KU LEUVEN

FACULTY OF SCIENCE

Population connectivity across a highway in large ungulates

Testing the power of redundancy analyses

Marc BIJNENS

Supervisor: Prof. dr. L. De Meester
KU Leuven

Co-supervisor: Joachim Mergeay
INBO, KU Leuven

Thesis presented in

fulfillment of the requirements

for the degree of Master of Science

in Biology

Academic year 2014-2015

Acknowledgements

I would like to express my sincere appreciation to everyone who helped me during my studies and at developing this master's thesis.

First of all, I wish to show my gratitude towards Prof. Dr. Luc De Meester for the trust and the opportunity I've been given.

Furthermore, I wish to thank Joachim Mergeay, for his guidance, insight and support, as well as his constructive attitude when proofreading this manuscript.

I would also like to acknowledge the cooperation I received from Dr. Alain Frantz in providing me with the addenda to his research.

I will not forget the contributions of my parents to my accomplishments, without their moral, logistic and financial support, I would not have been able to achieve this.

Of course I also wish to show special gratitude to my girlfriend/fiancée Katrijn Mellemans for all of her support. Without her patience, perseverance and encouragement, I would have never been able to start at a new field of study and succeed in completing my schooling.

Tienen, May 27th 2015 M.B.

Table of contents

1. Introduction

1.1. Landscape genetics

Since the dawn of mankind, humans have been influencing their environment. The urbanization process, that spread over Europe since 700 B.C., has drastically changed the way humans interact with the ecosystem. The evolution to densely populated urban zones and more scarcely populated rural areas, brought with it some geographical changes. Urbanized regions would have to be connected to each other for instance, causing transport systems to dissect the rural landscape (Antrop, 2004). The emergence of vast transportation systems throughout a landscape directly affects the ecosystem by destroying, removing or reconfiguring local landforms. Transportation networks, roads in particular, increasingly have indirect effects on the ecosystem as well. The nature of a road network makes it so that it renders vast areas of the landscape as road-affected. It is estimated that in the conterminous United States, approximately 83% of the land area lies within slightly more than 1km from a road (Coffin, 2007).

In attempts to counteract the effects of these road networks, different types of mitigation measures have been devised. Overpasses, culverts and underpasses have been built (Mata *et al.*, 2008). The changes in land use have also lead to the onset of a new discipline: “Road Ecology” (Coffin, 2007). Whilst road ecology focuses on all ecological processes affected by roads, it doesn’t quantify the influences roads actually have on the dispersion and interconnectivity of wild animal herds. Measuring these impacts shows to be rather difficult using simple observational techniques. That’s where genetics come into the picture. Monitoring the gene flow in and between populations provides tangible information about their interactions. However gathering useful data proved to be difficult in the past, as population genetics methods required discrete populations.

Landscape genetics made an end to this issue, as it doesn’t require a priori knowledge about discrete populations. In landscape genetics observed spatial genetic patterns are explained by using landscape variables (Manel *et al.*, 2003). It can in fact be used to test landscape connectivity and study gene flow (Holderegger and Wagner, 2008), not only based on geographical distances, but on estimates of functional connectivity. Landscape genetics also made it possible to investigate the environmental factors driving adaptation in wild and domestic species (Manel and Holderegger, 2013).

However, studies that investigate road effects using genetic data use only a fraction of the available molecular approaches (Balkenhol and Waits, 2009). In terms of statistical possibilities, things have also changed dramatically. Where in the past simple Chi-squared tests and multiple regressions were used to investigate the effects of landscape change on the decline in animal population sizes (van der Zee *et al.*, 1992), many more statistical techniques are available. Although redundancy analyses (RDA) (van den Wollenberg, 1977) have been used extensively in community ecology, especially when partitioning the explained variance into independent contributions of spatial and environmental effects (Borcard *et al.*, 1992, Borcard *et al.*, 2004, Cottenie, 2005). Their application in population genetics is still rather rare, despite the enormous potential of this approach to disentangle the drivers of population genetic patterns (Manel *et al.*, 2012, Orsini *et al.*, 2013).

1.2.Redundancy analysis

Redundancy analysis (RDA) is a technique used to explain variation in a response dataset by building a model out of explanatory variables. RDA combines multivariate multiple linear regression and principal component analysis (PCA). It works on two centred matrices, $\mathbf{X}$, containing explanatory variables, and $\mathbf{Y}$, containing response data. Each RDA consists of three or four steps (Borcard *et al.*, 2011).

- Each variable $\mathbf{y}$ is regressed on the explanatory matrix $\mathbf{X}$ to produce $\hat{\mathbf{y}}$, the fitted vector, and $\mathbf{y}_{\mathbf{res}}$, the residual vector. All vectors $\hat{\mathbf{y}}$ are assembled into a matrix $\hat{\mathbf{Y}}$**.**
- A PCA of the $\hat{\mathbf{Y}}$ matrix is computed, producing a vector of eigenvalues and a matrix $\mathbf{U}$ of canonical eigenvectors.
- Two types of ordination site scores can now be computed using $\mathbf{U}$. The first is to use $\mathbf{Y}$ to obtain a spatial ordination and the second is done using $\hat{\mathbf{Y}}$, this results in an ordination in the space of variables $\mathbf{Y}$ and $\mathbf{X}$ respectively.
- The fourth step is technically not a part of the RDA, but a PCA may also be performed on $\mathbf{Y}_{\mathbf{res}}$, resulting in an unconstrained ordination of the residuals.

Thus RDA results in a series of axes, linear combinations of the explanatory variables, that, in successive order, best explain the variation of the response matrix. These axes are all orthogonal to one another, therefore making RDA a constrained ordination procedure (Borcard *et al.*, 2011).

1.2.1. Distance-based redundancy analysis

Distance-based redundancy analysis was introduced as a method that aimed at “testing the significance of individual terms in a multifactorial analysis-of-variance model for multispecies response variables” (Legendre and Anderson, 1999). Although it was initially intended to be used to measure the response of multiple species to the same structured multifactorial experimental designs, it can also be used to measure the extent to which the genetic structure of a population is affected by a set of independent variables, such as spatial structures (Vangestel *et al.*, 2012). A db-RDA is an ordination method that typically consists of five steps (Legendre and Anderson, 1999):

- Calculating distances among replicates in the response dataset, using a distance measure of choice. These distances are stored in a matrix.
- Performing a principal coordinate analysis (PCoA) on this matrix.
- Select a priori relevant explanatory variables. In a landscape genetic study these can consist of the spatial information of the sampled individuals, or of specific landscape elements that are expected to influence the genetic structure. This depends entirely on the design of the experiment.

- Performing an RDA to analyse the relationship between the principal coordinates and the explanatory variables.
- Conducting a permutation test corresponding to the particular terms in the model.

In population genetics, db-RDA is used as a tool to find spatial genetic effects. Raw genetic data can not be used in RDA as such because different alleles on the same locus are interdependent. Transforming these genetic data into distance-based variables makes it possible to perform a RDA, which is exactly what db-RDA does.

1.2.2. Principal component analysis

PCA is a method of data complexity reduction. The technique transforms a given dataset from an original system of axes, defined by the variables, to a system of successive new axes that are all orthogonal to each other. The number of axes can be at most equal to the number of original variables. Each axis consecutively corresponds to the dimension of the highest variance in the scatter of points. This means that the earlier axes describe more of the total spreading of observations than the subsequent axes. This is represented in the form of eigenvalues (Pearson, 1901). The method can solely detect linear relationships, therefore linear regression is performed before using PCA in an RDA (Borcard *et al.*, 2011).

1.2.3. Principal coordinate analysis

PCoA is similar to PCA in that it transforms a dataset into a set of successive orthogonal axes. The importance of each axis is also measured by eigenvalues (Gower, 1966). The big difference is that PCoA can be performed on non-regressed data, as it is based on an association matrix. Providing the data put into the analysis consists of Euclidean distances, it will output Euclidean axes (Borcard *et al.*, 2011), making the result identical to PCA.

1.3.Moran’s eigenvector maps

Moran’s eigenvector map analyses represent a type of spatial eigenfunction analysis that attempts to model the spatial relations among points (Legendre and Legendre, 2012). A MEM contains a number of successive eigenvectors with their corresponding eigenvalues. These eigenvectors describe structures on a scale going from global for those with large positive eigenvalues to local for those with the negative eigenvalues. The absolute value of the eigenvalue also quantifies the intensity of spatial autocorrelation of the corresponding eigenvector. As can be seen in Fig. 1, eigenvectors with high positive eigenvalues, being the first eigenvectors in consecutive order, describe large scale trends in a dataset, whereas eigenvectors with negative eigenvalues, being the last eigenvectors in consecutive order, describe local trends (Dray *et al.*, 2006).

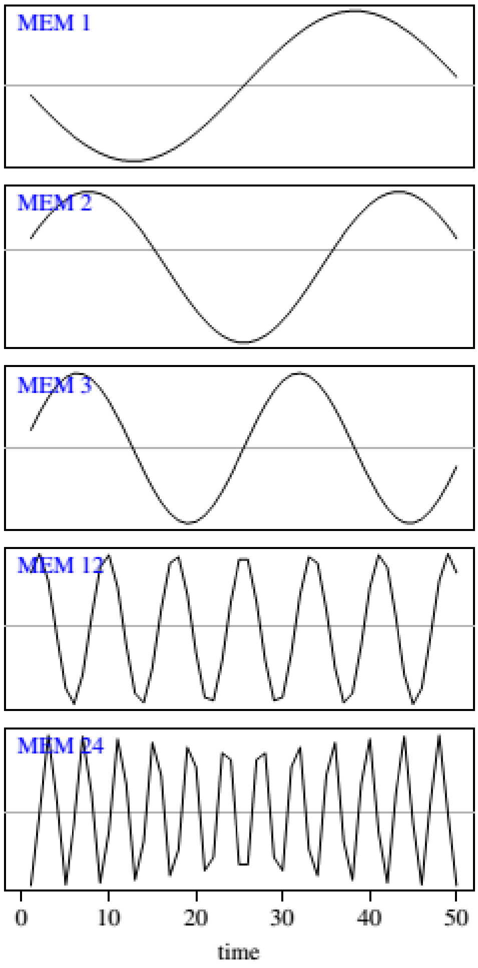

Fig. 1 A selection of distance based MEM eigenfunctions for a time series with 50 equispaced points, as taken from Legendre and Gauthier (2014).

Of course not all datasets will consist of equispaced points, datasets consisting out of more randomly placed samples will produces eigenvectors as can be observed in Fig. 2 (MEM1 and MEM15). The data in Fig. 2 being spatial distances allows it to be handled in a different way. The use of principal coordinates of neighbouring matrices (PCNM) is actually nothing more than a distance-based approach to MEM (Borcard *et al.*, 2011). The PCNM technique was introduced to detect and quantify spatial patterns over a wide range of scales in a dataset (Borcard and Legendre, 2002).

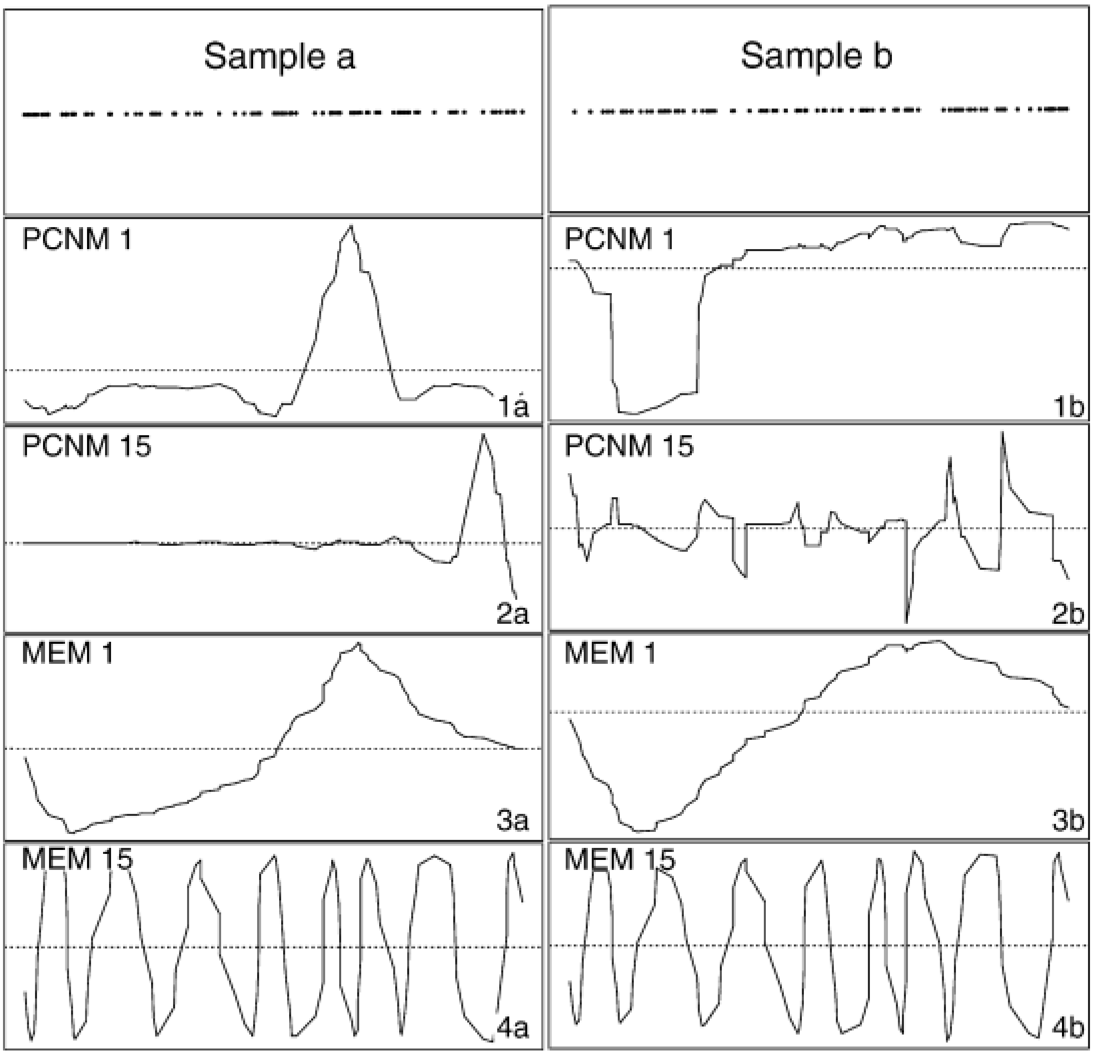

Fig. 2 Comparisons of the eigenvectors obtained for two different spatial weighting matrices. Two irregular samples of 100 sites randomly positioned along a straight line are considered (a and b). For each sample, the first (1) and fifteenth (2) eigenvectors obtained by the original PCNM approach are presented. The first (3) and fifteenth (4) Moran's eigenvectors (MEM) are also given for a spatial weighting matrix. As taken from (Dray *et al.*, 2006).

The PCNM method consists of four steps (Borcard *et al.*, 2011):

- The construction of a matrix containing Euclidean distances between sample sites.
- The truncation of the matrix to retain only distances between close neighbours. The truncation level can be chosen at will, although all sites have to be connected to at least one neighbour with a connection with equal or smaller length than the truncation level. Overall the truncation level is chosen to be the smallest possible value, while still considering the criterion. All other values are automatically set to an arbitrary "large" distance.
- The computation of a PCoA of the database with truncation values.
- The selection of eigenvectors that model positive spatial correlation.

These eigenvectors can then be used as explanatory variables in a multiple regression or RDA. Contrary to regular MEM, the relationship between the sign of the eigenvalues and the sign of the spatial correlation is not univocal. The value of Moran's I (Moran, 1950), however is a linear function of the eigenvalue in standard MEM eigenfunctions. Therefore Moran's I will be used as a criterion in selecting the eigenvectors that represent positive spatial correlation.

1.4.Forward selection

The downside of using MEM is the fact that it generates a lot of eigenvectors (or axes) that are not contributing to the power of the overall model. It is therefore necessary to objectively select only those eigenvectors that are necessary to achieve the highest possible level of explanation of variance by the model (Bellier *et al.*, 2007). However, the use of the classical forward selection method (Diehr and Hoflin, 1974) has two major flaws: Type I errors are highly inflated, meaning significant models are found when none should be found, and the amount of explained variance is overestimated.

The solution to the first problem is addressed by first performing a global test using all eigenvectors as explaining variables. Only if this test turns out to be significant, forward selection will be performed. To reduce the overestimation of explained variance, $R^2_{adj.}$ is used as an extra stop criterion (Blanchet *et al.*, 2008). Contrary to the original stop criterion, the significance level α, which depends on R^2 (Diehr and Hoflin, 1974), $R^2_{adj.}$ is influenced very little by the addition of unimportant variables. This means that the forward selection procedure stops after adding an extra eigenvector to the selected model made the $R^2_{adj.}$ of the selected model exceed the $R^2_{adj.}$ from the global test (Blanchet *et al.*, 2008). This double stop criterion will therefore assure that the forward selection procedure delivers a model without redundant variables, whilst still containing those variables important to the overall structure of the dataset (Borcard *et al.*, 2011).

1.5.History of wild boar (*Sus scrofa*) and red deer (*Cervus elaphus*) in Belgium

Belgian big game species are exclusively composed of deer (red (*Cervus elaphus*), roe (*Capreolus capreolus*) and fallow (*Dama dama*), mouflon (*Ovis amon mussimon*) and wild boar (*Sus scrofa*) (Prévot and Licoppe, 2013). Wild boars have been common across Europe for ages, although they were exterminated in many parts of Europe for at least some time in the past. Except for the northern territories and the high mountains, they can in theory be found across the entire European continent, even strongly cultivated ones (Lyneborg and den Hoed, 1972). Though they are found frequently in the Walloon part of Belgium and have recently expanded into Flanders as well, their populations are rather unbalanced, as the older animals are mostly hunted for (Verkem *et al.*, 2003). Wild boar populations managed to survive despite this prosecution, mostly thanks to their high fertility and endurance (Lyneborg and den Hoed, 1972).

Red deer can also be found across Europe. In contrast to wild boars however, they avoid strongly cultivated areas. Populations observed in Belgium today are probably resulting from introductions and human selection (Lyneborg and den Hoed, 1972). Red deer also seem to be heavily influenced by habitat fragmentation. They don't live in the Belgian and Dutch provinces of Limburg, though the habitat would appear to be suiting and populations of red deer live within range. This could probably be explained by the fact that red deer can't reach these areas without suitable corridors (Verkem *et al.*, 2003). The reason for past red deer introductions is their popularity as hunting game (Lyneborg and den Hoed, 1972).

During the last decades, both red deer and wild boar populations have grown in numbers and range (Prévot and Licoppe, 2013). The size of the Walloon wild boar population has even quadrupled over the past 30 years (Prévot and Morelle, 2012). Both species can be considered as non-migratory in southern Belgium (Prévot and Licoppe, 2013). However during their natal dispersion, young male individuals may travel for several kilometres. Most young males never migrate more than 10 km from their native territory though (Prévot and Morelle, 2012). Daily movement distances for wild boar in agro-forested landscapes range from 3 to 4 km on average up to 12 km at most (Morelle *et al.*, 2015). During their dispersal, red deer and wild boars face different challenges, as red deer could more easily jump agricultural fences, wild boars seem to cross motorways more easily (Prévot and Licoppe, 2013). Prévot and Morelle (2012) found that a number of tagged wild boars had crossed the E411 motorway during the course of their research. Through genetic analyses, it was shown that dispersion of roe deer is influenced by fenced transportation networks such as motorways (Coulon *et al.*, 2004, Coulon *et al.*, 2006, Kuehn *et al.*, 2007, Hepenstrick *et al.*, 2012)

1.6.Objectives of this master thesis

This thesis will focus on supplementing to the conclusions drawn by Frantz *et al.* (2012b) concerning the influence of the E411 motorway on gene flow in wild boar and red deer populations. Also I will try to compare distance-based redundancy analyses to previous results obtained through Bayesian cluster analyses (Binder, 1978). Bayesian cluster analyses are a derivative of the statistical clustering technique first introduced by Driver and Kroeber (1932). Here, genetic data of Walloon populations of wild boar and red deer will be used (Frantz *et al.*, 2012a). Since Frantz *et al.* (2012b) already performed several Bayesian cluster analyses, only the distance-based redundancy analyses will be performed. This will allow me to compare the results of both statistical methods to each other. In this comparison, special attention will be given to statistical power of each method and practical concerns whilst using them.

2. Materials and methods

2.1. Acquiring data and initial statistical standpoint

Frantz *et al*. (2012b) extracted DNA markers from tissue samples of 875 red deer (*Cervus elaphus*) and 325 wild boar (*Sus scrofa*). All samples were collected from harvested animals during legal hunts in roughly the same area in Wallonia, bisected by the E411 motorway. The E411 is a four lane motorway connecting Brussels with the Luxembourg border. It is fenced along stretches that are close to forested areas and was finished in the mid-1980s. The motorway does not feature any purpose-built wildlife passages along the section that was relevant for this study. Tunnels and underpasses for local traffic do exist. For the genotyping of the red deer samples, 13 microsatellite loci were used, whilst for the wild boar samples 14 microsatellite loci were utilized. Frantz *et al.* (2012b) performed a thorough data analysis in order to find if the motorway had a detectable and significant effect on the genetic differentiation within the sampled populations. Their analyses focused primarily on an elaborate implementation of tests for isolation by distance (IBD) (Wright, 1943) on the one hand, and Bayesian clustering methods on the other hand: STRUCTURE 2.3.1 (Pritchard *et al.*, 2000) was used with and without sampling location as prior, GENELAND 3.2.4 (Guillot *et al.*, 2005), which natively considers geographic coordinates, and BAPS 5.2 (Corander *et al.*, 2008), a program that implements a spatially explicit clustering method, were employed as well. These Bayesian clustering methods were then repeated on random subsamples taken out of the red deer dataset. This way these techniques were tested for their statistical power.

2.2. Statistical analysis

All statistical processes were performed using R 3.1.2 (R Core Team, 2014). Also R-packages "PCNM" (Legendre *et al.*, 2013), "vegan" (Oksanen *et al.*, 2014), "adegenet" (Jombart, 2008), "ade4" (Dray and Dufour, 2007) and "gap" (Zhao, 2007) were used. The required data files were retrieved from the Dryad repository (Frantz *et al.*, 2012a).

2.2.1. Redundancy analyses

Redundancy analysis is a multiple linear regression method whereby a matrix of response variables (here genotypic data) is regressed against a matrix of explanatory variables, thereby yielding a number of successive orthogonal axes that indicate how and how well the explanatory variables relate to the response variables. These axes correspond to the consecutive dimensions responsible for the highest variance observed in said dataset. The distance-based redundancy analysis is an adaptation of this method, starting from ecological or genetic distances among data points (yielding an association matrix) instead of directly using the sample values from a species times sampling site matrix (Legendre and Legendre, 2012). Several db-RDA were performed on both the red deer and the wild boar data. Genotypic information was first transformed to principal coordinates by performing a PCoA, preceded by the calculation of Euclidean differences between the samples. When using genotypic data, each allele at each locus can have the value 0, 1 or 2, yielding an intrinsic dependency among alleles. Moreover, when using highly diverse loci (such as with microsatellites), the genotypic data matrix has a lot of zero cells. This in turn makes PCA, PCoA or RDA methods unsuitable due to the so-called "double zero problem" (Legendre and Legendre, 2012). By transforming the raw genotypic data into an association matrix and computing a PCoA, this problem was circumvented. The resulting principal coordinates then served as response variables in all subsequent RDA, making these essentially distance-based RDA. The other two datasets, containing the position relative to the motorway and the geographic coordinates respectively, were used as explanatory variables. A schematic representation of all db-RDA that were performed on the red deer and wild boar datasets can be seen in Fig. 3.

To assess the role of the E411 motorway ("Motorway"), a db-RDA, using the position relative to the motorway as a categorical explanatory variable, was performed. This analysis provides information on the amount of genetic structure that is explained by separating all samples into two populations according to their position relative to the motorway.

In order to use the geographic coordinates in a RDA, they first need to be transformed into a distance matrix. This distance matrix was then be used to perform a principal coordinates of neighbouring matrices analysis (PCNM) (Borcard and Legendre, 2002), which is a special case of a distance-based MEM (Borcard *et al.*, 2011). This PCNM analysis gives a number of spatial predictors going from a very broad to a very fine scale. Each predictor gets a Moran's I value (Moran, 1950), which is a measure of spatial autocorrelation. Only those PCNM predictors with a positive value for Moran's I should be selected, as only they offer information on positive spatial autocorrelation. A db-RDA can then be performed, using the selected PCNM predictors as continuous explanatory variables. This analysis provides information on the amount of genetic structure that is explained by spatial genetic structure. Using the double stop criterion (Blanchet *et al.*, 2008), the spatial predictors, contributing to the overall ordering, can then be selected. This shows which predictors form the best spatial genetic explanations for the given dataset.

In a third analysis, the explanatory variables consist of the forward selected MEM-variables and the dummy variable Motorway. Again a db-RDA was performed, this analysis is to show the total genetic diversity that can be explained by the geographical structure of the dataset. This also allows to determine the amount of diversity that is explained by the PCNM predictors whilst also being explained by the location of the motorway. Finally another forward selection (Blanchet *et al.*, 2008) is performed on the dataset of explanatory variables. This way the importance of Motorway can be quantified, relative to the overall spatial genetic explanation.

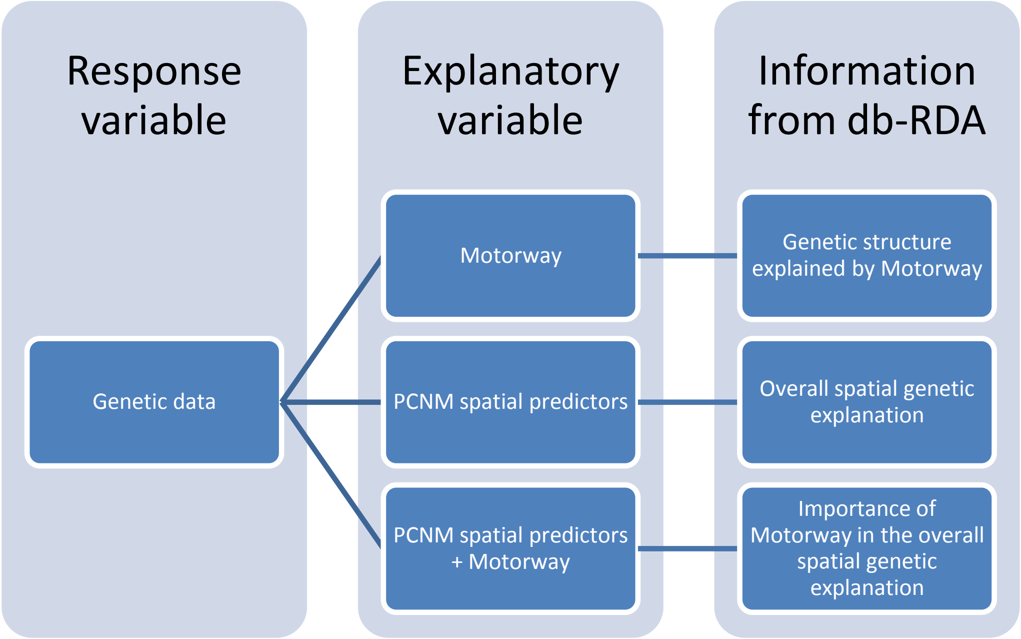
Response variable
Explanatory variable
Information from db-RDA
Genetic data
Motorway
PCNM spatial predictors
PCNM spatial predictors + Motorway
Genetic structure explained by Motorway
Overall spatial genetic explanation
Importance of Motorway in the overall spatial genetic explanation

Fig. 3 Schematic representation of all distance-based redundancy analyses performed on the red deer and wild boar datasets.

2.2.2. Permutation tests

In order to test for IBD, Frantz *et al.* (2012b), simulated multiple straight roads bisecting the study area and tested their explanatory value on the genetic datasets. Contrary to the aforementioned study, the roads simulated here are not completely random. All generated roads intersect the middle of the study area, with their angles arranged in such a way that together they fill a complete circle. For each simulated road, a db-RDA was performed as in section 2.2.1. This gives a measure for IBD in all directions across the dataset. To differentiate between the obtained values, a number of Chow tests (Chow, 1960) were performed, showing which of the generated roads has a significantly different explanatory value from the road with the highest and lowest value respectively (Ghilagaber, 2004). This provides information on the motorway as an isolating factor, as it can be coupled to the most approximate generated line.

2.2.3. Subsampling

To allow the deduction of a minimal sample size to detect a positive spatial genetic result, or in other words avoid a type II error, using Motorway as an explanatory variable, a power-analysis was performed (Kutner *et al.*, 2005). Frantz *et al.* (2012b) performed a similar test to determine the statistical power of the Bayesian clustering tools that were used. This also allows us to compare these Bayesian tools to the db-RDA used in this study. As the wild boar dataset never gave a positive result, Frantz *et al.* (Frantz *et al.*, 2012b) used only the red deer dataset in their power-analysis. Therefore we as well performed this analysis on the red deer dataset only. Several subsamples were taken semi-randomly from the red deer dataset, such that an equal amount of samples was taken at each side of the motorway. This way datasets were formed with 100, 200, 300 and 450 samples as was done by Frantz *et al.* (2012b). For each dataset size, ten replicates were created. Each of these replicates was then subjected to the procedures described in 2.2.1. To test the limits of the RDA even more, the dataset was subsampled further to 50, 25 and 10 samples.

2.2.4. Balancing the design of the wild boar dataset

As the wild boar dataset showed a rather unbalanced sampling design, it was downscaled to a more homogenous set in which the expected neutral spatial structure is as important in all directions. The original dataset spans around 100km from west to east and around 60km from north to south. After removing the samples that were furthest away from the motorway, a dataset spanning around 44 km from west to east and around 60km from north to south was retained. This dataset was then subjected to the procedure for finding effects of the position relative to the motorway described in 2.2.1. Since the dataset showed a large empty area in the middle, this process was repeated two times to generate a dataset for the top and bottom half of the remaining data. The downscaling criteria can be seen in Fig. 4.

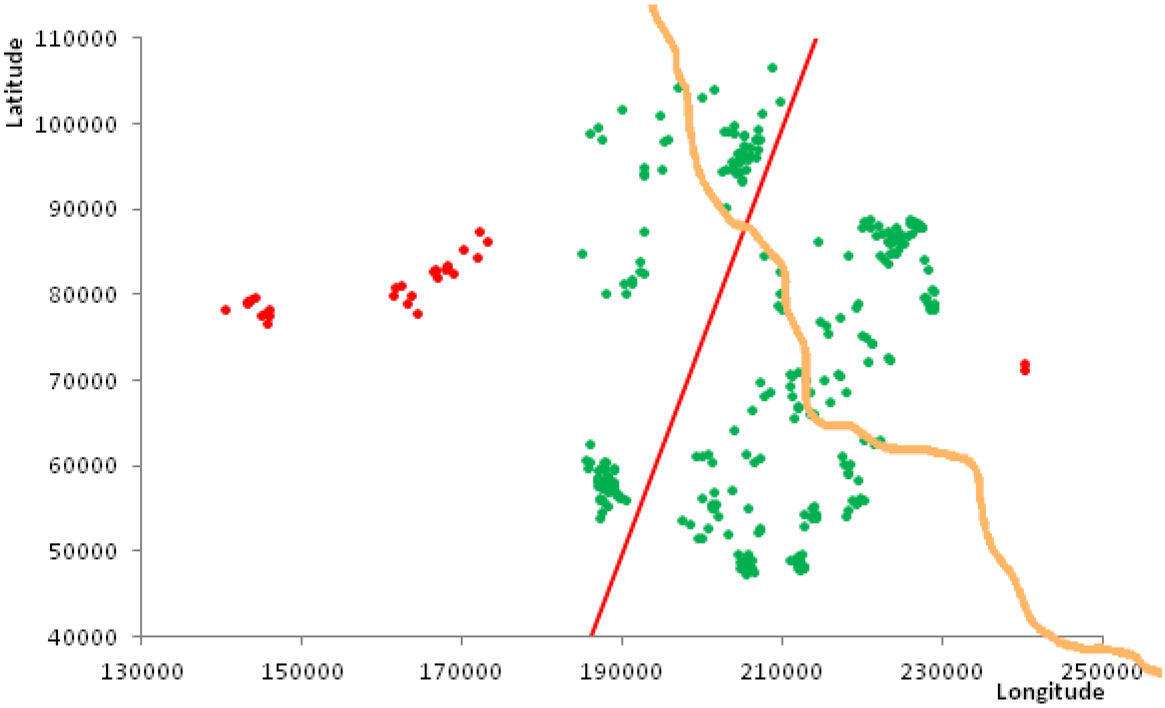

Fig. 4 Downscaling criteria for the wild boar dataset. The reduced dataset (Reduced) consisted of all samples shown in green, omitting the samples shown in red. This dataset was subsequently divided in two along the red line to produce a top (Top) and a bottom half (Bottom) of the reduced dataset. The motorway is depicted in orange.

2.2.5. Comparing the red deer database to the wild boar database

To enable a comparison between both the red deer and the wild boar dataset with as little interference from sampling effects as possible, a number of additional db-RDA were executed. To reduce sampling effects as much as possible, subsamples were taken from both datasets. In the subsampling procedure, special attention was given to taking an equal sample size in both subsamples and making sure each subsample consists of an equal amount of samples on both sides of the motorway. Furthermore the samples were all taken from the same 900km² area in the middle of the original research area (Fig. 5). This allowed for a more balanced comparison between the results for the red deer and the wild boar datasets. This comparison was done by following the procedures described in 2.2.1 and 2.2.2.

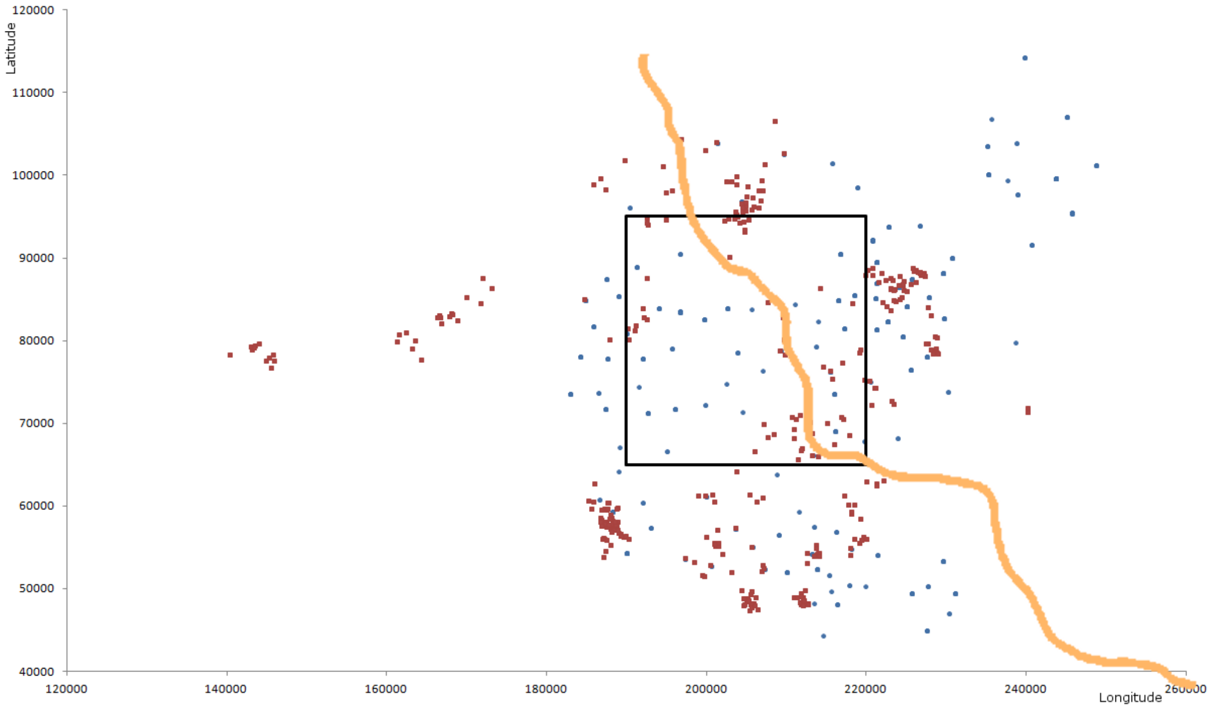

Fig. 5 Subsampling area used in comparing the red deer database to the wild boar database. Red deer samples are indicated as red squares, wild boar samples are drawn as blue dots. The black frame indicates the area used to take samples from. The motorway is indicated in orange.

2.3. Risk analysis

Since this paper only required thorough statistical work, no tangible dangers were encountered. All of the work was performed at a desk using a computer.

3. Results

3.1.Red deer dataset

3.1.1. db-RDA

All db-RDA performed on the red deer dataset yielded significant genetic structure. The spatial predictors resulting from the MEM analysis (55 variables with positive values for Moran's I out of 402 total variables) had a higher coefficient of determination than the position relative to the motorway (Table 1).

Table 1. Significance levels and coefficients of determination for all models generated using db-RDA on the red deer dataset.

Explanatory variables	Adjusted R^2	ANOVA p-value
1. Motorway	0.0174	0.001
2. Global MEM-model (55 variables)	0.0415	0.001
3. Selected MEMs (MEMs 1, 5, 2, 4, 10, 31, 6, 9)	0.0415	0.001
4. Motorway + selected MEMs	0.0414	0.001
5. Selected variables from 4 (Motorway + MEMs 1, 5, 2, 10, 31)	0.0400	0.001

Forward selection of the MEM predictors resulted in a reduced model containing eight MEM-variables. Variables 1, 5, 2, 4, 10, 31, 6 and 9 were selected in descending order of contribution to the overall value of $R^2_{adj.}$ (Table 2).

Table 2. Contributions of each individual MEM-variable to the overall $R^2_{adj.}$ of the model, after forward selection was conducted. Only MEM-variables were used to construct the global model.

MEM-variable	Contribution to $R^2_{adj.}$
1	0.015
5	0.0075
2	0.0067
4	0.0056
10	0.0024
31	0.0016
6	0.0013
9	0.0012

The first three RDA axes were significant canonical axes, which all correlated significantly with the position relative to the motorway (Fig. 6).

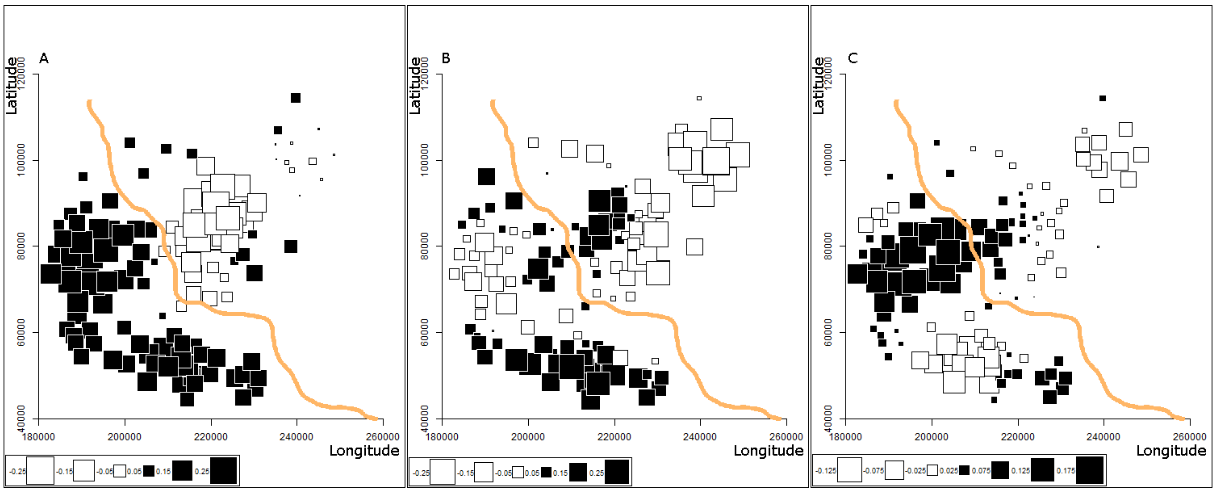

Fig. 6 Plot of the fitted scores of the three canonical axes (A-C) from the db-RDA using the selected MEM spatial predictors as explanatory variables for the red deer data. Linear regression of these fitted scores on the position relative to the motorway gives a p-value of $<2.2x10^{-16}$ for the first canonical axis (MEM1) (A), $p<2.2x10^{-16}$ for the second axis (MEM5) (B) and p=0.003604 for the third canonical axis (MEM2) (C). The motorway (E411) is indicated with an orange line.

Forward selection performed on the model using the MEM axes as well as Motorway as explanatory variables, resulted in a reduced model containing six axes. Motorway contributed the most to the overall value of R^2, followed by MEM axes 1, 5, 2, 10 and 31 (Table 3).

Table 3. Contributions of each individual variable to the overall $R^2_{adj.}$ of the model, after forward selection was conducted. Pre-selected MEM-variables and Motorway were used to construct the global model.

Variable	Contribution to $R^2_{adj.}$
Motorway	0.017
MEM 1	0.0076
MEM 5	0.0070
MEM 2	0.0051
MEM 10	0.0019
MEM 31	0.0010

3.1.2. Effect of simulated motorway

Since one can expect patterns of IBD along all spatial axes, I tested which spatial axis explained the genetic structure best, by simulating 100 motorways along the radius of a circle with centre in the centre of our study region and testing the strength of each regression. The lines with the highest and lowest value for R^2 are plotted in Fig. 7. The best fitting line (green in Fig. 7) fitting closely to the E411 yielded an R^2=0.0164 (p=0.001), whereas the worst, nearly perpendicular to the best, yielded an R^2=0.0017 (p=0.012).

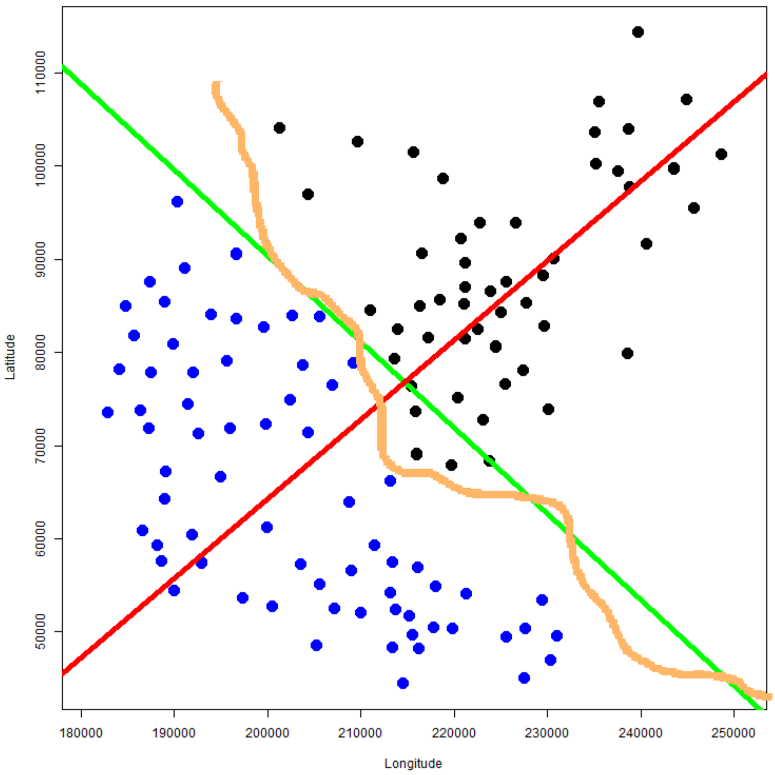

Fig. 7 Overview of the sampling location for the red deer dataset. Samples are shown in black and blue (indicating their location relative to the motorway), the motorway (E411) is displayed in orange. The green line corresponds to the generated road that showed the highest coefficient of determination. The red line corresponds to the generated road that showed the lowest coefficient of determination. Coefficients of determination were calculated on the results of db-RDA conducted using the generated roads as explanatory variables.

Fig. 8 shows a plot of all adjusted R^2 values. The value of the actual motorway is R^2=0.0174.

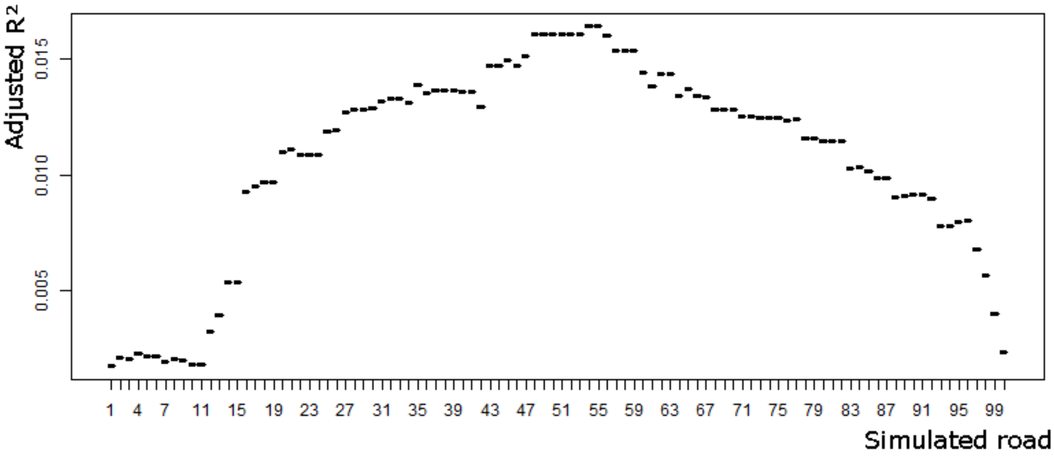

Fig. 8 Coefficients of determination for all 100 simulated motorways. Coefficients of determination were calculated on the results of db-RDA conducted using the generated lines as explanatory variables. Analysis performed on the red deer dataset.

Using Chow's test, all simulated roads were compared to the line with the highest and lowest R^2 value respectively. Fig. 9 shows the lines with the highest and lowest value for R^2 and their respective significantly differing lines.

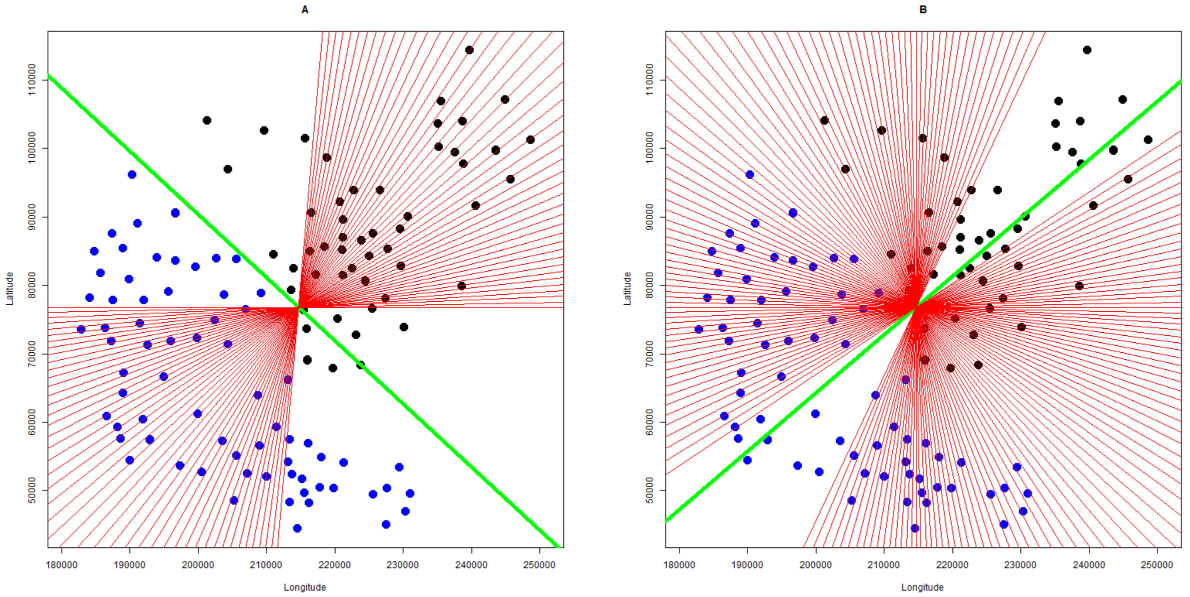

Fig. 9 Plots of the simulated roads with the highest (A) and the lowest (B) coefficient of determination for the red deer data (Fig. 8) in green, with their corresponding significantly differing roads in red. Blue dots correspond to samples taken on one side (SW) of the motorway, black dots originate from the other side (NE) of the motorway. Significant differences calculated using Chow's test.

3.1.3. Subsampling

To test the sensitivity of the db-RDA to sample size (determining the degrees of freedom of the analyses), subsampling was performed as in Frantz *et al.* (2012b) (Table 4). In contrast to Frantz *et al.* (2012b), we not only test for significance, but also determine the effect size of the test.

Table 4. Results of analyses of population genetic structure when reducing the size of the red deer data set. Results taken from Frantz *et al.* (2012b).

Method	Sample size	Frequency of inference of no substructure	Frequency of inference of two genetic clusters
STRUCTURE without location priors	450	0	10
	300	7	3
	200	9	1
	100	10	0
STRUCTURE with sampling location priors	450	0	10
	300	0	10
	200	0	10
	100	5	5
GENELAND	450	0	10
	300	0	10

	200	6	4
	100	10	0
BAPS	450	1	9
	300	9	1
	200	10	0
	100	10	0

In addition to having the dataset downsized to 450, 300, 200 and 100 samples, as was done by Frantz *et al.* (Frantz *et al.*, 2012b), the dataset was also further reduced to 50, 25 and 10 samples (Table 5). This was done to really test the correlation between sample size and statistical power for db-RDA.

Table 5. Results of RDA when reducing the size of the red deer data set.

Method	Sample size	Frequency of no significant effect of the motorway	Frequency of significant effect of the motorway
RDA	450	0	10
	300	0	10
	200	0	10
	175	0	10
	150	1	9
	100	1	9
	75	2	8
	50	6	4
	25	8	2
	10	7	3

3.2.Wild boar dataset

3.2.1. db-RDA

No db-RDA performed on the wild boar dataset yielded significant genetic structure. The spatial predictors resulting from the MEM analysis (57 variables with positive values for Moran's I out of 181 total variables) had a higher coefficient of determination than the position relative to the motorway (Table 6).

Table 6. Significance levels and coefficients of determination for all models generated using db-RDA on the wild boar dataset. No values were found for the selected variables, as according to the double-stop criterion (Blanchet *et al.*, 2008) forward selection can not be performed if the global model shows no significance.

Explanatory variables	Adjusted R²	ANOVA p-value
1. Motorway	0.000251	0.338
2. Global MEM-model (57 variables)	0.00616	0.416
3. Selected MEMs	NA	NA
4. Motorway + global MEM-model	0.00375	0.446
5. Selected variables from 4	NA	NA

3.2.2. Effect of simulated motorway

The permutation test was executed similarly to the one done on the red deer dataset. The lines with the highest and lowest value for R² were plotted in Fig. 10. The simulated road with the highest adjusted R² (0.00422) explained the genetic distances significantly, with an ANOVA p-value of 0.01. Whilst the road with the lowest adjusted R² (-0.00109) did not offer a significant explanation for the genetic distances (ANOVA p=0.845).

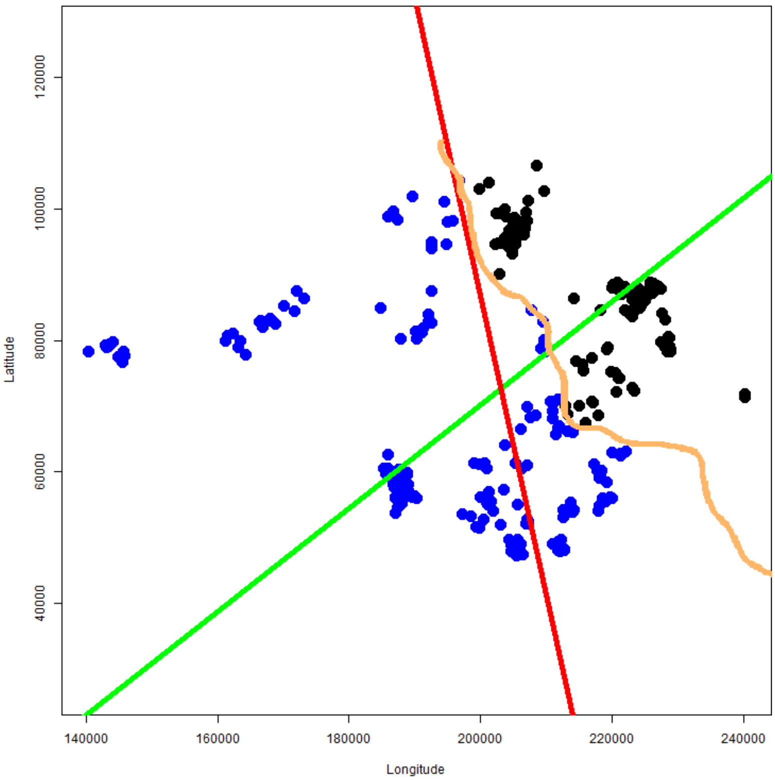

Fig. 10 Overview of the sampling location for the wild boar dataset. Samples are shown in black and blue (indicating their location relative to the motorway), the motorway (E411) is displayed in orange. The green line corresponds to the generated road that showed the highest coefficient of determination. The red line corresponds to the generated road that showed the lowest coefficient of determination. Coefficients of determination were calculated on the results of db-RDA conducted using the generated roads as explanatory variables.

Fig. 11 shows a plot of all adjusted R² values. The value of the actual motorway is R²=0.000251.

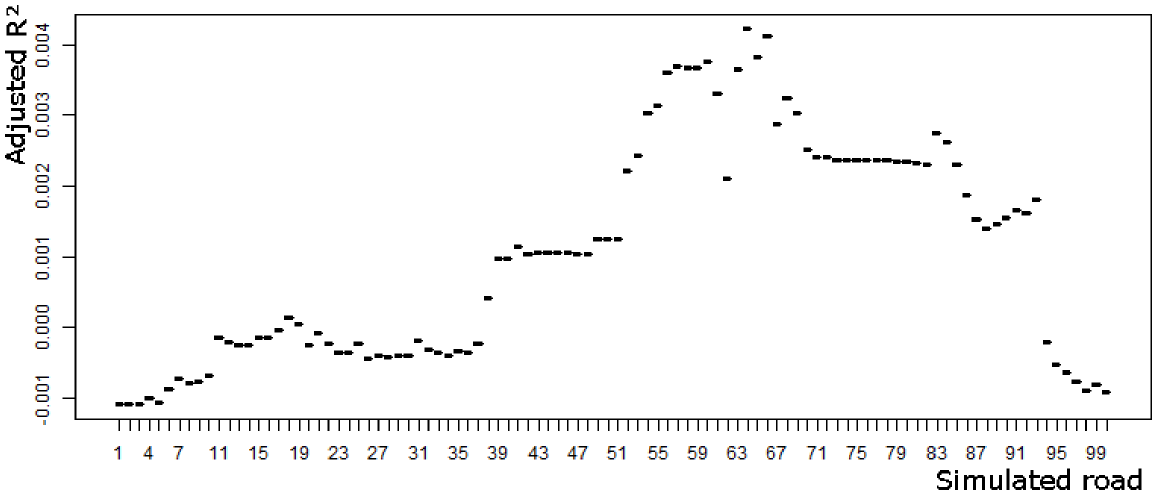

Fig. 11 Coefficients of determination for all 100 simulated roads for the wild boar dataset. Coefficients of determination were calculated on the results of db-RDA conducted using the generated lines as explanatory variables.

Using Chow's test, all generated roads were compared to the road with the highest and lowest R² value respectively. Fig. 12 shows the roads with the highest and lowest value for R² and their respective significantly differing roads.

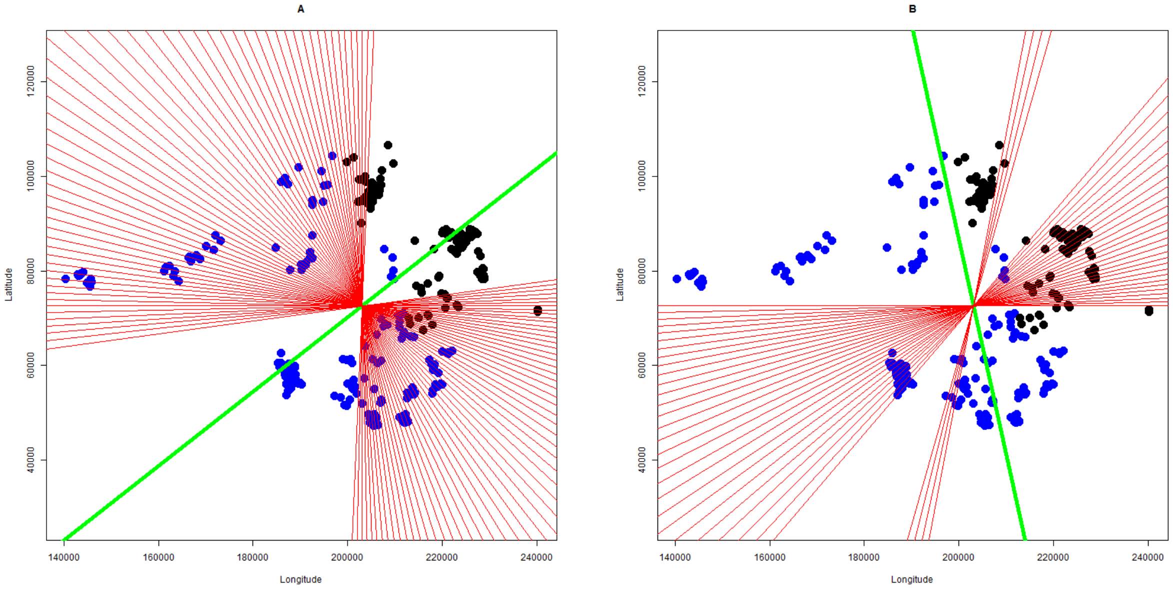

Fig. 12 Plots of the simulated roads with the highest (A) and the lowest (B) coefficient of determination for the wild boar dataset (Fig. 8) in green, with their corresponding significantly differing roads in red. Blue dots correspond to samples taken to the west of the motorway, black dots are taken to the east of the motorway. Significant differences calculated using Chow's test.

3.2.3. Balancing the spatial design

In order to find effects that are possibly hidden by the rather unbalanced design of the wild boar dataset, the dataset was split into more homogenously distributed subsets. Sample selection was done using the criteria shown in Fig. 13.

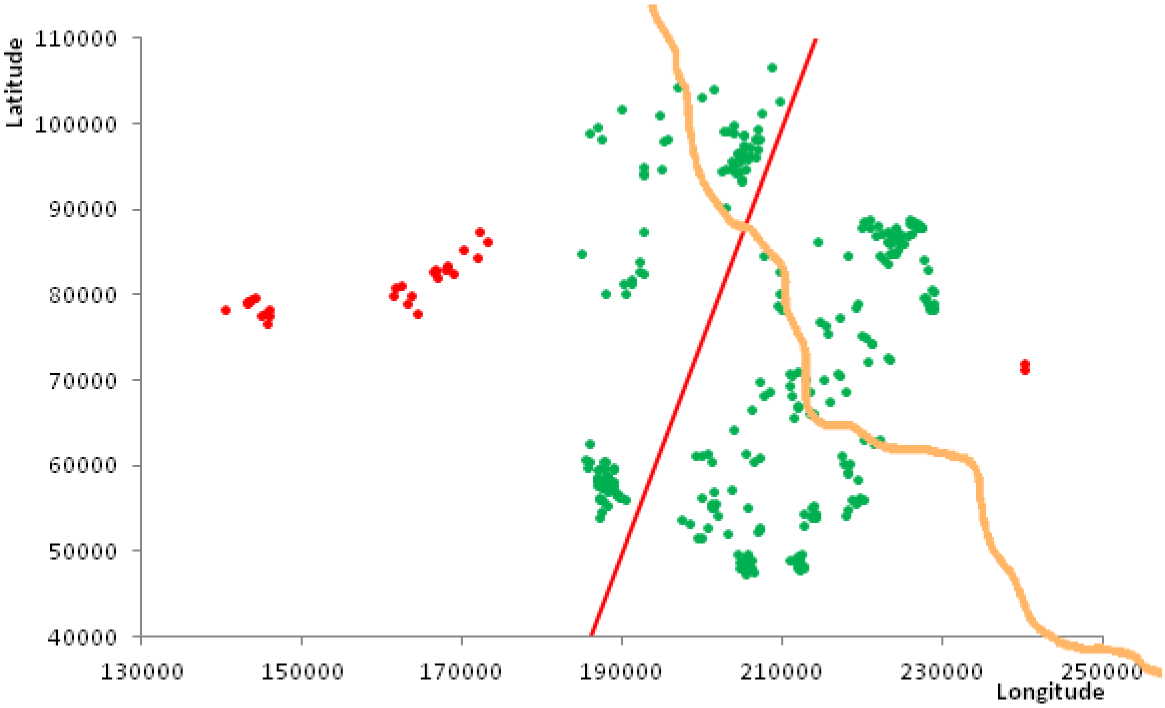

Fig. 13 Downscaling criteria for the wild boar dataset. The reduced dataset (Reduced) consisted of all samples shown in green, omitting the samples shown in red. This dataset was subsequently divided in two along the red line to produce a north-western (NW) and a south-eastern half (SE) of the reduced dataset. The motorway is depicted in orange.

None of the db-RDA performed on the newly formed datasets significantly explained the genetic distances between the samples. The effect size was highest in NW, although still being very low (Table 7).

Table 7. Significance levels and coefficients of determination generated using db-RDA on the reduced wild boar datasets. Results for complete dataset are included for comparison. All db-RDA were performed using the position relative to the motorway as explanatory variable.

Dataset	Adjusted R^2	ANOVA p-value
All (325) samples	0.000251	0.338
Reduced (294 samples)	-0.000259	0.497
NW (113 samples)	0.00676	0.051
SE (181 samples)	-0.00102	0.561

3.3.Comparing the red deer dataset to the wild boar dataset

Both datasets were reduced to datasets containing 50 samples, taken from the marked area in Fig. 5. These samples were obtained by randomly taking 25 samples within the marked area from each side of the motorway.

3.3.1. db-RDA

Using the subsampled datasets for red deer and wild boar, db-RDA were performed in order to explain the variation in genetic structure. Only Motorway in the deer dataset offered a significant explanation. Overall the models had higher significance and effect size for red deer than for wild boar (Table 8).

Table 8. Significance levels and coefficients of determination generated using db-RDA on the subsampled datasets of red deer and wild boar.

Explanatory variables	Adjusted R^2	ANOVA p-value
Motorway (deer)	0.0465	0.002
Motorway (boar)	0.00165	0.385
MEM (deer)	0.0431	0.07
MEM (boar)	0.000616	0.476

3.3.2. Effect of simulated motorway

The simulation tests were executed similarly to the ones done on the full datasets. For the red deer dataset, the simulated road with the highest effect size ($R^2_{adj.}$=0.0704, p=0.001) explained the genetic distances significantly. Whilst the simulated road with the lowest effect size ($R^2_{adj.}$=-0.00803, p=0.838) did not offer a significant explanation for the genetic distances. For the wild boar dataset, neither of these simulated roads explained the genetic distances significantly (highest effect size: $R^2_{adj.}$=0.00433, p=0.271; lowest effect size: $R^2_{adj.}$=-0.0111, p=0.957). Fig. 14 shows the orientation of the simulated roads with the highest and lowest effect sizes.

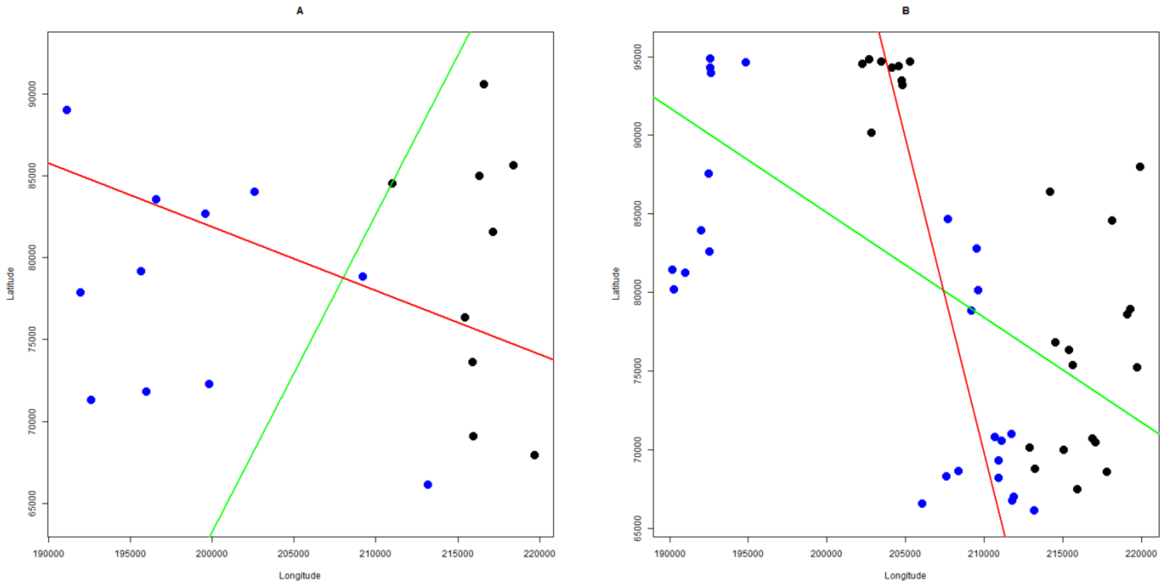

Fig. 14 Graphical representation of the generated roads with the highest and lowest effect sizes using the reduced datasets for red deer (A) and wild boar (B). The green lines show the generated roads with the highest effect sizes. The red lines correspond to the generated roads with the lowest effect sizes. Blue dots correspond to samples taken to the west of the motorway, black dots are taken to the east of the motorway.

Chow’s test was used to compare all simulated roads for the reduced red deer dataset to the simulated road that had the highest effect size (Fig. 15).

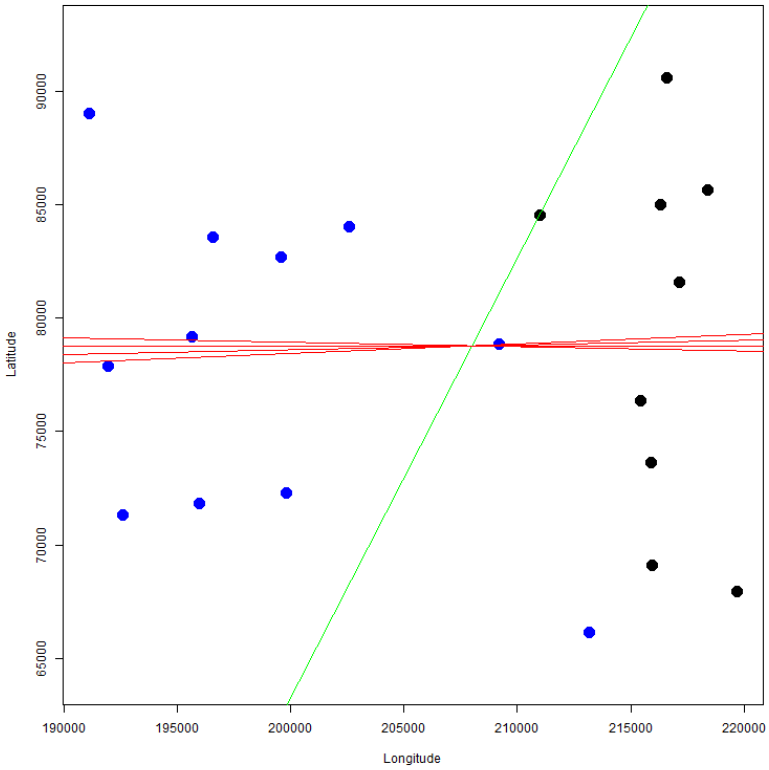

Fig. 15 **Plot of the simulated roads with the highest coefficient of determination for the reduced red deer dataset in green, with its corresponding significantly differing roads in red. Blue dots correspond to samples taken to the west of the motorway, black dots are taken to the east of the motorway. Significant differences calculated using Chow’s test.**

4. Discussion

I tried to supplement to the conclusions drawn by Frantz *et al.* (2012b), regarding the influence of the E411 motorway on gene flow in wild boar and red deer populations (Frantz *et al.*, 2012b). I also compared distance-based redundancy analyses to Bayesian cluster analyses using genetic data from wild boar and red deer populations in Wallonia (Frantz *et al.*, 2012a).

4.1.Red deer dataset

4.1.1. db-RDA

The results from the db-RDA performed on the red deer dataset confirm the findings of Frantz *et al.* (2012b). The E411 motorway has a significant influence on the spatial genetic structure of the red deer dataset (Table 1). The motorway explains 1.7% of the total genetic structure, where 4.2 % of this genetic structure is explainable by the overall spatial structure of the dataset. When looking at the breakdown of the MEM-variables (Table 2), it is apparent that most of the spatial genetic structure (1.5%) is to be found at a broad scale (MEM1). When Motorway is included in the model and forward selection is performed, Motorway is the first selected variable, indicating that it explains the genetic variance better than any of the spatial eigenvector functions (Table 3). Looking at the plot of the fitted scores of the three canonical axes (Fig. 6), the overlap in spatial structural information between Motorway and MEM1 is quite clear. This all leads me to conclude that the E411 motorway indeed impacts the genetic structure of the red deer population. This coincides with information about roe deer populations. Coulon *et al.* (2004) found that gene flow within a roe deer population is influenced by fragmentation of the wooded areas. The type of barrier that bisects this wooded area determines the magnitude of the impact that fragmentation has on the genetic structure of the population. Coulon *et al.* (2006) did not find an absolute barrier, but rather attributed it to an accumulation of factors (a fenced highway and several rivers and canals). Kuehn *et al.* (2007) on the other hand did not find a significant influence on genetic diversity at all, but they did find an influence of a fenced motorway on genetic divergence in roe deer. Hepenstrick *et al.* (2012), finally, did find that a fenced highway influences the spatial genetic structure of a roe deer population. They also found that an unfenced railway has absolutely no influence, whereas a wide river still shows to be somewhat of a barrier. This would mean that the fact that the E411 motorway acts as a gene flow barrier correlates to the fact that it is completely fenced off.

4.1.2. Effect of simulated motorways

To test if Motorway merely represents isolation-by-distance perpendicular to the axis of the E411, I performed simulations whereby 100 differently oriented highways were simulated. At least for red deer, the simulated motorways reveal a significant impact on the spatial genetic structure at the position of the motorway. The simulated motorway with the highest $R^2_{adj.}$=0.016 corresponds to the position of the real E411 motorway (Fig. 7). The fact that the real motorway has a slightly higher $R^2_{adj.}$=0.017 is explained by the fact that the real motorway is not completely straight in contrast to the simulated motorway. This fact also tells me that the effect is likely due to the E411 motorway, as a straight line that approaches the location of the motorway closely still has a lower $R^2_{adj.}$ than the original curved shape of the road. The values for $R^2_{adj.}$ also drop when turning away from the direction of the E411 (Fig. 8). This means that there is a clear trend towards higher genetic differences between samples across the motorway than between samples on the same side of the motorway. Looking at the results of the Chow's tests, it is clear that the motorway has a higher influence on spatial genetic structure than the spatial structures perpendicular to its direction (Fig. 9A). Perpendicular to the motorway, the $R^2_{adj.}$ has the significantly lowest value, meaning the influence of spatial structures on genetic structure clearly is the lowest in this direction (Fig. 9B). This all corroborates the hypothesis that the E411 motorway acts as a strong barrier to gene flow.

4.1.3. Subsampling

Subsampling the red deer dataset allowed me to determine the minimal amount of samples needed to find spatial genetic structure using distance-based redundancy analyses. Table 5 shows that the effect of the motorway remains significant in all ten pseudoreplicates of 175 samples or more. Even sample sizes of 100 yielded only a falsely non-significant effect (a type II error) in a single pseudoreplicate (10%). This contrasts with results of the Bayesian analyses of Frantz *et al.* (2012b), which failed to detect (depending on the clustering algorithm used) the true effect of the motorway in 50% to 100% of the pseudoreplicates consisting of 100 samples. This shows the superiority of redundancy analyses over Bayesian clustering methods.

4.2. Wild boar dataset

4.2.1. db-RDA

The db-RDA performed on the wild boar dataset did not return any significant spatial genetic effects. Neither the E411 motorway, nor the spatial distribution of the samples seems to significantly affect the genetic structure of the wild boar population. This can be due to a number of factors. An explanation could be found in sampling design or in the actual distribution of wild boar in the region. Fig. 10 shows a rather large gap between the north-western and south-eastern parts of the dataset, which may cause interferences in the db-RDA. Also, the spatial extent of sampling is wider (east to west) for wild boars than for red deer. However, both would yield a stronger spatial genetic structure, and this gap, if it had a strong effect on the genetic structure, should have been detected with the RDA. Since we didn't detect such a pattern, the sampling design is not a likely cause of this difference. Secondly, the wild boar dataset may be too small (N=325) to detect significant patterns. This is likely the case, but the effect ($R^2_{adj.}$) of the motorway is also much smaller anyway, even when the red deer dataset is downscaled to similar size as the wild boar dataset. This shows that the motorway has a much stronger impact on red deer than on wild boar. This leaves us with the third explanation, that wild boars are intrinsically more mobile (as indicated by the much shallower and non-significant overall neutral spatial genetic structure) and are less hindered by the motorway (as indicated by the lack of a significant effect of the motorway itself). On average a wild boar has a daily range of 1.3km², which goes up to 2.4km² in urbanized areas (Morelle *et al.*, 2015), which would allow for the occurrence of wild boars crossing the motorway. This is supported by Prévot and Morelle (2012), who found that tagged individuals were found on the other side of the E411 motorway after recapture.

4.2.2. Effect of simulated motorways

The highest value for $R^2_{adj.}$=0.00422 is found for the simulated motorway running through the gap dividing the wild boar dataset in two (Fig. 10). This indicates that this gap has a stronger influence on genetic structure than the E411 motorway. This is confirmed when Chow's test is used to compare the simulated motorways (Fig. 12). This means that the spatial gap inside the wild boar dataset could mask a possible effect of the E411 motorway. Fig. 11 shows that $R^2_{adj.}$ has a rather narrow range, meaning there is little difference in genetic impact from any spatial structure. The absolute values of $R^2_{adj.}$ are also quite low, indicating the spatial genetic structure in the wild boar dataset is not very profound.

4.2.3. Balancing the spatial design

To test the hypothesis that the absence of spatial genetic structure is to be explained by the way the sampling was done, the wild boar dataset was scaled down to form three new datasets (Fig. 13). Whilst none of the newly formed datasets yielded significant results, the results for NW are promising. With a p-value of 0.051, the $R^2_{adj.}$=0.00676 is as good as significant, tiny it may be. This would show a slight effect of the E411 motorway on the genetic structure of the wild boar population. Although the value for $R^2_{adj.}$ is very low compared to the value in the red deer dataset, it is higher than the value for $R^2_{adj.}$ found for the simulated motorway corresponding to the spatial gap in the dataset. This could mean there is a weak effect of the motorway on the genetic structure of the wild boar dataset. As shown in section 4.1.3, sample size is sufficient (113 samples) to be able to find significant results in the red deer dataset. This means that a potential effect in the wild boar dataset is in any case lower than the effect in the red deer dataset. There still is a possibility though, that the lack of significance in NW can be explained by its low sample size. The fact that SE and Reduced still distinctly show non-significant results, leads me to conclude that, although sample design might influence results for the wild boar dataset, the E411 motorway probably does not influence spatial genetic structure of the surrounding wild boar population by much.

4.3. Comparing the red deer dataset to the wild boar dataset

When comparing the results for the red deer dataset to those for the wild boar dataset, sampling design might play a role in explaining the differences found. To be able to avoid sampling design as a factor, both the red deer and the wild boar dataset were reduced to the same amount of samples (50), which were equally distributed on both sides of the motorway and came from the same area (Fig. 5). This way both datasets would be comparable in sampling design, leaving differences in statistical results dependent on the actual spatial genetic structure of the dataset. As I showed earlier (4.1.3), the optimal dataset for db-RDA consists of a minimum of 175 samples for red deer, with datasets consisting of down to 100 samples still performing reasonably well. These datasets only consisted of 50 samples as otherwise it would not have been possible to create datasets for both species with equal amounts of samples on both sides of the motorway in the same area. This means a lack of statistical power will almost certainly be a problem. However, for the sake of comparison, these analyses were retained.

4.3.1. db-RDA

The spatial predictors from the MEM analysis of the red deer dataset don't show a significant effect when the data are downscaled to 50 samples, but this can be attributed to the low sample size of the reduced dataset. Table 8 does clearly show higher values for $R^2_{adj.}$ for the red deer dataset using Motorway or MEM as explanatory variables. This implies that the spatial genetic effects are higher in the red deer population than they are in the wild boar population. The fact that for the red deer dataset only Motorway has a significant effect, underlines the significance of the impact of the E411 motorway on the spatial genetic structure of the red deer population.

4.3.2. Effect of simulated motorways

The simulated motorways confirm the fact that spatial genetic effects are higher in the red deer dataset. Only the red deer dataset shows significant effects, the orientation of the largest effects can be seen in Fig. 14. This orientation does not seem to correspond to the location of the motorway, however Chow's test shows me that the real orientation of the E411 motorway does not yield a significantly different effect from the highest effect detected in this reduced dataset. This confirms the existence of a spatial genetic effect on the red deer population caused by the E411 motorway. Although the dataset does seem a bit too small to make a conclusion about the wild boar population, there does not seem to be an important genetic effect on the wild boar population from the motorway.

4.3.3. Comparing the full-sized datasets

When simply comparing the results found for the complete datasets of red deer (3.1) and wild boar (3.2), it becomes quite clear that the spatial genetic structure in the red deer dataset is more apparent than that in the wild boar dataset. Not only is there a clear influence of the E411 motorway on the genetic structure of the red deer population, other spatial factors also play a significant role. For the wild boar dataset, the effects are less obvious. The fact that no significant effects were found might mean that no effects are to be found, but it might as well mean that sample size is too low to detect an effect. Considering the fact that no significant spatial effects were found at all, it is probable that this second hypothesis is true. However, in case a spatial genetic effect was to be found, provided that the sample size of the wild boar dataset was higher, this effect would be smaller than the one found in the red deer dataset. I have shown that the spatial genetic effects were still found in all pseudoreplicates of the red deer dataset when it was scaled down to 300 samples (Table 5), proving that these effects are bigger than the potential spatial genetic effects in the wild boar population (as the wild boar dataset consisted of 325 samples).

4.4.Comparing db-RDA to Bayesian cluster analyses

Revising the analyses done by Frantz *et al.* (2012b) and facing them with those I performed in this paper, it is possible to compare distance-based redundancy analyses to Bayesian cluster analyses. Overall I did not find substantial differences when it comes to conclusions, both types of analyses find a significant influence of the E411 motorway on the genetic structure of the red deer population, whilst no influence was found for the wild boar population. However, the subsampling procedure performed on the red deer database (2.2.3, 3.1.3 and 4.1.3) shows an important difference between both techniques. Looking at Table 4, it is clear cluster analyses suffer from type II errors when sample size is too low. For BAPS and STRUCTURE without priors the threshold lies around 450 samples, GENELAND still outputs reliable results down to 300 samples and STRUCTURE, using the sampling locations as priors, performs best with no type II errors at 200 samples. At 100 samples, however, only STRUCTURE with priors can still identify two genetic clusters, albeit with a type II error-percentage of 50%. Table 5 shows the value of db-RDA as a powerful statistical tool for finding spatial genetic effects. With the dataset down to 175 samples still no type II errors occur. At 100 samples 90% of the db-RDA can still find a significant genetic effect of the motorway. Even with a dataset that is reduced to 10 samples, db-RDA can still find this effect in 30% of the trials. This clearly shows that db-RDA can find spatial genetic effects a lot easier with smaller sample sizes. Another important advantage to using db-RDA, is that it offers a measure for the effect size, whereas Bayesian cluster analyses can only identify the presence or absence of the effect. Overall, this indicates that the redundancy analyses are, at least in this study, far superior to the more commonly used Bayesian clustering methods. Moreover, the RDA method, by estimating effect sizes, allows one to compare the strength of a genetic barrier across species or across studies. And finally, RDA does not require very powerful computers or long computing time, whereas Bayesian analyses such as STRUCTURE do.

5. Conclusion

In conclusion, my research demonstrates the power of redundancy analyses in landscape genetics compared to much more laborious Bayesian clustering approaches. As indicated by Frantz *et al.* (2012b), the E411 motorway does indeed affect gene flow in the red deer population that was monitored in the Walloon region of Belgium. The wild boar population that was monitored in the same area does not seem to undergo a similar influence from the motorway. However, the results for the wild boar population seem to imply possible type II errors due to sampling design. Focusing on a spatially better balanced subset of the wild boar data, a weak and marginally significant effect of the motorway was detected. Concerning the choice of statistical method for researching spatial genetic effects, distance-based redundancy analyses showed their value. Not only does db-RDA offer a measure for effect size, which enables the comparison of importance for individual barriers, it also outperforms Bayesian cluster analyses when it comes to finding spatial effects in small datasets.

6. Summary

Nature preservation has gained a lot of interest in the past decades. However quantifying the impact man has and finding the right measures to counteract the deterioration shows to be rather difficult. The first step in preserving natural populations of any kind of organism is identifying the factors that influence its life cycle. In western Europe, populations of large mammals mostly suffer from the consequences of habitat fragmentation. This fragmentation is caused not in the least by human transportation infrastructure. To counteract the deterioration of wildlife habitats, numerous countermeasures have been taken. Overpasses, culverts and underpasses have been built and their effectiveness has been shown (Mata *et al.*, 2008). However the magnitude of the impact transportation networks have on the interconnectivity of mammalian populations has not been fully quantified. Frantz *et al.* (2012b) have tried to identify the effect of the E411 motorway on red deer (*Cervus* elaphus) and wild boar (*Sus scrofa*) in Wallonia. However they failed to fully quantify the effects they observed. In this paper I found a clear effect of the motorway on the red deer population, explaining 1.7% of the genetic structure of the population. In wild boar there was no clear effect apparent, this might however be caused by sampling design. The statistical technique I used, db-RDA (Legendre and Anderson, 1999), outperformed the Bayesian cluster techniques used by Frantz *et al*. (2012b) in two ways: it offers a measure for effect size and it is a powerful tool for finding effects in small datasets.

7. Samenvatting

Natuurbehoud heeft sterk aan belangstelling gewonnen de afgelopen jaren. Het opmeten van de menselijke impact op de natuur en het vinden van de juiste maatregelen blijft echter moeilijk. Het behouden van natuurlijke populaties van eender welke soort organisme begint bij het aanwijzen van de factoren die hun levenscyclus beïnvloeden. In West-Europa leiden grote zoogdieren vooral onder de gevolgen van habitat fragmentatie. Deze versnippering hangt sterk samen met de bouw van transportnetwerken. Om de achteruitgang van natuurgebieden tegen te gaan, werden tal van maatregelen genomen. Allerhande wildcorridors werden aangelegd en hun efficiëntie werd bewezen (Mata *et al.*, 2008). De grootte van de impact die transportnetwerken hebben op de verbondenheid tussen zoogdierenpopulaties werd echter nog niet volledig bepaald. Frantz *et al.* (2012b) hebben het effect van de autosnelweg E411 op populaties van edelherten (*Cervus elaphus*) en wilde zwijnen (*Sus scrofa*) trachten te vinden. De grootte van dit effect hebben ze echter niet volledig kunnen bepalen. In deze Master-thesis vond ik een duidelijk effect van de autosnelweg, dat 1.7% van de totale genetische structuur van de populatie edelherten verklaarde. De populatie wilde zwijnen werd schijnbaar niet beïnvloed, al kan dit aan de manier van staalname liggen. De statistische methode die ik gebruikte, db-RDA (Legendre and Anderson, 1999), toonde zich beter geschikt dan de Bayesische clustertechnieken die gebruikt werden door Frantz *et al.* (2012b). Ten eerste biedt het een maat voor de gevonden effecten en ten tweede is deze techniek veel krachtiger waardoor effecten zelfs in kleine datasets gevonden kunnen worden.

8. References

ANTROP, M. 2004. Landscape change and the urbanization process in Europe. *Landscape and urban planning,* 67**,** 9-26.

BALKENHOL, N. & WAITS, L. P. 2009. Molecular road ecology: exploring the potential of genetics for investigating transportation impacts on wildlife. *Molecular Ecology,* 18**,** 4151-4164.

BELLIER, E., MONESTIEZ, P., DURBEC, J.-P. & CANDAU, J.-N. 2007. Identifying spatial relationships at multiple scales: principal coordinates of neighbour matrices (PCNM) and geostatistical approaches. *Ecography,* 30**,** 385-399.

BINDER, D. A. 1978. Bayesian Cluster Analysis. *Biometrika,* 65**,** 31-38.

BLANCHET, F. G., LEGENDRE, P. & BORCARD, D. 2008. Forward selection of explanatory variables. *Ecology,* 89**,** 2623-2632.

BORCARD, D., GILLET, F. & LEGENDRE, P. 2011. *Numerical Ecology with R,* New York, Springer Science+Business Media.

BORCARD, D. & LEGENDRE, P. 2002. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling,* 153**,** 51-68.

BORCARD, D., LEGENDRE, P., AVOIS-JACQUET, C. & TUOMISTO, H. 2004. Dissecting the spatial structure of ecological data at multiple scales. *Ecology,* 85**,** 1826-1832.

BORCARD, D., LEGENDRE, P. & DRAPEAU, P. 1992. Partialling out the spatial component of ecological variation. *Ecology,* 73**,** 1045-1055.

CHOW, G. C. 1960. Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica,* 28**,** 591-605.

COFFIN, A. W. 2007. From roadkill to road ecology: A review of the ecological effects of roads. *Journal of Transport Geography,* 15**,** 396-406.

CORANDER, J., SIRÉN, J. & ARJAS, E. 2008. Spatial modelling of genetic population structure. *Computational Statistics,* 23**,** 111-129.

COTTENIE, K. 2005. Integrating environmental and spatial processes in ecological community dynamics. *Ecology Letters,* 8**,** 1175-1182.

COULON, A., COSSON, J. F., ANGIBAULT, J. M. A., CARGNELUTTI, B., GALAN, M., MORELLET, N., PETIT, E., AULAGNIER, S. & HEWISON, A. J. M. 2004. Landscape connectivity influences gene flow in a roe deer population inhabiting a fragmented landscape: an individual-based approach. *Molecular Ecology,* 13**,** 2841-2850.

COULON, A., GUILLOT, G., COSSON, J. F., ANGIBAULT, J. M. A., AULAGNIER, S., CARGNELUTTI, B., GALAN, M. & HEWISON, A. J. M. 2006. Genetic structure is influenced by landscape features: empirical evidence from a roe deer population. *Molecular Ecology,* 15**,** 1669-1679.

DIEHR, G. & HOFLIN, D. R. 1974. Approximating the Distribution of the Sample R2 in Best Subset Regressions. *Technometrics,* 16**,** 317-320.

DRAY, S. & DUFOUR, A. 2007. The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software,* 22**,** 1-20.

DRAY, S., LEGENDRE, P. & PERES-NETO, P. R. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling,* 196**,** 483-493.

DRIVER, H. E. & KROEBER, A. L. 1932. Quantitative expression of cultural relationships. *University of California Publications in American Archaeology and Ethnology,* 31**,** 211-256.

FRANTZ, A. C., BERTOUILLE, S., ELOY, M., LICOPPE, A., CHAUMONT, F. & FLAMAND, M. 2012a. Data from: 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). Dryad Data Repository.

FRANTZ, A. C., BERTOUILLE, S., ELOY, M. C., LICOPPE, A., CHAUMONT, F. & FLAMAND, M. C. 2012b. 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.

FRANTZ, A. C., POPE, L. C., ETHERINGTON, T. R., WILSON, G. J. & BURKE, T. 2010. Using isolation-by-distance-based approaches to assess the barrier effect of linear landscape elements on badger (Meles meles) dispersal. *Molecular Ecology,* 19**,** 1663-1674.

GHILAGABER, G. 2004. Another Look at Chow's Test for the Equality of Two Heteroscedastic Regression Models. *Quality and Quantity,* 38**,** 81-93.

GOWER, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika,* 53**,** 325-338.

GUILLOT, G., ESTOUP, A., MORTIER, F. & COSSON, J. F. 2005. A Spatial Statistical Model for Landscape Genetics. *Genetics,* 170**,** 1261-1280.

HEPENSTRICK, D., THIEL, D., HOLDEREGGER, R. & GUGERLI, F. 2012. Genetic discontinuities in roe deer (Capreolus capreolus) coincide with fenced transportation infrastructure. *Basic and Applied Ecology,* 13**,** 631-638.

HOLDEREGGER, R. & WAGNER, H. H. 2008. Landscape genetics. *BioScience,* 58**,** 199-207.

JOMBART, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics,* 24**,** 1403-1405.

KUEHN, R., HINDENLANG, K. E., HOLZGANG, O., SENN, J., STOECKLE, B. & SPERISEN, C. 2007. Genetic effect of transportation infrastructure on roe deer populations (Capreolus capreolus). *Journal of Heredity,* 98**,** 13-22.

KUTNER, M. H., NACHTSHEIM, C. J., NETER, J. & LI, W. 2005. *Applied linear statistical models*, McGraw-Hill.

LEGENDRE, P. & ANDERSON, M. J. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs,* 69**,** 1-24.

LEGENDRE, P., BORCARD, D., BLANCHET, F. G. & DRAY, S. 2013. PCNM: MEM spatial eigenfunction and principal coordinate analyses. R package version 2.1-2/r109. http://R-Forge.R-project.org/projects/sedar/.

LEGENDRE, P. & GAUTHIER, O. 2014. Statistical methods for temporal and space–time analysis of community composition data(). *Proceedings of the Royal Society B: Biological Sciences,* 281**,** 20132728.

LEGENDRE, P. & LEGENDRE, L. 2012. *Numerical ecology,* Amsterdam, Elsevier.

LYNEBORG, L. & DEN HOED, G. 1972. *Wilde zoogdieren in Europa,* Amsterdam, Moussault's uitgeverij NV.

MANEL, S., GUGERLI, F., THUILLER, W., ALVAREZ, N., LEGENDRE, P., HOLDEREGGER, R., GIELLY, L., TABERLET, P. & CONSORTIUM, I. 2012. Broad-scale adaptive genetic variation in alpine plants is driven by temperature and precipitation. *Molecular Ecology,* 21**,** 3729-3738.

MANEL, S. & HOLDEREGGER, R. 2013. Ten years of landscape genetics. *Trends in Ecology & Evolution,* 28**,** 614-621.

MANEL, S., SCHWARTZ, M. K., LUIKART, G. & TABERLET, P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology and Evolution,* 18**,** 189-197.

MATA, C., HERVÁS, I., HERRANZ, J., SUÁREZ, F. & MALO, J. E. 2008. Are motorway wildlife passages worth building? Vertebrate use of road-crossing structures on a Spanish motorway. *Journal of Environmental Management,* 88**,** 407-415.

MORAN, P. A. P. 1950. Notes on Continuous Stochastic Phenomena. *Biometrika,* 37**,** 17-23.

MORELLE, K., PODGÓRSKI, T., PRÉVOT, C., KEULING, O., LEHAIRE, F. & LEJEUNE, P. 2015. Towards understanding wild boar Sus scrofa movement: a synthetic movement ecology approach. *Mammal Review,* 45**,** 15-29.

OKSANEN, J., BLANCHET, F. G., KINDT, R., LEGENDRE, P., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., STEVENS, M. H. H. & WAGNER, H. H. 2014. vegan: Community Ecology Package. R package version 2.2-1/r2913. http://R-Forge.R-project.org/projects/vegan/.

ORSINI, L., VANOVERBEKE, J., SWILLEN, I., MERGEAY, J. & DE MEESTER, L. 2013. Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology,* 22**,** 5983-5999.

PEARSON, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine,* 2**,** 559-572.

PRÉVOT, C. & LICOPPE, A. 2013. Comparing red deer (Cervus elaphus L.) and wild boar (Sus scrofa L.) dispersal patterns in southern Belgium. *European Journal of Wildlife Research,* 59**,** 795-803.

PRÉVOT, C. & MORELLE, K. 2012. Potentiel de dispersion du sanglier et historique de la colonisation de la plaine agricole en Wallonie. *Demain la chasse: comment reprendre l'initiative?* Wépion, Belgique: Forêt Wallonne asbl.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. 2000. Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-959.

R CORE TEAM 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

VAN DEN WOLLENBERG, A. 1977. Redundancy analysis an alternative for canonical correlation analysis. *Psychometrika,* 42**,** 207-219.

VAN DER ZEE, F. F., WIERTZ, J., TER BRAAK, C. J. F. & VAN APELDOORN, C. 1992. Landscape change as a possible cause of the badger Meles meles L. decline in The Netherlands. *Biological Conservation,* 61**,** 17-22.

VANGESTEL, C., MERGEAY, J., DAWSON, D. A., CALLENS, T., VANDOMME, V. & LENS, L. 2012. Genetic diversity and population structure in contemporary house sparrow populations alons an urbanization gradient. *Heredity,* 109**,** 163-172.

VERKEM, S., DE MAESENEER, J., VANDENDRIESSCHE, B., VERBEYLEN, G. & YSKOUT, S. 2003. *Zoogdieren in Vlaanderen: Ecologie en verspreiding van 1987 tot 2002,* Mechelen & Gent, Natuurpunt Studie & JNM-Zoogdierenwerkgroep.

WRIGHT, S. 1943. Isolation by Distance. *Genetics,* 28**,** 114-138.

ZHAO, J. H. 2007. gap: Genetic Analysis Package. *Journal of Statistical Software,* 23**,** 1-18.

9. Addendum

9.1. R-script for db-RDA, simulating motorways and subsampling

```
library(PCNM)
library(vegan)
library(adegenet)
library(ade4)
library(gap)

####################################
#Distance-based Redundancy Analysis Deer data
####################################

# Input data was retrieved from the dryad depository (Frantz et al., 2012a).
# This script is based on Borcard et al. (2011) "7.4.2.3 PCNM analysis of the mite data"

dat<-read.csv("Input_data_deer.csv",header=TRUE,sep=";",row.names=1)
pos<-read.csv("Input_pos_deer.csv",header=TRUE,sep=";",row.names=1)
road<-read.csv("Input_pos_0_1_deer.csv",header=TRUE,sep=";",row.names=1)

dat.dist<-vegdist(dat,method="euclidean",diag=TRUE,na.rm=TRUE)
dat.pcoa<-pcoa(dat.dist)
dat.ax<-dat.pcoa$vectors

road.rda<-rda(dat.ax,road)
anova(road.rda,step=1000)
road.r2a<-RsquareAdj(road.rda)$adj.r.squared

pos.dist<-dist(pos,method="euclidean",diag=TRUE)
View(as.matrix(pos.dist))

pos.PCNM<-PCNM(pos.dist)
plot(pos.PCNM$spanning,pos)
dmin<-pos.PCNM$thresh
nb.ev<-length(pos.PCNM$values)

select<-which(pos.PCNM$Moran_I$Positive==TRUE)
length(select)
pos.PCNM.plus<-as.data.frame(pos.PCNM$vectors)[,select]

dat.rda<-rda(dat.ax,pos.PCNM.plus)
anova(dat.rda,step=1000)
dat.r2a<-RsquareAdj(dat.rda)$adj.r.squared

dat.fwd<-forward.sel(dat.ax,as.matrix(pos.PCNM.plus),adjR2thresh=dat.r2a)
nb.sig.PCNM <- nrow(dat.fwd) #8
PCNM.sign <- sort(dat.fwd[,2]) #1  2  4  5  6  9 10 31
PCNM.red <- pos.PCNM.plus[,c(PCNM.sign)]

dat.rda2<-rda(dat.ax~.,data=PCNM.red)
anova(dat.rda2,step=1000)
dat.fwd.r2a<-RsquareAdj(dat.rda2)$adj.r.squared

axes.test <- anova(dat.rda2, by="axis", step=1000)
nb.ax <- length(which(axes.test[,4] <= 0.05))

dat.axes <- scores(dat.rda2, choices=c(1:nb.ax), display="lc", scaling=1)
png(filename="deer selected PCNM axes.png",width=1500,height=600,units="px")
par(mfrow=c(1,3),cex=1)
```

```
s.value(pos,
dat.axes[,1],grid=FALSE,include.origin=FALSE,xlim=c(172000,265000),ylim=c(40000,120000))
axis(1,pos=40000)
axis(2,pos=180000)
s.value(pos,
dat.axes[,2],grid=FALSE,include.origin=FALSE,xlim=c(172000,265000),ylim=c(40000,120000))
axis(1,pos=40000)
axis(2,pos=180000)
s.value(pos,
dat.axes[,3],grid=FALSE,include.origin=FALSE,xlim=c(172000,265000),ylim=c(40000,120000))
axis(1,pos=40000)
axis(2,pos=180000)
dev.off()

shapiro.test(resid(lm(dat.axes[,1] ~ ., data=road)))
dat.axis1.road <- lm(dat.axes[,1]~., data=road)
summary(dat.axis1.road)

shapiro.test(resid(lm(dat.axes[,2] ~ ., data=road)))
dat.axis2.road <- lm(dat.axes[,2]~., data=road)
summary(dat.axis2.road)

shapiro.test(resid(lm(dat.axes[,3] ~ ., data=road)))
dat.axis3.road <- lm(dat.axes[,3]~., data=road)
summary(dat.axis3.road)

road_PCNM<-
read.csv("Input_road_PCNM_fwd_deer.csv",header=TRUE,sep=";",row.names=1,dec=",")
road_PCNM.rda<-rda(dat.ax,road_PCNM)
anova(road_PCNM.rda,step=1000)
road_PCNM.r2a<-RsquareAdj(road_PCNM.rda)$adj.r.squared

road_PCNM.fwd<-forward.sel(dat.ax,as.matrix(road_PCNM),adjR2thresh=road_PCNM.r2a)
nb.sig.road_PCNM <- nrow(road_PCNM.fwd)
road_PCNM.sign <- sort(road_PCNM.fwd[,2])
road_PCNM.red <- road_PCNM[,c(road_PCNM.sign)]

road_PCNM.rda2<-rda(dat.ax~.,data=road_PCNM.red)
anova(road_PCNM.rda2,step=1000)
road_PCNM.fwd.r2a<-RsquareAdj(road_PCNM.rda2)$adj.r.squared

##################################
#Simulating motorways
##################################

# This script is loosely based on the random barrier script by Frantz et al. (2010).

n.roads<-100

mid.x<-mean(pos[,1])
mid.y<-mean(pos[,2])
mid<-data.frame(mid.x,mid.y)
names(mid)<-c("x","y")

circle.points<-data.frame(matrix(0,nrow=n.roads,ncol=2))
names(circle.points)<-c("x","y")

circle.points[1,1]<-mid.x+1
circle.points[1,2]<-mid.y
for(i in 2:n.roads){
  circle.points[i,1]<-circle.points[i-1,1]-(2/n.roads)
  ifelse(i>(n.roads/2)+1,circle.points[i,2]<-circle.points[i-1,2]-(2/n.roads),circle.points[i,2]<-
circle.points[i-1,2]+(2/n.roads))
}

rnd.roads<-data.frame(matrix(0,nrow=nrow(pos),ncol=n.roads))
rownames(rnd.roads)<-row.names(pos)
for(i in 1:nrow(pos)){
  for(j in 1:n.roads){
    k<-((circle.points[j,2]-mid[1,2])*pos[i,1]-(circle.points[j,1]-
mid[1,1])*pos[i,2]+circle.points[j,1]*mid[1,2]-circle.points[j,2]*mid[1,1])/sqrt((circle.points[j,2]-
mid[1,2])^2+(circle.points[j,1]-mid[1,1])^2)
    ifelse(k==0,rnd.roads[i,j]<-NA,ifelse(k<0,rnd.roads[i,j]<-0,rnd.roads[i,j]<-1))
  }
}

rnd.roads.r2a<-data.frame(matrix(0,nrow=1,ncol=n.roads))
for(i in 1:n.roads){
  rnd.roads.rda<-rda(dat.ax,rnd.roads[,i])
  capture.output(anova(rnd.roads.rda,step=1000),file="Random roads anova's.txt",append=TRUE)
  rnd.roads.r2a[1,i]<-RsquareAdj(rnd.roads.rda)$adj.r.squared
}
shapiro.test(as.numeric(rnd.roads.r2a))

rnd.roads.diff<-data.frame(matrix(0,nrow=n.roads,ncol=n.roads))
for(i in 1:n.roads){
  for(j in 1:n.roads){
    rnd.roads.diff[i,j]<-chow.test(rnd.roads[,i],dat.ax,rnd.roads[,j],dat.ax)[4]
  }
}
rnd.roads.diff.sign<-which(rnd.roads.diff<=0.05,arr.ind=TRUE)

road.max<-which.max(rnd.roads.r2a)
max<-max(rnd.roads.r2a)
road.min<-which.min(rnd.roads.r2a)
min<-min(rnd.roads.r2a)
road.max.coord<-circle.points[road.max,]
slope.max<-(mid[1,2]-road.max.coord[1,2])/(mid[1,1]-road.max.coord[1,1])
road.min.coord<-circle.points[road.min,]
slope.min<-(mid[1,2]-road.min.coord[1,2])/(mid[1,1]-road.min.coord[1,1])

png(filename="deer rnd roads sign diff max.png",width=800,height=800,units="px")
par(pty="s")
plot(pos,col=ifelse(road[row(pos),1]==0,"black","blue"),asp=1,pch=16,cex=2)
rnd.roads.diff.sign.max<-which(rnd.roads.diff.sign[,2]==road.max)
slope<-(mid[1,2]-circle.points[road.max,2])/(mid[1,1]-circle.points[road.max,1])
abline(a=mid[1,2]-slope*mid[1,1],b=slope,col="green")
for(i in 1:length(rnd.roads.diff.sign.max)){
  slope<-(mid[1,2]-circle.points[rnd.roads.diff.sign[rnd.roads.diff.sign.max[i],1],2])/(mid[1,1]-
circle.points[rnd.roads.diff.sign[rnd.roads.diff.sign.max[i],1],1])
  abline(a=mid[1,2]-slope*mid[1,1],b=slope,col="red")
}
dev.off()

png(filename="deer rnd roads sign diff min.png",width=800,height=800,units="px")
par(pty="s")
plot(pos,col=ifelse(road[row(pos),1]==0,"black","blue"),asp=1,pch=16,cex=2)
rnd.roads.diff.sign.min<-which(rnd.roads.diff.sign[,2]==road.min)
slope<-(mid[1,2]-circle.points[road.min,2])/(mid[1,1]-circle.points[road.min,1])
abline(a=mid[1,2]-slope*mid[1,1],b=slope,col="green")
```

```
for(i in 1:length(rnd.roads.diff.sign.min)){
  slope<-(mid[1,2]-circle.points[rnd.roads.diff.sign[rnd.roads.diff.sign.min[i],1],2])/(mid[1,1]-
circle.points[rnd.roads.diff.sign[rnd.roads.diff.sign.min[i],1],1])
  abline(a=mid[1,2]-slope*mid[1,1],b=slope,col="red")
}
dev.off()

rnd.roads.r2a.plot<-c(rnd.roads.r2a[,road.min:n.roads],rnd.roads.r2a[,1:road.min-1])
png(filename="deer R² rnd roads.png",width=800,height=400,units="px")
boxplot(rnd.roads.r2a.plot,names=NULL)
dev.off()

png(filename="deer rnd roads.png",width=800,height=800,units="px")
par(pty="s")
plot(pos,col=ifelse(road[row(pos),1]==0,"black","blue"),asp=1,pch=16,cex=2)
abline(a=mid[1,2]-slope.max*mid[1,1],b=slope.max,col="green")
abline(a=mid[1,2]-slope.min*mid[1,1],b=slope.min,col="red")
dev.off()

png(filename="deer all rnd roads.png",width=800,height=800,units="px")
par(pty="s")
plot(pos,col=ifelse(road[row(pos),1]==0,"black","blue"),asp=1,pch=16,cex=2)
for(i in 1:n.roads){
  slope<-(mid[1,2]-circle.points[i,2])/(mid[1,1]-circle.points[i,1])
  percentage<-(rnd.roads.r2a[1,i]-min)/(max-min)
  colour<-rgb(1-percentage,percentage,0)
  abline(a=mid[1,2]-slope*mid[1,1],b=slope,col=colour)
}
dev.off()

####################################
#Subsampling
####################################

# This script reiterates several key steps from the db-RDA script, using a number of smaller datasets.

n.datasets<-10
dataset.size<-10

west<-which(road==1)
east<-which(road==0)
red.roads.r2a<-data.frame(matrix(0,nrow=1,ncol=n.datasets))
dat.new.r2a<-data.frame(matrix(0,nrow=1,ncol=n.datasets))
write("\n",file="Reduced data road anova's.txt",append=TRUE)
write.table(dataset.size,file="Reduced data road
anova's.txt",append=TRUE,row.names=FALSE,col.names=FALSE)
write(" samples\n\n",file="Reduced data road anova's.txt",append=TRUE)
write("\n",file="Reduced data anova's.txt",append=TRUE)
write.table(dataset.size,file="Reduced data
anova's.txt",append=TRUE,row.names=FALSE,col.names=FALSE)
write(" samples\n\n",file="Reduced data anova's.txt",append=TRUE)
for(i in 1:n.datasets){
  select.west<-sample(west,dataset.size%/%2,replace=FALSE)
  select.east<-sample(east,dataset.size%/%2,replace=FALSE)
  dat.new.west<-dat[select.west,]
  dat.new.east<-dat[select.east,]
  dat.new<-rbind(dat.new.west,dat.new.east)
  pos.new.west<-pos[select.west,]
  pos.new.east<-pos[select.east,]
  pos.new<-rbind(pos.new.west,pos.new.east)
  road.new.west<-road[select.west,]
  road.new.east<-road[select.east,]
  road.new<-as.data.frame(c(road.new.west,road.new.east))

  dat.new.dist<-vegdist(dat.new,method="euclidean",diag=TRUE,na.rm=TRUE)

  dat.new.pcoa<-pcoa(dat.new.dist)
  dat.new.ax<-dat.new.pcoa$vectors

  road.new.rda<-rda(dat.new.ax,road.new)
  capture.output(anova(road.new.rda,step=1000),file="Reduced data road
anova's.txt",append=TRUE)
  red.roads.r2a[1,i]<-RsquareAdj(road.new.rda)$adj.r.squared

  pos.new.dist<-dist(pos.new,method="euclidean",diag=TRUE)

  pos.new.PCNM<-PCNM(pos.new.dist)
  plot(pos.new.PCNM$spanning,pos.new)
  dmin.new<-pos.new.PCNM$thresh
  nb.ev.new<-length(pos.new.PCNM$values)

  select<-which(pos.new.PCNM$Moran_I$Positive==TRUE)
  pos.new.PCNM.plus<-as.data.frame(pos.new.PCNM$vectors)[,select]

  dat.new.rda<-rda(dat.new.ax,pos.new.PCNM.plus)
  capture.output(anova(dat.new.rda,step=1000),file="Reduced data anova's.txt",append=TRUE)
  dat.new.r2a[1,i]<-RsquareAdj(dat.new.rda)$adj.r.squared
}
```

R-scripts used on the wild boar data are roughly identical, except for the amount of canonical axes, the file names and the exclusion of the subsampling script.

9.2.R-script for downscaling the wild boar database

```
####################################
#Downscaling the wild boar database
####################################

dat<-read.csv("Input_data_boar.csv",header=TRUE,sep=";",row.names=1)
pos<-read.csv("Input_pos_boar.csv",header=TRUE,sep=";",row.names=1)
road<-read.csv("Input_pos_0_1_boar.csv",header=TRUE,sep=";",row.names=1)

select.pos<-which(pos[,1]>180000)
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])
select.pos<-which(pos[,1]<235000)
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])

# The reduced database was formed using the script above. This database was then split up into two
# new databases using the script below. As described in 2.2.4 and Fout! Verwijzingsbron niet
gevonden..

select.pos<-data.frame(matrix(0,nrow=nrow(pos),ncol=1))
rownames(select.pos)<-row.names(pos)
for(i in 1:nrow(pos)){
  k<-((100000-50000)*pos[i,1]-(210000-190000)*pos[i,2]+210000*50000-
100000*190000)/sqrt((100000-50000)^2+(210000-190000)^2)
```

```
  ifelse(k==0,select.pos[i,1]<-NA,ifelse(k<0,select.pos[i,1]<-0,select.pos[i,1]<-1))
}

select.pos<-which(select.pos[,1]==1)
# To distinguish between the eastern and western part of the reduced database, the value indicated
# in red needs to be changed between 1 and 0 respectively.
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])
```

After creating the reduced databases, the script used for the complete database (see 9.1) was executed on the newly formed databases.

9.3. R-script for making an unbiased comparison between the red deer and the wild boar database

```
###################################
#Subsampling deer database for comparison
###################################

dat<-read.csv("Input_data_deer.csv",header=TRUE,sep=";",row.names=1)
pos<-read.csv("Input_pos_deer.csv",header=TRUE,sep=";",row.names=1)
road<-read.csv("Input_pos_0_1_deer.csv",header=TRUE,sep=";",row.names=1)

select.pos<-which(pos[,1]>190000)
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])
select.pos<-which(pos[,1]<220000)
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])
select.pos<-which(pos[,2]>65000)
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])
select.pos<-which(pos[,2]<95000)
pos<-pos[select.pos,]
dat<-dat[select.pos,]
road<-as.data.frame(road[select.pos,])

west<-which(road==1)
east<-which(road==0)
select.west<-sample(west,25,replace=FALSE)
select.east<-sample(east,25,replace=FALSE)
dat.new.west<-dat[select.west,]
dat.new.east<-dat[select.east,]
dat<-rbind(dat.new.west,dat.new.east)
pos.new.west<-pos[select.west,]
pos.new.east<-pos[select.east,]
pos<-rbind(pos.new.west,pos.new.east)
road.new.west<-road[select.west,]
road.new.east<-road[select.east,]
road<-as.data.frame(c(road.new.west,road.new.east))
```

The subsampled wild boar database was formed using the same script. Both new datasets were then subjected to the script described in 9.1.

FACULTY OF SCIENCE
Geel Huis
Kasteelpark Arenberg 11 bus 2100
3001 LEUVEN (HEVERLEE), BELGIUM
tel. + 32 16 32 14 01
fax + 32 16 32 19 95
www.kuleuven.be