Regression commonality analyses on hierarchical genetic distances

Jérôme Prunier, Marc Colyn, Xavier Legendre, Marie-Christine Flamand

Regression commonality analyses on hierarchical genetic distances

Jérôme G. Prunier[1,4], Marc Colyn[2], Xavier Legendre[3] and Marie-Christine Flamand[1]

[1]Institut des Sciences de la Vie, Université catholique de Louvain, Croix du Sud 4-5, L7.07.14, 1348, Louvain-la-Neuve, Belgium

[2]CNRS-UMR 6553, Université de Rennes 1, Station Biologique 35380, Paimpont, France

[3]Museum National d'Histoire Naturelle (MNHN), DPBZ, Réserve de la Haute Touche 36290, Obterre, France

[4]Station d'Écologie Théorique et Expérimentale, Université de Toulouse, CNRS, UPS, France

Corresponding author: J. G. Prunier, Institut des Sciences de la Vie, Université catholique de Louvain, Croix du Sud 4-5, L7.07.14, 1348, Louvain-la-Neuve, Belgium. E-mail: jerome.prunier@gmail.com

Decision date: 15-Nov-2016

.'

Abstract

Landscape genetics is emerging as an important way of supporting decision-making in landscape management, in response to the deterioration of matrix permeability due to habitat loss and fragmentation. In line with unremitting methodological developments in landscape genetics, a new analytical procedure was recently proposed as a way of evaluating the effects of landscape gradients on genetic structures. This procedure is based on the computation of inter-individual hierarchical genetic distances (HGD), a metric of genetic differentiation taking into account the hierarchical structure in populations as inferred from clustering algorithms. HGD can be used as dependent variables in multivariate regressions to assess the effects of various landscape predictors on spatial patterns of genetic differentiation. However, multicollinearity may obscure the interpretation of multivariate regressions. We illustrate how regression commonality analyses (CA), a detailed variance partitioning procedure that can be used to deal with multicollinearity issues, can thoroughly improve our understanding of landscape connectivity when HGD are used as a dependent variable, with the red deer (*Cervus elaphus*) as an empirical example. Using logistic regression commonality analyses on HGD, we showed that semi-natural open areas, transportation infrastructures and, to a lesser extent, urban areas and rivers, were associated with an increase in hierarchical genetic differentiation in red deer. Regressions based on HGD provided detailed results that could not have been obtained with regressions based on standard genetic distances, with notably additional insights as to the possible influence of linear features such as roads and highways on landscape connectivity. Furthermore, CA helped identify synergistic associations among variables as well as suppressors, thus resolving inconsistencies among hierarchical levels and revealing spurious correlations that may have gone unnoticed in the course of classical regression analyses. We thus recommend the use of regression commonality analysis on hierarchical genetic distances as a promising statistical tool for landscape geneticists.

Keywords: Cervus elaphus; commonality analysis; landscape connectivity; spatial genetics; spurious correlations; suppression; variance partitioning.

.’

Introduction

The viability of local populations strongly depends on the permeability of the landscape matrix to individual movements between resource patches (Dunning et al. 1992, Taylor et al. 1993). However, in anthropogenic landscapes, habitat loss, land-use conversion and fragmentation are known to alter matrix permeability and constitute critical concerns for the conservation of most organisms (Ricketts 2001, Fahrig 2003, Fischer and Lindenmayer 2007). Understanding how the landscape matrix influences dispersal movements of individuals is thus critical for wildlife conservation. Allowing the assessment of functional landscape connectivity (Tischendorf and Fahrig 2000), landscape genetics is emerging as an important way of supporting decision-making in landscape management (Segelbacher et al. 2010, Manel and Holderegger 2013, Mijangos et al. 2015). Two complementary approaches can be used to investigate the influence of landscape features on landscape genetic patterns, namely boundary-based methods (e.g. Monmonier 1973, Barbujani et al. 1989, Pritchard et al. 2000, Chen et al. 2007, Jombart et al. 2008), aiming at identifying genetic structures, and direct gradient analyses (*sensu* ter Braak and Prentice 2004) ; e.g. Smouse et al. 1986, Legendre et al. 1994, Legendre and Anderson 1999, Selkoe et al. 2010), aiming at statistically testing how observed genetic structures are influenced by landscape features.

Boundary-based methods and direct gradient analyses rely on distinct assumptions (Balkenhol et al. 2014). Boundary-based methods, stemming from the metapopulation paradigm (Levins 1969, Hanski 1999), assume that individuals can be grouped into spatially distinct panmictic populations and that gene flow between populations is low. On the contrary, direct gradient analyses assume that all individuals come from a single continuous population, with non-negligible, though heterogeneous, levels of gene flow across the landscape, a situation where no clear genetic structure may be identified. There is however a large range of situations between these two extremes (Harrison 1991). Individuals may indeed be distributed into spatially distinct habitats but may exhibit higher dispersal rates than expected under the strict metapopulation paradigm (Baguette 2004, Mayer et al. 2009), resulting in a stratified (or hierarchical) genetic structure, with connected subpopulations at the inferior level of the hierarchy being nested within less-connected populations at the superior level. Hierarchical genetic structures are likely to occur in many wildlife species. For instance, species inhabiting complex heterogeneous landscapes (Urban et al. 1987) with various landscape features impeding dispersal movements differently (e.g. Coulon et al. 2008, Ginson et al. 2015; see also Supplementary material Appendix 1), species with

.’

behavioural mating restrictions (Chapuisat et al. 1997, Bouzat and Johnson 2004) or with restricted gene flow (e.g. Giles et al. 1998) may exhibit such stratified genetic structures.

Hierarchical genetic structures are usually analysed through the use of Analysis of Molecular Variance (AMOVA; Excoffier et al. 1992) or Spatial Analysis of Molecular Variance (SAMOVA ; Dupanloup et al. 2002), the latter indirectly allowing the detection of zones of sharp genetic changes, that is, potential barriers to gene flow. Whilst these methods do not make any assumption about Hardy-Weinberg equilibrium within populations, they however require populations to be delineated a priori and are not appropriate when genetic data have been gathered according to an individual-based sampling scheme. In line with unremitting methodological developments in landscape genetics, Balkenhol et al. (2014) proposed a new analytical framework to thoroughly investigate the influence of landscape features on hierarchical genetic structures. This framework consists of three main steps: first, running a clustering algorithm to identify hierarchical landscape genetic structures; then, using the outputs of this hierarchical clustering to compute hierarchical genetic distances (HGD); finally, using these HGD as dependent variables in subsequent direct gradient analyses (see Supplementary material Appendix 1 for a validation of the method based on simulated data). Hierarchical genetic clustering is an iterative clustering procedure in STRUCTURE (Pritchard et al. 2000), a software that allows identifying the most likely number of panmictic genetic clusters (K) and quantifies the probability that an individual belongs to each cluster in the form of individual ancestry values (or Q-values). When a hierarchical genetic structure exists, using the ΔK statistic (Evanno et al. 2005) allows identifying the optimal K-value, say K_1, at the uppermost level H_1 of the genetic structure. To identify additional population substructures, analyses can then be repeated for each of the K_1 clusters inferred at the previous step, resulting in a set of optimal K-values K_2 and new individual Q-values at the inferior hierarchical level H_2 (Coulon et al. 2008, Balkenhol et al. 2014). This iterative process is ended when the inferred number of genetic groups is 1 (according to log-likelihood plots), that is, when inferred Q-values for all individuals are about $1/K$ (Pritchard et al. 2000).

As a boundary-based method, hierarchical clustering provides insightful visual information about spatial hierarchical genetic structures when clusters at each hierarchical level are superimposed on a map (e.g. Cárdenas et al. 2015). In addition, Balkenhol et al. (2014) proposed using Q-values obtained from this procedure to compute HGD. Hierarchical genetic distances are pairwise estimates of genetic

.’

dissimilarity among individuals, taking into account the hierarchical structure of the considered population. Using HGD in direct gradient analyses may allow statistically (rather than just visually) assessing the influence of landscape features on hierarchical genetic structures: it is expected to help detect subtle landscape effects that may be otherwise overlooked in classical approaches assuming that all individuals come from a single continuous population (Balkenhol et al. 2014). Several authors proposed similar approaches, based on the use of summary statistics derived from clustering algorithms (Murphy et al. 2008, Alvarado-Serrano and Hickerson 2016, Howell et al. 2016), although they were not designed to take stratified clustering into account. The use of HGD allowed disentangling the relative influence of various landscape features at various hierarchical levels in *Puma concolor*, highlighting the need for a multi-scale management of this endangered species in Idaho and Western Montana (USA). It is however, to our knowledge, the only example of the use of HGD in the current literature.

Nevertheless, statistically evaluating the effects of spatial features on observed genetic structures using direct gradient analyses has been shown to be plagued by even weak levels of multicollinearity, because of synergistic or antagonistic processes operating among non-independent predictors (Farrar and Glauber 1967, Courville and Thompson 2001, Smith et al. 2009, Dormann et al. 2013, Prunier et al. 2015), a situation that is likely to arise in most landscape genetic studies. Regression models based on HGD are no exception, and a thorough understanding of the correlation structure among predictors is thus crucial to avoid erroneous conclusions and subsequent inefficient or counterproductive conservation measures. To respond to this challenge, landscape geneticists can rely on regression commonality analysis (CA), a detailed variance partitioning technique that is particularly well suited in the case of multicollinearity (Nimon and Oswald 2013, Ray-Mukherjee et al. 2014, Prunier et al. 2015). Details about this statistical procedure are provided in the Material and Methods section.

In this study, we first aimed to highlight how the use of HGD rather than standard genetic distances as an input variable in direct gradient analyses can thoroughly improve our understanding of the influence of landscape features on spatial patterns of hierarchical genetic differentiation; secondly, we aimed to provide a detailed empirical illustration of how CA, a tool that is still at an early stage in landscape genetics (e.g., Renner et al. 2015, Gouskov et al. 2016), can help identify spurious correlations that may have gone unnoticed in the framework of a classical regression model. With the red deer (*Cervus elaphus*), a large ungulate in central France, as an empirical example, we used multiple

regressions on distance matrices (Smouse et al. 1986) and logistic regressions on distance matrices (Prunier et al. 2015) to assess the influence of spatial predictors on both inter-individual standard and hierarchical genetic distances. In each case, predictors' contribution to model fit was also evaluated in the light of CA. We expected spatial patterns of genetic differentiation in red deer to be shaped by anthropogenic landscape features such as transportation infrastructures (Jackson and Fahrig 2011), urban and agricultural areas (Frantz et al. 2006, Godvik et al. 2009). We also expected analyses based on HGD to outperform those based on classical genetic distances in the detection of such landscape effects.

Material and methods

Multicollinearity, suppression and commonality analysis

The interpretation of generalized linear models, including linear and logistic regressions, is often obscured by the presence of multicollinearity, that is, specific patterns of non-null bivariate correlations between predictors. For a given dependent variable, multicollinearity among predictors may indeed be responsible for a distortion of beta weights (i.e., standardized regression coefficients), as well as a distortion of their standard errors and marginal statistics used to test their significance. This phenomenon is known as suppression (Horst 1941, Conger 1974, Lewis and Escobar 1986, Cohen et al. 2003, Paulhus et al. 2004, Beckstead 2012). In any suppression situation, a predictor (say X1), having a null or low zero-order correlation with the dependent variable, purifies the relationship between another predictor (X2) and the dependent variable by removing the irrelevant variance that X2 shares with the dependent variable (i.e., the "criterion-irrelevant variance" of X2; Beckstead 2012). As a result, the presence of the suppressor X1 in the regression equation increases the predictive validity of X2 as well as the overall model fit, but at the cost of a more or less pronounced distortion of the beta weight assigned to X1, which is likely to cause serious difficulties for the proper interpretation of model parameters.

Since Conger (1974), three kinds of suppression are recognized: classical, cross-over and reciprocal suppression (see Paulhus et al. 2004 for a detailed review). In the case of classical suppression, a suppressor variable, although unrelated to the dependent variable (that is, with a null or negligible zero-order correlation), yet receives a high beta weight when it is included in the model. In the case of cross-over suppression, a suppressor has low (positive or negative) zero-order correlation with the dependent variable but yet receives a beta weight that is relatively large and of opposite sign when compared with its zero-order correlation. Finally, in the case of reciprocal suppression, two predictors are (for instance)

positively intercorrelated but also inversely correlated to the dependent variable: this configuration leads to an artificial boost in their respective beta weights. Given the possible important distortions of beta weights in such situations, identifying suppressors is crucial to avoid a flawed interpretation of generalized linear models.

To meet this challenge, researchers may consider the use of commonality analysis (CA), a detailed variance partitioning procedure that can provide decisive support in assessing the reliability of model parameters (beta weights, p-values) in face of multicollinearity (e.g. Newton and Spurrell 1967, Mood 1971, Creager 1971, Seibold and McPhee 1979, Nimon and Oswald 2013, Ray-Mukherjee et al. 2014, Prunier et al. 2015). Commonality analysis decomposes the fit index of a generalized linear model into commonality coefficients (or commonalities), including both unique (U) and common (C) effects. Commonalities are non-overlapping components of variance in the dependent variable that ensue from formulae involving the regression of the dependent variable over all possible subsets of predictors (Mood 1971). For a given predictor, the sum of unique and common effect, that is, the total contribution T of the predictor, corresponds to its squared zero-order correlation with the dependent variable (that is, T C).

Unique effects U, or first-order commonalities, quantify the amount of variance in the dependent variable that is uniquely explained by a particular predictor, that is, the amount of variance assigned to a predictor when it is entered last in the model. Common effects C indicate the amount of variance that can be jointly explained by two (second-order commonalities) or more predictors together (k^{th}-order commonalities). Positive common effects occur in the case of synergistic association (or redundancy; Paulhus et al. 2004) among correlated predictors. Situations where C is substantially larger than U indicate that a predictor only indirectly contributes to the variance in the dependent variable (or to the model fit) because of its high correlation with other predictors (Conger 1974) and that the observed significant beta weight is actually partially or totally artefactual. Note that the notions of direct or indirect contributions to model fit should not be confused with notions of direct or indirect causal links between variables (Wright 1921): as a general rule, CA provides unique opportunities to assess the reliability of model parameters (beta weights, p-values) in face of collinearity but may in no circumstances be considered as a way to inform causal relationships among variables (see Supplementary material Appendix 9).

While positive commonalities are indicative of redundancy, negative commonalities are considered as a way of identifying the loci and magnitude of classical or reciprocal suppression: this is probably the main contribution of CA over other variance partitioning procedures (Nimon and Oswald 2013) in which negative variance components are usually interpreted as zero (Legendre and Legendre 1998, Peres-Neto et al. 2006). Negative commonalties actually quantify the amount of predictive power that would be lost by other predictors if the (classical or reciprocal) suppressor variable was not included in the regression model (see Supplementary material Appendix 7 for an illustration). In the case of a classical suppressor, the sum of common contributions of the suppressor with other predictors is negative and counterbalances its unique contribution to the variance in the dependent variable (that is, T $C \approx 0$). In the case of a reciprocal suppressor, the predictor is involved in several (if not all) negative commonalities although the sum of its common contributions does not counterbalance its unique contribution to the variance in the dependent variable (that is, $T = U + C > 0$): it thus acts as partial suppressor (Nimon 2010), removing irrelevant predictive variance in spite of its relationship with the dependent variable. As cross-over suppressors may not be involved in any negative commonality and thus may not be identified through the sole investigation of negative commonalities. Hence, it is also crucial to investigate possible discrepancies between the sign of zero-order correlations (or the sign of structure coefficients; Courville and Thompson 2001) and the sign of beta weights to identify cross-over suppressors.

As an additional layer of consideration when assessing CA results, one may also investigate the sum of scaled total contributions, that is, the sum of total contributions T divided by model fit (in the case of ordinary least-square regression, scaled total contributions correspond to squared structure coefficients). This sum is supposed to be 100 % when predictors are orthogonal, but may exceed 100% in the case of synergistic association among variables, or on the contrary be less than 100% in the case of suppression (Nimon 2010). Simultaneously investigating zero-order correlations, beta weights and commonalities can thus help identify the location and magnitude of suppression, revealing possible spurious correlations (notably in cases of classical and cross-over suppression), and thus thoroughly improving the interpretation of multivariate models.

Study area and biological model

The study was carried out in Centre region (France), over an area of approximately 30000 km² (170 x 180 km). The study area is a typical agro-forestry landscape with two predominant landscape features (semi-natural open areas 69% and woods 25%; Fig. 1), a river network including the river Loire and its tributaries and secondary anthropogenic elements such as urban areas (mostly villages and small towns), roads and fenced highways (A20 and A71) equipped with ecopassages (Lesbarrères and Fahrig 2012). The spatial configuration of these anthropogenic features mainly stems from the historical development of urban areas along rivers. Roads layouts, connecting cities and villages, thus often coincide with the course of streams, possibly leading to collinearity issues among landscape features.

We focused on red deer (*Cervus elaphus*), an intensively managed game species in Europe (Apollonio et al. 2010, Zachos et al. 2016). Red deer are able to cross many kinds of terrain (Perez-Espona et al. 2009) and to move over thousands of metres (Prévot and Licoppe 2013), seasonal migrations and dispersal events in stags typically extend up to 50 km (Daniels and McClean 2003, Hamann et al. 2003, Jarnemo 2008). In continental Europe, resource supplementation is an important characteristic of habitat selection in red deer, with a trade-off between using shelter in wooded habitats during the day and foraging semi-natural open areas at night (Godvik et al. 2009, Allen et al. 2014). Nevertheless, the absence or the loss of connectivity between forest massifs due to resistant natural landscape features or human activities is expected to hinder gene flow in this species (Perez-Espona et al. 2009, Dellicour et al. 2011, Frantz et al. 2012, Pérez-González et al. 2012).

Following an individual-based sampling scheme (Prunier et al. 2013), we collected red deer tissue samples (ear or muscle biopsy) from adult individuals during 2012 and 2013 hunting seasons and retained 669 samples for analyses (347 males and 322 females; See Supplementary material Appendix 2 for details). When individual coordinates were not precisely recorded but provided at the scale of hunting areas (forest massifs), we randomly scattered samples inside each massif with one individual per square-kilometre (Graves et al. 2012).

Laboratory procedures

We extracted DNA using a chloroform-based extraction method (Doyle and Doyle 1990) and performed genotyping using 23 microsatellite loci in four multiplex (MP). MP1, MP2 and MP3 were the same as described in Dellicour *et al.* (2011). MP4 contained nine additional microsatellite loci: CelJP27,

OarFCB304, OarFCB5, RT1, T156, T193, T26, T268 and T501 (Perez-Espona et al. 2009). Polymerase chain reactions (PCR) were performed on a Verity thermocycler (Applied Biosystems, Warrington, UK) using Qiagen Multiplex Kit (Qiagen, Hilden, Germany) in a 5.2 µL volume containing 2.5 µL Qiagen multiplex PCR Master Mix, 1 µL of water, 0.5 µL of Q-solution, 0.5 µL of primer mix (between 0.07 and 0.16 µM of each primer) and 0.7 µL of genomic DNA at 15 ng.µL^{-1}. After initial denaturation at 94°C for 10 min, cycling conditions were the following: denaturation at 94°C for 30 s; annealing at 56 (MP1), 53 (MP2 and MP3) or 55°C (MP4) for 90 s; extension at 72 °C for 30 s. A total of 32 cycles were performed for MP1, 36 cycles for MP2 and MP3, and 30 cycles for MP4. Final incubation was at 72°C for 10 min. PCR products were separated using an ABI 3130xl Genetic Analyzer (Applied Biosystems, Warrington, UK) and genotypes were analysed using GENEMAPPER 4.1 (Applied Biosystems).

To assess the reliability of genotyping, we estimated the mean error rate per locus e_l (Pompanon et al. 2005) by blind replication of 48 out of 1033 samples (4.6%), collected as part of a general research program on *C. elaphus*. We checked the presence of null alleles per locus with MICROCHECKER 2.2.3 (Van Oosterhout et al. 2004) by analysing homozygote excess in four well-defined populations located in four distinct forest massifs (*Choeurs-Bommiers*, *Lancosme*, *Loches* and *Orléans*; Fig. 1). We also estimated the number of alleles per locus, observed and expected heterozygosity, and checked Hardy-Weinberg equilibrium and gametic disequilibrium with GENEPOP 4.2.1(Rousset 2008) after sequential Bonferroni correction to account for multiple related tests (Rice 1989).

Hierarchical genetic clustering

We followed the hierarchical genetic clustering procedure described in Coulon et al. (2008). We used STRUCTURE with the admixture model and the correlated allele frequency model, without prior population information. Runs were performed with a burn-in period of 200.000 and 200.000 subsequent Markov chain Monte Carlo (MCMC) repetitions. At each hierarchical level, the number *K* of clusters ranged from 1 to 10 and 5 runs were performed for each value. To ensure MCMC convergence in inferior hierarchical levels, we adjusted the standard deviation of *alpha*, the Dirichlet parameter for the degree of admixture, from 0.025 (default) to 0.5 each time *alpha* plots showed substantial fluctuations before the end of the burn-in. We used STRUCTURE HARVESTER (Earl and vonHoldt 2012) to obtain Log-likelihood plots and ΔK statistics and to infer the optimal *K*-value. We then performed 20 runs with this

optimal K-value and compiled the ten best runs using CLUMPP (Jakobsson and Rosenberg 2007) to get final averaged individual Q-values. Individuals were assigned to the cluster for which their Q-value was the higher and, provided this value was higher than 0.6, were considered for clustering at the inferior hierarchical levels until they were assigned to their final hierarchical cluster, that is, when the optimal K-value was 1. For each hierarchical level (H_1 to H_3; see results), we used Q-values to compute pairwise matrices of ancestry-based HGD following Balkenhol et al. (2014).

Landscape predictors

We computed several predictors from various landscape features likely to affect gene flow in red deer through habitat loss and/or fragmentation. First, we extracted non-forest land-cover elements from the Corine Land Cover 2006 dataset (European Environment Agency) and combined them into two feature classes: *open* (for semi-natural open areas such as meadows and crops) and *urban* (for urbanised areas). In the same way, we extracted linear elements from national maps (BD Topo from National Geographic Institute, France, 1/25 000) and combined them into three additional feature classes: *rivers*, main *roads* and *highways*. Note that we did not consider the A85 highway in this study as it was in use for only five years at the time of sampling: owing to the generation time in *Cervus elaphus* (estimated at 8.3 years in Kruuk et al. 2002), this lag time was regarded as too short for a possible genetic effect to be detectable (Landguth et al. 2010). All spatial data were managed and rasterized at a 200 m resolution using ARCGIS 10.0 and its extension SPATIAL ANALYST. We then computed five resistance surfaces based on spatial densities of each landscape feature. To do so, we overlaid a 1 km grid on each layer and calculated the percentage of each feature per square-kilometre (Balkenhol et al. 2014). In each layer, we rescaled pixel resistance values to range from 1 (null or extremely low densities) to 100 (maximal densities), except in the case of highway pixels, which systematically received a resistance value of 100. The five layers were finally used in CIRCUITSCAPE 3.5.8 (McRae and Shah 2009) to compute pairwise effective distances among individuals. Isolation-by-distance being incorporated into resistance surfaces, we did not include Euclidean distances as an additional explanatory variable (McRae 2006, Garroway et al. 2011, Peterman et al. 2014). We z-transformed all landscape predictors into z-scores (by subtracting the mean and dividing by the standard deviation) to facilitate the comparison of model parameters (Schielzeth 2010). We finally used Pearson's correlation coefficients *r* and variance inflation factors VIF

as measures of linear relationships among predictors (Dormann et al. 2013) to assess multicollinearity in our dataset.

Multiple linear regression and CA on standard genetic distances

As a first dependent variable, we used standard, non-hierarchical genetic distances based on the Bray-Curtis percentage dissimilarity metric (dataset *Bc*; Legendre and Legendre 1998). To assess the relationships between standard genetic distances and landscape predictors, we used multiple linear regressions on distance matrices (Smouse et al. 1986), a statistical procedure that is similar to classical multiple ordinary least-square regression, except that the significance of model fit (multiple R^2) and beta weights β is assessed through permutations of the dependent matrix (Legendre et al. 1994). Multiple linear regressions and CA were respectively conducted using packages *ecodist* (Goslee and Urban 2007) and *yhat* (Nimon et al. 2008) in R 3.1.2 (R Development Core Team 2014). We assessed significance of multiple R^2 and beta weights β through 1000 permutations after sequential Bonferroni correction (Holm 1979) and computed 95% confidence intervals around commonalities using a bootstrap procedure, with 1000 replicates based on a random removal of 30% of individuals without replacement (Peterman et al. 2014, Prunier et al. 2015). These confidence intervals can be used to identify predictors whose unique contribution to the variance in the dependent variable is non-null and robust to the random removal of a subset of individuals.

Multiple logistic regressions and CA on hierarchical genetic distances

As a second set of dependent variables, we used HGD computed for each identified hierarchical level. Z-transformed HGD violated the assumption of normality of residuals in ordinary least-square regression (see Supplementary material Appendix 4), notably because of a bimodal distribution of HGD at the first hierarchical level, with low HGD for pairs of individuals belonging to the same cluster and high HGD for pairs of individuals belonging to distinct clusters. This bimodal distribution is consistent with the metapopulation paradigm that underlies boundary-based methods such as STRUCTURE and assumes that individuals don't come from a single continuous population but can be grouped into spatially distinct panmictic populations. To handle such bimodal data, we used multiple logistic regressions on distance matrices (Prunier et al. 2015). In multiple logistic regressions, odds-ratio ψ, that

is, semi-standardized beta weights $\hat{\beta}$ raised to the exponent ($\psi = e^{\hat{\beta}}$), are used to evaluate the increase of the likelihood of a success (that is, the occurrence of the outcome of interest) with one standard deviation change in a given predictor. This approach is highly relevant here, as HGD can easily be translated as the probability that two individuals come from distinct clusters. Using zero as a probability threshold, we re-coded z-transformed HGD matrices into binary matrices with 0 for pairs of individuals with negative z-scores (thus belonging to the same cluster) and with 1 for pairs of individuals with positive z-scores ("success" of coming from distinct clusters or being genetically dissimilar). Note that we created binary matrices from the z-transformed ancestry-based HGD rather than from the original assignment of individuals to each cluster because HGD are computed so as to account for genetic structures already contained in superior levels of the hierarchy (Balkenhol et al. 2014). The two procedures may however lead to the same results at the first hierarchical level. All logistic regressions were performed using the *glm* function with *logit* link in R 3.1.2. We computed semi-standardised beta weights $\hat{\beta}$ following King (2007) with the mean predicted probability as a reference value and used the *NagelkerkeR2* function in package *fmsb* to compute the Nagelkerke's Index as a pseudo-R^2. Logistic CA were performed with the *cc4log* function provided in Roberts & Nimon (2012). We assessed significance levels of model fit and semi-standardised beta weights $\hat{\beta}$ through 1000 permutations after sequential Bonferroni correction (Holm 1979) and, as previously, computed 95% confidence intervals around commonalities using a bootstrap procedure, with 1000 replicates based on a random removal of 30% of individuals without replacement. R-scripts used to perform CA can be found here: http://www.jeromeprunier.eg2.fr/5.html.

Results

Genetic data and hierarchical genetic structures

No locus departed significantly from Hardy–Weinberg equilibrium after sequential Bonferroni correction (mean fixation index ranging from 0 to 0.125; Supplementary material Appendix 2) or showed evidence for null alleles, when considering populations from *Choeurs-Bommiers*, *Lancosme*, *Loches* and *Orléans* (Fig. 1). However, preliminary analyses on additional French populations (sampled as part of a general research program on *C. elaphus*; Colyn et al. 2015) indicated the presence of null alleles in CSM66 (data not shown), which was thus discarded to avoid any bias in subsequent analyses. Furthermore, we detected significant gametic disequilibrium between ILSTS06 and T26 on the one hand and between OarFCB5 and markers OarFCB304 and T501 on the other hand. We thus discarded

ILSTS06, less polymorphic than T26, as well as OarFCB5. Our final genetic dataset contained 20 loci. The effective number of alleles ranged from 2.75 to 9.5 (Supplementary material Appendix 3). The genotyping error rate, estimated by blind replications as the mean error rate per locus e_l, was < 4.5% in CSSM16, < 2.4% in CSPS115 and CSSM14 and 0% for the 17 other loci.

We identified three hierarchical genetic levels (Fig. 2). At the first level, individual assignments showed that one cluster (A) mostly consisted of individuals north of the Cher river and east of the A20 highway (including several forest massifs of which *Orléans*, *Choeurs-Bommiers* and *Châteauroux)*, while the second cluster (B) mostly consisted of individuals in the south-western part of the study area (including several forest massifs of which *Lancosme* and *Loches*; Fig. 1). We only identified nine cross-assigned individuals at this level, suggesting that exchanges between these two superior hierarchical clusters were possible but not frequent. Clusters A and B were further divided into several clusters at inferior levels of the hierarchy, with all individuals assigned to their final cluster at the third level (Fig. 2). We identified a total of seven final clusters, revealing important genetic substructuring in this dataset. The proportion of individuals with ancestry value ≥ 0.6 ranged from 88 to 97% across hierarchical levels, indicating that most individuals were assigned with satisfactory confidence at each level of the hierarchy.

Landscape predictors

Absolute values of Pearson's correlation coefficients, ranging from 0.238 to 0.689, that is, below the traditional threshold of 0.7, and VIF values, ranging from 1.225 to 2.393 (Table 1), together suggested moderate multicollinearity in this study (Dormann et al. 2013). Highest correlations occurred between *urban* and linear features *rivers* ($r = 0.561$), *roads* ($r = 0.539$) and *highways* ($r = 0.689$), in accordance with the historical implementation of urban areas along rivers and the subsequent creation of transportation infrastructures connecting cities (Fig. 1). All bivariate correlations among predictors were positive.

Multiple linear regressions and CA on standard genetic distances

When using standard genetic distances (*Bc*), the multiple linear regression was significant but only explained 3.85% of variance in the dependent variable (Table 2). Nevertheless, all predictors were highly significant. Only *rivers* had a negative beta weight (Supplementary material Appendix 5). However,

investigating commonalities allowed clarifying these results (Table 2). The sum of all negative commonalities showed that 5.46% of the regression effect was caused by suppression (Fig. 3) but the sum of scaled total contributions exceeded 100%, indicating that the regression was mostly biased by synergistic associations among predictors. Predictor *rivers* showed a discrepancy between the signs of β (β = -0.045) and zero-order correlation with the dependent variable (r = 0.062), specifically designating it as a cross-over suppressor (Table 2). Predictors *urban* and *roads* had negligible unique contributions ($U = 0.001$) when compared to the sum of their common contributions ($C \geq 0.016$) and only indirectly contributed to the variance in Bray-Curtis distances (*Bc*) because of shared contributions with *highways*, through second-order ([*urban,highways*]), third-order ([*open,urban,highways*] and [*urban,roads,highways*]), fourth-order ([*open,urban,roads,highways*]) and fifth-order positive commonalities (Fig. 3). Although the unique contributions of predictors *highways* and *open* were low (0.3% and 0.9% respectively), confidence intervals around their first-order commonalities indicated that they were robust contributors to the variance in standard genetic distances. Nevertheless, the global model fit was particularly low, suggesting that spatial patterns of genetic differentiation were poorly mirrored by the use of standard genetic distances.

Multiple logistic regressions and CA on hierarchical genetic distances

When using HGD, 98% of genotypes were assigned at the first hierarchical level. Logistic regression model was significant and explained 18.74% of variance in HGD (Table 3). All predictors were highly significant, with *rivers* the only variable associated with a negative semi-standardised beta weight $\hat{\beta}$ (Supplementary material Appendix 5). As previously, investigating commonalities allowed refining these results (Table 3). The sum of all negative commonalities showed that only 0.43% of the regression effect was caused by suppression while the sum of scaled total contributions, exceeding 100%, indicated that the regression was mostly biased by synergistic associations among predictors. Predictor *rivers* showed a discrepancy between the signs of $\hat{\beta}$ ($\hat{\beta}$ = -0.007) and zero-order correlation with the dependent variable (r = 0.181), specifically designating it as a cross-over suppressor variable. The total contribution of predictor *urban* (T 120) was almost totally explained by its common effects (U 0 while 119) because of its high positive correlation with other predictors, notably *highways* (r), through second-order ([*urban,highways*]), third-order ([*open,urban,highways*],

[*urban,rivers,highways*] and [*urban,roads,highways*]), fourth-order ([*open,urban,roads,highways*] and [*urban,rivers,roads,highways*]) and fifth-order positive commonalities (Fig. 4). Predictor *urban* thus only showed indirect contribution to the dependent variable. Although showing large common effects because of their high positive correlations with *highways* (e.g., [*open,highways*] 17%, [*roads,highways*] = and [*urban,roads,highways*] of model fit), predictors *open* and *roads* had non-negligible unique contributions U, respectively explaining 7.56% (U 014) and 6.10% ($U = 0.011$) of model fit (Fig. 4). Finally, *highways* appeared as the main contributor to the variance in genetic differentiation at this first hierarchical level, uniquely accounting for 12.84% of the variance explained by the logistic model ($U = 0.024$). The presence of highways and, to some extent, the density of semi-natural open areas and the density of roads, were thus the main contributors to the variance in genetic differentiation at the first level of the hierarchy, with other predictors either acting as cross-over suppressor (*rivers*) or indirectly contributing to model fit through a synergistic association with *highways* (*urban*).

Ninety-six per cent of genotypes were assigned at the second hierarchical level. The logistic regression model was significant and explained 9.41% of variance in HGD (Table 3). Only predictors *urban*, *roads* and *highways* were significant after sequential Bonferroni correction. All corresponding semi-standardised beta weights were positive (Supplementary material Appendix 5). The sum of all negative commonalities showed that only 1.38% of the regression effect was caused by suppression while the sum of scaled total contributions, exceeding 100%, indicated that the regression was mostly biased by synergistic associations among predictors. The non-significant predictor *open* was easily identified as a cross-over suppressor variable, as it showed a discrepancy between the signs of $\hat{\beta}$ ($\hat{\beta}$ = -0.013) and zero-order correlation with the dependent variable (r = 0.172). Predictor *highways* showed the highest unique effect (U), uniquely contributing to 13.39% of the variance explained by the logistic model. Predictors *urban* and *roads* showed little unique contribution (U 005) when compared to the sum of their common contributions (C 048): their respective contributions to the variance in the dependent variable were mostly indirect because of their high positive correlations with *highways*, notably through second-order ([*urban,highways*] and [*roads,highways*]), third-order ([*urban,rivers,highways*] and [*urban,roads,highways*]), fourth-order ([*urban,rivers,roads,highways*]) and fifth-order positive commonalities (Fig. 5). Nevertheless, their unique contributions were robust to the random removal of

30% of individuals and could be considered as non-negligible. The main contributors to the variance in inter-individual measures of genetic differentiation at the second hierarchical level were thus the densities of highways and, to a lesser extent, the densities of roads and urban areas.

Finally, 88% of genotypes were assigned at the third hierarchical level. The logistic regression model was still significant (p = 0.018) but only explained 1.21% of variance in HGD (Table 3). This low model fit may suggest that STRUCTURE overestimated the number of clusters and thus the number of hierarchical levels, for instance because of isolation-by-distance (Frantz et al. 2009). However, inferred clusters at the third hierarchical level were similarly observed when using spatial principal component analysis (Jombart et al. 2008), a clustering algorithm taking spatial autocorrelation into account (Supplementary material Appendix 6), thus supporting inferred clusters at this level. Only two predictors were significant after sequential Bonferroni correction: *rivers* and *roads* (Table 3; Supplementary material Appendix 5). Investigating commonalities clearly indicated that *roads* acted as a classical suppressor, as its unique contribution was almost totally counter-balanced by the sum of its common effects (see Supplementary material Appendix 7 for a detailed interpretation of this suppression effect). As a result, the sum of scaled total contributions was less than 100%, indicating that the regression effects were indeed confounded by suppression. The density of rivers thus appeared as the only contributor to the variance in measures of genetic differentiation at the third hierarchical level, accounting for 33% of model fit ($U/\mathrm{R}^2 = 0.004/0.012 = 0.33$). Although this result was biologically meaningful, model fit was low, suggesting that other spatial or historical processes (Epps and Keyghobadi 2015), not considered in this study, were also responsible for the observed genetic structure at this level.

Discussion

In this work, we used regression commonality analyses to assess the relative influence of various landscape features on standard and hierarchical pairwise measures of genetic differentiation in the red deer (*Cervus elaphus*), a large ungulate in central France. We confirmed that the use of HGD as a dependent variable in direct gradient analyses is a relevant approach to identify possible drivers of genetic differentiation when compared to the classical use of standard genetic distances. Importantly, we illustrated how regression commonality analyses, and notably logistic regression commonality analyses on HGD, could outperform classical regression models, actually providing additional insights as to the possible influence of linear features such as roads and highways on landscape connectivity, while

revealing spurious correlations resulting from multicollinearity among spatial predictors in empirical datasets.

Regression commonality analyses on hierarchical genetic distances

Logistic regressions on HGD provided additional detailed results that could not have been obtained in the course of classical linear regressions on standard genetic distances. Indeed, the only predictors that significantly contributed to the variance in standard measures of genetic differentiation were the density of semi-natural open areas, a widespread land-cover feature acting on the overall genetic structure, and, more marginally, the presence of highways. In general terms, the influence of discrete linear features acting as local barriers to gene flow, such as highways and roads, could only be confidently identified when using HGD. Boundary-based detection methods allow the detection of local sharp genetic variations while classical direct gradient analyses based on standard genetic distances, assuming that all sampled individuals come from a single continuous population, rather give insight into the importance of landscape permeability on overall genetic variation (Guillot et al. 2009). This observation may explain why, when considering standard measures of genetic differentiation (*Bc*), multiple linear regression explained less than 4% of variance in the dependent variable, that is, two to five times less than the amount of variance explained in multiple logistic regression based on HGD at the two first levels of the hierarchy, clearly indicating better model fit when accounting for hierarchical genetic patterns as inferred from hierarchical clustering (Balkenhol et al. 2014; see also Supplementary material Appendix 8).

This is an outstanding benefit of the use of direct gradient analyses on HGD over classical procedures when dealing with multi-scaled landscape genetic processes (Balkenhol et al. 2014), a situation likely to occur in many empirical datasets, whatever admixture levels among inferred clusters (provided clusters actually exist; e.g. Prunier et al. 2014). In our dataset, most individuals were confidently assigned to their cluster (Q-values > 0.6), while only few cougars showed high ancestry values in Balkenhol et al. (2014). In both cases though, direct gradient analyses on HGD provided a consistent overview of landscape connectivity, with new insights as to the possible influence of local discrete barriers to gene flow. Note however that inferred hierarchical genetic structures should always be considered with caution or compared to the outputs of other methods (see for instance Supplementary

material Appendix 6), as Bayesian clustering methods may sometimes lead to wrong inferences (Frantz et al. 2009, Puechmaille 2016).

Nevertheless, these conclusions could not have been drawn without the help of CA. In this study, CA provided a clear quantification of unique and common contributions of predictors to the variance in dependent variables and helped identify synergistic associations among variables as well as suppressors, thus resolving inconsistencies among hierarchical levels and revealing spurious correlations that may have otherwise gone unnoticed. For instance, when evaluating the influence of predictors on standard or hierarchical genetic distances without the help of CA, *rivers* would have been spuriously considered as a landscape feature facilitating gene flow, while it was on the contrary identified as a possible driver of genetic differentiation at the third hierarchical level. This predictor actually acted as a cross-over suppressor, purifying the relationship between other predictors and (standard or hierarchical) pairwise measures of genetic differentiation, but at the cost of spurious correlations with the dependent variables. Because of the specific configuration of the landscape in our study, stemming from the historical implementation of human activities along river valleys, it may also have been difficult to disentangle the relative influence of anthropogenic landscape features such as highways, roads and urban areas on spatial patterns of genetic differentiation without the help of commonalities. For instance, the apparent influence of urban areas on standard measures of genetic differentiation and HGD at the first level was likely artefactual, reflecting synergistic association between urban areas and highways: *urban* had actually little direct contribution to the dependent variables but a substantial amount of variance in *highways* was assigned to *urban* in the process of computing standard and semi-standardised beta weights. These spurious correlations, resulting from collinearity among predictors, were easily confounded by investigating zero-order correlations, beta weights and commonalities, making CA an essential statistical procedure to assist in the interpretation of direct gradient analyses in landscape genetics.

Landscape connectivity in red deer

Hierarchical genetic clustering in red deer allowed identifying two distinct clusters at the first hierarchical level, separated by the Cher valley on the one hand, and by the A20 highway and the vast adjacent agricultural plain on the other hand (Fig. 1). This observed structure may stem from the demographic history of populations, indicating that each cluster A and B could be considered a specific

management unit: individuals from cluster B may originate from remnant populations in *Lancosme* while populations in cluster A may ensue from regular restocking with individuals from the Domaine National de Chambord since the 1950's (Klein 1990, Dellicour et al. 2011, Colyn et al. 2015). Although testing for this hypothesis was beyond the scope of this study, such separate origins may indeed explain the observed genetic pattern at this first hierarchical level, with individuals from clusters A and B being confronted to a barrier zone where genetic exchanges are hindered by the presence of multiple resistant landscape features such as highways and adjacent open areas (Drescher et al. 2001) or anthropized river valleys. A similar pattern was for instance suspected in a French population of roe deer (*Capreolus capreolus*) on either side of the Garonne valley (Coulon et al. 2006). The visual identification of such genetic boundaries between clusters A and B was consistent with the statistical identification of highways, roads and semi-natural open areas as the main contributors to the variance in HGD at the first hierarchical level. Importantly, the recent creation of the A85 highway along the river Cher may reinforce this genetic pattern in the future. Nevertheless, the further identification of up to seven final clusters suggests limited gene flow between main forest massifs at all inferior hierarchical levels, thus highlighting the need for an in-depth understanding of landscape connectivity in the study area.

All considered predictors were associated with an increase in genetic differentiation to some degree. The observed pattern of genetic distances was first associated with semi-natural open areas, identified as a possible important driver of genetic structuring at the first hierarchical level. Individuals may be reluctant to cross open areas (crops, meadows) in the absence of nearby wooded patches where they can easily find shelter at night or refuge in case of anthropogenic disturbance (Godvik et al. 2009, Allen et al. 2014). Farming activities may be responsible for an important loss of connectivity between forest massifs and subsequent genetic structuring in this species. Genetic distances further matched the presence of highways and roads, at both the first and the second hierarchical levels. These findings are consistent with previous studies reporting higher road-mediated mortality due to collision in many mammals with high dispersal abilities (Roach et al. 2001, Epps et al. 2005, Perez-Espona et al. 2009) or showing that red deer may be reluctant to cross highways despite the presence of specific road-crossing structures (e.g. Frantz et al. 2012). Unsurprisingly, the A71 highway, separating clusters A1 and A2 at the second hierarchical level, was associated with a gentler genetic boundary than the A20 highway, separating clusters A and B at the first hierarchical level (Fig. 1 ; Colyn et al. 2015). Apart from the specific features required for red

deer to use wildlife over- and underpasses, the location of ecopassages, and notably distance to cover (Clevenger and Waltho 2005), is known to have a great influence on their ability to maintain or enhance connectivity (Malo et al. 2004, Lesbarrères and Fahrig 2012). While the A71 highway is equipped with evenly-spaced ecopassages surrounded by woods, the A20 highway is mainly located within a vast agricultural plain and is equipped with ecopassages connecting highly fragmented wooded patches (Fig. 1): though indubitably essential for dispersal events between *Lancosme* and *Châteauroux* forest massifs (Colyn et al. 2015), the effectiveness of these ecopassages may be limited by suboptimal local characteristics of landscape composition, thus hindering genetic exchanges on either side of the A20 highway. As in other related species (Wang and Schreiber 2001), human avoidance in red deer may lead to an increase in inter-individual genetic differentiation on either side of urban areas (Frantz et al. 2006). However, urban areas were only detected as possible contributors to the variance in genetic differentiation at the second hierarchical level, and this effect was mostly shared with other features such as roads and highways: the urban network was maybe too sparse in our study area to hinder red deer movements, except when associated with transportation infrastructures. Finally, rivers were associated with an increase in genetic differentiation at the third hierarchical level: although red deer are able to swim across large water bodies (Perez-Espona et al. 2009), rivers may constitute historical natural boundaries between forest massifs, thus shaping landscape genetic variation at this inferior hierarchical level.

All these results together suggest that the effects of habitat loss and fragmentation on landscape genetic patterns in red deer operate at multiple scales, that is, both among and within putative ecologically-relevant management units. This loss of connectivity probably stems from human activities (through farming and transportation infrastructures), notably when they are concentrated along valleys as they reinforce rivers as historical natural boundaries between forest massifs. Although the impact of highways may be alleviated by ecopassages allowing genetic exchanges between red deer populations, landscape composition in the direct vicinity of wildlife crossing structures may affect their efficiency (Lesbarrères and Fahrig 2012). Future studies should now be conducted to confirm these findings, to ascertain whether current mitigation measures could be improved in the specific context of the studied area, and to evaluate how the recent creation of the A85 highway along the river Cher (Fig. 1) could further hinder gene flow in this game species.

Conclusion

This study, which is, to our knowledge, the first application of the analytical framework proposed by Balkenhol et al. (2014) on an empirical dataset since the original publication, confirmed the relevance of direct gradient analyses based on HGD to disentangle the complex, hierarchical genetic structure in wildlife populations. It also exemplified the use of CA as an efficient way of assessing the reliability of model parameters (beta weights, p-values) in face of suppression or redundancy. We recommend the use of regression commonality analysis on hierarchical genetic distances as a promising statistical tool for landscape geneticists.

Acknowledgments

This work was funded by the Fondation François Sommer pour la Chasse et la Nature, the Société de Vènerie, the Fédération Régionale des Chasseurs du Centre and the Fédération Départementale des Chasseurs de L'Indre, and was supported by grants from the Public Service of Wallonia (PSW), General Directorate for Agriculture, Natural Resources and Environment. It was coordinated by A. Bouron (Fédération Régionale des Chasseurs du Centre) and V. Gicquel (Fédération Départementale des Chasseurs de L'Indre). We warmly thank all the people and institutions that collected samples and provided complementary genotypes, including J. Alvarado, A. Frantz, R. Kuehn, X. Legendre, H. Prot, F. Zachos, the Fédérations départementales des chasseurs de la région Centre, the Associations départementales des chasseurs de grand gibiers and the Région Wallone. We also thank F. Chaumont for providing research facilities, M.C. Eloy for her precious assistance with laboratory analyses, as well as K. Nimon, K. Saint-Pé, T. Keitt and four anonymous reviewers for insightful comments on first drafts of this manuscript.

REFERENCES

Allen, A. M. et al. 2014. The impacts of landscape structure on the winter movements and habitat selection of female red deer. - Eur. J. Wildl. Res. 60: 411–421.

Alvarado-Serrano, D. F. and Hickerson, M. J. 2016. Spatially explicit summary statistics for historical population genetic inference. - Methods Ecol. Evol. 7: 418–427.

Apollonio, M. et al. 2010. European ungulates and their management in the 21st century. - Cambridge University Press.

Baguette, M. 2004. The classical metapopulation theory and the real, natural world: a critical appraisal. - Basic Appl. Ecol. 5: 213–224.

Balkenhol, N. et al. 2014. A multi-method approach for analyzing hierarchical genetic structures: a case study with cougars *Puma concolor*. - Ecography 37: 552–563.

Barbujani, G. et al. 1989. Detecting regions of abrupt change in maps of biological variables. - Syst. Zool. 38: 376–389.

Beckstead, J. W. 2012. Isolating and Examining Sources of Suppression and Multicollinearity in Multiple Linear Regression. - Multivar. Behav. Res. 47: 224–246.

Bouzat, J. L. and Johnson, K. 2004. Genetic structure among closely spaced leks in a peripheral population of lesser prairie-chickens. - Mol. Ecol. 13: 499–505.

Cárdenas, L. et al. 2015. Hierarchical analysis of the population genetic structure in *Concholepas concholepas*, a marine mollusk with a long-lived dispersive larva. - Mar. Ecol.: 1–11.

Chapuisat, M. et al. 1997. Microsatellites Reveal High Population Viscosity and Limited Dispersal in the Ant *Formica paralugubris*. - Evolution 51: 475.

Chen, C. et al. 2007. Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study. - Mol. Ecol. Notes 7: 747–756.

Clevenger, A. P. and Waltho, N. 2005. Performance indices to identify attributes of highway crossing structures facilitating movement of large mammals. - Biol. Conserv. 121: 453–464.

Cohen, J. et al. 2003. Applied multiple regression/correlation analysis for the behavioral sciences. - L. Erlbaum Associates.

Colyn, M. et al. 2015. La génétique du paysage : origine et flux de dispersion des populations de cerfs en région Centre. - Faune Sauvage 307: 37–43.

Conger, A. J. 1974. A Revised Definition for Suppressor Variables: a Guide To Their Identification and Interpretation. - Educ. Psychol. Meas. 34: 35–46.

Coulon, A. et al. 2006. Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. - Mol. Ecol. 15: 1669–1679.

Coulon, A. et al. 2008. Congruent population structure inferred from dispersal behaviour and intensive genetic surveys of the threatened Florida scrub-jay (*Aphelocoma cœrulescens*). - Mol. Ecol. 17: 1685–1701.

Courville, T. and Thompson, B. 2001. Use of Structure Coefficients in Published Multiple Regression Articles: Beta is not Enough. - Educ. Psychol. Meas. 61: 229–248.

Creager, J. A. 1971. Orthogonal and Nonorthogonal Methods for Partitioning Regression Variance. - Am. Educ. Res. J. 8: 671.

Daniels, M. and McClean, C. 2003. Red deer calf tagging programmes in Scotland - an analysis. - J. Br. Deer Soc. 12: 420–123.

Dellicour, S. et al. 2011. Population structure and genetic diversity of red deer (*Cervus elaphus*) in forest fragments in north-western France. - Conserv. Genet. 12: 1287–1297.

Dormann, C. F. 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. et al. 2001. Farmland prices determinants. Paper presented at the American Agricultural Economics Association Annual Meeting.

Dunning, J. B. et al. 1992. Ecological processes that affect populations in complex landscapes. - Oikos 65: 169–175.

Dupanloup, I. et al. 2002. A simulated annealing approach to define the genetic structure of populations. - Mol. Ecol. 11: 2571–2581.

Earl, D. A. and vonHoldt, B. M. 2012. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. - Conserv. Genet. Resour. 4: 359–361.

Epps, C. W. and Keyghobadi, N. 2015. Landscape genetics in a changing world: disentangling historical and contemporary influences and inferring change. - Mol. Ecol. 24: 6021–6040.

Evanno, G. et al. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. - Mol. Ecol. 14: 2611–2620.

Excoffier, L. et al. 1992. Analysis of Molecular Variance Inferred from Metric Distances among DNA Haplotypes - Application to Human Mitochondrial-DNA Restriction Data. - Genetics 131: 479–491.

Fahrig, L. 2003. Effects of Habitat Fragmentation on Biodiversity. - Annu. Rev. Ecol. Evol. Syst. 34: 487–515.

Farrar, D. E. and Glauber, R. R. 1967. Multicollinearity in Regression Analysis: The Problem Revisited. - Rev. Econ. Stat. 49: 92.

Fischer, J. and Lindenmayer, D. B. 2007. Landscape modification and habitat fragmentation: a synthesis. - Glob. Ecol. Biogeogr. 16: 265–280.

Frantz, A. C. et al. 2006. Genetic structure and assignment tests demonstrate illegal translocation of red deer (*Cervus elaphus*) into a continuous population. - Mol. Ecol. 15: 3191–3203.

Frantz, A. C. et al. 2009. Using spatial Bayesian methods to determine the genetic structure of a continuously distributed population: clusters or isolation by distance? - J. Appl. Ecol. 46: 493–505.

Frantz, A. C. 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*). - Mol. Ecol. 21: 3445–3457.

Garroway, C. J. et al. 2011. Using a genetic network to parameterize a landscape resistance surface for fishers, *Martes pennanti*. - Mol. Ecol. 20: 3978–3988.

Giles, B. E. et al. 1998. Restricted gene flow and subpopulation differentiation in *Silene dioica*. - Heredity 80: 715–723.

Ginson, R. et al. 2015. Hierarchical analysis of genetic structure in the habitat-specialist Eastern Sand Darter (*Ammocrypta pellucida*). - Ecol. Evol. 5: 695–708.

Godvik, I. M. R. et al. 2009. Temporal scales, trade-offs, and functional responses in red deer habitat selection. - Ecology 90: 699–710.

Goslee, S. C. and Urban, D. L. 2007. The ecodist package for dissimilarity-based analysis of ecological data. - J. Stat. Softw. 22: 1–19.

Gouskov, A. et al. 2016. Fish population genetic structure shaped by hydroelectric power plants in the upper Rhine catchment. - Evol. Appl. 9: 394–408.

Graves, T. A. et al. 2012. The influence of landscape characteristics and home-range size on the quantification of landscape-genetics relationships. - Landsc. Ecol. 27: 253–266.

Guillot, G. et al. 2009. Statistical methods in spatial genetics. - Mol. Ecol. 18: 4734–4756.

Hamann, J.-L. et al. 2003. Les apports du marquage pour la gestion du Cerf élaphe. - Bull Mens Natl Chasse 260: 30–36.

Hanski, I. 1999. Metapopulation Ecology. - Oxford University Press.

Harrison, S. 1991. Local extinction in a metapopulation context : an empirical evaluation. - Biol. J. Linn. Soc. 42: 73–88.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. - Scand. J. Stat. 6: 65–70.

Horst, P. 1941. The role of the predictor variables which are independent of the criterion. - Soc. Sci. Res. Counc. 48: 431–436.

Howell, P. E. et al. 2016. Contiguity of landscape features pose barriers to gene flow among American marten (*Martes americana*) genetic clusters in the Upper Peninsula of Michigan. - Landsc. Ecol.: 1–12.

Jackson, N. D. and Fahrig, L. 2011. Relative effects of road mortality and decreased connectivity on population genetic diversity. - Biol. Conserv. 144: 3143–3148.

Jakobsson, M. and Rosenberg, N. A. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. - Bioinformatics 23: 1801–1806.

Jarnemo, A. 2008. Seasonal migration of male red deer (*Cervus elaphus*) in southern Sweden and consequences for management. - Eur. J. Wildl. Res. 54: 327–333.

Jombart, T. et al. 2008. Revealing cryptic spatial patterns in genetic variability by a new multivariate method. - Heredity 101: 92–103.

King, J. E. 2007. Standardized Coefficients in Logistic Regression. Paper presented at the Annual meeting of the Southwest Educational Research Association.

Klein, F. 1990. La réintroduction du cerf *Cervus elaphus*. - Rev Ecol Terre Vie Issue Suppl 5: 131–134.

Kruuk, E. B. et al. 2002. Antler size in red deer: heritability and selection but no evolution. - Evol. Int. J. Org. Evol. 56: 1683–1695.

Landguth, E. L. et al. 2010. Quantifying the lag time to detect barriers in landscape genetics. - Mol. Ecol. 19: 4179–4191.

Legendre, P. and Legendre, L. F. J. 1998. Numerical Ecology. - Elsevier Science B.V, Amsterdam.

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

Legendre, P. et al. 1994. Modeling brain evolution from behavior - A permutational regression approach. - Evolution 48: 1487–1499.

Lesbarrères, D. and Fahrig, L. 2012. Measures to reduce population fragmentation by roads: what has worked and how do we know? - Trends Ecol. Evol. 27: 374–380.

Levins, R. 1969. Some demographic and genetic consequences of environmental heterogeneity for biological control. - Bull. Entomol. Soc. Am. 15: 237–240.

Lewis, J. W. and Escobar, L. A. 1986. Suppression and enhancement in bivariate regression. - Statistician 35: 17–26.

Malo, J. E. et al. 2004. Can we mitigate animal–vehicle accidents using predictive models? - J. Appl. Ecol. 41: 701–710.

Manel, S. and Holderegger, R. 2013. Ten years of landscape genetics. - Trends Ecol. Evol. 28: 614–621.

Mayer, C. et al. 2009. Patchy population structure in a short-distance migrant: evidence from genetic and demographic data. - Mol. Ecol. 18: 2353–2364.

McRae, B. H. 2006. Isolation by resistance. - Evolution 60: 1551–1561.

McRae, B. H. and Shah, V. B. 2009. Circuitscape User Guide. Online. The University of California, Santa Barbara. Available from: http://www.circuitscape.org.

Mijangos, J. L. et al. 2015. Contribution of genetics to ecological restoration. - Mol. Ecol. 24: 22–37.

Monmonier, M. 1973. Maximum-difference barriers: An alternative numerical regionalization method. - Geogr. Anal. 5: 245–261.

Mood, A. M. 1971. Partitioning variance in multiple regression analyses as a tool for developing learning models. - Am. Educ. Res. J. 8: 191–202.

Murphy, M. A. et al. 2008. Representing genetic variation as continuous surfaces: an approach for identifying spatial dependency in landscape genetic studies. - Ecography 31: 685–697.

Newton, R. G. and Spurrell, D. J. 1967. A Development of Multiple Regression for the Analysis of Routine Data. - Appl. Stat. 16: 51.

Nimon, K. 2010. Regression commonality analysis: demonstration of an SPSS solution. - Mult. Linear Regres. Viewp. 36: 10–17.

Nimon, K. F. and Oswald, F. L. 2013. Understanding the Results of Multiple Linear Regression: Beyond Standardized Regression Coefficients. - Organ. Res. Methods 16: 650–674.

Nimon, K. et al. 2008. An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. - Behav. Res. Methods 40: 457–466.

Paulhus, D. L. et al. 2004. Two replicable suppressor situations in personality research. - Multivar. Behav. Res. 39: 303–328.

Perez-Espona, S. et al. 2009. Genetic diversity and population structure of Scottish Highland red deer (*Cervus elaphus*) populations: a mitochondrial survey. - Heredity 102: 199–210.

Pérez-González, J. et al. 2012. Population structure, habitat features and genetic structure of managed red deer populations. - Eur. J. Wildl. Res. 58: 933–943.

Peterman, W. E. et al. 2014. Ecological resistance surfaces predict fine-scale genetic differentiation in a terrestrial woodland salamander. - Mol. Ecol. 23: 2402–2413.

Pompanon, F. et al. 2005. Genotyping errors: causes, consequences and solutions. - Nat. Rev. Genet. 6: 847–846.

Prévot, C. and Licoppe, A. 2013. Comparing red deer (*Cervus elaphus L.*) and wild boar (*Sus scrofa L.*) dispersal patterns in southern Belgium. - Eur. J. Wildl. Res. 59: 795–803.

Pritchard, J. K. et al. 2000. Inference of population structure using multilocus genotype data. - Genetics 155: 945–959.

Prunier, J. G. 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. - Mol. Ecol. 22: 5516–5530.

Prunier, J. G. et al. 2014. A 40-year-old divided highway does not prevent gene flow in the alpine newt *Ichthyosaura alpestris*. - Conserv. Genet. 15: 453–468.

Prunier, J. G. et al. 2015. Multicollinearity in spatial genetics: Separating the wheat from the chaff using commonality analyses. - Mol. Ecol. 24: 263–283.

Puechmaille, S. J. 2016. The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem. - Mol. Ecol. Resour. 16: 608–627.

R Development Core Team 2014. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing.

Ray-Mukherjee, J. et al. 2014. Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. - Methods Ecol. Evol. 5: 320–328.

Renner, S. C. et al. 2015. Using multiple landscape genetic approaches to test the validity of genetic clusters in a species characterized by an isolation-by-distance pattern. - Biol. J. Linn. Soc. in press.

Rice, W. R. 1989. Analysing tables of statistical tests. - Evolution 43: 223–225.

Ricketts, T. H. 2001. The matrix matters: Effective isolation in fragmented landscapes. - Am. Nat. 158: 87–99.

Roberts, J. K. and Nimon, K. 2012. A Software Solution for Conducting Logistic Commonality Analysis. Paper presented at the Annual meeting of the Southwest Educational Research Association.

Rousset, F. 2008. GENEPOP ´007: a complete re-implementation of the GENEPOP software for Windows and Linux. - Mol. Ecol. Resour. 8: 103–106.

Schielzeth, H. 2010. Simple means to improve the interpretability of regression coefficients: Interpretation of regression coefficients. - Methods Ecol. Evol. 1: 103–113.

Segelbacher, G. et al. 2010. Applications of landscape genetics in conservation biology: concepts and challenges. - Conserv. Genet. 11: 375–385.

Seibold, D. R. and McPhee, R. D. 1979. Commonality analysis: A method for decomposing explained variance in multiple regression analyses. - Hum. Commun. Res. 5: 355–365.

Selkoe, K. A. 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. - Mol. Ecol. 19: 3708–3726.

Smith, A. C. et al. 2009. Confronting collinearity: comparing methods for disentangling the effects of habitat loss and fragmentation. - Landsc. Ecol. 24: 1271–1285.

Smouse, P. E. et al. 1986. Multiple-regression and correlation extensions of the mantel test of matrix correspondence. - Syst. Zool. 35: 627–632.

Taylor, P. D. et al. 1993. Connectivity is a vital element of landscape structure. - Oikos: 571–573.

ter Braak, C. J. F. and Prentice, I. C. 2004. A Theory of Gradient Analysis. - In: Advances in Ecological Research. Elsevier, pp. 235–282.

Tischendorf, L. and Fahrig, L. 2000. On the usage and measurement of landscape connectivity. - Oikos 90: 7–19.

Urban, D. L. et al. 1987. Landscape Ecology. - BioScience 37: 119–127.

Van Oosterhout, C. et al. 2004. micro-checker: software for identifying and correcting genotyping errors in microsatellite data. - Mol. Ecol. Notes 4: 535–538.

Wang, M. and Schreiber, A. 2001. The impact of habitat fragmentation and social structure on the population genetics of roe deer (*Capreolus capreolus L.*) in Central Europe. - Heredity 86: 703–715.

Wright, S. 1921. Correlation and Causation. - J. Agric. Res. 20: 557–585.

Zachos, F. E. et al. 2016. Genetic Structure and Effective Population Sizes in European Red Deer (*Cervus elaphus*) at a Continental Scale: Insights from Microsatellite DNA. - J. Hered. 107: 318–326.

Supplementary material (Appendix EXXXXX at <www.oikosoffice.lu.se/appendix>). Appendix 1–9

Table Legends

Table 1. Variance Inflation Factors (VIF) and matrix of Pearson's correlation coefficients among predictors.

Predictors	VIF	Pearson's correlation matrix			
		open	*urban*	*rivers*	*roads*
open	1.225				
urban	2.393	0.370			
rivers	1.518	0.238	0.561		
roads	1.531	0.335	0.539	0.294	
highways	2.158	0.385	0.689	0.500	0.520

Table 2. Both typical multiple linear regression results and additional parameters derived from CA. Typical multiple linear regression results include: set of predictors considered in the linear regression model (Pred), model fit (Multiple R^2; ∗∗∗: P-value <0.001), Pearson's correlation coefficient between predictors and the dependent variable (*r*), beta weights (β) and p-values (*p*). P-values in bold indicate significant predictors after sequential Bonferroni correction. Additional parameters include unique effect (U), common effect (C), total contribution (T) and scaled total contribution (% of R^2) of each predictor. The sum of scaled total contributions is also provided. Predictors in bold indicate main contributors to model fit according to CA (see text for details).

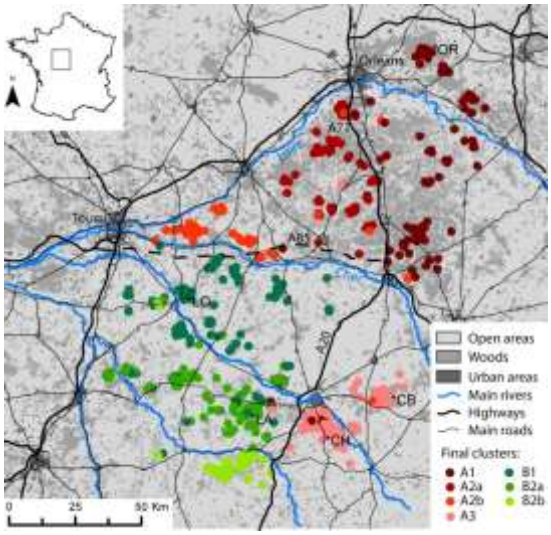
Pred	Multiple R	r		p	U	C	T	% of R^2
open		0.156	0.103	**<0.001**	0.009	0.016	0.025	63.4%
urban		0.144	0.058	**<0.001**	0.001	0.020	0.021	54.3%
rivers	3.85 % ***	0.062	-0.045	**<0.001**	0.001	0.002	0.003	9.9%
roads		0.129	0.035	**<0.001**	0.001	0.016	0.017	42.9%
highways		0.155	0.079	**<0.001**	0.003	0.021	0.024	62.1%
							Sum :	232.6%

Table 3. For each dataset H_1 to H_3, both typical multiple logistic regression results and additional parameters derived from CA. Typical multiple logistic regression results include: set of predictors considered in the logistic regression model (Pred), model fit (Pseudo-R^2; ∗∗∗: P-value <0.001; ∗: P-value <0.05), Pearson's correlation coefficients between predictors and untransformed values of the dependent variables (*r*), semi-standardised beta weights ($\hat{\beta}$; computed using the mean predicted probability as a reference value), odds-ratio (ψ) and p-values (*p*). See legend in Table 2 for other details.

Dataset	Pred	Pseudo-R	r	$\hat{\beta}$		p	U	C	T	% of R^2
H_1	***open***	18.74 % ***	0.249	0.066	1.069	**<0.001**	0.014	0.065	0.079	42.1%
	urban		0.291	0.028	1.028	**<0.001**	0.001	0.119	0.120	64.0%
	rivers		0.181	-0.007	0.9993	**<0.001**	0.000	0.044	0.044	23.6%
	roads		0.285	0.068	1.070	**<0.001**	0.011	0.097	0.108	57.7%
	highways		0.337	0.124	1.132	**<0.001**	0.024	0.128	0.152	81.1%
									Sum :	**268.5%**
H_2	*open*	9.41 % ***	0.172	-0.013	0.987	0.125	0.001	0.011	0.012	12.6%
	urban		0.293	0.044	1.045	**<0.001**	0.003	0.069	0.072	76.7%
	rivers		0.197	0.005	1.005	0.344	0.000	0.032	0.032	33.6%
	roads		0.266	0.043	1.044	**<0.001**	0.005	0.048	0.053	56.3%
	highways		0.329	0.084	1.088	**<0.001**	0.013	0.070	0.083	87.3%
									Sum :	**266.5%**
H_3	*open*	1.21 % *	-0.015	0.010	1.010	0.226	0.000	0.001	0.001	7.4%
	urban		-0.007	-0.014	0.986	0.261	0.000	0.000	0.001	4.9%
	rivers		0.050	0.034	1.034	**0.012**	0.004	0.002	0.006	47.0%
	roads		-0.064	-0.036	0.965	**0.009**	0.004	-0.003	0.001	8.2%
	highways		0.017	0.032	1.032	0.060	0.002	0.001	0.003	23.9%
									Sum :	**91.4%**

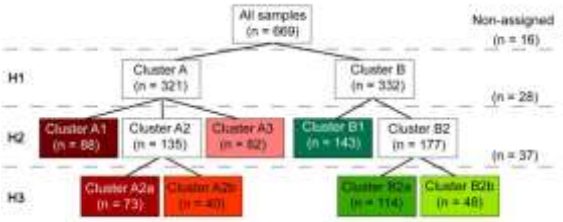
Figure Legends

Figure 1. Main spatial characteristics of the study area. Coloured dots indicate the localisation of the 588 individuals confidently assigned to their final hierarchical cluster (Q-values ≥ 0.6). Colours stand for the seven final hierarchical clusters. Asterisks (*) indicate noticeable forest massifs: *Choeurs-Bommiers* CB, *Châteauroux* CH, *Lancosme* LA, *Loches* LO and *Orléans* OR.



.’

Figure 2. Hierarchical splits of clusters inferred with STRUCTURE from the first to the third hierarchical level, with *n* the number of samples assigned to each cluster. The number of non-assigned individuals at each hierarchical level (Q-values < 0.6) is given on the right-hand side of the panel. Colours of the seven final clusters are the same as in Figure 1.



.’

Figure 3. Plot of the 31 commonality coefficients computed in dataset *Bc*, including both unique and common effects. Coefficients represent the percentage of variance explained in the dependent variable by each set of predictors. Ninety-five per cent confidence intervals were computed using a bootstrap procedure, with 1000 replicates based on a random removal of 30 % of individuals without replacement. The sum of coefficients equals the model fit index. %Total, summing to 100%, represents the percentage of variance explained in predicted values (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). O: *open*; U: *urban*; Ri: *rivers* ; Ro: *roads*; H: *highways*.

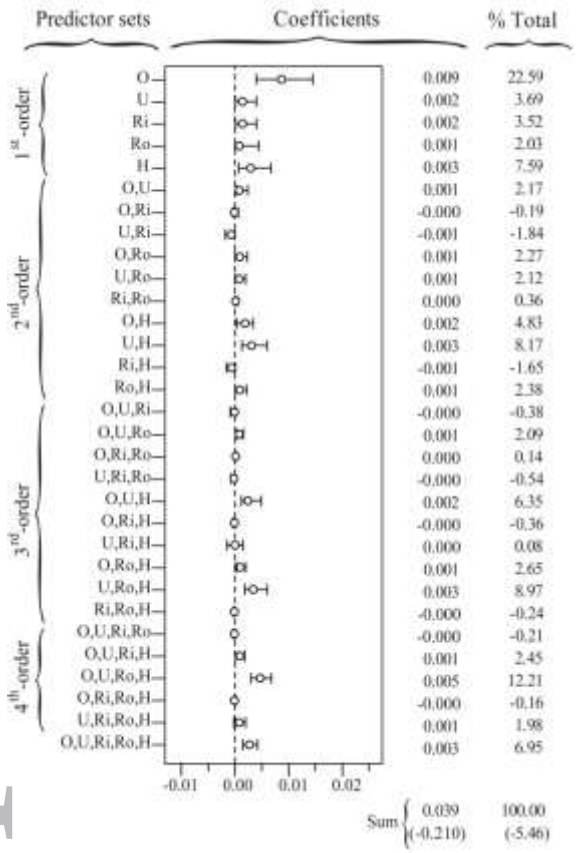
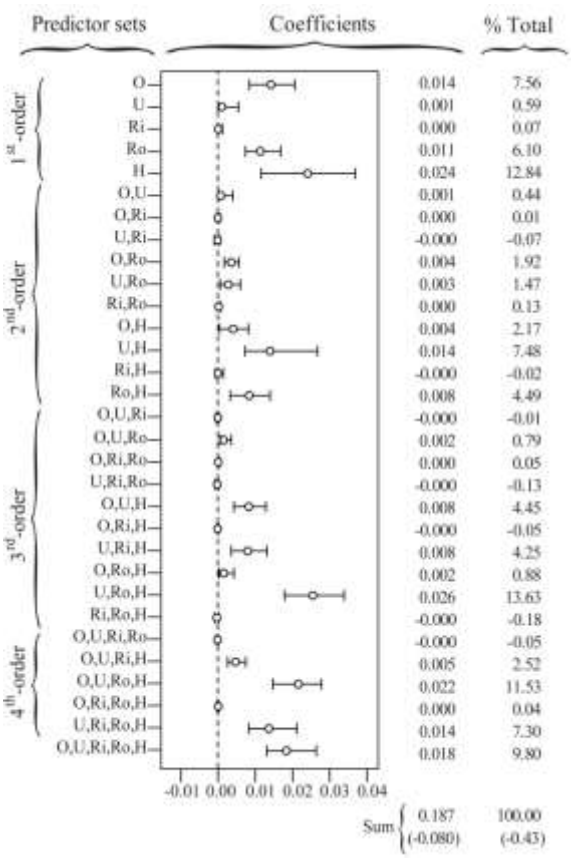
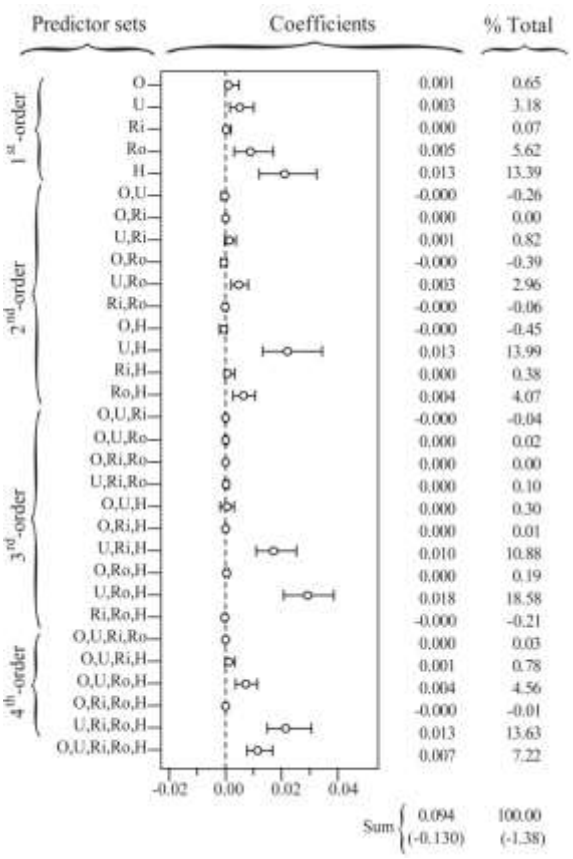


Figure 4. Plots of the 31 commonality coefficients computed in dataset H_1, including both unique and common effects. See legend in Figure 3.



.’

Figure 5. Plots of the 31 commonality coefficients computed in dataset H_2, including both unique and common effects. See legend in Figure 2.



.’