Mammalian Biology 83 (2017) 64–70

Contents lists available at ScienceDirect

Mammalian Biology

journal homepage: www.elsevier.com/locate/mambio

Original investigation

The R package "phuassess" for assessing habitat selection using permutation-based combination of sign tests

Lorenzo Fattorini[a], Caterina Pisani[a,*], Francesco Riga[b], Marco Zaccaroni[c]

[a] Department of Economics and Statistics, University of Siena, P.zza S. Francesco 8, 53100, Siena, Italy
[b] Institute for Environmental Protection and Research (ISPRA), Via Ca' Fornacetta, 9, 40064, Ozzano dell'Emilia, Italy
[c] Department of Biology, University of Florence, Via Madonna del Piano 6, 50019, Sesto Fiorentino, Italy

ARTICLE INFO

Article history:
Received 6 May 2016
Accepted 13 December 2016
Handled by Heiko G. Rödel
Available online 14 December 2016

Keywords:
R software
Lepus europaeus
Multiple testing approach
Nonparametric procedure
Proportional habitat use

ABSTRACT

To bypass the methodological problems involved in the use of compositional data analysis by Aebischer, Robertson and Kenward (1993), the assessment of proportional habitat use is approached in a multiple-testing framework exploiting data from radio-tagged animals. The habitat use is assessed separately for each habitat type by means of the sign test to compare the proportion of use vs the proportion of availability. The resulting *p*-values are combined in an overall test statistic whose significance is determined permuting sample observations. The "phuassess" package for the R software, available from Comprehensive R Archive Network (CRAN), allows to straightforwardly perform the assessment. The package is presented and described and, in order to exemplify its application, habitat selection is assessed in a population of European brown hares *(Lepus europaeus)* settled in central Italy.

Introduction

Regression-based resource selection functions are currently the most widely used method for analyzing habitat selection. The spatial units partitioning the study region (e.g. pixels or quadrats) are conceived as resources and covariates associated with these units (e.g. elevation, forest cover, habitat type) are adopted as predictors. In this framework a resource selection function (RSF) is any function of the covariates that is proportional to the probability of use of a spatial unit (Manly et al., 2007). As such, RSF models proceed at spatial level, i.e. taking spatial units as sampling units. Then, if presence/absence data are recorded, 0 is assigned to units where the species is absent and 1 to units where present, while in the case of presence/available data, 1 is assigned to units where the species is present and 0 is assigned to all the possible units, or to a sample of them. In both cases, the prevailing models are binomial generalized linear models, usually logistic regression (Boyce et al., 2002). Although concerns arise with the use of telemetry data collected on temporal scale that leads to serious challenge to the usual independence assumptions, several techniques can help to handle this problem. Among them, Johnson et al. (2008) adopt a weighted distribution approach that incorporates autocorrelation by modelling animal movements, while Koper and Manseau (2009) adopt a generalized estimating equation including an additional variance component to accommodate correlated data. Alternative approaches to RSF models are discrete choice models (e.g. McCracken et al., 1998; McDonald et al., 2006) and categorical regression models (Kneib et al., 2011). These alternatives allow for the simultaneous analysis of several habitats. Multinomial logit models use only the observed locations as sampled units, thus bypassing the problem of generating absence data, while in categorical regression models the habitat type is treated as the response variable rather than presence/absence. All these advanced methodologies perform a profound analysis of habitat selection, providing explanations and predictions of this complex phenomenon.

However, the first and probably the most simple question to be addressed in habitat selection studies is if animals use habitats proportionally to their availability (the so-called proportional habitat use, henceforth referred to as PHU) or if there is significant evidence that they prefer – or avoid – some of them. As pointed out by Johnson (1980), PHU can be assessed at different levels by comparing: 1) the proportion of each habitat within the animal home range versus the available proportion within the study area, which is referred to as the Johnson's second order selection; 2) the proportion of animal radio-locations within each habitat versus the corresponding proportion within the home range, which is referred to as the Johnson's third order selection.

* Corresponding author.
E-mail address: caterina.pisani@unisi.it (C. Pisani).

http://dx.doi.org/10.1016/j.mambio.2016.12.003
1616-5047/

In this framework, Aebischer et al. (1993) propose the use of compositional data analysis. The approach by Aebischer et al. (1993) has the merits of viewing habitat selection analysis as the assessment of statistical hypotheses regarding the animal population under study. As such, it proceeds at animal level taking animals instead of radio locations as sampled units, and avoiding correlations problems by combining all the observations of one animal in one proportion. Then, proportions are transformed by log-ratios in order to avoid their unit-sum constraint and familiar multivariate tests are performed to compare used vs available habitats, but also to assess differences for sex, age, season or to relate habitat use to a set of covariates. Testing can be performed both parametrically, supposing a multivariate normal distribution for the log-ratios, i.e. an additive logistic distribution for the original proportions (Fang et al., 1990), and nonparametrically, by means of permutational procedures. In this sense the paper by de Valpine and Harmon-Threatt (2013) may be viewed as an extension of the compositional analysis obtained by using the more flexible Dirichlet-multinomial distribution for proportions.

However, turning to the first, simple question of comparing proportions of use vs proportions of availability, Fattorini et al. (2014) outline serious methodological drawbacks of compositional analysis. These mainly arise from the use of log-ratio transform of proportions in order to allow the application of standard multivariate techniques. Indeed the equivalence in the expectations of the log-ratio transforms does not entail equivalence in the expectations of original proportions, in such a way that the assessment performed on log-ratios does not actually assess PHU. Moreover, the procedure necessitates of arbitrary reconstructions of sample data in presence of zeroes, and it tacitly presumes that data are symmetrically distributed. As a possible consequence, in some situations the actual rejection rates of the PHU hypothesis turn out to be dramatically greater than the nominal level at which assessments are performed. That has been proven in a wide simulation study by Fattorini et al. (2014).

In order to overcome these drawbacks, Fattorini et al. (2014) propose an alternative formulation of the PHU hypothesis expressed in terms of the number of animals for which the proportion of use exceeds the proportion of availability, in such a way to assess habitat selection for each habitat type. Each single hypothesis is assessed by means of the sign test. The p-values resulting from each single test are combined for assessing the overall PHU hypothesis by means of the permutational procedure by Pesarin (2001) (see also Pesarin and Salmaso, 2010). At the end, the permutation test gives rise to an overall p-value.

The procedure is completely nonparametric, in the sense that it does not necessitate of any assumption about the probability distribution generating the sample data. If the overall PHU hypothesis is rejected, the p-values of each partial hypothesis are considered and the whole set of habitat types is partitioned into the set of preferred, avoided and proportionally used habitats. Contrary to Aebischer et al. (1993), habitats causing the rejection of PHU hypothesis can be detected. A further but less formal ordering is also performed between preferred habitats as well as between avoided habitats, while proportionally used habitats are obviously considered equivalent. These orderings are obtained from all the pair comparisons performed in the set of preferred and avoided habitats. The procedure is not rigorous, in the sense that problems involved by the use of multiple testing are neglected in this phase. Finally, it should be noticed that animals may have different periods of observation (e.g. animals which are partially missing because died). In this case proportions are not equally distributed among individuals because variances vary among them. This is a very serious problem for any parametric approach, such as the compositional analysis based on multivariate normality of log-ratios, but is not so for the permutational approach, because equality of individual distributions is not required in this case.

Notwithstanding the methodological problems involved in the use of compositional analysis for assessing PHU – never pointed out before Fattorini et al. (2014) – the procedure is widely applied and its use does not seem to decrease. Indeed, in accordance to Google Scholar, the number of citations of Aebischer et al. (1993) was 117 from 1993 to 1998, 467 from 1999 to 2004, 758 from 2005 to 2009 and 719 from 2010 to now. Probably, one reason for the current wide application of compositional analysis is the availability of the function `compana` (Calenge, 2007) in the R (R Core Team, 2012) package “adehabitatHS”, which automatically performs the procedure by Aebischer et al. (1993).

In this paper the R package “phuassess”, which automates the novel procedure based on the sign test, is described while its documentation is reported in the Supplementary material. It is hoped that the possibility of automatically adopt the combination of the sign test for assessing PHU will encourage the use of this procedure for avoiding the inflated type 1 errors and the arbitrary reconstruction of data which may be involved when using the procedure by Aebischer et al. (1993).

To exemplify the application of the novel procedure in the R environment and to evidence the differences with the procedure based on compositional analysis, habitat selection is assessed for a population of European brown hares (*Lepus europaeus*) settled in central Italy. Data are collected in 2008 from a sample of 14 individuals tagged with GPS collars.

Material and methods

Study area

The study was carried out in winter-spring 2008 in the protected area of Spicciano (Tuscany, Central Italy), encompassed 610.94 ha of heterogeneous farmland habitat and hilly landscape (380 m a.s.l.), positioned at 11°11′E, 43°33′N. Climate is of Mediterranean type, with mild winters and hot and dry summers (annual temperature range was 1.3°–29.7 °C). Actual land use data were recorded by means of field surveys. Vegetation was classified into 7 categories: *woodland* (30%), *scrub land and hedges* (7%), *winter cereals* (17%), *extensive fruit crops* (i.e. vineyards and orchards with inter-row cover-crops) (25%), *intensive fruit crops* (i.e. vineyards and orchards without inter-row cover-crops) (5%), *meadows* (5%), *fallow fields* (11%) (Figure A of the Supplementary material). Classification and analysis of GIS data were performed with ArcGis 10.1 (ESRI, Redlands, California). In the study area we estimated hare density by means of spotlight counts (Langbein et al., 1999), obtaining a value of 24.6 hares per 100 ha.

Data collection

Hares were captured by means of nets placed across opening near hedgerows and other sites suitable for hare resting. Captured individuals were fitted with GPS collars (Tellus mini—Televilt, weighting 74 g) scheduled to acquire animal location every 2 h for 98 days. Hares were released in the same place of capture. Location errors of GPS collars were evaluated at about 15 m, in such a way that circles of radius 15 m centred on the recorded positions were likely to cover most true locations.

The data set was collected adopting the study design II and III (Thomas and Taylor, 1990; Manly et al., 2007). We obtained individual localizations for each marked hare. We determined the population reference area building a minimum convex polygon obtained by pooling the localizations of all the individuals plus a buffer zone outlining the polygon of width 256 m. The radius of the