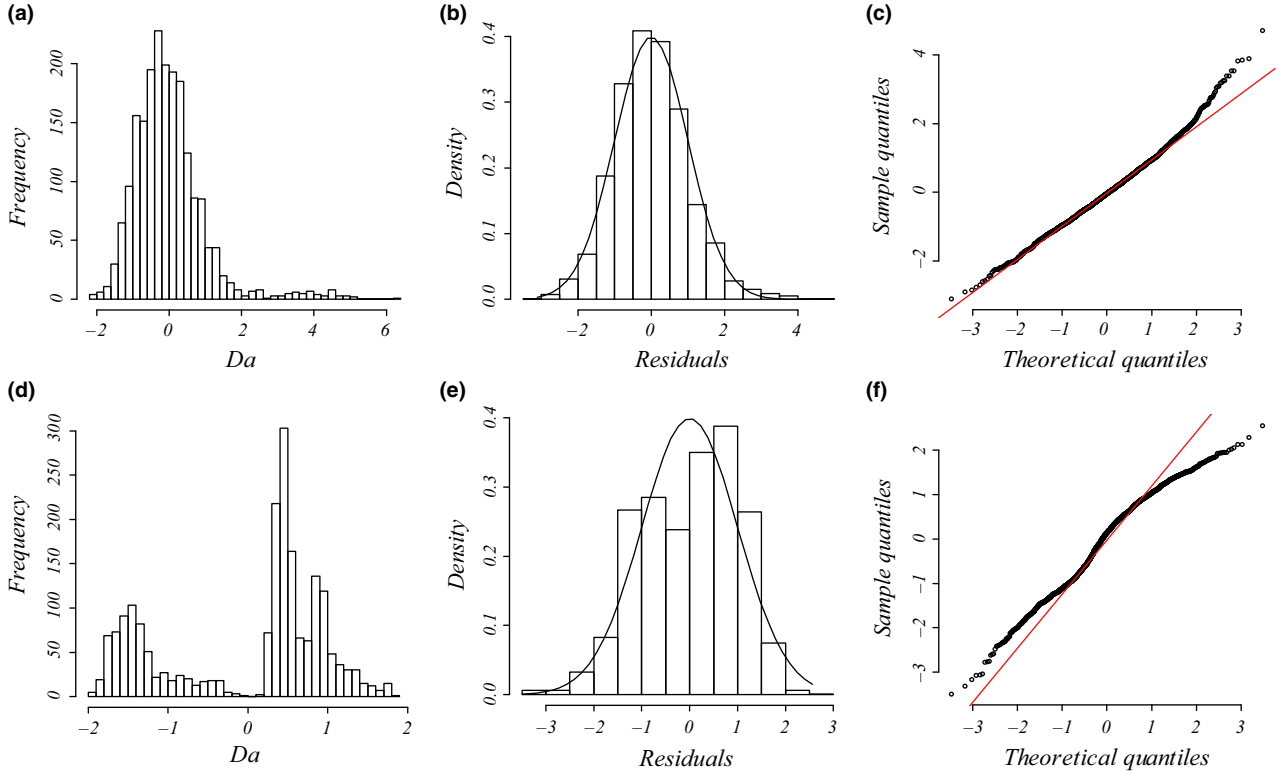


Fig. 2 Distribution of the dependent variable *Da* in data sets I (a) and II (d). Histograms of studentized residuals resulting from MRDM in data sets I (b) and II (e). Normal Q–Q plots of studentized residuals resulting from MRDM in data sets I (c) and II (f).

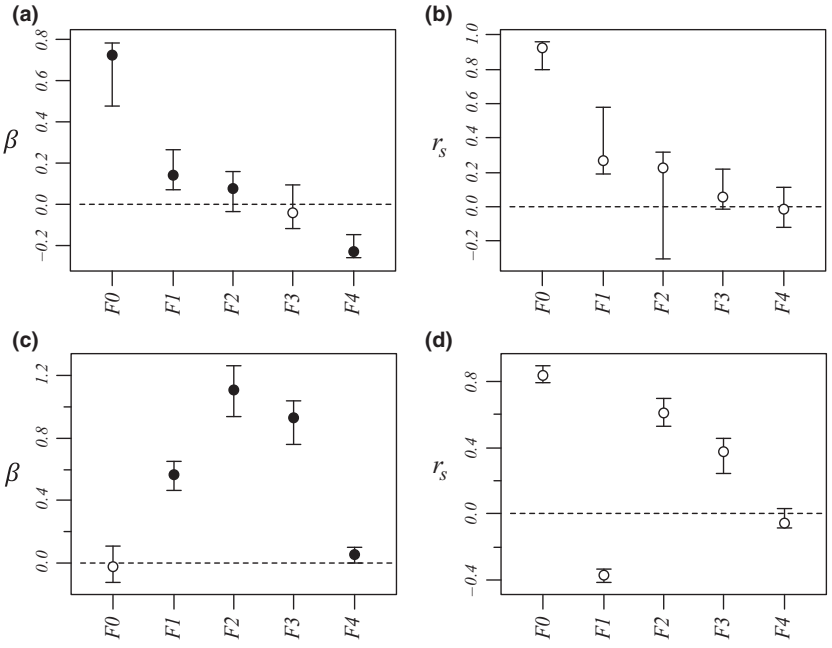


Fig. 3 Plots of beta weights β and structure coefficients r_s in data set I (a, b) and data set III (c, d) with 95% bootstrap confidence intervals computed on the basis of 1000 replicates with a random selection of 58 out of 64 populations without replacement. Significant beta weights after sequential Bonferroni correction are indicated by filled circles.

metrics across bootstrap replicates, confirming that multicollinearity may vary with the spatial configuration of sampled points. This bootstrap procedure was also used to determine whether the observed differences between pairs of metrics (beta weights and structure coefficients) were robust to the removal of a few sampled points (Appendix S2, Supporting information).

When the maximal cost distance was set to 500 m (data set II), genetic distances showed a multimodal distribution, thus violating the assumptions of normality of residuals in OLS regression (Fig. 2d–f). This kind of distribution usually requires the use of nonparametric, rank-based regression methods, assessing the monotonic rather than the linear relationships between the dependent variable and predictors (Van Strien *et al.* 2012; Dormann *et al.* 2013; Balkenhol *et al.* 2014). However, as variance-partitioning procedures such as CA cannot be applied in the case of nonparametric regressions, we transformed all genetic distances into binary variables and performed logistic regression on distance matrices (LRDM). Multivariate logistic regressions are used to predict the likelihood of a success (e.g. the probability $\hat{p}$ that $Y = 1$) given a set of predictors. When the dependent variable Y is expressed as the log odds of a success (logit transformation), the logistic regression equation is simply a linear combination of predictors where semi-standardized beta weights $\hat{\beta}$ are estimated using maximum-likelihood procedures (Smith & McKenna 2013). Odd-ratios ψ, that is, semi-standardized beta weights raised to the exponent ($\Psi = e^{\hat{\beta}}$), can then be used to evaluate the increase of the likelihood of a success with a one standard deviation change in X. Logistic regressions are uncommon in landscape genetics (but see for instance Weigel *et al.* 2013) but can be useful to handle nonlinear data. Using zero as a threshold, z-transformed genetic distances were thus recoded into binary data with 0 for pairs of individuals with negative z-scores and 1 for pairs of individuals with positive z-scores ('success' of being genetically dissimilar). Logistic regression was then performed using the *glm* function with a *logit* link in R 3.1.0, with all predictors being included in the model. Semi-standardized beta weights $\hat{\beta}$ were computed following King (2007) with the mean predicted probabilities as a reference value. To evaluate overall model fit, we used the Nagelkerke's Index as a pseudo-R^2, with a range (from 0 to 1) identical to the range of OLS multiple R^2 (Roberts & Nimon 2012; Smith & McKenna 2013). Nagelkerke's Index was computed with the *NagelkerkeR2* function in package *fmsb,* while logistic CA was performed with the *cc4log* function provided in Roberts & Nimon (2012). Structure coefficients were not computed here, as they are specific to linear regressions. Because binary data came from pairwise genetic distances and thus could not be considered as independent, significance levels of model fit and predictors were estimated using a randomization procedure similar to the one used in MRDM (Legendre *et al.* 1994). Rows and columns of the binary matrices were randomly permuted 10 000 times, and logistic regressions were performed on each permuted matrix to create a theoretical distribution of pseudo-R^2 and beta weights $\hat{\beta}$ under the null hypothesis of random pairwise genetic distances. Observed pseudo-R^2 and $\hat{\beta}$ were then compared to theoretical

distributions at a significance level of 0.05 with sequential Bonferroni correction (Holm 1979). As previously, 95% confidence intervals around commonalities were computed using a bootstrap procedure, with 1000 replicates based on a random selection of 58 out of 64 populations without replacement (Peterman *et al.* 2014).

First illustration: data set I

Absolute zero-order Pearson's correlations among predictors in landscape *A* ranged from 0.064 to 0.452, while VIF ranged from 1.165 to 2.61 (Fig. 1, panel a_4). As expected, predictors associated with the three predominant, and mutually exclusive, land-cover features (f_1, f_2 and f_3) were negatively correlated and showed the highest VIF. Nonetheless, bivariate correlations and VIF were below traditional thresholds ($|r| < 0.7$ and VIF < 10, respectively), suggesting at first glance little multicollinearity in this data set.

Multivariate regression model was significant and explained 52.10% of variance in the dependent variable (Table 2). Except f_3, all predictors were significant after sequential Bonferroni correction. Comparing absolute values of β allowed ranking predictors from the most to the least influent in the following order: f_0, f_4, f_1, f_2 and f_3 (Fig. 3a), although the difference between β_2 and β_3 was not robust to the random removal of a few sampling points (Appendix S2, Supporting information). For instance, the dependent variable increased by 0.721 standard deviation with a one standard deviation change in f_0, all other predictors being hold constant. The linear feature f_4 was the only significant predictor responsible for a decrease in genetic distances ($\beta_4 = -0.228$). Most researchers would content themselves with such results, maybe looking for a rationale to explain how a linear feature such as roads (f_4) could enhance gene flow. In our simulations though, the continuous feature f_0 was the only driver of gene flow, and all other significant correlations were thus spurious ones.

Examining structure (r_s) and squared structure (r_s^2) coefficients can help quantify the direct effect of predictors on the dependent variable (Nathans *et al.* 2012). The ranking of predictors based on absolute values of structure coefficients, that is, rescaled zero-order correlations (Box 1), diverged from the ranking based on beta weights (Fig. 3b), with f_4 actually showing negligible r_s value (Table 2). Furthermore, pairwise differences between r_{s2}, r_{s3} and r_{s4} were not robust to the random removal of a few sampling points (Appendix S2, Supporting information). The actual direct contribution of f_4 to model fit was null ($r_{s4}^2 = 0$), which in this case means that a substantial proportion of the variance in one or several other predictors was assigned to f_4 in the process of computing beta weights, specifically designating it as a suppressor variable. Nevertheless, the spurious effects associated with f_1 and f_2 could not be explained. Structure coefficients indicate with no doubt that biological interpretations based only on β may be erroneous (Courville & Thompson 2001), despite low levels of collinearity (VIF < 3), but they cannot inform about which predictors jointly share variance in predicting the dependent variable or in what quantity.

The actual synergistic or antagonistic processes operating among predictors can be assessed by CA. Figure 4 provides the 31 commonality coefficients, including both unique and common effects. Commonalities indicate the percentage of variance in the dependent variable that is uniquely explained by each predictor (unique effects) or set of predictors (common effects). With CA being a variance-partitioning procedure, the sum of commonalities equals the model fit index R^2 (here, $R^2 = 0.521$). *%Total* values are obtained by dividing commonalities by the fit index: they sum to 100 and represent the percentage of explained variance in model fit. CA indices reported in Table 2 can be derived from Fig. 4, by directly reading in the case of unique effects *U* or by summing all commonalities involving a given predictor in the case of common effects *C*. Total effects

Table 2 MRDM results in data set I. Typical MRDM results and additional parameters derived from CA: predictors (*pred*), model fit index (*multivariate* R^2; ***: *P*-value <0.001), beta weights β and *P*-values *P*, structure and squared structure coefficients (r_s and r_s^2), and finally, unique, common and total contributions of predictors to the variance in dependent variable (*U*, *C* and *T*)

Pred	Multiple R^2	β	*P*	r_s	r_s^2	*U*	*C*	*T*
***F*₀**		0.721	**<0.001**	0.926	0.858	0.314	0.133	0.447
f_1	52.10%	0.138	**<0.001**	0.267	0.071	0.009	0.028	0.037
f_2	***	0.075	**0.008**	0.223	0.050	0.002	0.024	0.026
f_3		−0.044	0.103	0.056	0.003	0.001	0.001	0.002
f_4		−0.228	**<0.001**	−0.011	<0.001	0.045	−0.045	<0.001

P-values in bold indicate significant predictors after sequential Bonferroni correction. Predictors in bold indicate main unique contributors to model fit according to CA (see text for details).

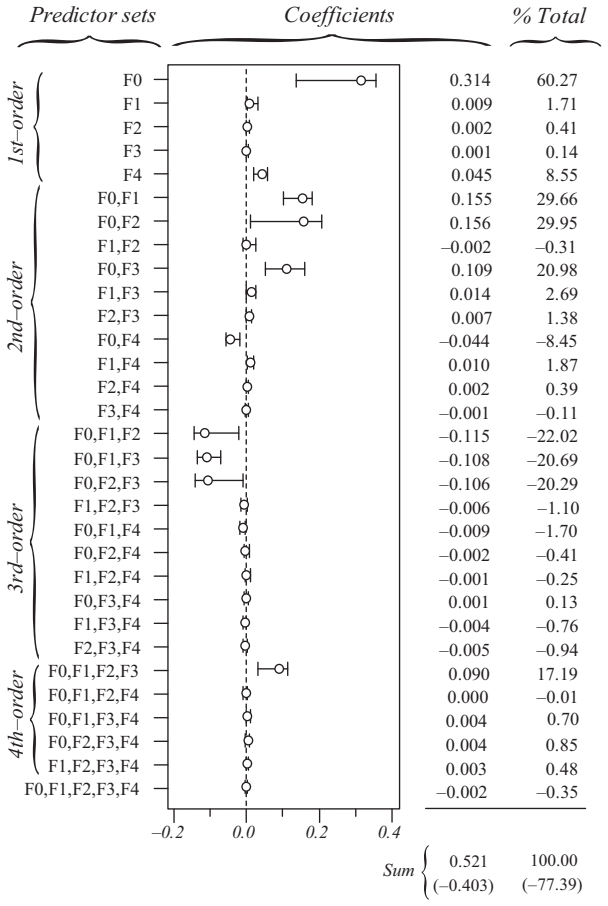


Fig. 4 The 31 commonality coefficients computed in data set I, including both unique and common effects. *Coefficients* represent the percentage of variance explained in the dependent variable *Y* by each set of predictors. Ninety-five per cent confidence intervals were computed using a bootstrap procedure, with 1000 replicates based on a random selection of 58 out of 64 populations without replacement. The sum of coefficients equals the model fit index. *%Total*, summing to 100%, represents the percentage of variance explained in predicted values $\hat{Y}$ (that is, in model fit) by each set of predictors. In brackets: percentage of variance explained in the dependent variable (respectively, in model fit) that is due to suppression (sum of all negative commonalities).

T are obtained by summing *U* and *C*. Note that *T* values may also be computed by squaring zero-order correlations *r* (Fig. 1, panel a_4), indicating that CA actually encompasses several informative diagnostic indices (see Box 2).

The sum of all negative commonalities showed that 77.4% of the regression effect was caused by suppression. All predictors were associated with negative commonalities (Fig. 4), thus suggesting a complex suppression situation in this data set (Lewis & Escobar 1986). The importance of f_0 was confirmed, as this predictor showed the highest unique contribution U_0 to the variance in *Da*. Reported coefficients U_0 are to be interpreted as follows: the continuous feature f_0 uniquely contributed to 31.4% of the total variance in the dependent variable and to 60.27% of the 52.1% of variance explained by the regression model. Unique contributions of f_1, f_2 and f_3 were actually negligible ($U < 1\%$). Positive second- and fourth-order commonalities involving these predictors ($[f_0f_1]$, $[f_0f_2]$, $[f_0f_3]$ and $[f_0f_1f_2f_3]$) were partially (or almost totally) counterbalanced by third-order negative commonalities ($[f_0f_1f_2]$, $[f_0f_1f_3]$ and $[f_0f_2f_3]$; Fig. 4), resulting in positive sums *C* of commonalities if f_1 ($C_1 = 0.028$) and f_2 ($C_2 = 0.024$), and negligible one in f_3 ($C_3 = 0.001$; Table 2). All these commonalities involved f_0: predictors f_1 and f_2 thus only contributed to the regression model because of their resultant synergistic association with f_0. Negative commonalities involving f_1, f_2 and f_3 were produced because correlations among land-cover predictors had opposite signs (Fig. 1, panel a_4): these predictors acted as partial suppressors, suppressing irrelevant variance in f_0, which thus showed a larger beta weight β_0 than if the correlations had been in the same direction. Finally, the unique contribution of f_4 to the dependent variable ($U_4 = 0.045$) was almost totally counterbalanced by the negative second-order commonality $[f_0f_4]$ (−0.044; Fig. 4), resulting in a null total contribution ($T_4 = 0$; Table 2). The predictor f_4 was actually unrelated to the dependent variable *Da* ($r_4 = -0.008$) and acted as a suppressor variable: about 4.5% of irrelevant variance in f_0 was assigned to f_4 in the process of computing beta weights, increasing the overall model fit but also resulting in a significantly non-null value for β_4.

To sum up, while the classical interpretation of beta weights would have yielded erroneous conclusions, CA correctly identified f_0 as the true driver of gene flow. Other significant predictors either acted as suppressor (f_4) or partial suppressors (f_1 and f_2), the latter indirectly contributing to model fit through their synergistic association with f_0.

Second illustration: data set II

In the second data set, the exact same predictors were used, but the dependent variable was computed in the context of restricted dispersal (Table 1). The logistic model was significant and accounted for 51.56% of (pseudo-) variance in *Da* (Table 3). Three predictors were significant after sequential Bonferroni correction: f_0, f_3 and f_4. Predictor f_0 was identified as the main predictor ($\hat{\beta}_0 = 0.445$), with an odd-ratio $\psi_0 = 1.56$. This means that populations were 1.56 times more likely to show high pairwise genetic distances ($Da > 0$) with a one standard deviation change in f_0 values. Predictor f_4 was ranked second, with $\hat{\beta}_4 = 0.218$ and $\psi_4 = 1.24$, while predictor f_3 showed a negative weight

Table 3 LRDM results in data set II. Typical LRDM results and additional parameters derived from CA: predictors (*pred*), model fit index (*Pseudo-R^2*; ***: *P*-value <0.001), semi-standardized beta weights $\hat{\beta}$ (computed using the mean predicted probability of 0.663 as a reference value), odd-ratios ψ and *P*-values *P*, and finally, unique, common and total contributions of predictors to the variance in dependent variable (*U*, *C* and *T*)

Pred	Pseudo-R^2	$\hat{\beta}$	ψ	*P*	*U*	*C*	*T*
$\boldsymbol{F_0}$		0.445	1.560	**<0.001**	0.182	0.253	0.434
f_1	51.56%	0.022	1.022	0.297	<0.001	0.011	0.011
f_2	***	0.037	1.038	0.187	0.001	0.036	0.037
f_3		−0.122	0.886	**0.001**	0.013	−0.012	0.001
f_4		0.218	1.244	**<0.001**	0.021	0.280	0.301

P-values in bold indicate significant predictors after sequential Bonferroni correction. Predictors in bold indicate main unique contributors to model fit according to CA (see text for details).

($\hat{\beta}_3 = -0.122$). This means that populations were more likely to be genetically similar ($\psi_3 < 1$) with a one standard deviation change in f_3 values (Fig. 5). This last result was, however, inconsistent with the zero-order correlation $r_3 = 0.05$ (Fig. 1, panel a_4) that indicated a slightly positive relationship between f_3 and untransformed values of *Da*. Furthermore, the continuous feature f_0 was actually the only true driver of gene flow in landscape *A*.

CA allowed clarifying these results. The sum of all negative commonalities showed that 42% of the regression effect was caused by suppression (Fig. 6). As previously, all predictors were associated with negative commonalities (Fig. 6), thus suggesting a complex suppression situation in this data set (Lewis & Escobar 1986). The importance of f_0 was confirmed, as this predictor had the largest unique contribution to the variance in *Da* ($U_0 = 18.2\%$; Table 3) and accounted for 35.2% of the regression effect (Fig. 6). Predictor f_3 was involved in second-, third-, fourth- and fifth-order non-null commonalities (Fig. 6) but was easily identified as a suppressor variable, as its unique contribution U_3 was almost totally counterbalanced by the sum of its common contributions C_3 (Table 3), thus providing a better understanding of the inconsistency between a

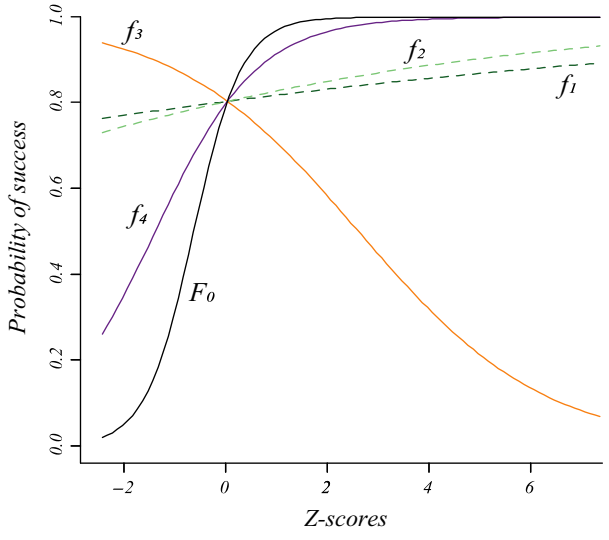


Fig. 5 Predicted probability of success in function of each *z*-transformed predictor in data set II using LRDM. Nonsignificant predictors are in dashed lines. The true driver of gene flow (F_0) is in capital letters.

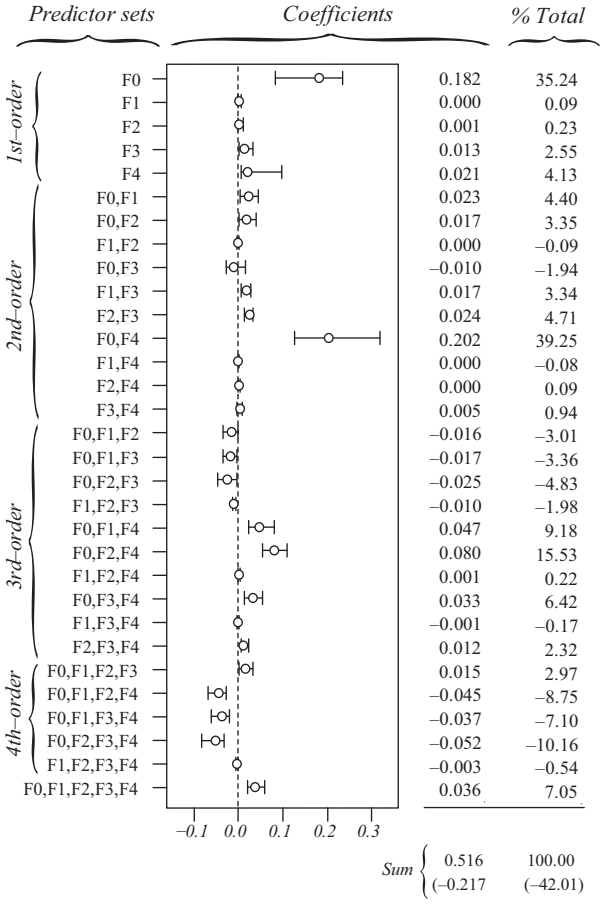


Fig. 6 The 31 commonality coefficients computed in data set II, including both unique and common effects. See legend in Fig. 4.

slightly positive zero-order correlation r_3 and a negative semi-standardized beta weight $\hat{\beta}_3$. While the unique contribution of f_4 was low ($U_4 = 0.02$), the sum of all commonalities involving f_4 was high ($C_4 = 0.28$), notably because of the second-order commonality $[f_0f_4]$ that accounted for 39.25% of model fit (Fig. 6). Predictor f_4 was positively correlated with predictor f_0 ($r = 0.301$; Fig. 1, panel a_4) and thus mainly contributed to the regression model because of its high common contribution with f_0. Finally, and as expected, nonsignificant predictors f_1 and f_2 showed negligible unique contributions U and low common contributions C, the latter resulting from a trade-off between positive third-order commonalities ($[f_0f_1f_4]$ and $[f_0f_2f_4]$) and negative third- and fourth-order commonalities (Fig. 6). Note that because it is involved in all the largest negative commonalities ($[f_0f_1f_2]$, $[f_0f_1f_3]$ and $[f_0f_2f_3]$), predictor f_0 may also be considered as a partial suppressor. As previously, while the classical interpretation of semi-standardized beta weights and odd-ratios would have yielded erroneous conclusions, CA correctly identified f_0 as the true driver of gene flow. Other significant predictors either acted as suppressor (f_3) or indirectly contributed to model fit through a synergistic association with f_0 (f_4).

Third illustration: data set III

Absolute zero-order Pearson's correlations among predictors in landscape *B* ranged from 0.089 to 0.766, while VIF ranged from 1.072 to 14.573 (Fig. 1, panel b_4), suggesting potential multicollinearity issues in this example. The most problematic predictor was f_2 (VIF > 10), as it was highly correlated with f_0 ($r > 0.7$; Fig. 1, panel b_4). This predictor would usually be excluded from the model but was conserved here for illustration purpose. As in landscape *A*, predictors associated with the three predominant land-cover features (f_1, f_2 and f_3) were negatively correlated.

The linear model was significant and accounted for 61.58% of variance in *Da* (Table 4). Except for f_0, all predictors showed positive and significant beta weights after sequential Bonferroni correction. Comparing absolute values of beta weights allowed ranking predictors from the most to the least influent in the following order: f_2, f_3, f_1, f_4 and f_0 (Fig. 3c), although the difference between β_0 and β_4 was not robust to the random removal of a few sampling points (Appendix S2, Supporting information). However, the ranking of predictors based on beta weights was inconsistent with the ranking based on structure coefficient r_s (Fig. 3d). Indeed, predictor f_0, though nonsignificant, showed the highest direct contribution to the dependent variable ($r_{s0} = 0.834$), while f_1 actually appeared negatively correlated with *Da* (Table 4).

The sum of all negative commonalities indicated a really complex suppression situation here, as 92.8% of the regression effect was caused by suppression (Fig. 7). In particular, negative commonalities identified suppression involving predictors f_0, f_1, f_2 and f_3, through second-order ($[f_1f_2]$, $[f_1f_3]$, $[f_2f_3]$) and third-order commonalities ($[f_0f_1f_2]$, $[f_0f_1f_3]$, $[f_0f_2f_3]$). These negative commonalities were partially counterbalanced by positive second-order ($[f_0f_1]$, $[f_0f_2]$, $[f_0f_3]$), third-order ($[f_1f_2f_3]$) and fourth-order commonalities ($[f_0f_1f_2f_3]$), resulting in positive common contributions to the dependent variable in f_0, f_1 and f_2 ($C_0 = 42.8\%$; $C_1 = 2.7\%$; $C_2 = 14.3\%$) and negative common contribution in f_3 ($C_3 = -6.6\%$; Table 4). With such a negative common contribution but at the same time a non-null total contribution to model fit ($T_3 = 8.5\%$), predictor f_3 was identified as a partial suppressor. This partial suppression effect notably ensued from the specific pattern of bivariate correlations in this data set, f_3 being positively correlated with *Da* but negatively correlated with all other predictors (Fig. 1, panel b_4). Predictor f_1 may also be considered as a partial suppressor variable: there was a mismatch in the sign of beta weight β_1 and structure coefficient r_{s1}, but the sum C of all commonalities involving f_1 was still positive ($C_1 = 0.027$). The influence of these predictors was thus to be interpreted with much caution. Nevertheless, a few conclusions could still be drawn from this analysis. First, it is worth observing that unique contributions allowed ranking land-cover predictors f_1, f_2 and f_3 in an order consistent with simulated process (that is, $f_3 > f_2 > f_1$; Table 4), although neither the magnitude nor

Table 4 MRDM results in data set III

Pred	Multiple R^2	β	*P*	r_s	r_s^2	*U*	*C*	*T*
F_0		−0.019	0.644	0.834	0.696	<0.001	0.428	0.428
F_1	61.58%	0.568	**<0.001**	−0.373	0.139	0.059	0.027	0.086
F_2	***	1.107	**<0.001**	0.607	0.368	0.084	0.143	0.227
F_3		0.926	**<0.001**	0.372	0.138	0.151	−0.066	0.085
f_4		0.056	**<0.001**	−0.057	0.003	0.003	−0.001	0.002

See legend in Table 2.

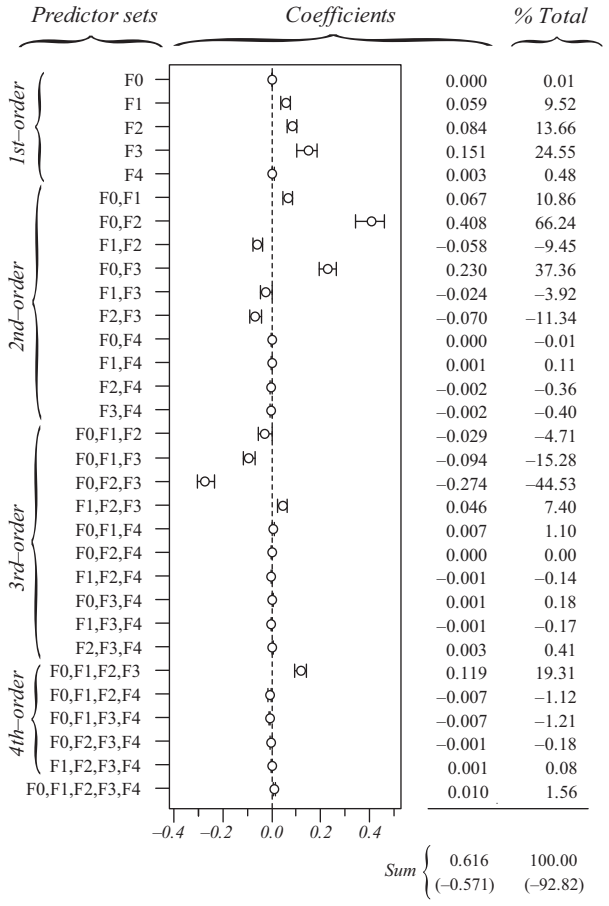


Fig. 7 The 31 commonality coefficients computed in data set III, including both unique and common effects. See legend in Fig. 4.

the sign of reported beta weights could be fully trusted. Second, the influence of predictor f_4 on gene flow could be immediately ruled out: although characterized by a significant beta weight, f_4 had both little unique and common contributions to the dependent variable ($U_4 = 0.003$; $C_4 = -0.001$; Table 4), resulting in a negligible total contribution to model fit ($T_4 = 0.2\%$), in accordance with simulated process. Finally, although nonsignificant, predictor f_0 had actually the highest total contribution to model fit ($T_0 = 42.8\%$). This contribution was completely explained by common effects of f_0 with other predictors ($C_0 = 42.8\%$), especially through the positive second-order commonality [f_0f_2] which accounted for 66.24% of model fit (Fig. 7). Because of high collinearity among these two predictors ($r = 0.766$; Fig. 1, panel b_4), the variance in f_0 was mainly assigned to f_2 in the process of computing beta weights. Excluding f_2 from the model made the percentage of suppression drop from 92.8% to <9% and allowed correctly identifying the main contributors to genetic distances while ruling out f_4 (data not shown). But by doing so, f_2 could obviously not be identified as one of the drivers to gene flow.

Advantages of commonality analyses

As a preliminary remark, note that because of multicollinearity among spatial features, zero-order correlations can be non-null despite no true causal relationship between the dependent variable and spatial predictors (e.g. f_1 in data set I or f_4 in data set II; Nathans *et al.* 2012), thus confirming the utility of multivariate procedures over univariate ones in resistance model optimization (Balkenhol *et al.* 2009; Spear *et al.* 2010; Graves *et al.* 2013). Whatever the model fit index used to identify the best resistance surface from a set of alternative surfaces (e.g. Perez-Espona *et al.* 2008; Shirk *et al.* 2010; Dudaniec *et al.* 2013; Peterman *et al.* 2014), multicollinearity among spatial features may indeed be responsible for the selection of suboptimal or even spurious univariate models (Spear *et al.* 2010). For instance, Perez-Espona *et al.* (2008) found that inland lochs and rivers might facilitate gene flow in Scottish red deers: this result may be biologically realistic, but may also ensue from artefactual correlations due to spatial multicollinearity among landscape features. However, the three provided illustrations also confirmed that the typical interpretation of multivariate regressions based on the ranking of significant beta weights is flawed by even weak levels of multicollinearity among predictors (Mac Nally 2002; Dormann *et al.* 2013), with the risk of erroneous conclusions, misdirected research and inefficient or counterproductive conservation measures. The investigation of commonalities can circumvent these flaws to some extent and provide insightful information about the relative influence of landscape predictors on the dependent variable in direct gradient analyses.

First, CA can be used to clarify the relative importance of predictors in the case of synergistic association among variables (Ray-Mukherjee *et al.* 2014). For instance, f_4 in data set II showed little unique contribution U to model fit, especially when compared with the unique contribution of f_0. The computation of the semi-standardized beta weight in f_4 was actually dictated by collinearity between f_4 and f_0 so that $\hat{\beta}_4$ reflected the indirect rather than the real contribution of f_4 to model fit. Similarly, f_0 in data set III had negligible unique contribution to model fit: this predictor mostly explained variance in the dependent variable when in synergistic association with other predictors. In both cases, the classical interpretation of significant beta weights was flawed by positive collinearity among predictors. In an empirical situation, that is, in the absence of any information about potential drivers of gene flow, deciding whether a predictor with little unique

contribution is or is not responsible for the observed biological response is left to the appreciation of investigators. Disentangling the relative contribution of such explanatory variables in a predictive perspective would actually require replications in different multicollinearity contexts (Anderson *et al.* 2010; Short Bull *et al.* 2011; Prunier *et al.* 2013). Nevertheless, this ability of CA to provide a clear quantification of unique contributions of predictors to model fit is valuable for interpreting results of multivariate regressions.

Second, CA can reveal spurious correlations that may have gone unnoticed in the framework of a typical interpretation of significant beta weights. The investigation of commonalities in data sets I and II confirmed that suppression may occur despite moderate levels of multicollinearity (Ray-Mukherjee *et al.* 2014). Researchers are thus probably often confronted to this kind of spurious effects although they may not be aware that they are dealing with a suppression situation (Lewis & Escobar 1986). This is especially true when the sign and the magnitude of beta weights make sense with regard to biological expectations. Should this not be the case, a convincing rationale is usually proposed to justify any unexpected outcome (Graves *et al.* 2013). By fully clarifying how variables contribute to prediction, CA provides a way to identify suppressors and to foil spurious correlations. Two predictors were identified as suppressors in data sets I (f_4) and II (f_3). When significant, these variables were thought to facilitate dispersal, although they were not part of the simulating process. However, the total contributions T of these two predictors, that is the amount of explained variance in the dependent variable irrespective of collinearity, were low, unique effects being counterbalanced by their negative commonalities. Suppression situations are to be identified for a proper interpretation of regression models, but they are not necessarily to be avoided (Lewis & Escobar 1986; Pandey & Elliott 2010). Once spurious correlations have been identified, suppression may eventually improve the detectability of influential predictors by removing the part of their variance that is irrelevant to predict the dependent variable (Nimon & Reio 2011; Ray-Mukherjee *et al.* 2014). It is yet essential to bear in mind that although some variables may act as explicit suppressors, suppression situations can also be particularly complex (data set III; Lewis & Escobar 1986): the interpretation of commonalities is thus always to be carried out with much caution (Ray-Mukherjee *et al.* 2014).

Finally, and contrary to procedures such as stepwise regression, CA is independent of variable order and thus replicable for a given model (Lewis 2007; Ray-Mukherjee *et al.* 2014). By allowing a thorough understanding of the relative contribution of spatial predictors to genetic variability, CA can help identify the most parsimonious set of predictors from a given full model (Kraha *et al.* 2012), thus making it a valuable complementary tool to model selection procedures.

Limitations and future applications

CA appears as a promising procedure in spatial genetics. However, this is not a panacea. First, although investigating commonalities allow identifying the location and magnitude of nonindependence among predictors (Nimon 2010), it cannot solve multicollinearity issues *per se* (Dormann *et al.* 2013). Indeed, CA does not provide any corrected or weighted parameter that may eventually be used in a predictive perspective. CA is also limited to parametric models: although it is hitherto available for use with linear regressions, logistic regressions, as well as mixed-effect models (Roberts & Nimon 2012; Ray-Mukherjee *et al.* 2014), commonalities cannot be interpreted in the context of nonlinear, rank-based regressions. LRDM may sometimes constitute an interesting alternative to nonparametric models (for instance in the case of multimodal distributions; see data set II), but may not properly represent all nonlinear relationships.

Furthermore, the interpretation of CA becomes more difficult as the number of predictors increases, because the number of commonalities expands exponentially with the number of predictors (Ray-Mukherjee *et al.* 2014; see Box 3). In our illustrations, we only considered a set of five predictors, a reasonable number when considering current spatial genetic studies (Zeller *et al.* 2012). Nevertheless, the number of commonalities would have increased from 31 to 63 with a single additional variable in full regression models. Although interpreting and reporting commonalities for a higher number of predictors may be highly informative (Nathans *et al.* 2012), the number of predictors may sometimes be too high or suppression situations too complex for all explanatory variables to be conserved in the regression model (e.g. data set III). CA thus does not resolve the problem of variable preselection (Dormann *et al.* 2013). Rather than using a predefined correlation threshold to avoid multicollinearity, investigators should investigate multicollinearity patterns and consider both the ecology of studied species and local landscape characteristics to select a set of nonredundant and biologically relevant predictors.

Finally, commonalities are specific to a given model, because of particular patterns of bivariate correlations. In a complex multicollinearity context, all predictors may, to some extent, remove irrelevant variance from some other variables (see Figs 4, 6 and 7; Lewis & Escobar 1986) so that adding or removing predictors is likely to modify commonalities, with the emergence of

new suppression situations. This issue has no consequence on direct gradient analyses when they are performed in an explanatory perspective as a single full model is explored. In a predictive perspective though, regressions are often performed on a set of nested or alternative models and model fit indices are then compared to identify the model structure showing the best predictive power. As model fit indices are influenced by suppression situations (Paulhus *et al.* 2004), suppressors may be retained in the final best model, with the risk of erroneous conclusions. Ideally, commonalities should be systematically computed and thoroughly inspected to identify, in each model, both suppressors and main contributors to model fit (see Blair *et al.* 2013 for such a model selection procedure using VIF). This framework may though be unrealistic when the number of models is large, notably in resistance model optimization. As a consequence, CA cannot fully address the issues described by Graves *et al.* (2013) about the current low performance of predictive analyses in landscape genetics, although multicollinearity is in all likelihood part of the problem. In model selection or resistance model optimization, commonalities should at least be investigated in the final best-fitted multivariate model, to avoid any erroneous interpretation of multivariate regressions. When models are nested, CA may also be used as a preliminary tool to assess commonalities in the full model, providing initial indications as to the relative contributions of predictors to the dependent variable.

We illustrated the use of CA in spatial genetics and showed that an in-depth understanding of multivariate regression results could be achieved when local collinearity structure was taken into account. Providing information about the unique contribution of predictors to genetic variability while easily revealing spurious correlations, CA is a promising tool in spatial genetics, especially as commonalities are easily computed in linear and logistic regressions, with functions and scripts now available for use in various statistical softwares such as R, SPSS or SAS (Nimon *et al.* 2008, 2010; Nimon 2010; Kraha *et al.* 2012; Roberts & Nimon 2012). CA may assume even greater value if used to assist the interpretation of regressions on data collected in different multicollinearity contexts, that is, coming from distinct spatial or temporal replicates (Anderson *et al.* 2010; Short Bull *et al.* 2011; Dormann *et al.* 2013), or on the contrary to assist the interpretation of regressions on data collected in distinct but co-occurring species (Storfer *et al.* 2010). CA may also facilitate the comparison of empirical and simulated data in the framework of landscape genetic model validation (Shirk *et al.* 2012). We thus strongly urge spatial geneticists to systematically investigate commonalities in explanatory full models or in best-fitted predictive models. Further methodological developments are now needed to determine how CA could be used to enhance the reliability of resistance model optimization procedures in spatial genetics. A particular attention should also be paid to the validity of regression CA in the specific framework of maximum-likelihood population-effects (MLPE) mixed models (Clarke *et al.* 2002), an increasingly used statistical tool in spatial genetics (Selkoe *et al.* 2010; Van Strien *et al.* 2012; Prunier *et al.* 2013). Finally, future studies should be conducted to assess how the location and magnitude of multicollinearity among predictors (including both synergistic associations and suppression situations) could be investigated using current variation-partitioning procedures in the framework of constrained ordination techniques (Borcard *et al.* 1992; Peres-Neto *et al.* 2006).

Acknowledgements

This work was supported by A. Bouron from the *Fédération Régionale des Chasseurs du Centre*, V. Giquel-Chanteloup from the *Fédération Départementale des Chasseurs de L'Indre*, as well as the *Société de Vènerie* and the *Fondation François Sommer pour la Chasse et la Nature*. We warmly thank two anonymous reviewers for insightful comments on a first draft of this manuscript.

References

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

Anderson CD, Epperson BK, Fortin MJ *et al.* (2010) Considering spatial and temporal scale in landscape-genetic studies of gene flow. *Molecular Ecology*, **19**, 3565–3575.

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

Angers B, Magnan P, Plante M, Bernatchez L (1999) Canonical correspondence analysis for estimating spatial and environmental effects on microsatellite gene diversity in brook charr (*Salvelinus fontinalis*). *Molecular Ecology*, **8**, 1043–1053.

Azen R, Budescu DV (2003) The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, **8**, 129–148.

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

Balkenhol N, Holbrook JD, Onorato D *et al.* (2014) A multimethod approach for analyzing hierarchical genetic structures: a case study with cougars *Puma concolor*. *Ecography*, **37**, 001–012.

Barbujani G, Oden NL, Sokal RR (1989) Detecting regions of abrupt change in maps of biological variables. *Systematic Zoology*, **38**, 376–389.

Beckstead JW (2012) Isolating and examining sources of suppression and multicollinearity in multiple linear regression. *Multivariate Behavioral Research*, **47**, 224–246.

Blair C, Jimenez-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**, e57433.

Bolliger J, Lander T, Balkenhol N (2014) Landscape genetics since 2003: status, challenges and future directions. *Landscape Ecology*, **29**, 361–366.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

ter Braak CJF, Prentice IC (1988) A theory of gradient analysis. *Advances in Ecological Research*, **18**, 271–317.

Braunisch V, Segelbacher G, Hirzel AH (2010) Modelling functional landscape connectivity from genetic population structure: a new spatially explicit approach. *Molecular Ecology*, **19**, 3664–3678.

Campbell KT, Tucker ML (1992) The use of commonality analysis in multivariate canonical correlation analysis. *Annual meeting of the Southwest Educational Research Association*, Houston, Texas.

Capraro RM, Capraro MM (2001) Commonality analysis: understanding variance contributions to overall canonical correlation effects of attitude toward mathematics on geometry achievement. *Multiple Linear Regression Viewpoints*, **27**, 16–23.

Castillo JA, Epps CW, Davis AR, Cushman SA (2014) Landscape effects on gene flow for a climate-sensitive montane species, the American pika. *Molecular Ecology*, **23**, 843–856.

Chen C, Durand E, Forbes F, Francois O (2007) Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. *Molecular Ecology Notes*, **7**, 747–756.

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.

Courville T, Thompson B (2001) Use of structure coefficients in published multiple regression articles: beta is not enough. *Educational and Psychological Measurement*, **61**, 229–248.

Cushman SA, Landguth EL (2010) Spurious correlations and inference in landscape genetics. *Molecular Ecology*, **19**, 3592–3602.

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.

Cushman SA, Wasserman TN, Landguth EL, Shirk AJ (2013) Re-evaluating causal modeling with Mantel tests in landscape genetics. *Diversity*, **5**, 51–72.

Dormann CF, Elith J, Bacher S *et al.* (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 27–46.

Drescher K, Henderson J, McNamara K (2001) Farmland prices determinants. *American Agricultural Economics Association Annual Meeting*, Chicago, Illinois.

Dudaniec RY, Spear SF, Richardson JS, Storfer A (2012) Current and historical drivers of landscape genetic structure differ in core and peripheral salamander populations. *PLoS One*, **7**, e36769.

Dudaniec RY, Rhodes JR, Wilmer JW *et al.* (2013) Using multilevel models to identify drivers of landscape-genetic structure among management areas. *Molecular Ecology*, **22**, 3752–3765.

Dyer RJ, Nason JD, Garrick RC (2010) Landscape modelling of gene flow: improved power using conditional genetic distance derived from the topology of population networks. *Molecular Ecology*, **19**, 3746–3759.

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

Epperson BK, McRae BH, Scribner K *et al.* (2010) Utility of computer simulations in landscape genetics. *Molecular Ecology*, **19**, 3549–3564.

Frantz AC, Bertouille S, Eloy MC *et al.* (2012) Comparative landscape genetic analyses show a Belgian motorway to be a gene flow barrier for red deer (*Cervus elaphus*), but not wild boars (*Sus scrofa*). *Molecular Ecology*, **21**, 3445–3457.

Garroway CJ, Bowman J, Wilson PJ (2011) Using a genetic network to parameterize a landscape resistance surface for fishers, *Martes pennanti*. *Molecular Ecology*, **20**, 3978–3988.

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

Graham MH (2003) Confronting multicollinearity in ecological multiple regression. *Ecology*, **84**, 2809–2815.

Graves TA, Wasserman TN, Ribeiro MC *et al.* (2012) The influence of landscape characteristics and home-range size on the quantification of landscape-genetics relationships. *Landscape Ecology*, **27**, 253–266.

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

Guarnizo CE, Cannatella DC (2014) Geographic determinants of gene flow in two sister species of tropical Andean frogs. *Journal of Heredity*, **105**, 216–225.

Guillot G, Rousset F (2013) Dismantling the Mantel tests. *Methods in Ecology and Evolution*, **4**, 336–344.

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

Hadi AS, Ling RF (1998) Some cautionary notes on the use of principal components regression. *American Statistician*, **52**, 15–19.

Holderegger R, Wagner HH (2008) Landscape genetics. *BioScience*, **58**, 199–207.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Holzhauer SIJ, Ekschmitt K, Sander AC, Dauber J, Wolters V (2006) Effect of historic landscape change on the genetic structure of the bush-cricket *Metrioptera roeseli*. *Landscape Ecology*, **21**, 891–899.

Jombart T, Devillard S, Dufour AB, Pontier D (2008) Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, **101**, 92–103.

Keller D, Holderegger R, van Strien MJ (2013) Spatial scale affects landscape genetic analysis of a wetland grasshopper. *Molecular Ecology*, **22**, 2467–2482.

King JE (2007) Standardized Coefficients in Logistic Regression. *Annual meeting of the Southwest Educational Research Association*, San Antonio, Texas.

King RS, Baker ME, Whigham DF *et al.* (2005) Spatial considerations for linking watershed land cover to ecological indicators in streams. *Ecological Applications*, **15**, 137–153.

Kraha A, Turner H, Nimon K, Zientek LR, Henson RK (2012) Tools to support interpreting multiple regression in the face of multicollinearity. *Frontiers in Psychology*, **3**, 1–10.

Landguth EL, Cushman SA (2010) CDPOP: a spatially explicit cost distance population genetics program. *Molecular Ecology Resources*, **10**, 156–161.

LeBreton JM, Ployhart RE, Ladd RT (2004) A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods*, **7**, 258–282.

Legendre P, Anderson MJ (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1–24.

Legendre P, Fortin MJ (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**, 831–844.

Legendre P, Legendre LFJ (1998) *Numerical Ecology*. Elsevier Science, Amsterdam.

Legendre P, Lapointe FJ, Casgrain P (1994) Modeling brain evolution from behavior – a permutational regression approach. *Evolution*, **48**, 1487–1499.

Lewis M (2007) Stepwise versus hierarchical regression: pros and cons. *Annual Meeting of the Southwest Educational Research Association*, San Antonio, Texas.

Lewis JW, Escobar LA (1986) Suppression and enhancement in bivariate regression. *The Statistician*, **35**, 17–26.

Lichstein JW (2007) Multiple regression on distance matrices: a multivariate spatial analysis tool. *Plant Ecology*, **188**, 117–131.

Luximon N, Petit EJ, Broquet T (2014) Performance of individual vs. group sampling for inferring dispersal under isolation-by-distance. *Molecular Ecology Resources*, **14**, 745–752.

Mac Nally R (2000) Regression and model-building in conservation biology, biogeography and ecology: the distinction between and reconciliation of 'predictive' and 'explanatory' models. *Biodiversity and Conservation*, **9**, 655–671.

Mac Nally R (2002) Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables. *Biodiversity and Conservation*, **11**, 1397–1401.

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

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

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

McRae BH, Shah VB (2009) *Circuitscape User Guide*. ONLINE. The University of California, Santa Barbara. Available at: http://www.circuitscape.org.

Monmonier M (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geographical Analysis*, **5**, 245–261.

Mood AM (1971) Partitioning variance in multiple regression analyses as a tool for developing learning models. *American Educational Research Journal*, **8**, 191–202.

Nanninga GB, Saenz-Agudelo P, Manica A, Berumen ML (2014) Environmental gradients predict the genetic population structure of a coral reef fish in the Red Sea. *Molecular Ecology*, **23**, 591–602.

Nathans LL, Oswald FL, Nimon K (2012) Interpreting multiple linear regression: a guidebook of variable importance. *Practical Assessment, Research & Evaluation*, **17**, 1–19.

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, **19**, 153–170.

Neter J, Wasserman W, Kutner MH (1990) *Applied Linear Statistical Models*, 3rd edn. Irwin, Chicago.

Newton RG, Spurrel DJ (1967) A development of multiple regression for the analysis of routine data. *Applied Statistics*, **16**, 51–64.

Nimon K (2010) Regression commonality analysis: demonstration of an SPSS solution. *Multiple Linear Regression Viewpoints*, **36**, 10–17.

Nimon KF, Oswald FL (2013) Understanding the results of multiple linear regression: beyond standardized regression coefficients. *Organizational Research Methods*, **16**, 650–674.

Nimon K, Reio TG (2011) Regression commonality analysis: a technique for quantitative theory building. *Human Resource Development Review*, **10**, 329–340.

Nimon K, Lewis M, Kane R, Haynes RM (2008) An R package to compute commonality coefficients in the multiple regression case: an introduction to the package and a practical example. *Behavior Research Methods*, **40**, 457–466.

Nimon K, Henson R, Gates M (2010) Revisiting interpretation of canonical correlation analysis: a tutorial and demonstration of canonical commonality analysis. *Multivariate Behavioral Research*, **45**, 702–724.

Nimon K, Henson R, Roberts K (2013a) Using fit indices to holistically assess forms of predictor importance in multilevel models. Paper presented at the *Annual meeting of the American Psychological Association*, Honolulu, HI.

Nimon K, Oswald FL, Roberts JK (2013b) Interpreting regression effects. R package version 2.0-0. URL: http://cran.r-project.org/web/packages/yhat/index.html.

O'Brien RM (2007) A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity*, **41**, 673–690.

Orsini L, Mergeay J, Vanoverbeke J, De Meester L (2013) The role of selection in driving landscape genomic structure of the waterflea *Daphnia magna*. *Molecular Ecology*, **22**, 583–601.

Pandey S, Elliott W (2010) Suppressor variables in social work research: ways to identify in multiple regression models. *Journal of the Society for Social Work and Research*, **1**, 28–40.

Paulhus DL, Robins RW, Trzesniewski KH, Tracy JL (2004) Two replicable suppressor situations in personality research. *Multivariate Behavioral Research*, **39**, 303–328.

Pedhazur EJ (1997) *Multiple Regression in Behavioral Research*, 3rd edn. Harcourt Brace, Orlando, Florida.

Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology*, **87**, 2614–2625.

Perez-Espona S, Perez-Barberia FJ, McLeod JE *et al.* (2008) Landscape features affect gene flow of Scottish Highland red deer (*Cervus elaphus*). *Molecular Ecology*, **17**, 981–996.

Perez-Espona S, McLeod JE, Franks NR (2012) Landscape genetics of a top neotropical predator. *Molecular Ecology*, **21**, 5969–5985.

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

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

Pilot M, Jedrzejewski W, Branicki W *et al.* (2006) Ecological factors influence population genetic structure of European grey wolves. *Molecular Ecology*, **15**, 4533–4553.

Pratt JW (1987) Dividing the indivisible: using simple symmetry to partition variance explained. In: *Proceedings of the Second International Tampere Conference in Statistics* (eds Pukkila T, Puntanen S), pp. 245–260. University of Tampere, Tampere.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Prunier JG, Kaufmann B, Fenet S *et al.* (2013) Optimizing the trade-off between spatial and genetic sampling efforts in patchy populations: towards a better assessment of functional connectivity using an individual-based sampling scheme. *Molecular Ecology*, **22**, 5516–5530.

Prunier JG, Kaufmann B, Lena JP *et al.* (2014) A 40-year-old divided highway does not prevent gene flow in the alpine newt *Ichthyosaura alpestris*. *Conservation Genetics*, **15**, 453–468.

R Development Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org/. Accessed on 22 August 2014

Ray-Mukherjee J, Nimon K, Mukherjee S *et al.* (2014) Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution*, **5**, 320–328.

Rioux Paquette S, Talbot B, Garant D, Mainguy J, Pelletier F (2014) Modelling the dispersal of the two main hosts of the raccoon rabies variant in heterogeneous environments with landscape genetics. *Evolutionary Applications*, **7**, 856–868.

Roberts JK, Nimon K (2012) A Software Solution for Conducting Logistic Commonality Analysis. *Annual meeting of the Southwest Educational Research Association*, New Orleans.

Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, **1**, 103–113.

Segelbacher G, Cushman SA, Epperson BK *et al.* (2010) Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics*, **11**, 375–385.

Selkoe KA, Watson JR, White C *et al.* (2010) Taking the chaos out of genetic patchiness: seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. *Molecular Ecology*, **19**, 3708–3726.

Shirk AJ, Wallin DO, Cushman SA, Rice CG, Warheit KI (2010) Inferring landscape effects on gene flow: a new model selection framework. *Molecular Ecology*, **19**, 3603–3619.

Shirk AJ, Cushman SA, Landguth EL (2012) Simulating pattern-process relationships to validate landscape genetic models. *International Journal of Ecology*, **2012**, 1–8.

Short Bull RA, Cushman SA, Mace R *et al.* (2011) Why replication is important in landscape genetics: American black bear in the Rocky Mountains. *Molecular Ecology*, **20**, 1092–1107.

Smith TJ, McKenna CM (2013) A comparison of logistic regression pseudo R^2 indices. *Multiple Linear Regression Viewpoints*, **39**, 17–26.

Smith AC, Koper N, Francis CM, Fahrig L (2009) Confronting collinearity: comparing methods for disentangling the effects of habitat loss and fragmentation. *Landscape Ecology*, **24**, 1271–1285.

Spear SF, Balkenhol N, Fortin MJ, McRae BH, Scribner K (2010) Use of resistance surfaces for landscape genetic studies: considerations for parameterization and analysis. *Molecular Ecology*, **19**, 3576–3591.

Stine RA (1995) Graphical interpretation of variance inflation factors. *American Statistician*, **49**, 53–56.

Storfer A, Murphy MA, Spear SF, Holderegger R, Waits LP (2010) Landscape genetics: where are we now? *Molecular Ecology*, **19**, 3496–3514.

Thompson B (2006) *Foundations of Behavioral Statistics: An Insight-Based Approach*, 1st edn. The Guildford Press, New York.

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

Vangestel C, Mergeay J, Dawson DA *et al.* (2012) Genetic diversity and population structure in contemporary house sparrow populations along an urbanization gradient. *Heredity*, **109**, 163–172.

Vigneau E, Devaux MF, Qannari EM, Robert P (1997) Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics*, **11**, 239–249.

Wang IJ (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**, 3403–3411.

Wedding LM, Lepczyk CA, Pittman SJ, Friedlander AM, Jorgensen S (2011) Quantifying seascape structure: extending terrestrial spatial pattern metrics to the marine realm. *Marine Ecology Progress Series*, **427**, 219–232.

Weigel DE, Connolly PJ, Powell MS (2013) The impact of small irrigation diversion dams on the recent migration rates of steelhead and redband trout (*Oncorhynchus mykiss*). *Conservation Genetics*, **14**, 1255–1267.

Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP (2006) Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, **75**, 1182–1189.

With KA (1997) The application of neutral landscape models in conservation biology. *Conservation Biology*, **11**, 1069–1080.

Worthington Wilmer J, Elkin C, Wilcox C *et al.* (2008) The influence of multiple dispersal mechanisms and landscape structure on population clustering and connectivity in fragmented artesian spring snail populations. *Molecular Ecology*, **17**, 3733–3751.

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

J.G.P. designed the study, performed modelling and simulating the work and wrote the manuscript; J.G.P. and K.N. analysed output data; K.N., M.C., X.L. and M.C.F. provided edits to the study.

Data accessibility

Spatial data sets I, II and III, as well as corresponding R-scripts used to get provided results. Dryad Digital Repository: doi:10.5061/dryad.86gm0.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 Details on CDPOP simulations.

Appendix S2 Differences between predictor metrics in data sets I (a) and III (b).