How much information is there in arthropod data about the landscape around sampling sites? Exploring a theory based on site-specificity of organisms

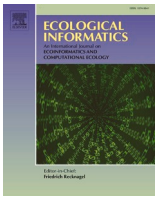
Musters, C.J.M.; Snoo, G.R. de

Ecological Informatics 81 (2024) 102645

Contents lists available at ScienceDirect

Ecological Informatics

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



How much information is there in arthropod data about the landscape around sampling sites? Exploring a theory based on site-specificity of organisms

C.J.M. Musters [a,*], G.R. de Snoo [a,b]

[a] Leiden University, Institute of Environmental Sciences (CML), P.O. Box 9518, 2300 RA Leiden, the Netherlands
[b] Netherlands Institute of Ecology (NIOO-KNAW), P.O. Box 50, 6700 AB Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Arthropods
Reversed normalized brier score
Spatial scale
Landscape characteristics
Random Forest

ABSTRACT

The importance of local versus landscape drivers of biodiversity is presently intensely discussed, which raises the question what information ecological sampling can provide about the relative importance of these factors and how the amount of information is distributed over different spatial scales. Here, we have tried to assess the amount of the information in sets of arthropod samples on four landscape characteristics, i.e., the percentage arable land, semi-natural area, urban area, and edge density, at spatial scales varying from 100 m to 3000 m around sample sites. A large, existing dataset of different studies from all over Europe was used for that. Random Forests were used for predicting the characteristic classes of the surrounding area. The accuracy of the predictions, calculated as the reversed Normalized Brier score, was used as measure of the amount of information. The results showed that, at least in Europe, the amount of information is different between edge density on the one hand, and arable land, semi-natural area, and urban area on the other hand. In case of edge density, the information decreased from 100 m to 250 m around the sample site, then increased to get a hump-shape between 250 and 3000 m, with the maximum amount at 1750 m. In case of the other three landscape characteristics, the information decreased from 100 m to 1000 m, and then stayed equal or slightly increased. These results could be explained by assuming that organisms present at a sample site are either site-specific, or non-site-specific. Site-specific organisms are thought to enable predictions of characteristics at the small scales, while non-site-specific organisms are thought to indicate characteristics of larger scales. The results implied that, for study designs, it is important to be aware of the type of processes that result in the presence of species at sample sites. For effective conservation measures for arthropods, the results showed that landscapes at a spatial scale of at least 9.6 km^2 should be taken in consideration in Europe.

1. Introduction

The relative importance of local versus landscape drivers of biodiversity is presently intensely studied and discussed, because of the decline in biodiversity and the search for effective nature conservation measures (Akter et al., 2023; Cardoso et al., 2009; Estrada-Carmona et al., 2022; Gallé et al., 2022; Gonthier et al., 2014; Harvey et al., 2022; Köthe et al., 2022; Marja et al., 2022; Martin et al., 2019; Petit and Landis, 2023; Schweiger et al., 2005; Tscharntke et al., 2012; Tscharntke et al., 2021). In studying the relationship between organisms and their environment, ecologists make a distinction between the effect of local characteristics of the environment on the abundance of organisms and the effect of environmental characteristics of the surroundings of the location - the landscape - on that abundance. Marja et al. (2022) concluded in their meta-analysis that "increasing landscape complexity primarily enhances species richness".

The results of these kind of studies will depend on the amount of information that ecological datasets contain on the local versus landscape characteristics of the areas around the sample sites. If information on either of the spatial levels is lacking or spars, the interpretation of the results of these studies might be problematic. Our study tries to assess the relative amount of information present in datasets based on samples of arthropod communities. It starts by developing a concept on how information in samples might be distributed over different spatial scales.

* Corresponding author.
E-mail address: musters@cml.leidenuniv.nl (C.J.M. Musters).

https://doi.org/10.1016/j.ecoinf.2024.102645
Received 13 October 2023; Received in revised form 13 May 2024; Accepted 14 May 2024
Available online 17 May 2024

In order to be able to correlate biodiversity with one or more spatial or temporal gradients of environmental characteristics of the study site, ecological field studies typically collect information on the set of organisms present at a set of sites at a set of moments in time. However, it has long been known that it is not true that each organism that is present at a site has an equal chance of being sampled. Many studies have shown that the organisms that were collected during ecological field work depended on sampling technique, perceptibility of the organisms, behavior and activity of the organisms, season, hours of the day, weather (temperature, precipitation, wind), and the biotic and abiotic characteristics of the sample site, including diversity therein (e.g., Hohbein and Conway, 2018; Jouveau et al., 2022; McNamara Manning and Bahlai, 2021; Thomas et al., 1998; Wardhaugh, 2014; Yi et al., 2012).

Apart from this, the set of organisms present at a site is not only affected by the characteristics of the site at that moment, but it is also affected by its past and the area surrounding it. Many studies showed that biodiversity is affected by both local and landscape characteristics (e.g., Gonthier et al., 2014; Marja et al., 2022; Petit and Landis, 2023). Recent studies showed that the size of the area around a sample site that affects a sample of arthropods might be larger than usually expected (Musters et al., 2021, 2022).

In general, a sample, even when it contains only information on a small part of the organisms present at the site, will show an expert that it was taken from, for example, a pond, forest, grassland, or desert. On the other hand, its species composition will also show the expert from which species pool, i.e., on which continent, it was taken. So, a sample will contain information on the sample site as well as on the surrounding area of the site at the very large level of scale. But how much information do samples contain on the landscape, at different levels of scales?

In this study, we tried to unravel for the first time the difference in relative amount of information on local versus landscape characteristics that sets of arthropod samples contain. Studies that consider difference in spatial scale while assessing information on landscapes are very rare (Bouasria et al., 2023). Since arthropods have been collected with all kind of techniques in all kind of landscapes, arthropod samples are well suited as examples for answering our research question. As to the concepts that were used to indicate the spatial scale, 'local' was used for the area around the sample site with a radius smaller than 100 m, 'region' for the area with a radius larger than 200 km, and 'landscape' for the areas in between, which deviates only slightly from the categories of Pearson and Dawson (2003).

For assessing the amount of information in a dataset, we predicted the characteristics of the landscape around the sample site. Since a prediction based on samples will be more accurate when the samples contain more information on the landscape characteristics, we considered the amount of information equivalent to the accuracy of the prediction. Based on this, different levels of spatial scale were compared on their differences in amount of information present in the datasets.

Four different landscape characteristics were predicted: percentage arable land, percentage semi-natural area, percentage urban area, and edge density. Percentage arable land is often regarded as the reverse of landscape complexity, which has been defined as the percentage non-arable land in an area (Tscharntke et al., 2012). Semi-natural areas included hedges, grassy margins, unmanaged grassland, shrubs, and fallows (Martin et al., 2019). Edge density is measured as the total length of the edges between crop fields and their surroundings, including crop/crop and crop/non-crop edges, divided by area. It is a metric for landscape configuration: it is large when crop fields and non-crop patches are small, and vice versa (Martin et al., 2019). One could argue that edge density is actually a more accurate metric of landscape complexity than the often-used percentage of non-arable land (Martin et al., 2019). Landscape complexity is regarded as the key predictor in the theory of the effectiveness of nature conservation policy in agricultural intensive areas (see Fig. 1 in Tscharntke et al., 2012). It plays a central role in the above described discussion about the relative importance of local versus landscape drivers of biodiversity.

The non-parametric technique of Random Forest (RF) was used for the predictions (Breiman, 2001; Prasad et al., 2006). This enabled optimal use of all the information that is present in a dataset, without the need for a priori variable selection or for specific, parametric assumptions on the relationship between the composition of organisms in samples and the characteristics of the landscape around the sample sites (Breiman, 2001; Fox et al., 2017).

2. Theory

To understand what kind of information might be present in a set of samples, we assume that this information is stored in a dataset: a matrix with the cases (sample) as the rows and the dependent variables (characteristics of the landscape) and the independent variables (taxonomic units) as columns. Each sample is collected at a different site. This dataset is used to study the relationship between the set of organisms in the samples and the characteristics of the area around the sample sites. So, not the sample site itself, but the area around it is the object of this study. This area has a predefined size indicated by the radius of a circle around the site.

Since organisms can be regarded as goal-oriented entities (Godfrey-Smith, 2014; Musters et al., 2023; Thompson, 1987), a distinction can be made between two fundamentally different types of reasons for an organism to be present at sample sites: an organism can be present either because it was intended to be at the site, or because it was not specifically intended to be there. We will call the first kind of organisms the 'site-specific' organisms and the second kind the 'non-site-specific' organisms.

A *site-specific organism* is present at a site because of *intended behavior of the organism or its parents*. Examples are that the organism hatched or metamorphized at the site and is somehow attached to it, that it is looking for food or mates, that the site provides a hiding place for predators or shelter against bleak weather. Caterpillars on host plants, hover flies on flowers, predators on plants where the prey is abundant, parasites on hosts, the presence of certain spiders in the vegetation that has the structure needed for constructing their webs, or the presence of scavengers on dead bodies, are examples. These kind of reasons for species to be present at a site are deterministic, because they depend on organisms being attracted to, or trying to avoid specific biotic or abiotic characteristics of the site. Therefore, sites that differ in their biotic or abiotic characteristics can be recognized by differences in the presence of the site-specific organisms in samples taken from there. Now, when our study object, the area, is small, chance will be high that the site characteristics are present in the complete area. A small area may easily lay completely within a woodlot, a field or road margin, a crop field, or a grassland. But when areas increase in size, the probability increases that

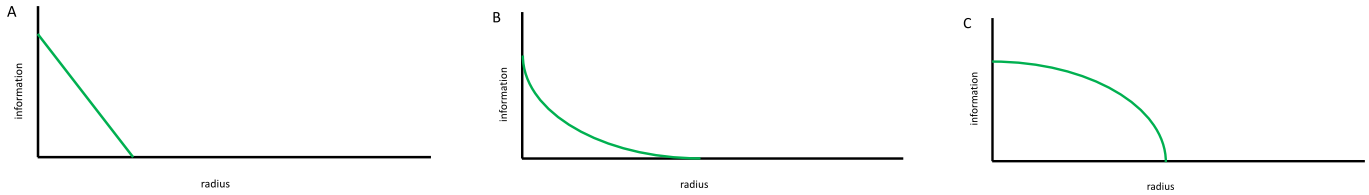


Fig. 1. Hypotheses on the relationship between radius of area around a sample site and the amount of information based on the site-specific organisms sampled.

other characteristics than those of the site become present in the areas. One could say that the site characteristics become diluted with increasing size of an area. As a consequence, the presence of site-specific organisms in the sample becomes less informative on the characteristics of the complete area. In other words, the amount of information that site-specific organisms give on the complete area around the sample site decreases with increasing area size (Fig. 1).

A *non-site-specific organism* is present at a site *without any specific ecological reasons*. Examples are that the organism happens to be passing by, that the wind blew it to the site, or that another organism brought it there. These types of events are stochastic, because they are independent of the biotic or abiotic characteristics of the site. Non-site-specific organisms are present in the samples because they were unintendedly transported to the sample site, for example by wind, water, or biota, including man. Alternative names for these organisms could be 'tourists' or 'vagrants', but we do not use these terms because they are also used in other contexts, which can lead to confusion. For example, tourist species may also be regarded as species that attract tourists (Yang et al., 2022) and vagrant species may be discussed in migration theory of birds (Gilroy and Lees, 2003). The transportation may take place over large distances, but the chance that an organism reaches a sample site from a place where it actually is site-specific, decreases probably quickly with distance (Östrand and Anderbrant, 2003). One can assume that each organism has during a certain life stage a more or less fixed distance over which it has a relatively high chance of being unintendedly transported by stochastic processes, the 'transportation distance'. An organism will only have a chance of being present in a sample when its original site is within this transportation distance. So, when an organism is present in a sample, it is known that its original site is present within an area with the radius of the transportation distance of the organism around the sample site. One could say that the presence of a non-site-specific organism in a sample gives one unite of information: the original site is present within a radius equal to its transportation distance. The more of these non-site-specific organisms are present in a set of samples, the more information on the surroundings of sample sites is present in the dataset. But since each organism has its own transportation distance, the amount of information on the total area around the sample site depends on the transportation distances of the organisms sampled. If many organisms in the samples have a more or less equal transportations distance of, say, 1000 m, and only a few have a distance of 5000 m, than more information is available in the dataset on the area with radius of 1000 m around the sample site than on the area with radius of 5000 m. So, the amount of information in the samples correlates to the distribution of the transportation distances of the organisms. In Fig. 2, a normal distribution and two related distributions are supposed to occur in arthropods (Jopp and Reuter, 2005; Thomas et al., 2003).

The complete amount of information in the dataset must be result of the mixture of site-specific and non-site-specific organisms in the set of samples. But which organisms are site-specific, and which not, may be largely unknown. However, in the above we have developed two kinds of hypotheses on the relationship between the amount of information and area size, one for the site-specific organisms and one for the non-site-specific organisms. And knowing that the dataset will be a mixture, we can now combine these two kinds of hypotheses to a set of hypotheses on the relationship between amount of information and area size in the complete dataset. The main problem for combining is that the scale of the x- and y-axis in Figs. 1 and 2 may not be the same. It is possible that the amount of information from either site-specific organisms or non-site-specific organisms is negligible low, so that either Fig. 2 or Fig. 1 reflects the relationship between amount of information and radius of the complete dataset. But when both types of information are relevant, many different hypotheses on that relationship can be generated. Fig. 3 gives an illustration. It combines Fig. 1A and Fig. 2A, assuming different scales of the y-axis: when the maximum amount of information from the site-specific organisms is larger than that of the non-site-specific species, when it is more or less equal, and when the maximum amount from the non-site-specific organisms is larger (Fig. 3).

We assume that in sets of arthropods samples collected in agricultural, urban, and semi-natural areas, the distribution of information on the characteristics of the landscape at different levels of spatial scales is more or less according to one of the graphs in Fig. 3. Any strong empirical deviation of the distribution from these graphs must be regarded as a rejection of the above theory.

Presently, it is unknown how information in samples is distributed over the size of area surrounding sample sites. In this study, we explored this distribution for the first time. Our way of doing that was by trying to predict the characteristics of the surrounding area at different spatial scales based on the samples, and evaluate these predictions. The better the prediction, the more information the samples contain on the surrounding area at that spatial scale.

3. Material and methods

To be able to perform the exploration of the distribution of information, a set of datasets was needed of samples of organisms present at sites of which characteristics of the surrounding area are known. Martin et al. (2019) collected such datasets from all over Europe. Also, a way of predicting the characteristics of an area from samples was needed that make optimal use of all available information within the sample. The Random Forest does exactly that (Breiman, 2001). For classification, it delivers a Brier score, which is a strictly proper score of the accuracy of a prediction and, therefore, can be regarded as a measure of the amount of the information that samples contain on the characteristics of an area (Brier, 1950; Ishwaran and Lu, 2019).

3.1. Data

Martin et al. (2019) brought the data together from 59 European landscape studies (doi: https://doi.org/10.5061/dryad.6tj407n). All studies were from agricultural areas, but the landscapes included varied strongly, from Scandinavian to Mediterranean, from low lands to mountainous, and from small-scale and closed to large-scale and open landscapes. The research units are samples characterized by site variables, crop variables, landscape variables, sampling variables, and arthropod data measured as abundancies of Operational Taxonomic Units (OTUs). For more details see Martin et al. (2019). From all studies in the dataset, we selected the data of the sample techniques for which 10 or more samples and 17 or more OTUs were available. To ensure that

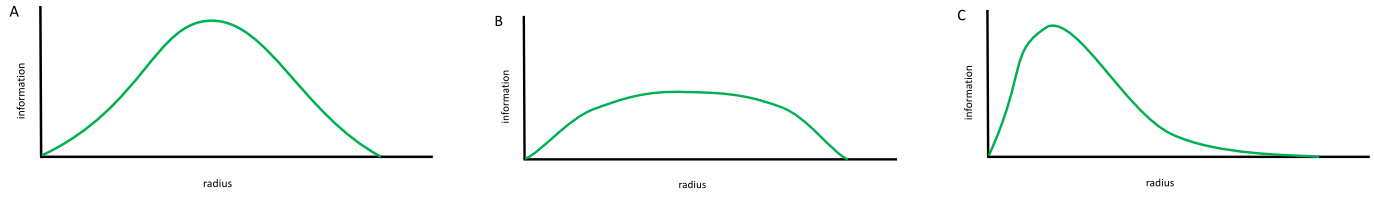


Fig. 2. Hypotheses on the relationship between radius of area around a sample site and the amount of information based on the non-site-specific organisms sampled. A: the transportation distance of organisms has a normal probability distribution over radius; B: the transportation distance has a near uniform probability distribution with a low mean probability and a large standard deviation; C: the transportation distance has a skewed probability distribution.

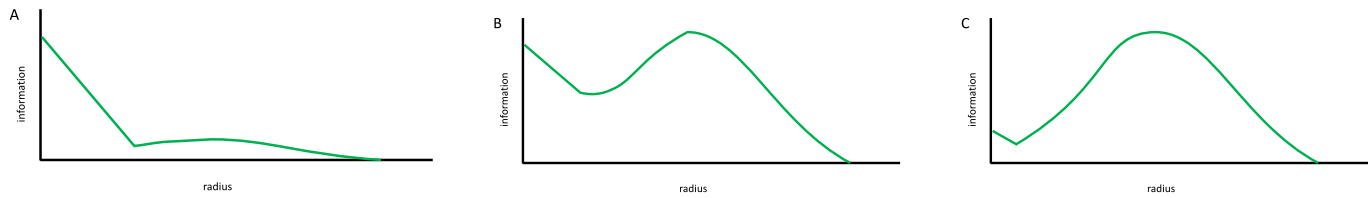


Fig. 3. Hypotheses on the relationship between radius of area around a sample site and the amount of information based on combining Fig. 1A and 2A. A: the maximum information from site-specific organisms is larger than that of non-site-specific ones; B: the maximum information from site-specific organisms is equal to that of non-site-specific ones; C: the maximum information from site-specific organisms is lower than that of non-site-specific ones.

our analyses were not biased by extremely large abundances for some OTUs, and zeroes in others, we applied a Hellinger transformation on all abundances (Borcard et al., 2011).

3.2. Landscape characteristics

The aim of this study was to analyze the amount of information in arthropod samples for predicting landscape characteristics with the abundance of arthropod OTUs in samples. Three types of landscape were to be predicted: the location of the sample site, the region of the sample site, and landscape around the site with varying radii. The information on type of landscape was taken from the data of Martin et al. (2019). For the location of the site, four local classes were used: grassland, crop field, orchard, and field margin. For the region, the locations of the studies were divided into four regional classes: Eastern Europe (longitude $\geq 10°$), Mid Northern Europe (longitude $<10°$ and $\leq 0°$; latitude $\geq 49°$), Mid Southern Europe (longitude $<10°$ and $\leq 0°$; latitude $<49°$), and West Europe (longitude $<0°$). For characterizing the landscape of the areas per radius, four landscape variables were selected: percentage arable land, percentage semi-natural area, percentage urban area, and edge density (km per ha) as defined by Martin et al. (2019). Two other available variables, percentage forest and percentage water, were not taken into consideration because of high amount of zero scores in these variables (Fig. S1, Supplementary Information). The values of each variable were categorized into five, approximately logarithmic, landscape classes from very low to very high (Table S1). The distribution of all samples over the landscape classes is shown in Fig. S2.

3.3. Statistical analyses

For predicting the location, the region, and the four landscape characteristics in the radii around the sample site, Random Forest (RF) for classification was applied. RFs consist of a large number of decision trees that classify the cases according the classes of the response variable (Breiman, 2001). Each tree of the forest uses a random subset of cases and predictive variables out of the complete dataset as a learning set. Using RF instead of a single classification tree prevents over-fitting (Breiman, 2001; Strobl et al., 2009). RFs are especially fit for handling datasets in which the number of predictive variables is large compared to the number of cases (Fox et al., 2017; Strobl et al., 2009). They do not have a problem with handling non-linear relationships between the predictive variables and the response variable (Strobl et al., 2009). Moreover, since on every node it is decided which predictive variable should be used for dividing the remaining set of cases, interactions between predictive variables are also taken into consideration. RF uses the internal 'Out Of Bag' (OOB) technique to assess the accuracy of the classification: each decision tree of the forest takes a different random subsets from the dataset as training set, the 'Bag' cases, and the classification of the cases outside that subset, the OOB cases, are used to test the trained decision trees and assess the classification error of the RF (Breiman, 2001; Fox et al., 2017).

We used the function *rfsrc()* with its default settings of the *randomForestSRC* package in R (Ishwaran et al., 2008; Ishwaran and Kogalur, 2007, 2022; R Core Team, 2022) for the classification. The in Section 3.2 *Landscape characteristics* defined classes of location, region, and the four landscape characteristics were the dependent or response variables, and the Hellinger transformed abundance or present/absent of OTUs were the independent or predictive variables. Each sample was considered as an independent case. Each RF consisted of 500 decision trees.

The classification of the samples into local and regional categories was done five times, one for every sample technique of which more than three separate studies were available. These five sample techniques are the pitfall, pan trap, sweep net, transect count, and trap nest, all as defined by Martin et al. (2019).

The classification of the samples into the classes of the four landscape characteristics was performed for every combination of study, sample technique, and radius separately. This resulted in 184 RFs of arable land, 174 RFs of semi-natural areas, 169 RFs of urban areas, and 168 RFs of edge density of which the accuracy of prediction of landscape characteristic were available for further analysis. That the number of RFs per landscape characteristic is not equal is because in some combination of study, sample technique, and radius the landscape characteristics fall all in the same class.

The function *rfsrc()* uses the classification errors of the RF to calculate the Normalized Brier score. The Brier score is a strictly proper scoring function that calculates the accuracy of probabilistic predictions (Brier, 1950). Normalizing the Brier score makes it independent of the number of categories of the response variable (Ishwaran and Lu, 2019). The Normalized Brier score has a theoretical value range between 0 (perfect accuracy) and 1 (zero accuracy). For clarity sake, that score was transformed, by subtracting it from one, into the 'reversed Normalized Brier score' (rNBs) that runs from 0 (zero accuracy) to 1 (perfect accuracy). Because of the bootstrapping parts of the RF procedure, classification may result in negative rNBs values. The distribution of these negative values can be regarded as half of the variance distribution of zero accuracy. In accordance with Strobl et al. (2009), the negative variance distribution was used to estimate a threshold value for accuracies higher than zero (Musters and van Bodegom, 2018). rNBs's higher than the absolute 0.025 quantile of the rNBs-distribution were regarded as significantly different from zero.

Our further analysis, the actual analysis of the amount of information, started with the exploration of the correlation between rNBs and the sample technique, number of OTUs, number of samples, longitude of study, latitude of study, maximum distance between sample site (calculated with https://geo.javawa.nl/coordcalc/), mean characteristic of the landscape, number of characteristic classes, and radius around the sample site. Based on the significance of the correlation between rNBs and the linear, quadratic, or cubic equation with these independent variables, and a backward stepwise model selection, a minimum Linear Mixed Model (LMM) was constructed for describing the relationship between dependent variable rNBs and independent variable landscape scale, i.e., radius, corrected for confounding variables. For a formal model selection, starting with a complete model including all independent variables and their interactions, our number of cases of calculated rNBs were too small (Anderson, 2008). Because it could not be assumed that the different rNBs calculated from the same dataset of a study were independent observations, the random effect variable was the study wherein the samples were collected. The LMM was fit with the

lmer() function of the *lme4* package in R (Bates et al., 2015). Graphs were made with the *scatterplot()* function of the *car* package, the *emmip()* function of the *emmeans* package (Lenth, 2022), and several functions of the *ggplot2* package (Wickham, 2016), all in R (R Core Team, 2022). Spatial autocorrelation of the rNBs of the pitfall studies with radius 500 m, the most common type of study in our dataset, was checked by calculation Moran's I with the function *moranI()* of the package *lctools* of R (Kalogirou, 2020), using a weight that selected the 3 nearest studies and the *p*-value of the randomized z-score.

4. Results

4.1. General results

The dataset contained 40 European studies that could be used to predict the local and regional class of the sample sites based on the five sampling techniques of which >3 studies were available. The accuracies of these predictions, in terms of the reversed Normalized Brier score (rNBs), were between 0.64 and 0.91 (Table 1).

The dataset contained 44 European studies that met the criteria (see Section 3.1. *Data* and *3.2. Landscape characteristics* in Material and methods) for predicting the four landscape characteristic classes of the areas surrounding the sampling sites. The maximum distance between the sample sites within these studies is on average 89.7 km, but strongly left skewed (Fig. S3). The total number of datasets that allowed prediction of the four landscape characteristic classes, i.e., all radii with known landscape characteristics, added over all sampling methods and all studies, was 185. In most of these datasets the radii had >different classes per landscape characteristic, but in some cases only one class was available and therefore did not show differentiation in the landscape characteristic. This resulted in a variable number of cases per landscape characteristic that allowed us to grow a RF to predict the landscape class of the surrounding area. Therefore, the number of rNBs calculations of the prediction differed between the four landscape characteristics (Table 2). The average rNBs over all four landscape characteristics was 0.206 (n = 697), but showed a left-skewed distribution (Fig. 4). Assuming that rNBs's higher than the absolute 0.025 quantile of the distribution, which was −0.238, are significantly different from zero, 247 predictions had an rNBs significantly higher than zero, which is 35.4% of all predictions (Table 2). The percentage of non-zero rNBs per radius variated between 24.3% and 42.1% (Table 3).

4.2. First analysis of the effect of scale on the accuracy of predictions

When plotting the rNBs against radius, the landscape characteristic seemed to show different relationships (Fig. 5). The correlations between the four landscape characteristics were not strong, with that between arable land and semi-natural area being the strongest ($r = -0.67$; Table S2).

The relationship between radius and rNBs per landscape characteristic is shown in Fig. 6. Based on Fig. 3, a cubic equation was used to first describe these relationships. In Fig. 7 is per landscape characteristic the effect given of the cubic relationship between radius and rNBs, estimated with an LMM without any confounding fixed effect variables and Study as random effect variable. When the predictions were based on present/absent of OTUs, instead of abundance, the accuracy of the predictions was slightly lower for all landscape characteristic (Table 2), this was consistent over spatial scales (Fig. S4), and the form of the relationship between the cubic relationship between radius and rNBs remained in all landscape characteristics the same (Fig. S5).

Table 1
Accuracy of the prediction of local and regional class per sampling technique. rNBs: reversed Normalized Brier score.

Sampling	Datasets	OTUs	Samples	rNBs Location	rNBs Region
Pitfall	15	1406	841	0.809	0.893
Pantrap	4	549	298	0.866	0.907
Sweepnet	10	485	449	0.800	0.798
Transect	7	334	356	0.644	0.734
Trapnest	4	249	208	0.710	0.725

4.3. Analyses of confounding variables

The rNBs of all predictions together had a quadratic relationship with both longitude and latitude resulting in relative low scores in Mid-Europe (Fig. S6 and S7). No spatial autocorrelation was detected between the rNBs of the pitfall studies at a radius of 500 m, the most common studies (Moran's $I = -0.0675$, Expected $I = -0.0769$, p-value = 0.954). Obviously, the mean landscape characteristic measured within a study (meanLandscape) was different between the four characteristics (Landscape) (Fig. S1). Sampling technique (Sample.tech) did hardly affect rNBs, except for pitfalls that had relative high scores (Fig. S8). The number of OTUs per study varied between 17 and 565 (mean = 100.2), the number of sampling sites between 10 and 160 (mean = 44.2). Neither the number of OTUs, nor that of sampling sites affected the rNBs (Fig. S9 and S10). The maximum distance (MaxDist) between sample sites within a study showed a cubic relationship with rNBs (Fig. S11). The rNBs was lowest when the lowest landscape characteristic class (lowest.cc) was 2 and highest when the number of classes (n.cc) was 2 (Fig. S12, S13, and S14).

4.4. Best models

The LMM with all the rNBs as dependent variable and the cubic log (radius) interacting with the four landscape characteristics, including all significant confounding independent variables as discussed above, showed that the relation between the radius and rNBs was indeed significantly different between the four landscape characteristics (Fig. S15). Especially predicting edge density deviated from predicting the other landscape characteristics in its effect on the relationship between rNBs and radius. Because of that, we also constructed an LMM for predicting all the landscape characteristics except edge density. This model showed that predicting arable land, semi-natural area, and urban area did not differ in their relationship between the radius and rNBs (Fig. S16).

Next, two minimal best fitting LMM were backwardly stepwise selected, one for the relationship between the rNBs and the radius when predicting edge density classes (Table 4) and one for that relationship when predicting the three other landscape characteristic classes (Table 5). These models showed a cubic relationship between rNBs and radius when predicting edge density (Fig. 7D) and a quadratic relationship when predicting the other three landscape characteristics (Fig. 8).

5. Discussion

The results showed that, as expected, arthropod samples contained ample information on the location and the region of the sample sites. The results also showed that arthropod samples indeed contain information that can be used to predict the characteristics of a landscape from the small (radius 100 m) to the large (radius 3000 m) landscape scale (Fig. 7). However, the information was in about 65% of the cases not enough to be accurate (Table 2). It may seem that this percentage is high, but one should take into consideration that in order to predict the class of the landscape accurately 1) a study should have done in a landscape that has enough variety in landscape characteristic to be predicted and 2) the study should have collected a large enough dataset, that is, it should have sampled enough sites and should have collected and identified enough OTUs. The studies of the landscapes within 250 m from the sample site had the lowest percentage of non-zero rNBs (24%), that of 100 m from the sample site the highest (42%; Table 3).

Table 2
Datasets and predictions per landscape characteristic variable. Mean landscape is in percentage for arable land, semi-natural area, and urban area, and in km per ha for edge density. Percentage of predictions that result in a rNBs that is significantly higher than zero are between brackets. P/A: present/absent.

Landscape characteristic	Datasets	Mean landscape	Predictions	Sign. pred.	Mean rNBs Abund.	Mean rNBs P/A
Arable land	185	23.8	184	44 (23.9)	0.113	0.091
Semi-natural area	185	9.4	174	54 (31.0)	0.206	0.184
Urban area	185	4.8	170	71 (41.8)	0.250	0.239
Edge density	185	0.3	169	79 (46.7)	0.263	0.247
All	740	–	697	247 (35.4)	0.206	0.188

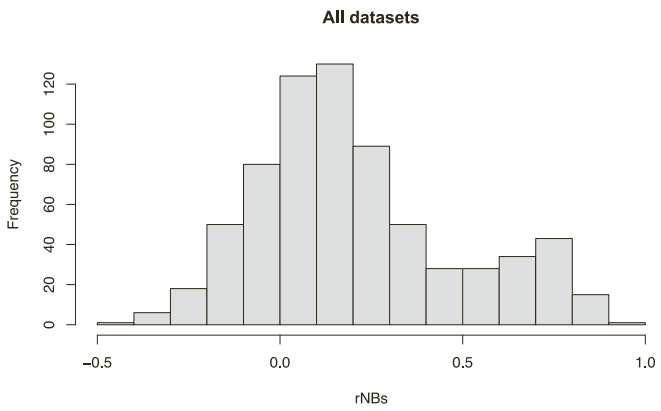


Fig. 4. The distribution of rNBs values over all 695 predictions of the landscape characteristic class.

Our most remarkable general result, though, is that the relationship of the spatial scale of the surrounding area and the amount of information in the samples is different for the prediction of *edge density* on the one hand, and *arable land, semi-natural area,* and *urban area* on the other hand.

For predicting *edge density*, a clear cubic relationship was found between the spatial scale of the surrounding area and the amount of information for that spatial scale (Fig. 7D). Of the scales studied, the information on the surrounding area in the 250 m radius around the sample site was the lowest and that in the ca. 1750 m radius was estimated to be the highest.

When predicting *arable land, semi-natural area,* and *urban area*, a quadratic relationship was found between the spatial scale of the surrounding area and the amount of information (Fig. 8). The highest amount of information was found on the landscapes in 100 m around the sample site. The information decreased to the scale of about 1000 m around the sample site and leveled off to an almost equal rNBs, but slightly increasing amount of information over the larger scales up to 3000 m. In this respect, no difference between the three landscape

Table 3
Distribution of predictions and non-zero rNBs over the radii. A: arable land; S: semi-natural area; U: urban area; E: edge density.

Radius (m)	Datasets	Predictions				Non-zero rNBs				Percentage				Mean perc.
		A	S	U	E	A	S	U	E	A	S	U	E	
100	44	43	40	34	40	14	19	18	26	32.6	35.0	55.9	45.0	42.1
250	44	44	43	39	44	7	12	15	14	15.9	16.3	30.8	34.1	24.3
500	44	44	43	44	41	9	12	20	13	20.5	20.9	27.3	48.8	29.4
1000	29	29	26	29	25	5	6	11	15	17.2	19.2	20.7	44.0	25.3
2000	14	14	13	14	10	5	4	5	6	35.7	38.5	28.6	50.0	38.2
3000	10	10	9	10	9	4	1	2	5	40.0	44.4	10.0	22.2	29.2
All	185	184	174	170	169	44	54	71	79	23.9	25.3	31.8	42.0	

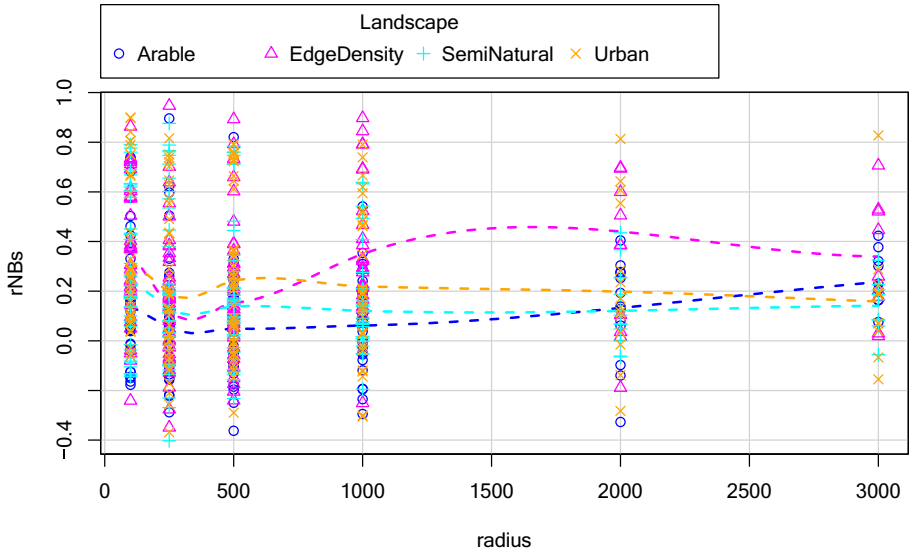


Fig. 5. Smoothed nonlinear regression lines of the raw relationship between rNBs and radius of the four landscape characteristics. Lines were constructed with the default settings of the function *scatterplot()* of the package *car* in R.

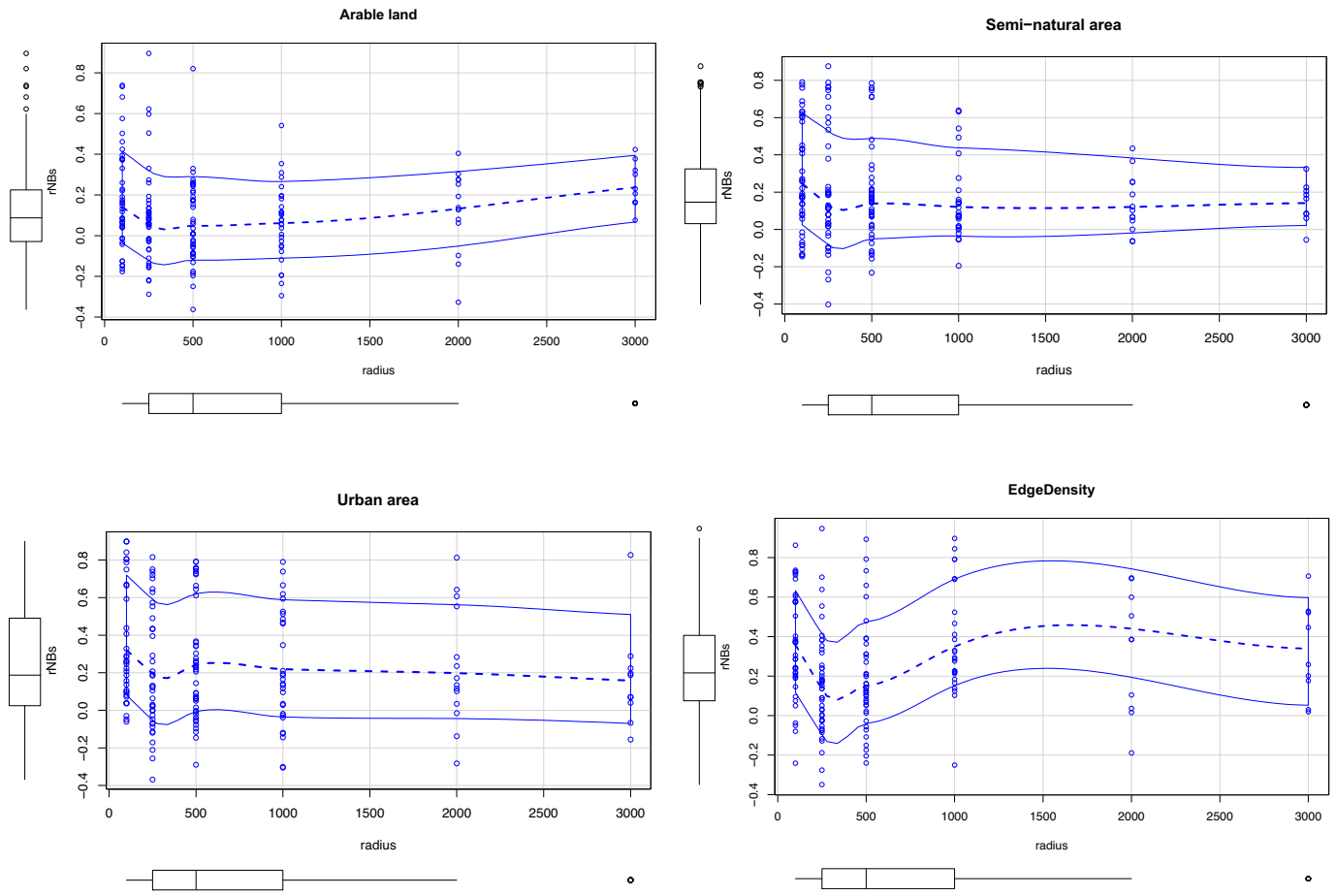


Fig. 6. Smoothed nonlinear regression lines and variance (colored area) of the raw relationship between rNBs and the radius of the four landscape characteristics. Lines were constructed with the default settings of the function *scatterplot()* of the package *car* in R.

characteristics was detected.

5.1. Linking the results to a theory of site-specificity

In the Theory section, we hypothesized that the distribution over levels of spatial scale of the amount of information from the site-specific organisms is fundamentally different from that from the non-site-specific organisms. Our results give us no reasons to reject our hypotheses.

First, the results concerning the prediction of *edge density* in the landscape around sample sites showed a high resemblance with the distribution of information in Fig. 3B, that assumed a more or less equal maximum amount of information of the site-specific and the non-site-specific organisms (compare Fig. 7D and Fig. 3B).

Second, when predicting *arable land, semi-natural area,* or *urban area* in the landscape around sample sites, the results showed a resemblance with the distribution of information in Fig. 3A, that assumes a relative high amount of information of the site-specific organisms and low information of the non-site-specific organisms, except that no decrease of information is detected at large levels of scale. This could be caused by the fact that the extend of this study is limited to levels of scale up to 3000 m around the sample site and the decrease is taken place at much larger scales, which is also suggested in previous studies (Musters et al., 2021, 2022). But it could also be caused by the small amount of information on these landscape characteristics at larger levels of spatial scale in the samples, so that any sign of a decrease is undetectable.

The relatively large amount of information that seems to be available on the edge density in landscapes as compared to the low amount on arable land, semi-natural area, and urban area is striking, especially since the latter three are landscape features that may occupy large areas, while edges have no area themselves, but reflect the configuration of the landscape. And even under the assumption that edges are landscape elements themselves, with a width of, say, 10 m, the area occupied by them is 10 to 100 times smaller than that of the other three elements studied.

According to our hypotheses on the distribution of information on spatial scale levels, our result suggests that organisms that inform us on edge density have an average transportation distance of ca. 1750 m, while that of organisms informing on arable land, semi-natural area, and urban area show no transportation distance within 3000 m. As stated before, this latter result could be due to the small amount of information on these landscape characteristics at larger levels of spatial scale in the samples, so that any sign of a transportation distance is undetectable.

This leaves us with the question why in arthropod samples there is ample information on differences in the landscape configuration at higher levels of spatial scale, while there is limited information on the difference in area of what, for arthropods, seems to be highly relevant types of land use, i.e., arable, semi-natural, and urban, at such scales? Martin et al. (2019) give several reasons why higher edge density may lead to higher survival of the populations of organisms due to increased opportunities for exchange between different patches, which again may result in recue effects, resource complementation and supplementation, and asynchronization with competitors and predators. However, higher survival of populations might also be expected due to larger semi-natural area and lower survival due to larger arable or urban areas (Aguirre-Gutiérrez et al., 2015; Mei et al., 2023; Sánchez-Bayo and Wyckhuys, 2019; Svenningsen et al., 2024). But why are these latter effects hardly traceable at large scales in our datasets? Source-sink

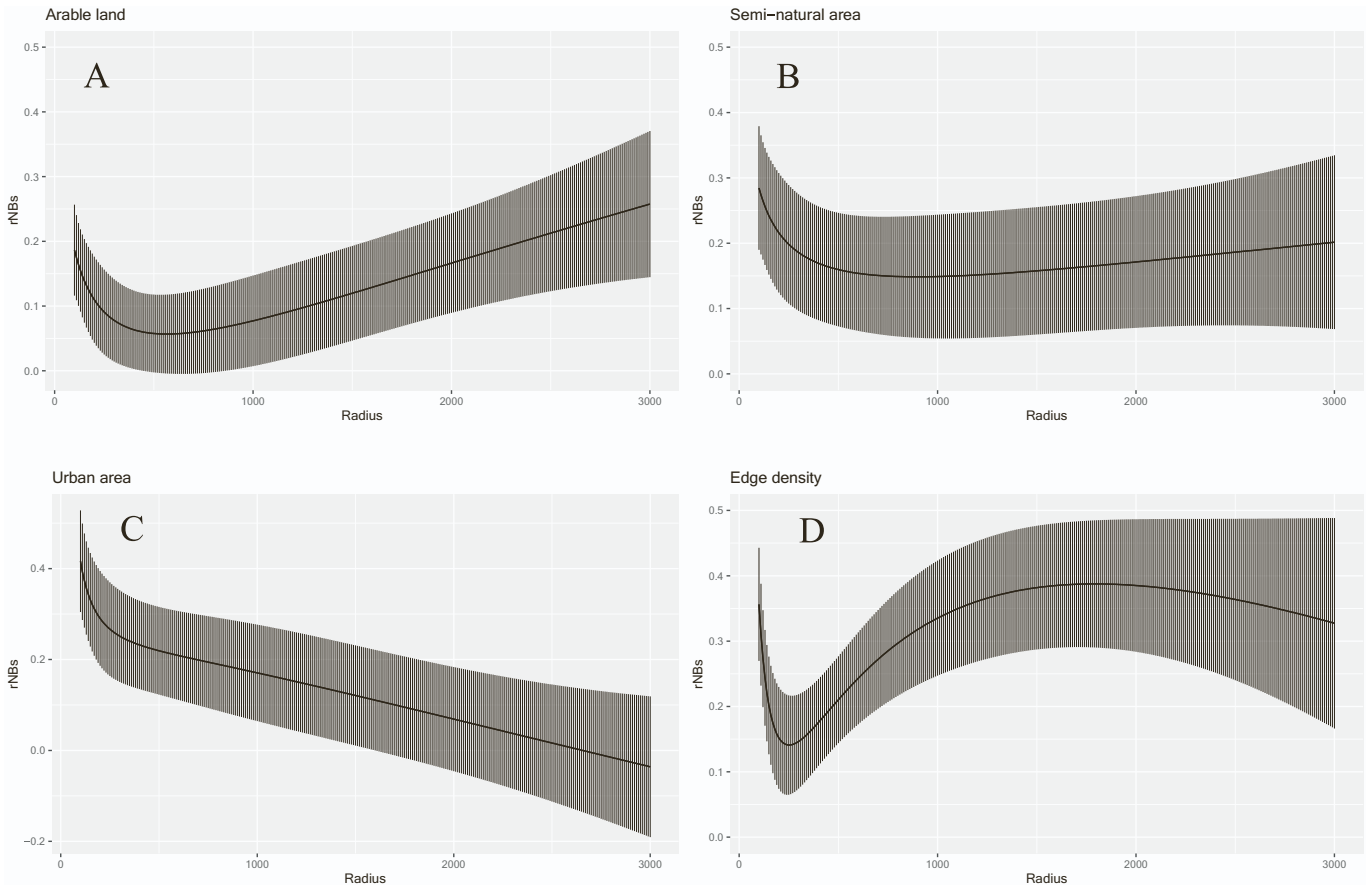


Fig. 7. Relationship and confidential interval (grey area) of the cubic relationship between radius and rNBs, estimated with an LMM without any confounding fixed effect variables and Study as random effect variable per landscape characteristic, calculated by the function *emmip()* of the *emmeans* package in R.

Table 4
Estimation of the effect of the fixed effect independent variables in the best fitting LMM model of rNBs for predicting edge density classes. Conditional R^2 of the complete model: 0.385; marginal R^2: 0.218; $n = 169$.

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	17.6500	4.5980	127.6	3.839	0.000	***
log10(radius)	−19.6200	5.2960	125.1	−3.705	0.000	***
I(log10(radius)^2)	7.1250	2.0030	124.4	3.558	0.001	***
I(log10(radius)^3)	−0.8438	0.2482	124.1	−3.399	0.001	***
I(meanLandscape^2)	1.7750	0.6807	62.6	2.608	0.011	*
I(meanLandscape^3)	−1.1140	0.5028	98.7	−2.216	0.029	*
Sample.techmalaise	−0.0643	0.1779	19.9	−0.361	0.722	
Sample.techpansweep	0.1972	0.2387	30.6	0.826	0.415	
Sample.techpantraps	0.0452	0.0900	98.4	0.502	0.617	
Sample.techpitfalls	0.1630	0.0663	86.9	2.458	0.016	*
Sample.techsuction	0.1660	0.1202	111.8	1.382	0.170	
Sample.techsurvey	−0.0523	0.1772	148.4	−0.295	0.768	
Sample.techtransect	0.0782	0.0932	82.7	0.839	0.404	
Sample.techtrapnests	0.0617	0.1099	29.5	0.562	0.579	
MaxDist	0.0083	0.0036	22.6	2.322	0.030	*
I(MaxDist^2)	−0.0001	0.0000	23.5	−2.333	0.029	*
I(MaxDist^3)	0.0000	0.0000	23.9	2.299	0.031	*
lowest.cc	−0.1124	0.0395	133.6	−2.842	0.005	**

theory (Pulliam, 1988) can help here: although, theoretically, non-site-specific organism can come from both source and sink populations, one might expect them more likely to come from sources than from sinks. This would result in a higher presence of organisms from sources in the datasets, and thus a higher amount of information about landscape features in the vicinity of the sample sites that support source populations. We think that these issue needs further research.

5.2. Further considerations

Tscharntke et al. (2012) described eight hypotheses for the way landscape characteristics affect biodiversity patterns and ecological processes, and gave ample references to support them. Of these we applied one for *biodiversity patterns* and one for *population dynamics* in our reasoning.

The *hypothesis for biodiversity patterns* says that the size of the landscape-wide species pool moderates local biodiversity (Tscharntke

Table 5
Estimation of the effect of the fixed effect independent variables in the best fitting LMM model of rNBs for predicting all landscape characteristic classes except edge density. Conditional R^2 of the complete model: 0.548; marginal R^2: 0.414; $n = 528$.

	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	1.99298	0.30509	494.4	6.532	0.000	***
log10(radius)	−0.64739	0.21838	493.3	−2.965	0.003	**
I(log10(radius)^2)	0.10788	0.04134	497.5	2.609	0.009	**
meanLandscape	0.00578	0.00136	516.0	4.258	0.000	***
LandscapeSemiNatural	0.20762	0.03779	503.7	5.494	0.000	***
LandscapeUrban	0.22020	0.03844	513.2	5.728	0.000	***
lowest.cc	−0.16674	0.02548	516.0	−6.544	0.000	***
I(highest.cc^2)	0.01459	0.00226	502.8	6.458	0.000	***
n.cc	−0.37839	0.07185	505.5	−5.267	0.000	***
I(n.cc^2)	0.02136	0.00975	503.0	2.19	0.029	*
meanLandscape:LandscapeSemiNatural	−0.00939	0.00258	513.2	−3.639	0.000	***
meanLandscape:LandscapeUrban	−0.00489	0.00387	508.2	−1.262	0.207	

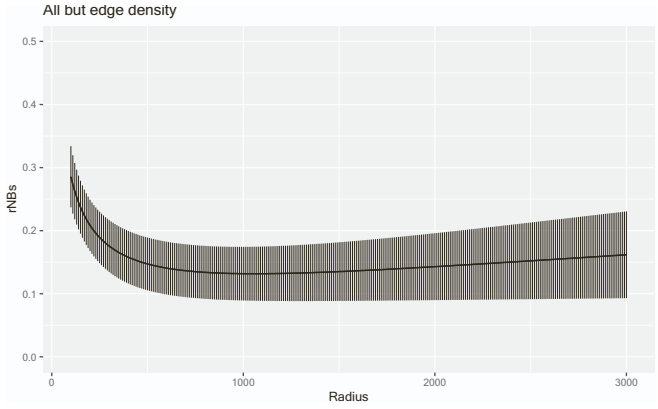


Fig. 8. Relationship and confidential interval (grey area) of the quadratic relationship between radius and rNBs predicting all the landscape characteristics except edge density, estimated with an LMM without any confounding fixed effect variables and Landscape:Study:Sample.tech as random effect variable, calculated by the function *emmip()* of the *emmeans* package in R.

et al., 2012). A simplified version of this hypothesis was used to explain the high amount of information on the regions from which the samples were taken (Table 1). It assumed that the four regions in which was divided Europe, viz. Eastern Europe, Mid Northern Europe, Mid Southern Europe, and West Europe, are large enough to contain distinctly different species pools. As far as we know, little research has been done on the absolute size of regions to be able to consider them as containing different species pools of arthropods. We also used the hypothesis for supporting one of our assumption about the datasets. The analysis of the information on landscape characteristics in sets of samples used the accuracy of predicting the landscape characteristic classes as measure of information. For that, it was assumed that all samples per study were taken from the same species pool. Distances between sample sites within one study were never larger than 385 km, but usually much smaller (Fig. S3). Of aquatic arthropods it has been shown that they may not be spatially limited at large scales (De Bie et al., 2012), even up to 300 km (Viana et al., 2015), suggesting that there is no reason to reject our assumption.

The *hypothesis for population dynamics* says, among other things, that spillover of organisms across habitats influences the landscape-wide community structure (Tscharntke et al., 2012). This hypothesis was used for the assumption that non-site-specific organisms may be present in samples. It was specified by assuming that organisms have, at a certain moment and place of sampling, a certain transportation distance. This assumption is support by literature discussed by Tscharntke et al. (2012), such as Schmidt et al. (2007), that shows that landscape-wide dispersal differs considerable among species and that species have specific spatial scales at which they respond to landscape complexity. Of course, the transportation distance may depend on certain characteristics of the landscape, often summarized in the 'connectivity' of a landscape. But it seems unlikely that the connectivity of the landscapes that were studied here, is different for the organisms that inform us on edge density from the organisms that do so on arable land, semi-natural area, or urban area. For that reason, we think that our results for the latter landscape characteristics is mainly due to low amounts of non-site-specific organisms that carry information on these characteristics in the samples.

5.3. Recommendations for study designs

The results and interpretation of the results have many consequences for the design of landscape studies of arthropod communities.

First, when a study is aimed at finding relationships between species abundancy and *local* biotic and abiotic characteristics, the size of the focus area should be chosen small, probably smaller than with a radius of 100 m (0.031 km^2). Also, it should be taken into consideration that part of organisms sampled are non-site-specific and do not inform on the local characteristics. These non-site-specific organisms cause noise in the dataset and may veil the relationship that is studied. They might even unjustly suggest context dependency of the results. Choosing a sample site and sampling time that minimalizes the chance for non-site-specific organisms could improve the research results.

Second, when the study is aimed at finding the effects of the *surrounding landscape* on the arthropod community, the theory suggests that the size of the surrounding area should be limited by the typical transportation distances of the species studied. According to our results, this size should be at least an area of a circle with radius of ca. 1750 m, which is 9.6 km^2, when studying arthropods. Also, it may be helpful for such studies to choose the sample site such that stochastic processes can easily transport organisms to it and to choose the time of the year such that these processes have a high chance of taken place. For monitoring systems that aim to follow the changes in arthropod abundancies as a result of landscape changes, the density of monitoring sample sites should be chosen in accordance with the size of the surrounding areas that is most cost effective. For arthropod in temperate agricultural landscapes the results of this study would mean that sample sites need not be closer than 3500 m to each other.

Third, a small, but consistently lower, amount of information was found when present/absent instead of abundance of OTUs were used. Obviously, abundancy gives more information on landscape characteristics than presence/absence, but the loss of information seems small. From a study efficiency point of view, it might be justified to limit the processing of samples to the assessment of the presence of OTUs.

Fourth, the confounding variables that turned out to be included in our best models showed that the amount of information available in an arthropod dataset may depend on the mean area or density of the landscape characteristic, the sampling technique, the distance between sample sites, and the way the depended variable is categorized in

classes.

Fifth, this study used samples taken from arthropod communities, without distinction between different arthropod groups. However, when the relationships between species abundancy and local characteristics are the focus of a study, one might consider to aim at arthropod groups that are known to be highly dependent on site-specific characteristics, i.e., species that are often referred to as specialists. And, when the focus is on the characteristics of the surrounding landscape, one might consider to aim at groups that are easily transported by wind, water, or biota.

Sixth, our results suggested that arthropod samples may contain little information on arable land, semi-natural area, or urban area in the landscape at scales beyond 500 m around the sample site. Research aimed at studying this might lead to non-informative results, unless it is based on large datasets.

Finally, this study uses terrestrial arthropods as a proxy for organisms taken from a local community for studying the amount of information in samples on the surroundings area. The results suggested that for a complete insight in that, spatial scales beyond the radius 3000 m should have been included in the study. Moreover, the method can also be applied on other animal groups, from aquatic invertebrates up to birds and mammals, and maybe even on plants and micro-organisms, as long as spillover of organisms can be assumed (Tscharntke et al., 2012). Bouasria et al. (2023) used regression RFs for studying the influence of spatial scales on model predictions of biomass.

In general, awareness of the processes that bring organisms to sample sites in relation to the aim of the research could improve the study design and, as a consequence, the research results.

5.4. Relevance for nature conservation

Although this study was not aimed at nature conservation, its results are relevant for that. The 'Intermediate Landscape Complexity Theory' of Tscharntke et al. (2012) says that nature conservation measures are most cost effective when they are done in agricultural landscapes of intermediate complexity. But the theory does not tell what the absolute size of the areas should be that one could consider as belonging to one landscape. The study showed that, for arthropods, this size should in Europe be at least of a radius of 1750 m, that is of 9.6 km^2, around the location of nature conservation measures, but other recent research show that it might be more (Akter et al., 2023; Cardoso et al., 2009; Evans et al., 2016; Musters et al., 2022).

Author's contribution

CM conceived the ideas, developed the theoretical concept, performed the statistical analyses, and wrote the first draft. GdS and CM discussed the ideas, improved the theoretical concepts and drafts, and agreed to publish. Neither of them had conflicts of interests.

CRediT authorship contribution statement

C.J.M. Musters: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **G.R. de Snoo:** Writing – review & editing, Supervision.

Data availability

Data are already available

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2024.102645.

References

Aguirre-Gutiérrez, J., Biesmeijer, J.C., van Loon, E.E., Reemer, M., WallisDeVries, M.F., Carvalheiro, L.G., 2015. Susceptibility of pollinators to ongoing landscape changes depends on landscape history. Divers. Distrib. 21, 1129–1140.

Akter, S., et al., 2023. Continent-wide evidence that landscape context can mediate the effects of local habitats on in-field abundance of pests and natural enemies. Ecol. Evol. 13, e9737.

Anderson, D.R., 2008. Model Based Inference in the Life Sciences: A Primer on Evidence. Springer, New York, USA.

Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67, 1–48.

Borcard, D., Gillet, F., Legendre, P., 2011. Numerical Ecology with R. Springer, New York, USA.

Bouasria, A., Bouslihim, Y., Gupta, S., Taghizadeh-Mehrjardi, R., Hengl, T., 2023. Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents. Ecol. Inform. 78, 102294.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brier, G.W., 1950. Verification of forecasts expressed in terms of probability. Mon. Weather Rev. 78, 1–3.

Cardoso, P., Aranda, S.C., Lobo, J.M., Dinis, F., Gaspar, C., Borges, P.A.V., 2009. A spatial scale assessment of habitat effects on arthropod communities of an Oceanic Island. Acta Oecol. 35, 590–597.

De Bie, T., et al., 2012. Body size and dispersal mode as key traits determining metacommunity structure of aquatic organisms. Ecol. Lett. 15, 740–747.

Estrada-Carmona, N., Sánchez, A.C., Remans, R., Jones, S.K., 2022. Complex agricultural landscapes host more biodiversity than simple ones: a global meta-analysis. PNAS 119, e2203385119.

Evans, T.R., Mahoney, M.J., Cashatt, E.D., Noordijk, J., de Snoo, G.R., Musters, C.J.M., 2016. The impact of landscape complexity on invertebrate diversity in interiors and fields in an agricultural area. Insects 7, 7.

Fox, E.W., Hill, R.A., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., Weber, M.H., 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environ. Monit. Assess. 189, 316.

Gallé, R., et al., 2022. Landscape-scale connectivity and fragment size determine species composition of grassland fragments. Basic Appl. Ecol. 65, 39–49.

Gilroy, J.J., Lees, A.C., 2003. Vagrancy theories: are autumn vagrants really reverse migrants? Br. Birds 96, 427–438.

Godfrey-Smith, P., 2014. Philosophy of biology. Princeton University Press, Princeton, USA.

Gonthier, D.J., et al., 2014. Biodiversity conservation in agriculture requires a multi-scale approach. Proc. R. Soc. B 281, 20141358.

Harvey, J.A., et al., 2022. Scientists' warning on climate change and insects. Ecol. Monogr. 2022, e1553.

Hohbein, R.R., Conway, C.J., 2018. Pitfall traps: a review of methods for estimating arthropod abundance. Wildl. Soc. Bull. https://doi.org/10.1002/wsb.928.

Ishwaran, H., Kogalur, U.B., 2007. Random Survival Forests for R, 7. R News, pp. 25–31.

Ishwaran, H., Kogalur, U.B., 2022. Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). R package version 3.1.1.

Ishwaran, H., Lu, M., 2019. Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Stat. Med. 38, 558–582.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S., 2008. Random survival forests. Ann. Appl. Stat. 2, 841–860.

Jopp, F., Reuter, H., 2005. Dispersal of carabid beetles - emergence of distribution patterns. Ecol. Model. 186, 389–405.

Jouveau, S., et al., 2022. Carabid activity-density increases with forest vegetation diversity at different spatial scales. Insect Conserv. Divers. 13, 36–46.

Kalogirou, S., 2020. Lctools: Local Correlation, Spatial Inequalities, Geographically Weighted Regression and Other Tools. R Package Version 0.2-8.

Köthe, S., et al., 2022. Improving insect conservation management through insect monitoring and stakeholder involvement. Biodivers. Conserv. https://doi.org/10.1007/s10531-022-02519-1.

Lenth, R., 2022. Emmeans: Estimated Marginal Means, Aka Least-Squares Means. R package version 1.8.1-1.

Marja, R., Tscharntke, T., Batáry, P., 2022. Increasing landscape complexity enhances species richness of farmland arthropods, agri-environment schemes also abundance - a meta-analysis. Agric. Ecosyst. Environ. 326, 107822.

Martin, E.A., et al., 2019. The interplay of landscape composition and configuration: new pathways to manage functional biodiversity and agroecosystem services across Europe. Ecol. Lett. 22, 1083–1094.

McNamara Manning, K., Bahlai, C.A., 2021. Experimental calibration of trapping methods for addressing bias in arthropod biodiversity monitoring. bioRxiv. https://doi.org/10.1101/2021.12.06.471448 preprint.

Mei, Z., et al., 2023. Inconsistent responses of carabid beetles and spiders to land-use intensity and landscape complexity in north-western Europe. Biol. Conserv. 283, 110128.

Musters, C.J.M., van Bodegom, P.M., 2018. Analysis of species attributes to determine dominant environmental drivers, illustrated by species decline in the Netherlands since the 1950s. Biol. Conserv. 219, 68–77.

Musters, C.J.M., Evans, T.R., Wiggers, J.M.R., van 't-Zelfde, M., de Snoo, G.R., 2021. Distribution of flying insects across landscapes with intensive agriculture in temperate areas. Ecol. Indic. 129, 107889.

Musters, C.J.M., Wiggers, J.M.R., de Snoo, G.R., 2022. Distribution of ground-dwelling arthropods across landscapes with intensive agriculture in temperate areas. Ecol. Indic. 140, 109042.

Musters, C.J.M., DeAngelis, D.L., Harvey, J.A., Mooij, W.M., van Bodegom, P.M., de Snoo, G.R., 2023. Enhancing the predictability of ecology in a changing world: a call for an organism-based approach. Front. Appl. Math. Stat. 9, 1046185.
Östrand, F., Anderbrant, O., 2003. From where are insects recruited? A new model to interpret catches of attractive traps. Agric. For. Entomol. 5, 163–171.
Pearson, R.G., Dawson, T.P., 2003. Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Glob. Ecol. Biogeogr. 12, 361–371.
Petit, S., Landis, D.A., 2023. Landscape-scale management for biodiversity and ecosystem services. Agric. Ecosyst. Environ. 347, 108370.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9, 181–199.
Pulliam, H.R., 1988. Sources, sinks, and population regulation. Am. Nat. 132, 652–661.
R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Sánchez-Bayo, F., Wyckhuys, K.A.G., 2019. Worldwide decline of the entomofauna: a review of its drivers. Biol. Conserv. 232, 8–27.
Schmidt, M.H., Thies, C., Nentwig, W., Tscharntke, T., 2007. Contrasting responses of arable spiders to the landscape matrix at different spatial scales. J. Biogeogr. 2007 https://doi.org/10.1111/j.1365-2699.2007.01774.x.
Schweiger, O., et al., 2005. Quantifying the impact of environmental factors on arthropod communities in agricultural landscapes across organizational levels and spatial scales. J. Appl. Ecol. 42, 1129–1139.
Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348.
Svenningsen, C.S., Peters, B., Bowler, D.E., Dunn, R.R., Bonn, A., Tøttrup, A.P., 2024. Insect biomass shows a stronger decrease than species richness along urban gradients. Insect Conserv. Divers. 17, 182–188.
Thomas, C.F.G., Parkinson, L., Marshall, E.J.P., 1998. Isolating the components of activity-density for the carabid beetle *Pterostichus melanarius* in farmland. Oecologia 116, 103–112.
Thomas, C.F.G., Brain, P., Jepson, P.C., 2003. Aerial activity of linyphiid spiders: modelling dispersal distances from meteorology and behaviour. J. Appl. Ecol. 40, 912–927.
Thompson, N.S., 1987. The misappropriation of teleonomy. Perspect. Ethol. 7, 259–274.
Tscharntke, T., et al., 2012. Landscape moderation of biodiversity patterns and processes - eight hypotheses. Biol. Rev. 87, 661–685.
Tscharntke, T., Grass, I., Wanger, T.C., Westphal, C., Batáry, P., 2021. Beyond organic farming – harnessing biodiversity-friendly landscapes. TREE 36, 919–930.
Viana, D.S., et al., 2015. Assembly mechanisms determining high species turnover in aquatic communities over regional and continental scales. Ecography 39, 281–288.
Wardhaugh, C.W., 2014. The spatial and temporal distributions of arthropods in forest canopies: uniting disparate patterns with hypotheses for specialisation. Biol. Rev. 89, 1021–1041.
Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, USA.
Yang, J., Su, K., Zhou, Z., Huang, Y., Hou, Y., Wen, Y., 2022. The impact of tourist cognition on willing to pay for rare species conservation: Base on the questionnaire survey in protected areas of the Qinling region in China. Glob. Ecol. Conserv. 33, e01952.
Yi, Z., Jinchao, F., Dayuan, X., Weiguo, S., Axmacher, J.C., 2012. A comparison of terrestrial arthropod sampling methods. J. Resourc. Ecol. 3, 174–182.