Imperial College London

The drivers of vascular plant species richness in UK broad leaved semi-natural woodlands

Petra Guy
August 2018

Word count 5882

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Science/Research at Imperial College London

Formatted in the journal style of The Journal of Ecology.
Submitted for the MRes/MSc in Computational Methods in Ecology and Evolution

The data for this project came from the CEH Woodland survey dataset. The concept for the work initially came from Simon Smart at CEH who helped me throughout the study with advice, feedback and encouragement.

The data was provided in Access database format. Significant data processing was required to obtain the necessary variables in the correct format for this analysis.

The use of zeta diversity within the project was suggested by Simon Smart, I developed the idea of using it to create a metric for heterogeneity, this is a new technique that was explored in this work.

Many thanks to my supervisor, Colin Prentice who has provided continuous input, comments, suggestions and immediate replies to my many emails.

Thanks to my course tutors, Samraat Pawar and James Rosindell, for their patient assistance.

Summary

1. The loss of plant biodiversity in the UK is a major concern, with a fifth of UK species endangered or vulnerable according to the latest IUCN Red List. The government's 25 Year Plan for the environment aims to halt this loss and build new habitats, including new woodlands. To ensure that biodiversity loss is halted in current woodlands and gain is maximized in new ones, we need to better understand the drivers of biodiversity.

2. This work restricts itself to an examination of vascular plant species diversity. First, the effects of environmental heterogeneity and abiotic factors were examined, to ascertain the main drivers of species richness in 103 semi-natural, broad-leaved woodlands across the UK. Second, species-area relationships were examined at two scales, to explore the relationship between the drivers of species richness and the exponent of the canonical species-area curve.

3. Habitat heterogeneity increased species richness, as did the exposure of the woodlands to surrounding natural habitats. Higher levels of soil organic matter, and the progression of woodlands to later successional stages, decreased species richness. Richness was also seen to have a unimodal response to soil pH with a peak around pH6.

4. At both scales, increased habitat heterogeneity led to an increase in the exponent of the species-area curve, whereas increased soil organic matter led to a decrease. At the larger scale, succession led to a decrease in the exponent.

5. *Synthesis.* Habitat heterogeneity measures included the presence of coppicing, open areas such as rides and riparian zones, the standard deviation of tree density, and the difference between species assemblages in different plots in the woodland. Results suggest that to maximize biodiversity woodlands should be managed; open areas should be maintained; and a variety of habitats should be encouraged. In addition, the increase in richness with exposure to surrounding natural habitats suggests that woodlands should, if possible, be large and connected. This implies that if possible, current woodlands should be increased in size, rather than creating new smaller disconnected sites.

Key words. Woodland biodiversity, species richness, zeta diversity, species area curves.

Introduction

Biodiversity worldwide continues to decline [Butchart et al., 2010, Tittensor et al., 2014] and the UK is no exception to this trend [Hayhow et al., 2016]. The IUCN Red List for the UK shows 22% of plant taxa as critically endangered, endangered or vulnerable [Stroh et al., 2014]. Biodiversity is important for ecosystem function [Gaston & Spicer, 2014], with woodlands performing specific additional functions including the improvement of water quality, carbon sequestration [Slee and Kyle, n.d] and the provision of habitats for other taxa. [Amar et al., 2006, Fuller & Warren, 1993].

The aims of the new 25-Year Environment Plan [Defra, 2018] include achieving "thriving plants and wildlife", planting 180,000 hectares of new woodland and ensuring that existing woodlands are better managed. However, more woodland is not necessarily the answer to biodiversity loss if we cannot ascertain why our current woodlands are not thriving.

This study was aimed at finding the main drivers of vascular plant species richness in UK broad-leaved semi-natural woodlands to provide research-based evidence on how best to manage, maintain and increase their plant biodiversity. This research is the only work of its kind based on a large, UK-specific dataset.

SUMMARY OF FACTORS SHOWN TO AFFECT SPECIES RICHNESS

pH AND SOIL ORGANIC MATTER

Studies of beech and hornbeam woodlands by Koojiman & Cammeraat [2010] report lower species richness with lower pH and increased soil organic matter (SOM). This was true for both taxa, suggesting that the effect was not the result of the physical barrier created by slowly degrading beech litter. Cornwell & Grubb [2003] showed that the greatest richness in woodlands occurred on nutrient rich soils, reporting a unimodal response with a peak around Ellenberg N of 7 for shade loving species. Since most nutrients are available around pH5/6 this implies that a unimodal richness response in woodlands should be expected.

LATITUDE

Ohlemuuler & Wilson, [2000] showed a decrease in richness with latitude in New Zealand for woody species, but this variation was not seen in herbaceous species. In studies on forests stretching from the equator to 60° latitude, Gillmann [2014] found that net primary production increases toward the equator and concludes that, since vascular plant richness is correlated with net primary production,

richness will also increase. However, this effect may not be evident given the modest change in latitude in our UK data.

ISOLATION

The 2010 Lawton report [Lawton et al., 2010] summarised the need for more connectivity within our landscape. Many authors, in many different settings, have shown that fragmentation reduces biodiversity. Brudvig et al [2009] showed that plants which depend on animals for dispersal increased when corridors connected experimental patches of regenerating pine forest. Thiele et al. [2018] studied plant species richness in agricultural, fragmented landscapes and concluded that connectivity increased the richness of grassland and wetland plants. Hannay et al. [1997] used the amount of forested area around a woodland to quantify isolation and showed that richness decreased as the amount of surrounding woodland decreased. Petit et al., [2004] showed that fragmentation constrains the dispersal of ancient woodland species in British lowland woodlands. Dzwonko & Loster [1988] showed that woods that had been isolated for longer in the agricultural landscapes in the Carpathian foothills had lower species richness.

HETEROGENEITY

Habitat heterogeneity is known to be important for biodiversity [Gardener, 2010]. It has been shown to increase richness when it takes the form of woodland management, [Boch et al., 2013, Schmidt, 2005] rides, number of soil types, the number of different habitats the woodland was exposed to [Honnay et al., 1999] and windthrow [Smart et al., 2014].

In this work I studied the drivers of vascular plant species richness in two ways. Firstly, using a range of variables to quantify abiotic factors and heterogeneity, I studied the effect of these variables on the species richness of a fixed area within 103 different woodlands.

Habitat heterogeneity is not trivial to quantify and has been expressed in many ways. Hannay et al., [1999] used length of rides, number of soil types and a ratio of perimeter and area. Schmidt [2005] used the presence of logging trails, and Baldi & Sadler [2008] used the number of CORINE land cover codes. I therefore explored different methods of expressing this factor, both using previous standard procedures and a new approach called zeta diversity [Hui & McGeogh, 2014]. Zeta diversity is the average number of species shared between different plots and is therefore a measure of the difference in species assemblages.

Secondly, I looked at the scale dependence of species richness as expressed in the canonical species area curve, (SAC), $S = cA^z$, and the effect of habitat heterogeneity and environmental conditions on the exponent, z. Some authors contend that the value of the exponent has no biological meaning [Tjorve & Tjorve, 2017]. Others suggest that the shape of the curve can be attributed to macro-ecological factors such as the balance between immigration and extinction [Li, 2002], community structure [Sugihara, 1979] and speciation [Hubbell, 1997].

Some authors consider that knowledge of z is informative as it relates to beta diversity and can therefore provide information about the area over which the species area curve is constructed, [Storch et al., 2014]. In this work I examined the combined influence of local abiotic factors and woodland heterogeneity on the exponent of curves constructed over two different scales. This is similar to the work of Shen et al [2009] and Baldi, [2008]. Shen considered the effect of tree density, soil nutrients, pH, slope, elevation and aspect and heterogeneity and concluded that dispersal and habitat heterogeneity jointly are required to explain the shape of species area curves. Baldi concluded that habitat heterogeneity, as quantified by the number of habitat types as shown through CORINE land cover maps, was more important than area for expressing species richness.

Material and Methods

DATA

The data were taken from the Centre for Ecology and Hydrology's woodland survey database [Smart et al., 2013]. 103 woodlands across the UK, but excluding Northern Ireland, were surveyed between 2000 and 2003. See figure 1 for locations. Sixteen randomly placed, nested plots were located in each woodland, see figure 2.

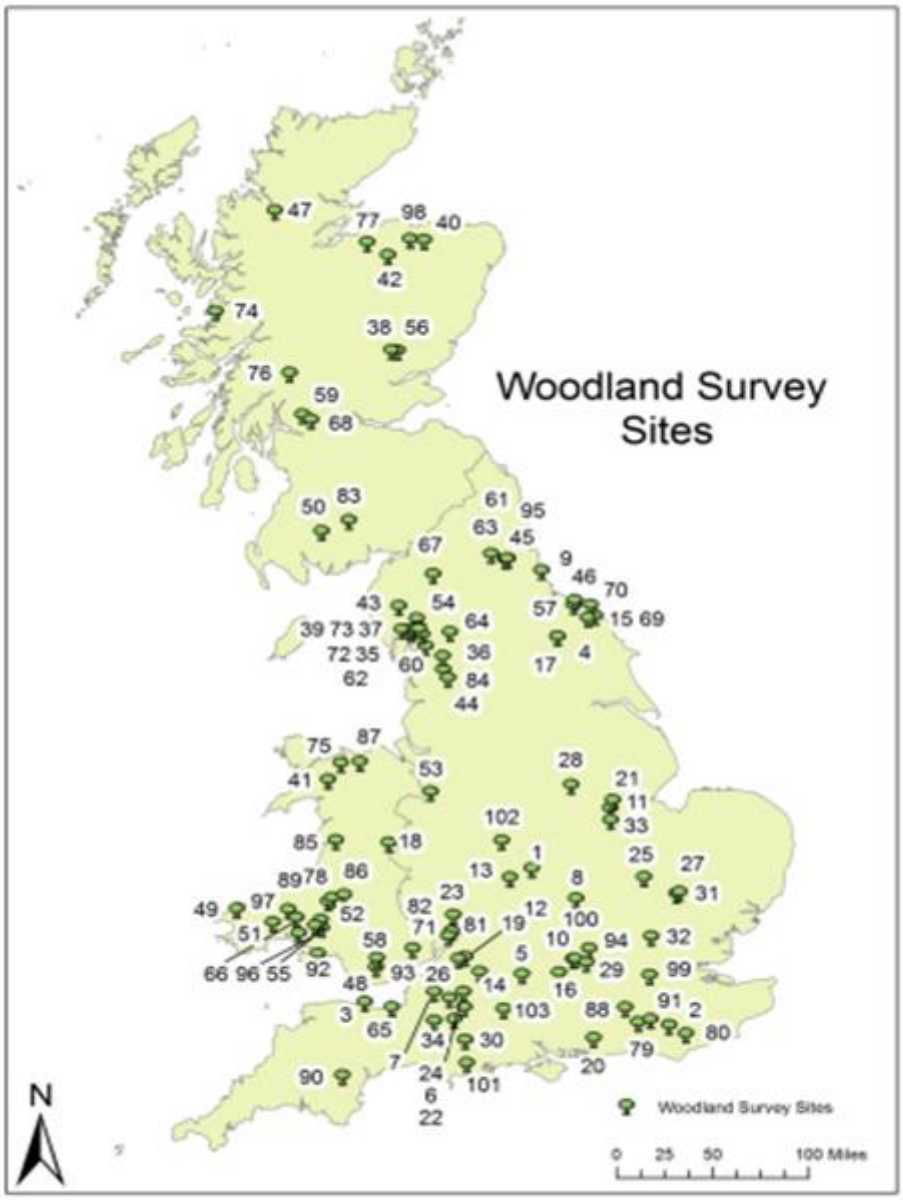

Figure 1. Locations of sites in the CEH woodland survey dataset, [Smart et al., 2003]

All flora was recorded within each plot. Additional plot data were collected, including diameter at breast height of trees, live basal area of trees and shrubs, soil pH, soil organic matter content, National

Vegetation Classification (NVC) code, and major soil group. Additional woodland information was also recorded, such as whether there were signs of woodland management, the presence of riparian zones, and open areas such as glades or rides.

Table 1 summarises the information taken from the dataset and details how it was transformed to create predictor variables. These variables were selected as previous authors have shown that they affect species richness.

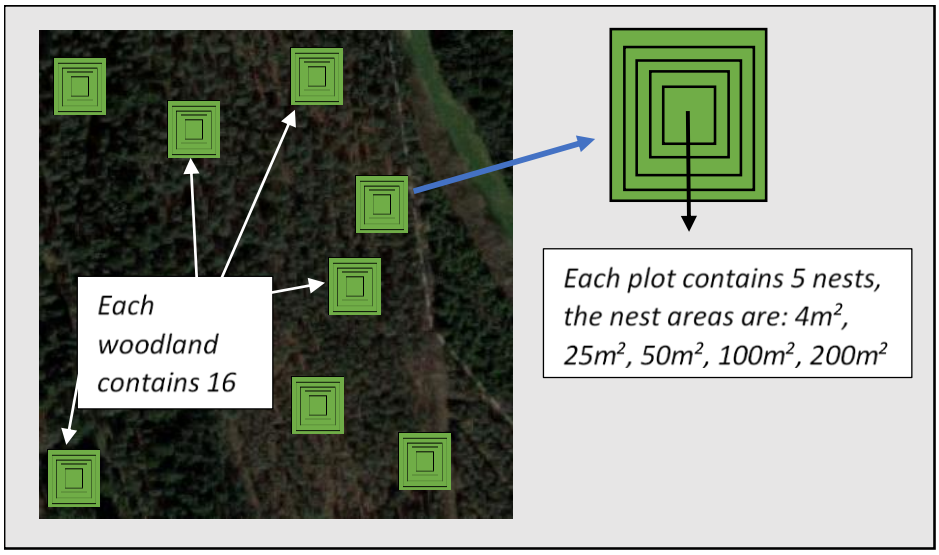

Figure 2. 16 plots were randomly placed withing each wood. Each plot contained 5 nests with area increasing from $4m^2$ to $200m^2$

Table 1 Summary of predictors created from the variables in the dataset

Variable	Description
Positive Heterogeneity Index (PHI)	A count of unique codes recorded for each woodland that have been shown by previous authors to be positively correlated with vascular plant species richness. The greater the PHI, the more elements within the woodland that could increase richness.
Northing	Latitude
Buffer	The proportion of land around the wood (to 3500m from the edge) that is not arable or urban. The larger the buffer, the more the surrounding area can be considered as positive for species richness. This variable therefore equates to the inverse of isolation.
Number of major soil groups (MSG)	The number of major soil group codes in each woodland.
Number of NVC codes (NVC)	The number of different NVC codes in each woodland. The NVC code represents the species assemblage. The number of different NVC codes therefore describes the number of different plant assemblages. This is equivalent to the number of CORINE land cover codes used by Baldi [2008]. The number of codes does not necessarily correlate with richness since it is possible that each code could represent a low richness assemblage. But since NVC codes represent a different species assemblage with different abiotic factors, if there are more codes it is more likely that the woodland has greater heterogeneity.
pH	The pH for the plot
Soil Organic Matter content (SOM)	The soil organic matter content for a plot
Live Basal Area (LBA)	The live basal area of all saplings, trees and shrubs recorded in a plot
Diameter at breast height (DBH)	The diameter at breast height of all trees recorded in each plot
Tree Density (TD)	Count of trees per plot
Area ratio	The ratio of area to perimeter

The richness for each woodland was calculated to be the unique number of species in the total 16 plots surveyed. The variables within the data base were at different scales, plot level and woodland level. Plot level variables, such as pH, DBH or SOM, were scaled appropriately to the richness. This was achieved by taking means and standard deviations. The means represent a summary of the entire woodland conditions and the standard deviations represent heterogeneity within the woodland. The variables are then of two types, those that represent average site conditions and those that represent heterogeneity and are summarised in Table 2.

Table 2. Summary of average site variables and variables which represent heterogeneity

Average site variables	**Heterogeneity variables**
Mean pH	Standard deviation pH
Mean SOM	Standard deviation SOM
Mean DBH	Standard deviation DBH
Mean LBA	Standard deviation LBA
Mean TD	Standard deviation TD
Buffer	Number NVC codes
Area ratio	Number MSG
PHI	

MODELLING

The modelling was designed to answer two main types of question. First, which variables had most influence on species richness? Second, which variables had most influence on the exponent of the species area curve?

SPECIES RICHNESS MODELS

Two methodologies are ideal for examining which variables most affect richness; model averaging [Symonds & Moussalli, 2010] and decision trees [De'ath & Fabricius, 2000]. Using these two different approaches allows different relationships to be found. Model averaging will detect a linear response, whereas decision trees can find non-linear relationships. [James et al., 2013]

MODEL AVERAGING

Model-averaging uses Akaike Information Criteria (AICc) to rank a set of competing models. The AICc is a measure of the goodness of fit and parsimony of the model [Burnham & Anderson, 2010]. Rather than a single model, a set of models are accepted [Grueber et al., 2011] and parameters averaged across the set.

Since highly correlated predictor variables are undesirable [Freckleton, 2011] the correlation coefficients of all variables were examined; where standard deviations and means had been taken of the same variable, these were highly correlated. The data were therefore split into two different

datasets, as detailed in Table 3, such that all variables within each dataset had Spearman and Pearson correlation coefficients below 0.54. Modelling was then carried out on both datasets. In addition, the variance inflation factors of the models were examined and were all below 2.

Table 3 Variables were split into two data sets to avoid correlations.

Mean data set	**Stand deviation data set**
Number MSG	Number MSG
Mean DBH	Sd mean DBH
Northing	Northing
Buffer	Buffer
Mean SOM	Sd SOM
PHI	PHI
Area ratio	Area ratio
Mean LBA	Sd LBA
Mean pH	Sd pH

Burnham & Anderson [2010] advocated the selection of predictor variables based on prior knowledge to avoid over-parameterization. I followed that recommendation.

Once the predictor variables were selected the dataset was standardized as suggested by Gelman, [2008]. This gives parameter effects sizes that can be compared directly; two standard deviations of change in the predictor variable will result in the effect size change in the response. A linear model was created using all the variables and the dredge function in the MuMin package in R [Barton, 2018] used to create a set of models using all permutations of the variable set. These are ranked according to the AICc. The model with the lowest AICc is considered the model that best represents the data. This method will always produce a best model, but that model is only best relative to all the other models generated. It is therefore important to ascertain the goodness of fit of the initial linear model used [Symonds & Moussalli, 2010]. The goodness of fit of the model using all the parameters was ascertained by examining the fitted values against the residuals, which showed no pattern, and by considering the value of R^2.

All models within an AICc of between two to seven can be considered as equally good [Burnham et al., 2010] and used to create a top-model set. The parameter effect sizes are then averaged across this model set and a 95% confidence interval obtained. If this confidence interval does not cross zero, these predictors are considered significant, [Kimberley et al., 2014]. In this work models within AICc of 2 ($\delta < 2$) of the highest ranked model were initially selected. The effect of increasing this value was then examined. If more variables were introduced which had confidence intervals that crossed zero, the lower value was used to avoid including redundant models [Grueber et al., 2011]. The relative importance of each effect size was also obtained, [Burnham, 2015]. Relative importance is a measure

of the probability that a variable is present in the model set. A relative importance of 1 means that a variable appears in every model in the model set.

DECISION TREES

Decision trees use recursive binary splits to divide the data into regions such that the root mean square of the response variable is minimized, [James et al., 2013]. (See supplementary material for more details). Decision trees are suited to analysing complex ecological datasets because they can fit non-linear relationships and are not sensitive to scale, [De'Ath & Fabricius, 2000; Elith et al., 2008]. They have been shown to be successful in separating ancient woodland indicators from other woodland plants, [Kimberley et al., 2013], using environmental variables to explain species richness of vertebrate groups [Mouchet et al., 2015] and to predict sites likely to contain invasive species, [Cutler et al., 2007]. In this analysis I used two tree models; random forest (RF) and gradient boosted machines (GBM) using the randomForest, [Liaw & Weiner, 2002] and gbm, [Ridgeway, 2017] packages in R.

To extract the variables which most influence species richness two metrics were used. The randomForest package provides variable importance, (VI). [Friedman & Meulman, 2003; Cutler et al., 2007]. VI is calculated after the model is generated. Each variable is randomized in turn, the model is used for prediction and the percentage increase in mean square error (MSE) is calculated. If there is large increase in MSE, then that variable can be considered as important to the model. A rank of the percent increase in MSE is used to rank the variables. The variable with the highest increase in MSE has the highest VI and is most important for predicting the response. Gradient boosted machines use relative influence, (RI), [Natekin & Knoll, 2013]. The percentage decrease in MSE is averaged for each variable over each time a variable was chosen to make a binary split in the data. The variable which has greatest RI is best at splitting the data and predicting the response.

Tree based methods sometimes do not give stable results for VI and RI, [Nicodemus et al., 2007]. Therefore, each model was repeated 100 times and the variables ranked by the average values of VI and RI. The value of VI and RI is relative and therefore subjective; in this work I selected variables which had RI and VI above 10%

USING ZETA DIVERSITY TO CREATE A METRIC FOR HETEROGENEITY

Habitat heterogeneity is important for species richness, but it can be expressed in different ways. In this work I used several methods to quantify heterogeneity: PHI, number of NVC codes, and standard

deviations of abiotic variables. In addition, I will now discuss how I used zeta diversity to create a new diversity metric.

Zeta diversity is the average number of species shared between multiple plots, [Hui & McGeoch, 2014] and can be used to show compositional change in species assemblages, [McGeogh et al., 2017] and model drivers of species turnover, [Latombe et al., 2017].

ζ_1 is the average number of species in each plot, ζ_2 is the average number of species shared between every combination of two plots, (see supplementary material for details). If plots contain the same species, then $\zeta_1 - \zeta_2 = 0$. If all plots contain different species then $\zeta_1 - \zeta_2 = \zeta_1$.

A homogeneous woodland is more likely to contain species that occur in many plots, whereas a heterogeneous woodland is more likely to have plots containing different species. This suggests a method of comparing heterogeneity between woodlands using the following ratio,

$$\zeta_r = \frac{\zeta_1 - \zeta_2}{\zeta_1}, \qquad 1$$

The more homogeneous the woodland, the closer the value of ζ_r to zero.

Calculation of ζ_2 uses a pairwise average, therefore it is comparable to the use of multisite dissimilarity metric to measure heterogeneity, [Baselga et al., 2007; Baselga, 2012].

The zeta values for each woodland were calculated using the zetadiv function in R, [Latombe et al., 2018]. The ζ_r coefficient was calculated and added to the mean and standard deviation datasets.

SCALE DEPENDENCE OF SPECIES RICHNESS

This part of the work involved calculating species area curves across the nests from $4m^2$ to $200m^2$ and accumulating species across the plots from $200m^2$ to $3200m^2$, as shown in Figure .

At the smaller scale, linear mixed effects models were used with a log/log transform to the area and species richness, with plot as the random variable. The gradient of these models was then extracted.

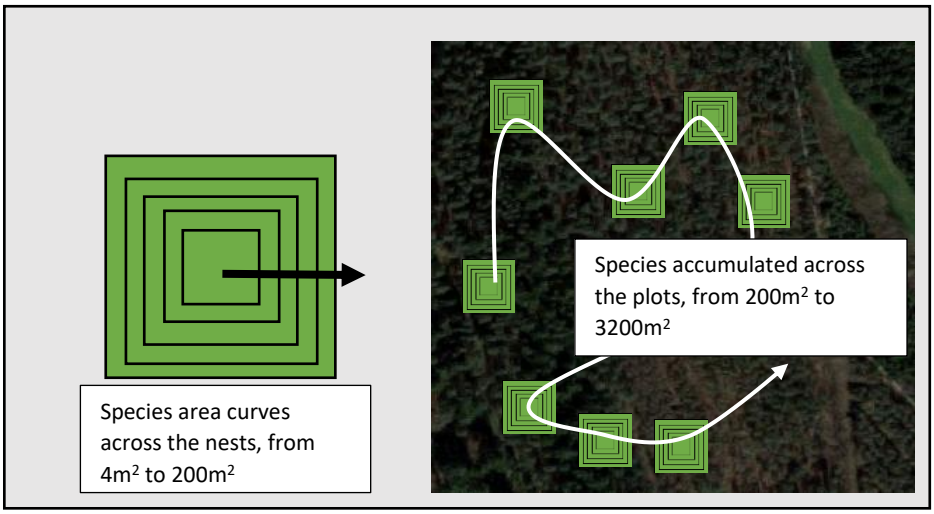

Figure 3. Two scales of species area dependence. At the nest scale, species area curves were calculated for areas from 4m² to 200m². At the plot scale species were accumulated across the 16 plots from 200m² to 3200m²

At the larger scale a spatially explicit method was used to construct species accumulation curves. [Scheiner, 2003]. This involved calculating the maximum and minimum slopes of all curves and taking their average. The slope for the average curve was extracted as above.

The z values at the two scales were then modelled with the same variables as used to model species richness using model averaging.

Results

SPECIES RICHNESS

Figure 4 shows the model-averaged effect sizes and confidence intervals for mean and standard deviation datasets, R^2 for the model are shown in table 4. Figures 5 and 6 show the results from the decision tree algorithms. All models are compared in table 5.

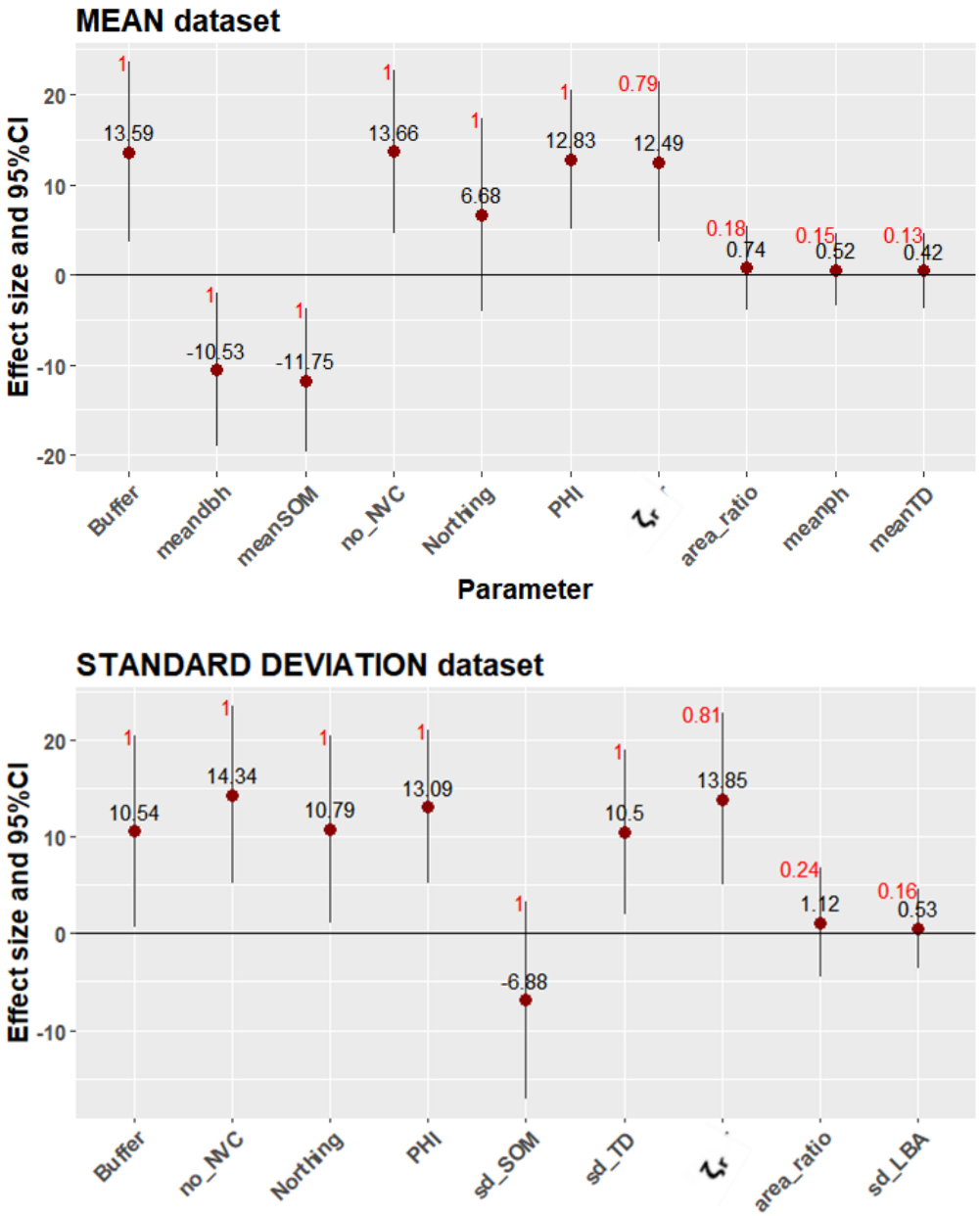

Figure 4. Model averaged effect sizes for woodland richness modelled with all variables using δ < 2 for the top model set. Variable importance is in red. The points show the average effect size and the bars show the 95% confidence intervals. Buffer, number NVC codes, PHI, ζr, and standard deviation of tree density are significant and important with a positive effect on richness. Mean SOM and mean DBH are significant and important with a negative effect. ζr is the least important effect, occurring in 80% of the models, all other effects occur in all models in the top model set

Table 4. R^2 values for model-averaged parameters when used for prediction from both mean and standard deviation datasets. The full model values are those for the linear model which includes all parameters. The model-averaged values are just those parameters which were contained in the average model set

Dataset	Full model R^2	Full model Adj R^2	Model Averaged R^2	Model Averaged Adj R^2
Mean	0.47	0.4	0.46	0.46
Standard deviation	0.45	0.38	0.44	0.44

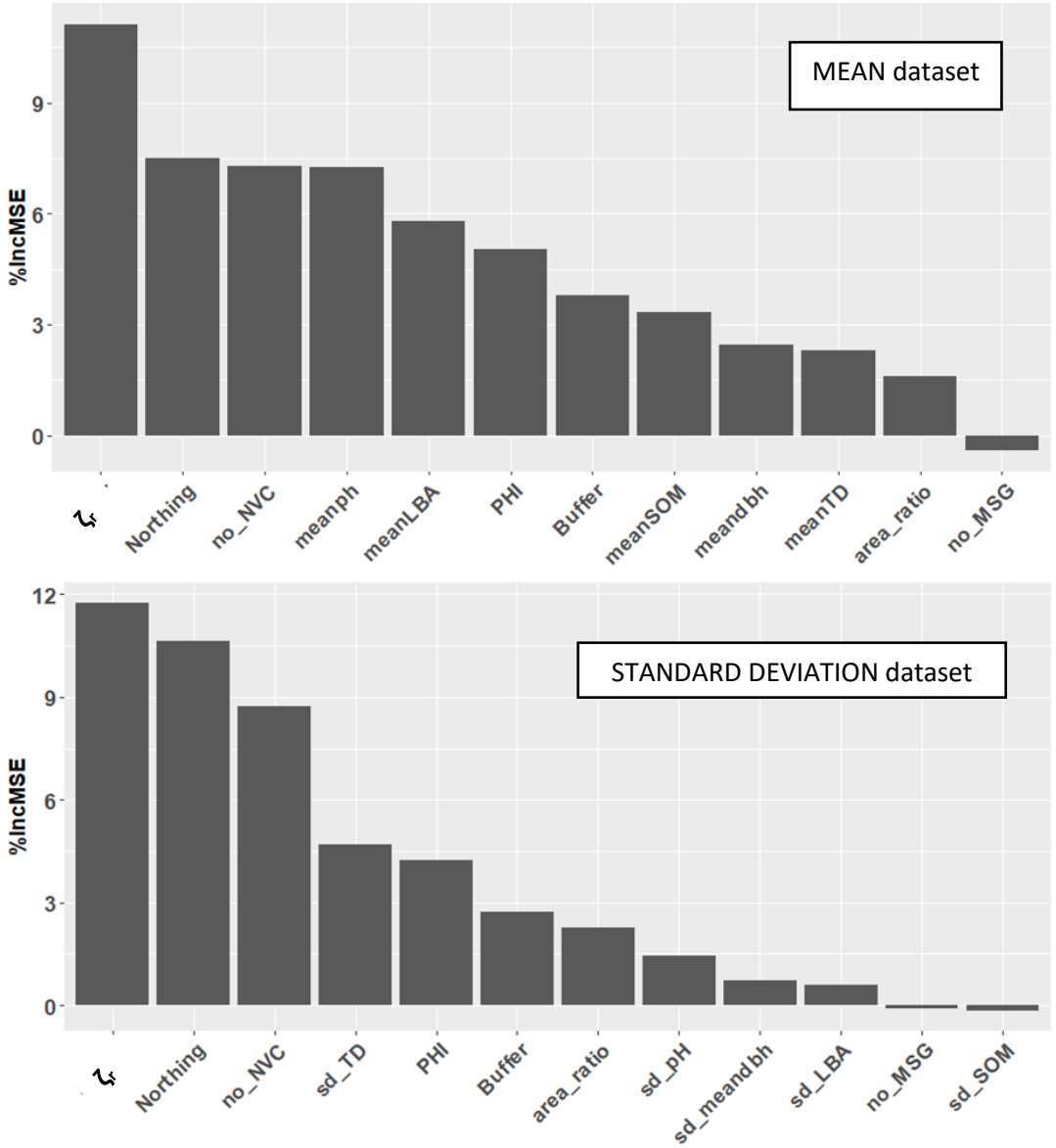

Figure 5. %IncMSe for the random forest algorithm. ζr and Northing are the most important variables with %Inc MSE >10

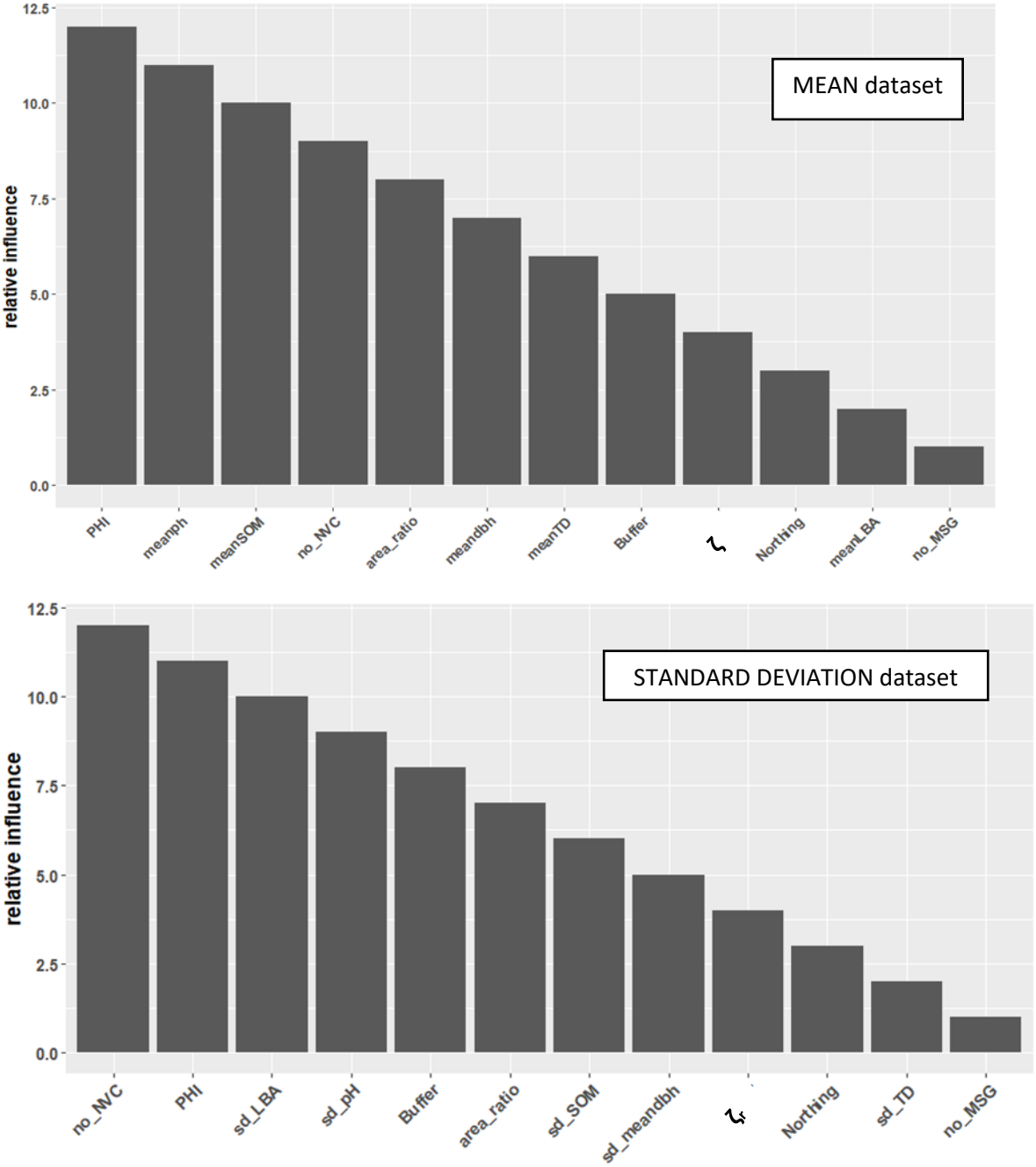

Figure 6. Relative influence for the gradient boosted machine. PHI, meanpH, and number of NVC codes are the most important variables with relative influence >10

Table 5. Summary of variables shown to affect richness from the three different models. The root-mean-squared error (rmse) was similar for all the models; the lowest was 18 for model averaging. Both decision tree methods had lower rmse for the training set and were therefore over-fitted, however the rmse of the test set was comparable to the average model.

Variable	Random Forest	GBM	Model Averaging
rmse	Test set 21	Test set 21	18
rmse train	12	10	NA
ζ_r	Y		Y
Northing	Y		Y
Number NVC		Y	Y
Sd TD			Y
PHI		Y	Y
Buffer			Y
Mean DBH			Y (–ve)
Mean SOM			Y (–ve)
Mean pH		Y	

Table 5 shows that, compared to model-averaging, the decision trees selected fewer variables and they did not add any new variables apart from mean pH. The decision trees also had greater rmse, although the rmse for the test set was similar to that of the average model. The results for the decision trees were not consistent on each of the 100 runs, although the results were averaged. This suggests that the decision trees were not successfully splitting the data and may not be appropriate for this dataset. However, it is interesting that mean pH was an important variable in the GBM. Since the response of richness to pH was expected to be unimodal, it was unlikely that the average model would select this variable. The GBM has therefore been shown to be useful in detecting this non-linear response.

Figure 7 shows the distributions of the nest and plot *z* values. Figures 8 and 9 give the average model effect sizes. Table 6 gives the R^2 for the full and average models.

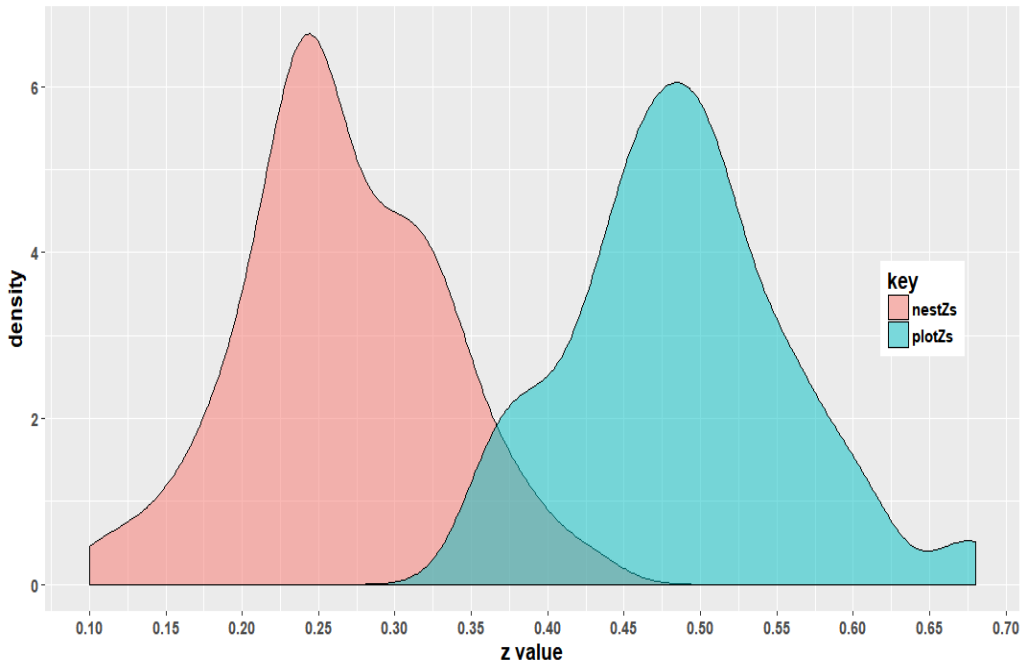

Figure 7. Distribution of z values at two scales. The pink distribution shows nest scale – 4m² to 200m ². These were obtained from a linear mixed effects model of log richness with log area with plot as random and are distributed around 0.25. The blue distribution shows plot scale – 200m² to 3200m². They were obtained by accumulating species using a spatially explicit model then using a linear model of log richness with log area. At the larger scale the values are distributed around 0.5.

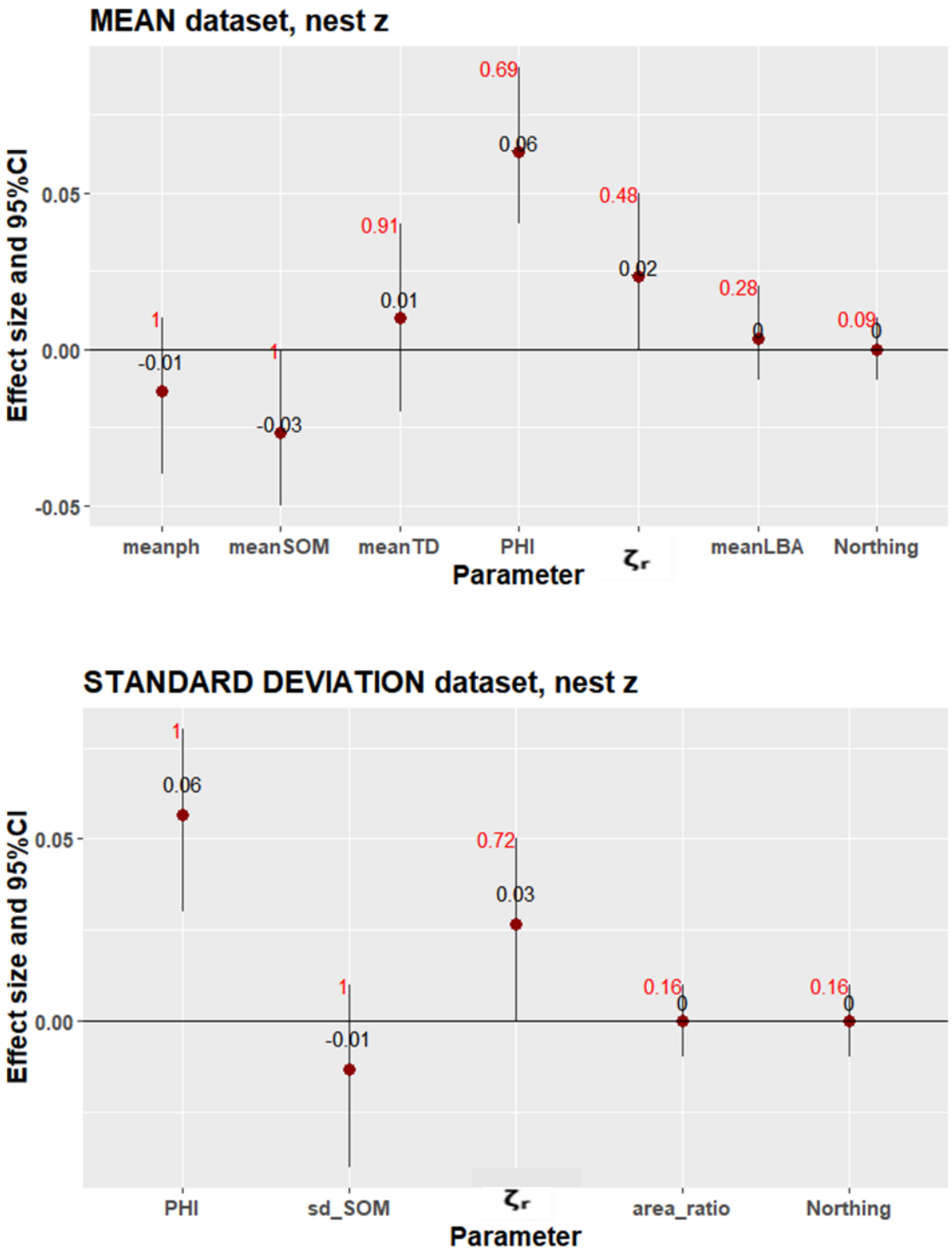

Figure 8. Model averaged effect sizes for nest z modelled with all variables using δ < 1.5 for the top model set. Variable importance is in red. The points show the average effect size and the bars show the 95% confidence intervals. PHI and ζr have a positive effect on nest z while meanSOM has a negative effect.

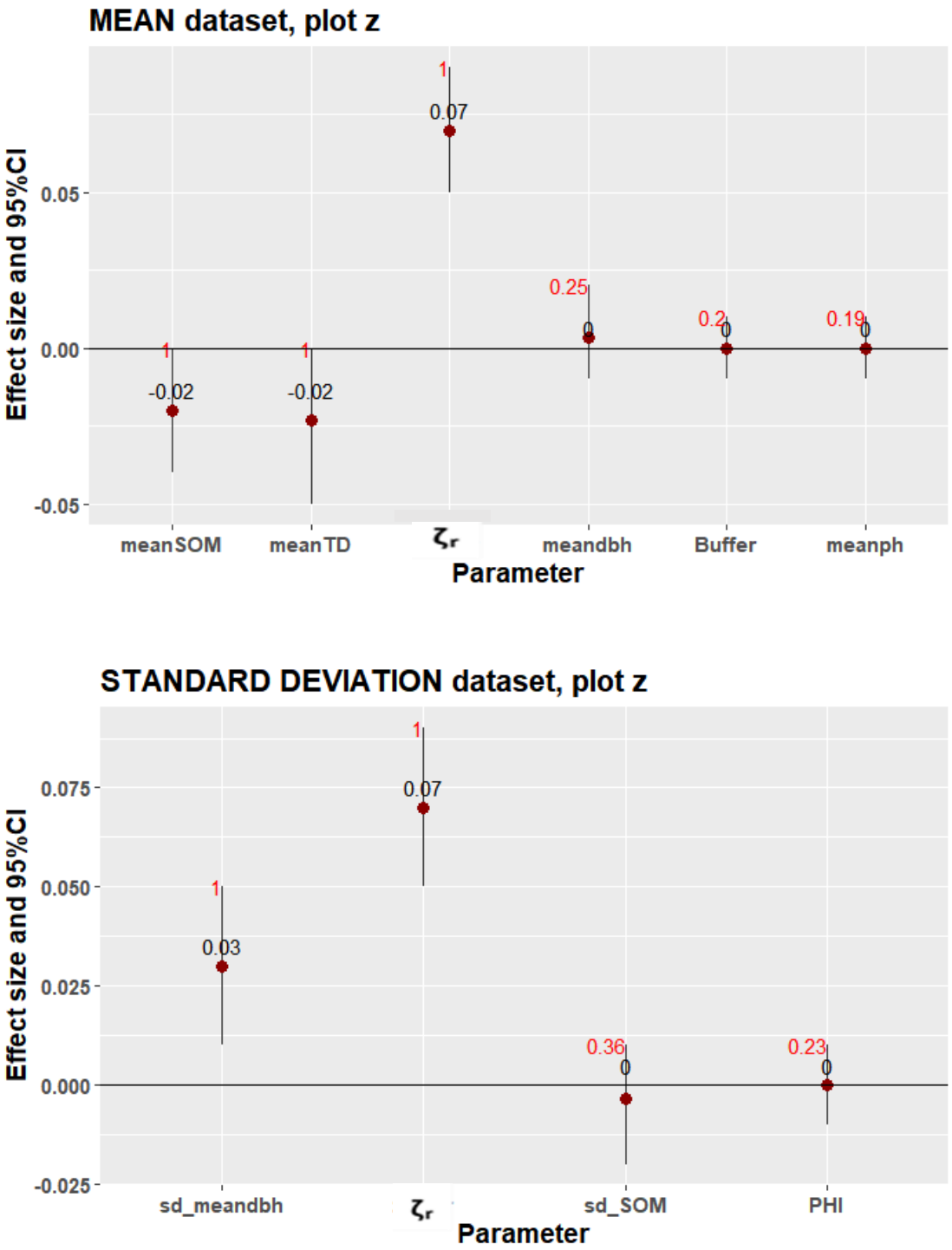

Figure 9. Model averaged effect sizes for plot z modelled with all the variables using δ < 1.5 for the top model set. Variable importance is in red. The points show the average effect size and the bars show the 95% confidence intervals. ζr and sd_meanDBH have a positive effect and mean SOM and mean TD have a negative effect.

Table 6. R^2 values for average model nest z and plots z parameters when used for prediction. Only around 30% of the variance of nest z is explained by the parameters. For plot z around 45% of the variance is explained.

Z values	Dataset	Full model R^2	Full model Adj R^2	Model Averaged R^2	Model Averaged Adj R^2
Nest zs	Mean	0.33	0.24	0.31	0.3
	Standard deviation	0.27	0.17	0.26	0.25
Plot zs	Mean	0.47	0.4	0.46	0.44
	Standard deviation	0.46	0.39	0.45	0.44

Discussion

VARIABLES AFFECTING SPECIES RICHNESS

HETEROGENEITY

Model-averaging and decision trees showed nine variables as affecting richness. Four of these represent heterogeneity: ζr, number of NVC codes, PHI and standard deviation of tree density. Of these, number of NVC codes, PHI and standard deviation of tree density are present in every model in the top model set and ζr is present in 80% of the models. In the decision trees, ζr was the most important variable.

The greater PHI, the more evidence there is of woodland management, open areas and riparian zones. This implies that all these features are required to maximise richness. The number of NVC codes and ζr quantify the number of different species assemblages. High PHI and different species assemblages are more likely to occur if environmental conditions differ across the woodland – for example, different levels of soil moisture or different light levels. The use of woodland management techniques such as coppicing to increase richness is well known, [Fuller & Warren, 1993, Gardener; 2010], but this is only one aspect of heterogeneity. Although light-demanding species increase richness, woodland specialist and shade-bearing species often favour a lack of disturbance, [Kimberley et al., 2013; Honnay et al., 1999; Schmidt, 2005], and therefore management should also incorporate assessing areas that should not be worked.

Plant diversity is also only one aspect of biodiversity within woodlands, which includes insects, bryophytes and fungi, all of which can benefit from undisturbed areas [Paillett et al., 2009]. Other measures could also be considered, where suitable, to increase riparian zones, such as reducing drainage, [Natural England and RSPB, 2014].

USE OF ζ_r AS HETEROGENEITY METRIC

Results presented here suggest that ζr is a useful way to quantify heterogeneity. In addition, ζr has other positive benefits over using number of NVC codes or PHI. Correct NVC classification is not always straightforward and PHI requires additional work surveying the site and could be subjective or fail to include important features. In contrast ζ_r is easy to calculate from species lists, thus requiring no additional effort in the field.

MEAN-DBH AND MEAN-LBA

Model-averaging shows meanDBH to have a negative effect on richness. This is to be expected, as an increase in meanDBH implies succession in the woodland [Packham et al., 2001; Smart et al., 2014] which leads to increased homogeneity [Keith et al., 2009], again indicating the importance of woodland management schemes.

BUFFER

Buffer size had a significant positive effect on richness in both datasets and occurred in every model in the top model set. The size of the buffer relates to the amount of the woodland that is exposed to natural habitats and therefore has two potential effects. Firstly, the larger the buffer the more connected the woodland; habitat fragmentation is known to have a negative impact on biodiversity, [Quine, 2011, Watts et al 2008]. Secondly, the larger the buffer, the more protected the woodland from agrochemicals and airborne environmental pollutants – particularly around the edges [Draaijers, et al., 1988] – which are linked to reduced species richness [Bobbink et al., 1998; Pallet et al., 2015, Simkin et al., 2016; Stevens et al., 2015]. Previous research has suggested that buffer regions should be included around woodlands for various reasons, including protection from anthropogenic influences [McWilliam et al., 2010] and vehicle pollution [Bignal et al., 2008]. It is not possible within this dataset to ascertain the mechanism for the increase in richness with buffer, but this work agrees with previous authors; larger more connected woodlands tend to have greater plant species richness.

NORTHING

Northing is selected as an important variable in the RF algorithm and model-averaging shows it be significant in the standard deviation dataset, with a positive effect in richness. This is in contradiction to previous authors, [Gilman et al., 2014; Ohlemuller & Wilson, 2000]. A bivariate plot of Northing with meanTD (figure 8) shows a negative correlation, suggesting that the increased richness could be due to the addition of light-demanding species.

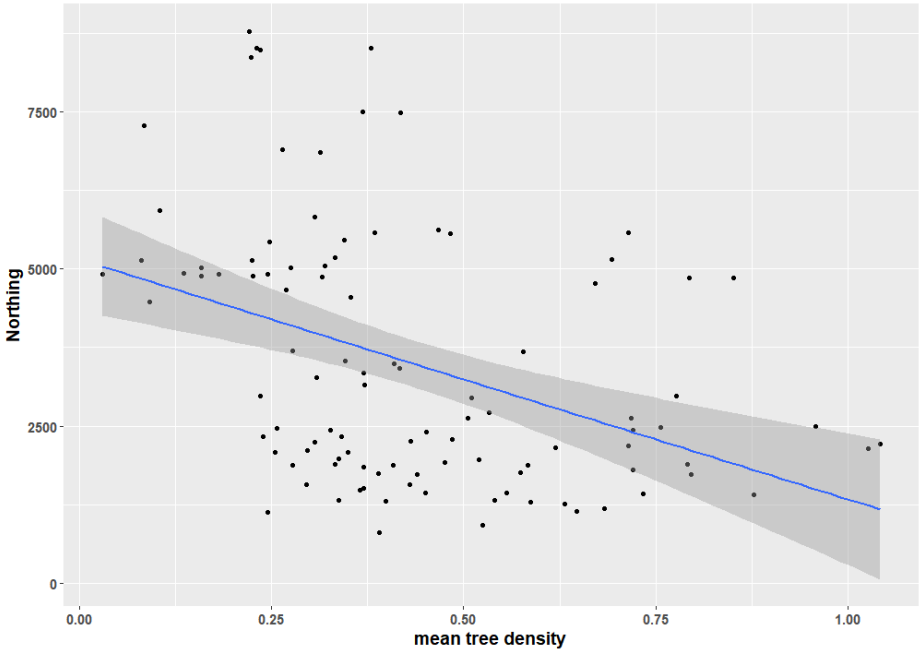

Figure 8. Scatter plot of northing with mean tree density suggests that the northern woodlands are more open. This could allow the inclusion of more light-loving species assemblages and hence add to species richness.

The Scottish woodlands might also consist of plant communities which tend to be richer, but figure 9 shows that this is not the case. The communities are very similar in the north and south. However, for the same community, richness tends to be greater in Scotland, (figure 10). This suggests that other variables, not availabe in this dataset, may be responsible for the increased richness in Scottish woodlands, or that a combination of positive predictors is more likely to co-occur in the more northern woodlands.

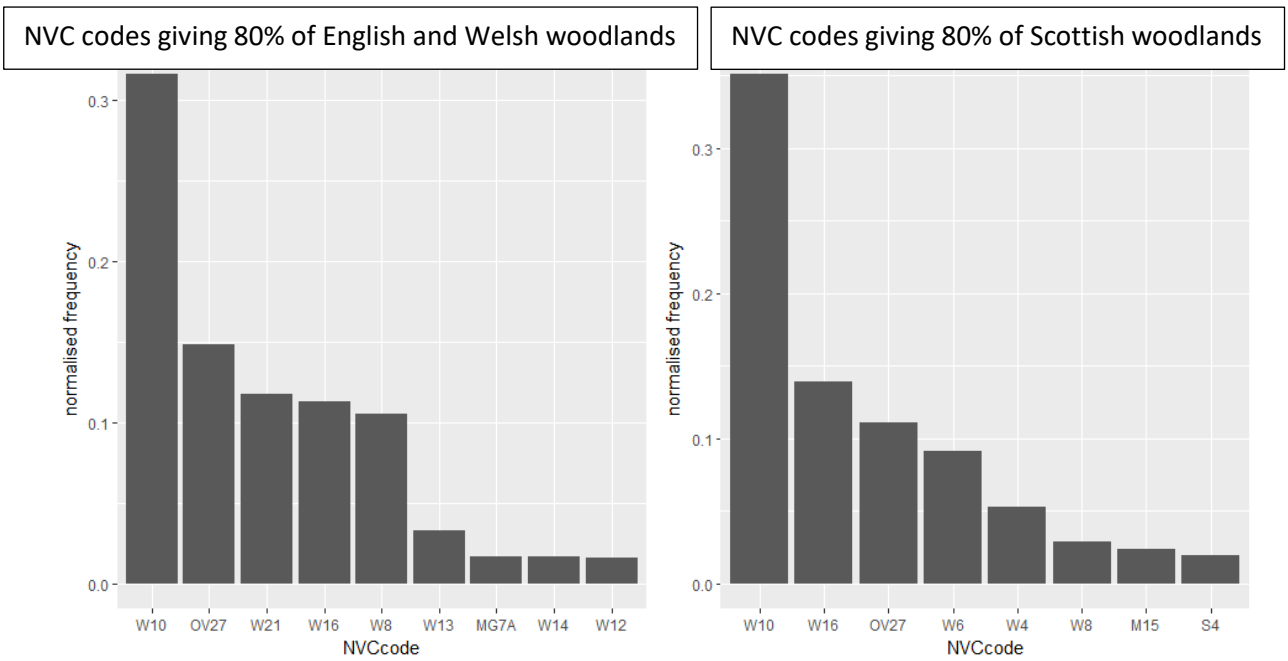

Figure 9. Distribution of NVC codes in Scotland and England and Wales. In both areas 30% of communities are W10, Quercus robur woodlands. These communities have one of the widest range of species richnesses of all NVC communities

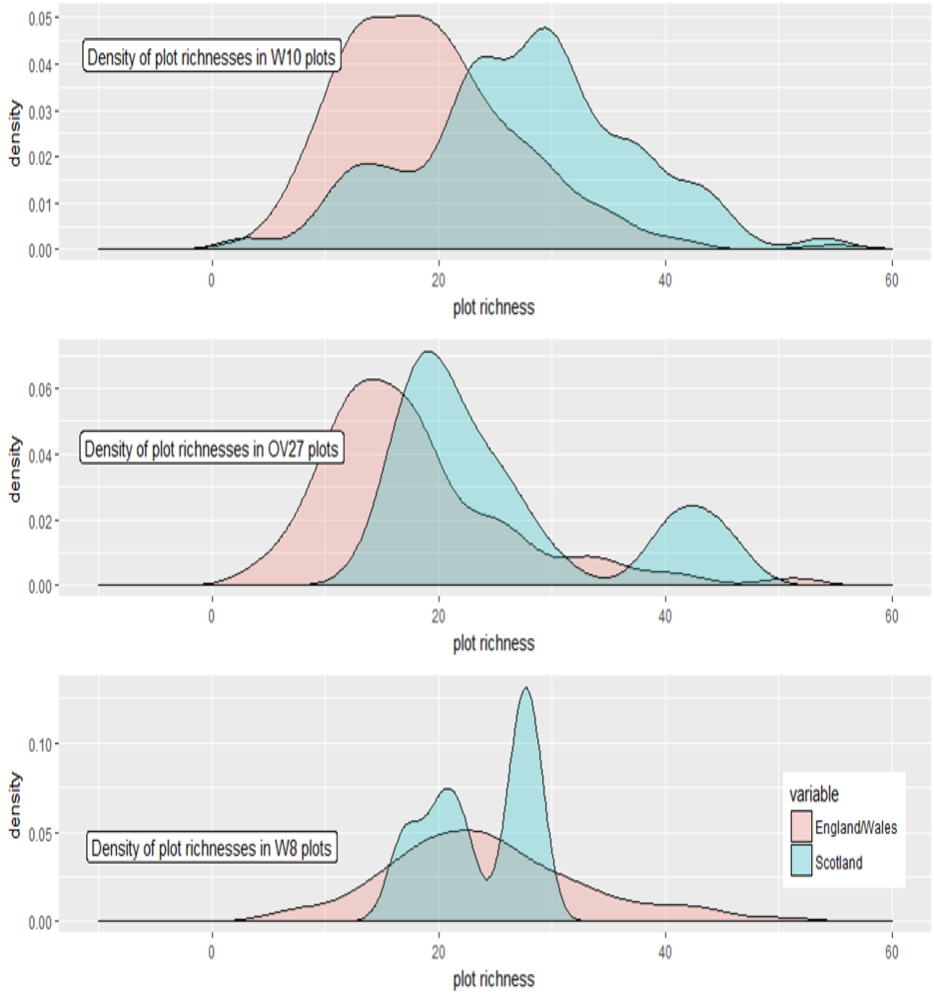

Figure 10. Density of plot richness for the most common NVC codes in Scotland and England and Wales. Given the same NVC community, richness tends to be greater in Scotland, particularly for the common W10 and OV27 communities.

SOIL ORGANIC MATTER

SOM content was shown to have a significant and important negative effect on species richness in agreement with Koojiman & Cammeraat [2010]. Organic soils have low pH and low pH soils in woodlands results in low species richness, [Cornwell & Grubb, 2003]. However, in this data, very few of the plots were located on organic soils. The six NVC codes in the Table 7 represent 82% of all plots surveyed.

Table 7 Soil types for the most common NVC codes in this data

NVC code	% of plots	Soil type taken from JNCC [refs]
W6	5%	Eutrophic moist soil with substantial deposition of mineral matter or on flood fen peat
W8	9.5%	Dry calcareous soils
W21	10%	Various
W16	11.5%	Very acidic, oligotrophic, pH<4, sandy podzolic
OV27	14%	Various
W10	32%	Base poor brown earths

Of these, only W6, 5% of the total, are potentially on peat, but W6 is a high nutrient, high richness community. It is therefore unlikely that the reduced richness with SOM is caused by reduced richness expected on organic soils.

The reduced richness could be due to low pH on other soil types since increased SOM has been shown to decrease pH, [Russell, 1960; Williams & McDonald, 1960]. On the other hand, adding SOM can also increase pH, depending on the initial acidity of the soil and the nature of the litter [Ritchie & Dolling, 1985]. This seems to be the case in this data, (figure 11).

Figure 12 shows that all the common NVC communities tend to show reduced richness as SOM increases, regardless of soil type. Given also that the NVC code represents a community with similar physical conditions, such as nutrient levels, soil pH and light levels – while the response to richness with increased SOM is seen across all communities – this finding implies that the variation in richness with SOM is unlikely to be due solely to one of those physical conditions, but perhaps to another factor not measured in this dataset.

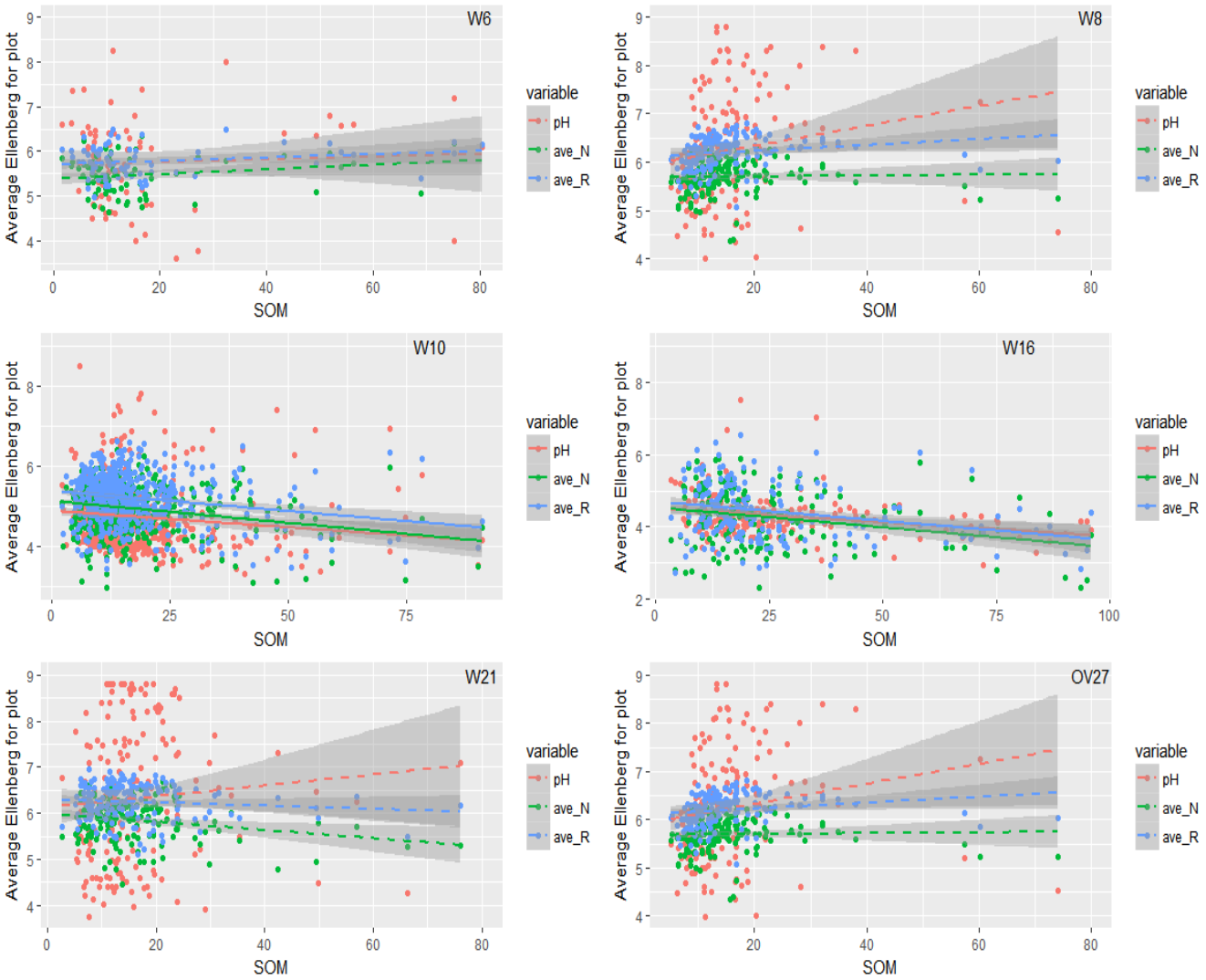

Figure 11. Variation in Ellenberg N and R and pH for plots within the six most common NVC communities found in this data. The Ellenberg values and pH decrease with increasing SOM in W10 and W16 communities which occur on acidic soils, but this is not seen in the remaining communities.

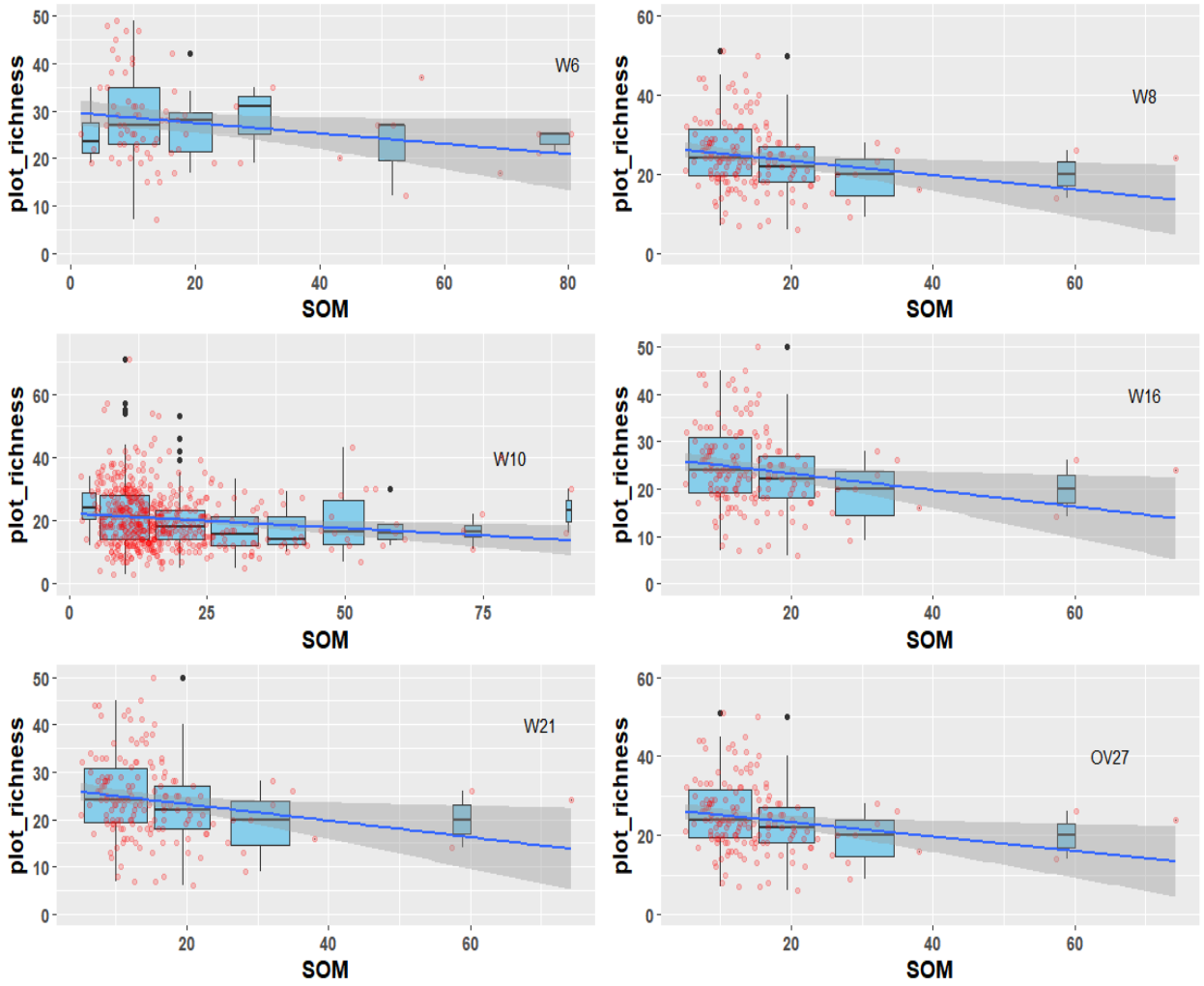

Figure 12. Plot richness with SOM for the six most common NVC communities in this data. The blue boxplots show the range in richness when grouped, the red dots are the raw ungrouped data. The variation in richness is high when SOM is low, but tends to be low for high values of SOM.

Increased nitrogen deposition can both increase SOM [Fang et al., 2014; Zak et al., 2016] and reduce richness [Bobbink et al., 1998; Pitcairn et al., 2002; Simkin et al., 2016; Stevens et al., 2016]. A smaller buffer could allow increased nitrogen deposition, [Draaijers et al., 1988; Kennedy & Pitman, 2004], which might be reflected in the Ellenberg N values [Maskell et al., 2010]. Therefore, if woodlands with smaller buffers showed both increased Ellenberg N values and increased SOM, nitrogen deposition might be the mechanism behind the change in SOM and the reduced richness. Figure 14 shows that while buffer is negatively correlated with Elllenberg N values, in support of this hypothesis, in contradiction, it is positively correlated with SOM content.

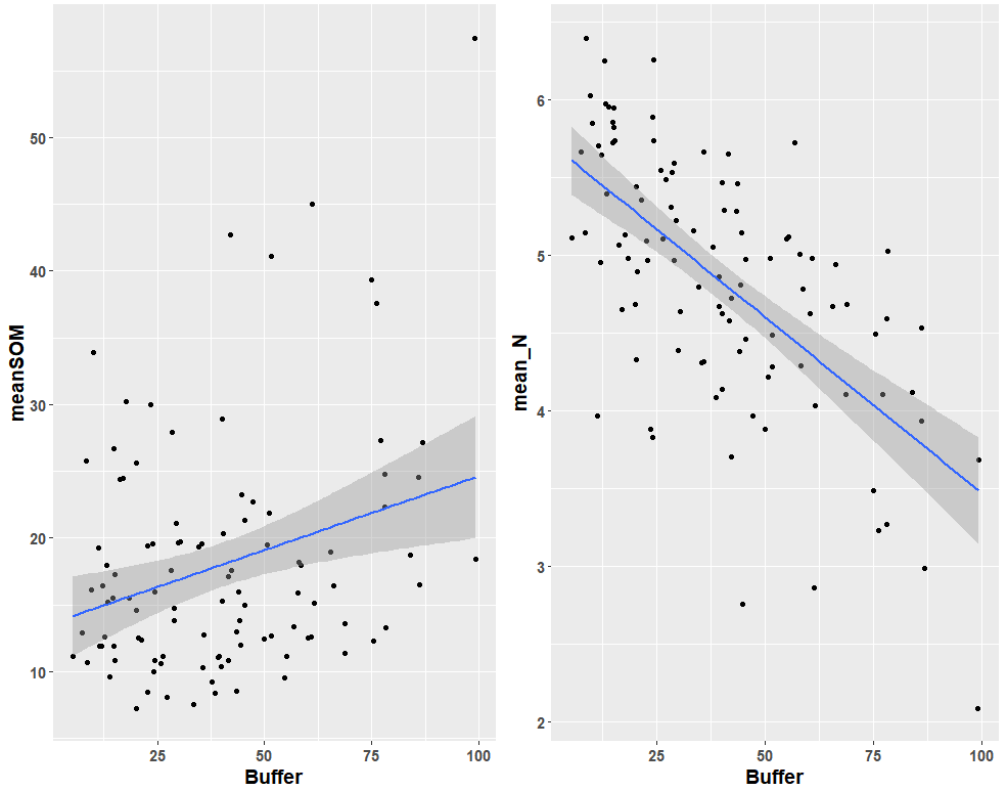

Figure 13. Change in mean SOM and Ellenberg N values with Buffer size. If nitrogen deposition was leading to increased SOM, then woodlands with smaller buffers would have greater SOM and larger Ellenberg N. The graph on the right shows that smaller buffers correlates with higher Ellenberg N, but not to greater mean SOM.

In summary, whilst SOM is negatively correlated with species richness, the causal mechanism is not evident.

SOIL pH

The GBM selected mean pH as important for predicting species richness. It was not expected that the average model would select the variable as the expected response of richness to pH was unimodal and the average model was linear. Although quadratic terms could be used in the linear model, this is not recommended [Symonds & Moussalli, 2010]. Figure 14 demonstrates that in this data plot richness does have a unimodal response to plot pH, in agreement with Cornwell & Grubb, [2003]

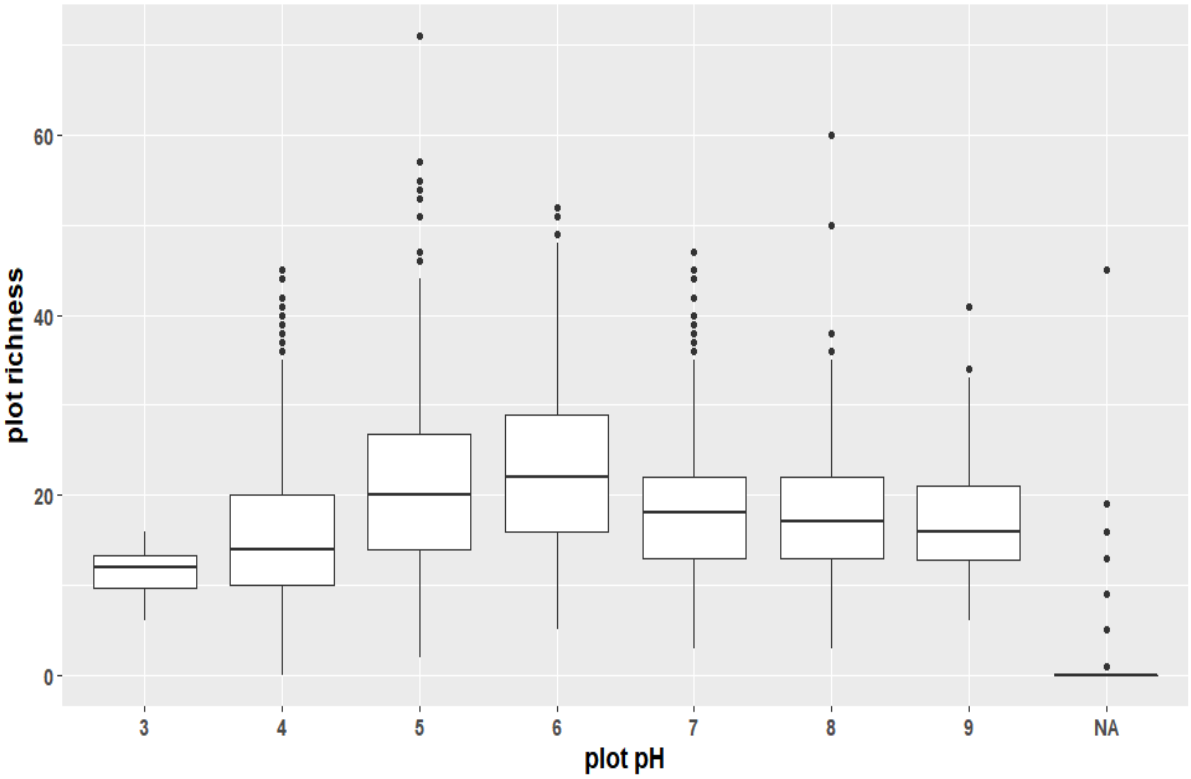

Figure 13. Plot richness response to pH can be seen to be unimodal with a peak around pH6.

NEST AND PLOT Z

The fitted *z* values were distributed around 0.25 and 0.5 for smaller and larger scales respectively, similar to those found by Crawley & Harral, [2001] and at the smaller scale reflect those suggested by Preston, [1962].

At smaller scales PHI and ζr, have a significant positive effect on *z* while SOM has a negative effect. The negative effect of SOM is due to its effect on species richness. If the species richness of a plot is reduced, fewer new species will be encountered as area is increased and therefore *z* will be reduced.

Neither dataset explained much variance in *z*: 31% and 26% for the mean and standard deviation datasets respectively.

At larger scales heterogeneity as quantified by ζr and sd-meanDBH have a significant positive effect on z. MeanSOM and meanTD have a significant negative effect, again, these variables are expected to reduce woodland species richness, and therefore reduce the value of *z*.

The average model parameters explain 46% and 45% of the variance in plot z for the mean and standard deviation datasets respectively.

At both scales heterogeneity has a positive effect on the value of z while SOM and meanTD, both factors which reduce species richness, have a negative effect.

VARIABLES OMITTED

When used for prediction, the average model parameters only explained 46% of the variance in richness, probably due to the exclusion of other important variables.

Woodland age and past land use have been shown to affect species composition [Hermy, 1994] and richness [Dzwonko & Loster, 1989; Peterken & Game 1984]. Ancient woodlands are more likely to contain long-lived, poorly dispersing perennials whose migration to new woodland sites is limited [Kimberley et al., 2015]. This potentially limits richness of new woodlands, particularly if they are isolated. Ancient cultivation has also been shown to induce a species composition gradient and thereby increase richness [Dupouey et al., 2002].

Plant available phosphorus has been shown to reduce species richness of native plants in alder forests [Hrivnak et al., 2015]. Phosphorus eutrophication favours competitive species particularly at higher light levels, and this can result in the exclusion of other species [Keersmaeker et al., 2004].

The amount of light reaching the woodland floor influences richness, with greater richness at higher light levels, [Kenedy & Pitman, 2004]. However, eutrophication together with high light levels encourages competitive species which tend to reduce richness, [Cornwell & Grubb, 2003, Keersmaeker et al., 2004]. Conversely, low pH in conjunction with low light levels can result in very little ground cover, [Cornwell & Grubb, 2003]

Rainfall varies greatly across the UK and has been shown to correlate positively with woody plant species richness [O'Brien, 1998; Richerson & Lum, 1980]. In western Scotland between 1981 and 2019 the mean monthly rainfall was above 50mm every month and reached a maximum of 140mm. In the southeast for the same period, the mean monthly rainfall was below 50mm for 10 out of 12 months with a maximum of only 60mm, [MetOffice, 2018]

SYNTHESIS

Habitat heterogeneity was found to be important for species richness in UK broadleaved woodlands. Heterogeneity in these models included coppicing, riparian zones and open areas, suggesting that all these factors need to be considered if woodland plant diversity is to be increased. Isolation was negatively correlated with species richness and therefore any possibility to increase the size and connectivity of new woodlands should be a priority. SOM was also negatively correlated with species richness, although the mechanism for this was unclear. Richness was shown to have a unimodal

response to soil pH with a peak around pH6. Diameter at breast height also had a negative effect on species richness, implying that steps to manage the succession of woodlands should be applied. The northern woodlands were found to be richer than those in the south, but again, the reason for this was not evident.

The use of zeta diversity to quantify heterogeneity was found to be a useful, easy to produce, objective metric and a significant predictor of richness.

At two scales heterogeneity was found to have positive effect on the exponent, *z*, of the species area curve; factors that reduce species richness were seen to reduce the value of *z*.

References

Amar, Arjun & M Hewson, C & M Thewlis, R & Smith, Ken & Fuller, Robert & Lindsell, Jeremy & Conway, Greg & Butler, Simon & A Macdonald, M. (2006). What's Happening to Our Woodland Birds?

Báldi, A. (2008) Habitat heterogeneity overrides the species–area relationship. *Journal of Biogeography*, 35 (4), 675–681.

Barton, K., (2018). MuMIn: Multi-Model Inference. R package version 1.42.1. https://CRAN.R-project.org/package=MuMIn

Baselga, A., Jimenez-Valverde, A., Niccolini, G. (2007) A multiple-site similarity measure independent of richness. *Biology Letters*, 3 (6), 642–645.

Baselga, A. (2013) Multiple site dissimilarity quantifies compositional heterogeneity among several sites, while average pairwise dissimilarity may be misleading. *Ecography*, 36 (2), 124–128.

Bignal, K.L., Ashmore, M.R., Headley, A.D. (2008) Effects of air pollution from road transport on growth and physiology of six transplanted bryophyte species. *Environmental Pollution*, 156 (2), 332–340.

Bobbink, R., Hornung, M. & Roelofs, J.G.M. (1998) The effects of air-borne nitrogen pollutants on species diversity in natural and semi-natural European vegetation. *Journal of Ecology*, 86 (5), 717–738.

Boch, S., Prati, D., Müller, J., Socher, S., et al. (2013) High plant species richness indicates management-related disturbances rather than the conservation status of forests. *Basic and Applied Ecology*, 14 (6), 496–505.

Brudvig, L.A., Damschen, E.I., Tewksbury, J.J., Haddad, N.M., et al. (2009) Landscape connectivity promotes plant biodiversity spillover into non-target habitats. *Proceedings of the National Academy of Sciences,* 106 (23), 9328–9332.

Burnham, K.P., Anderson, D.R., Model selection and multimodel inference: a practical information-theoretic approach. New York: Springer; 2010

Burnham, K.P., Anderson, D.R., Huyvaert, K.P. (2010) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behavioral Ecology and Sociobiology*, 65 (1), 23–35.

Burnham, K, P., (2015) Multimodel inference: understanding AIC relative variable importance values. Available from: https://sites.warnercnr.colostate.edu/anderson/wp-content/uploads/sites/26/2016/11/AICRelativeVariableImportanceWeights-Burnham.pdf. [Accessed 10th August 2018]

Chiarucci, A., Bacaro, G., Rocchini, D. & Fattorini, L. (2008) Discovering and rediscovering the sample-based rarefaction formula in the ecological literature. *Community Ecology*, 9 (1), 121–123.

Chiarucci, A., Bacaro, G., Rocchini, D., Ricotta, C., Palmer, M.W., Scheiner, S.M., (2009), Spatially constrained rarefaction: incorporating the autocorrelated structure of biological communities into sample-based rarefaction, *Community Ecology*, 10(2), pp209 - 214

Cornwell, W.K., Grubb, P.J. (2003) Regional and local patterns in plant species richness with respect to resource availability. *Oikos*, 100 (3), 417–428.

Crawley, M.J. (2001) Scale Dependence in Plant Biodiversity. *Science*, 291 (5505), 864–868.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., et al. (2007) Random Forests For Classification In Ecology. *Ecology*, 88 (11), 2783–2792.

De'ath, G., Fabricius, K.E. (2000) Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology*, 81 (11), 3178.

Defra, (2018), The 25 Year Environment Plan, Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/693158/25-year-environment-plan.pdf, [Accessed 8th August 2018]

Draaijers, G., Ivens, W., Bleuten, W. (1988) Atmospheric deposition in forest edges measured by monitoring canopy throughfall. *Water, Air, and Soil Pollution*, 42 (1-2).

Dupouey, J.L., Dambrine, E., Laffite, J.D., Moares, C. (2002) Irreversible Impact of past Land Use on Forest Soils and Biodiversity. *Ecology*, 83 (11), 2978.

Dzwonko, Z. & Loster, S. (1989) Distribution of Vascular Plant Species in Small Woodlands on the Western Carpathian Foothills. *Oikos*, 56 (1), 77.

Elith, J., Leathwick, J.R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology*, 77 (4), 802–813.

Fang, H.J., Cheng, S.L., Yu, G.R., Yang, X.M., et al. (2014) Nitrogen deposition impacts on the amount and stability of soil organic matter in an alpine meadow ecosystem depend on the form and rate of applied nitrogen. *European Journal of Soil Science*, 65 (4), 510–519.

Freckleton, R.P. (2010) Dealing with collinearity in behavioural and ecological data: model averaging and the problems of measurement error. *Behavioural Ecology and Sociobiology*, 65 (1), 91–101

Friedman, J.H., Meulman, J.J., (2003) Multiple additive regression trees with application in epidemiology. *Statistics in Medicine*, 22 (9), 1365–1381.

Fuller, R.J., Warren, M.S., (1993), Coppiced woodlands: their management for wildlife, Available from: http://jncc.defra.gov.uk/pdf/pubs93_Coppicedwoodlands.pdf [Accessed 8th August 2018]

Gardener, T. (2010), Monitoring forest biodiversity, Earthscan Publications, UK

Gaston, K.J., Spicer, J.I. (2004) Biodiversity: an introduction. Blackwell.

Gelman, A. (2008) Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine,* 27 (15), 2865–2873.

Gillman, L.N., Wright, S.D., Cusens, J., Mcbride, P.D., et al. (2014) Latitude, productivity and species richness. *Global Ecology and Biogeography,* 24 (1), 107–117.

Grueber, C.E., Nakagawa, S., Laws, R.J. & Jamieson, I.G. (2011) Multimodel inference in ecology and evolution: challenges and solutions. *Journal of Evolutionary Biology*, 24 (4), 699–711.

Hayhow DB, Burns F, Eaton MA, Al Fulaij N, August TA, Babey L, Bacon L, Bingham C, Boswell J, Boughey KL, Brereton T, Brookman E, Brooks DR, Bullock DJ, Burke O, Collis M, Corbet L, Cornish N, De Massimi S, Densham J, Dunn E, Elliott S, Gent T, Godber J, Hamilton S, Havery S, Hawkins S, Henney J, Holmes K, Hutchinson N, Isaac NJB, Johns D, Macadam CR, Mathews F, Nicolet P, Noble DG, Outhwaite CL, Powney GD, Richardson P, Roy DB, Sims D, Smart S, Stevenson K, Stroud RA,

Walker KJ, Webb JR, Webb TJ, Wynde R and Gregory RD (2016) State of Nature 2016. The State of Nature partnership.

Hermy, M. (1994) Effects of former land use on plant species diversity and pattern in European deciduous woodlands. *Biodiversity, Temperate Ecosystems, and Global Change*, pp123–144.

Honnay, O., Hermy, M., Coppin, P. (1999) Effects of area, age and diversity of forest patches in Belgium on plant species richness, and implications for conservation and reforestation. *Biological Conservation,* 87 (1), pp73–84.

Hrivnák, R., Slezák, M., Jarčuška, B., Jarolímek, I., et al. (2015) Native and Alien Plant Species Richness Response to Soil Nitrogen and Phosphorus in Temperate Floodplain and Swamp Forests. *Forests*. 6 pp 3501-3513

Hubbell, S.P. (1997) A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, 16 (5).

Hui, C. & Mcgeoch, M.A. (2014) Zeta Diversity as a Concept and Metric That Unifies Incidence-Based Biodiversity Patterns. *The American Naturalist,* 184 (5), pp684–694.

James, G., Witten, D., Hastie, T., Tibshirani, R., An Introduction to Statistical Learning: with Applications in R. New York, Springer Verlag; 2013.

Keersmaeker, L.D., Martens, L., Verheyen, K., Hermy, M., et al. (2004) Impact of soil fertility and insolation on diversity of herbaceous woodland species colonizing afforestations in Muizen forest (Belgium). *Forest Ecology and Management*, 188 (1-3), pp291–304.

Keith, S.A., Newton, A.C., Morecroft, M.D., Bealey, C.E., et al. (2009) Taxonomic homogenization of woodland plant communities over 70 years. *Proceedings of the Royal Society B: Biological Sciences*, 276 (1672), pp3539–3544.

Kennedy, F. &Pitman, R. (2004) Factors affecting the nitrogen status of soils and ground flora in Beech woodlands. *Forest Ecology and Management,* 198 (1-3), pp 1–14.

Kimberley, A., Blackburn, G.A., Whyatt, J.D., Kirby, K., et al. (2013) Identifying the trait syndromes of conservation indicator species: how distinct are British ancient woodland indicator plants from other woodland species? *Applied Vegetation Science*, 16 (4), pp667–675

Kimberley, A., Blackburn, G.A., Whyatt, J.D. & Smart, S.M. (2014) Traits of plant communities in fragmented forests: the relative influence of habitat spatial configuration and local abiotic conditions. *Journal of Ecology*, 102 (3), pp 632–640.

Kimberley, A., Blackburn, G.A., Whyatt, J.D. & Smart, S.M. (2015) How well is current plant trait composition predicted by modern and historical forest spatial configuration? *Ecography*,] 39 (1), pp67–76.

Kooijman, A.M. & Cammeraat, E. (2010) Biological control of beech and hornbeam affects species richness via changes in the organic layer, pH and soil moisture characteristics. *Functional Ecology*, 24 (2), pp469–477.

Latombe, G., Hui, C., Mcgeoch, M.A. (2017) Multi-site generalised dissimilarity modelling: using zeta diversity to differentiate drivers of turnover in rare and widespread species. *Methods in Ecology and Evolution*, 8 (4), pp431–442.

Latombe, G., McGeoch, M.A., Nipperess, D.A., Hui, C., (2018). zetadiv: Functions to Compute Compositional Turnover Using Zeta Diversity. R package version 1.1.1. https://CRAN.R-project.org/package=zetadiv

Lawton, J.H., Brotherton, P.N.M., Brown, V.K., Elphick, C., Fitter, A.H., Forshaw, J., Haddow, R.W., Hilborne, S., Leafe, R.N., Mace, G.M., Southgate, M.P., Sutherland, W.J., Tew, T.E., Varley, J., & Wynne, G.R. (2010) Making Space for Nature: a review of England's wildlife sites and ecological network. Report to Defra.

Liaw, A., Wiener, M., (2002). Classification and Regression by randomForest. R News 2(3),18--22.

Maskell, L.C., Smart, S.M., Bullock, J.M., Thompson, K., Stevens, C.J., (2010) Nitrogen deposition causes widespread loss of species richness in British habitats, *Global Change in Biology*, 16, pp 671-679

McGeoch, M.A .,Latombe, G.,Andrew, N.R., Nakagawa,S.,Nipperess, D.A., Roige,M., Marzinelli, E.M., Campbell, H.A., Verges,A Thomas, T.,Steinberg,P. D.,Selwood, K. E.,Hui,C.,(2017), The application of zeta diversity as a continuous measure of compositional change in ecology. Biorxiv, Available from www.biorxiv.org/content/early/2017/11/09/216580.full.pdf+html Accessed [12th August 2018]

Mcwilliam, W., Eagles, P., Seasons, M. & Brown, R. (2010) The housing-forest interface: Testing structural approaches for protecting suburban natural systems following development. *Urban Forestry & Urban Greening,* 9 (2), pp 149–159.

MetOffice, (2018), Weather and climate data, Available from: https://www.metoffice.gov.uk/climate/uk/regional-climates/ws#rainfall Accessed [12th August 2018]

Mouchet, M., Levers, C., Zupan, L., Kuemmerle, T., et al. (2015) Testing the Effectiveness of Environmental Variables to Explain European Terrestrial Vertebrate Species Richness across Biogeographical Scales. *Plos One,* 10 (7).

Natekin, A. Knoll, A. (2013) Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*. [Online] 7.

Natural England and RSPB (2014) Climate change adaptation manual, Available from: http://publications.naturalengland.org.uk/publication/5629923804839936, Accessed [12th August 2018]

Nicodemus, K.K., Wang, W., Shugart, Y. (2007) Stability of variable importance scores and rankings using statistical learning tools on single-nucleotide polymorphisms and risk factors involved in gene × gene and gene × environment interactions. *BMC Proceedings,* 1 (Suppl 1).

Ohlemuller, R., Wilson, J. (2000) Vascular plant species richness along latitudinal and altitudinal gradients: a contribution from New Zealand temperate rainforests. *Ecology Letters*, 3 (4), pp 262–266.

O'Brien, E., (1998), Water energy dynamics, climate, and prediction of woody plant species richness: an interim model, *Journal of Biogeography*, 25, pp 379-398

Oksanen, J., F., Blanchet, G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, R, P., O'Hara, O, B., Simpson, L, G., Solymos, P., Stevens, H, H., Szoecs, E., Wagner, H. (2018). vegan: Community Ecology Package. R package version 2.5-1. https://CRAN.R-project.org/package=vegan

Packham, J.R., Harding, D.J.L., Hilton, G.M., Stuttard, R.A., (2001), Functional ecology of woodlands and forests, The Netherlands, Kluwer.

Paillet, Y., Bergès, L., Hjältén, J., Ódor, P., Avon,C., Bernhardt-Romermann, M., Bijlsma, R., Bruyn, L., Fuhr, M., Grandin, U., Kanka, R., Lundin, L., Luque, S., Magura, T., Matesanz, S., Meszaros, I., Sebastia, M., Schmidt, W., Standovar, T., Tothmerez B., Uotila, A., Valladares, F., Vellak,K., Virtanen, R., (2010) Biodiversity Differences between Managed and Unmanaged Forests: Meta-Analysis of Species Richness in Europe. *Conservation Biology*, 24 (1), pp101–112.

Pallett, D., Pescott, Schäfer, S. (2016) Changes in plant species richness and productivity in response to decreased nitrogen inputs in grassland in southern England. *Ecological Indicators,* pp 6873–81.

Palpurina, S., Wagner, V., von Wehrden, H., Hájek, M., Horsák, M., Brinkert, A., Hölzel, N., Wesche, K., Kamp, J. , Hájková, P., Danihelka, J., Lustyk, P., Merunková, K., Preislerová, Z., Kočí, M., Kubešová, S., Cherosov, M., Ermakov, N., German, D., Gogoleva, P., Lashchinsky, N., Martynenko, V., Chytrý, M. and Grytnes, J. (2017), The relationship between plant species richness and soil pH vanishes with increasing aridity across Eurasian dry grasslands*. Global Ecol and Biogeography*, 26: pp 425-434.

Petit, S., Griffiths, L., Smart, S.S., Smith, G.M., et al. (2004) Effects of area and isolation of woodland patches on herbaceous plant species richness across Great Britain. *Landscape Ecology*, 19 (5), pp463–472.

Peterken, G.F. & Game, M. (1984) Historical Factors Affecting the Number and Distribution of Vascular Plant Species in the Woodlands of Central Lincolnshire. *The Journal of Ecology*,72 (1), 155.

Pitcairn, C., Skiba, U., Sutton, M., Fowler, D., et al. (2002) Defining the spatial impacts of poultry farm ammonia emissions on species composition of adjacent woodland ground flora using Ellenberg Nitrogen Index, nitrous oxide and nitric oxide emissions and foliar nitrogen as marker variables. *Environmental Pollution,* 119 (1), pp 9–21.

Preston, F.W. (1962) The Canonical Distribution of Commonness and Rarity: Part I. *Ecology*, 43 (2), 185.

Quine, C. P., C. Cahalan, A. Hester, J. Humphrey, K. Kirby, A. Moffat, and G. Valatin. (2011). Ch 8 Woodlands. In: R. Watson and S. Albon, editors. UK National Ecosystem Assessment. UNEP & Defra, pp 242-293

Richerson, P.J. & Lum, K.-L. (1980) Patterns of Plant Species Diversity in California: Relation to Weather and Topography. *The American Naturalist*, 116 (4), pp 504–536.

Ridgeway, G with contributions from others (2017). gbm: Generalized Boosted Regression Models. R package version 2.1.3. https://CRAN.R-project.org/package=gbm.

Ritchie, G. &; Dolling, P. (1985) The role of organic matter in soil acidification. *Australian Journal of Soil Research,*23 (4), 569.

Russell, J. (1960) Soil fertility changes in the long-term experimental plots at Kybybolite, South Australia. II. Changes in phosphorus. *Australian Journal of Agricultural Research,* 11 (6), 926.

Scheiner, S.M., (2003), Six types of species area curves, *Global Ecology and Biogeography*, 12, pp 441 - 447.

Schmidt, W., (2005), Herb layer species as indicators of biodiversity of managed and unmanaged beech forests, *Snow and landscape research.* 79 (1,2).

Shen, G., Yu, M., Hu, X.-S., Mi, X., et al. (2009) Species–area relationships explained by the joint effects of dispersal limitation and habitat heterogeneity. *Ecology,* 90 (11), pp 3033–3041.

Simkin, S.M., Allen, E.B., Bowman, W.D., Clark, C.M., et al. (2016) Conditional vulnerability of plant diversity to atmospheric nitrogen deposition across the United States. *Proceedings of the National Academy of Sciences,* 113 (15), pp 4086–4091.

Slee, B., Kyle, C., Biodiversity and Woodland Ecosystems, Available from: https://www.hutton.ac.uk/sites/default/files/files/Biodiversity%20&%20Woodland%20Ecosystems(1).pdf., [Accessed 8th August 2018]

Smart, S.M., Ashmore, M.R., Hornung, M., Scott, W.A., et al. (2004) Detecting the Signal of Atmospheric N Deposition in Recent National-Scale Vegetation Change Across Britain. *Water, Air, & Soil Pollution,* 4 (6), pp 269–278.

Smart, S.M., Kirby, K., Cornway, P., Wood, C.M., (2013), Woodlands survey of Great Britain 1971 - 2001, Available from: http://nora.nerc.ac.uk/id/eprint/504075/1/N504075CR.pdf, [Accessed 10th August 2018]

Smart, S.M., Ellison, A.M., Bunce, R.G.H., Marrs, R.H., et al. (2014) Quantifying the impact of an extreme climate event on species diversity in fragmented temperate forests: the effect of the October 1987 storm on British broadleaved woodlands. *Journal of Ecology*, 102 (5), pp 1273–1287.

Stevens, C.J., Payne, R.J., Kimberley, A. Smart, S.M., (2016) How will the semi-natural vegetation of the UK have changed by 2030 given likely changes in nitrogen deposition? *Environmental Pollution,* 208 Part B pp 879–889.

Storch, D., Kiel, P, Kuni, W.E., (2014), Scaling communities and biodiversity, available from: http://www.cts.cuni.cz/~storch/publications/Storch_et_al_2014_Scaling_communities_biodiversity.pdf, [Accessed 10th August 2018]

Stroh, P.A., Leach, S.J., August, T.A., Walker, K.J., Pearman, D.A., Rumsey, F.J., Harrower, C.A., Fay, M.F., Martin, J.P., Pankhurst, T., Preston, C.D. & Taylor, I. (2014). A Vascular Plant Red List for England. Botanical Society of Britain and Ireland, Bristol.

Symonds, M.R.E., Moussalli, A. (2010) A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. *Behavioral Ecology and Sociobiology*. [Online] 65 (1), pp 13–21.

Sugihara, G., (1979), Minimal community structure: an explanation of species abundance patterns, *The American Naturalist*, 116(6) pp 770-787.

Tittensor, D., Walpole, M., Hill, S., Boyce, D., Britten, L., Burgess, G., Butchart, H. M., Leadley, W., Regan, P., Alkemade, E., Baumung, R., Bellard, R., Bouwman, C., Bowles-Newark, A., Chenery, N.M., Christensen, W., Cooper, V., Crowther, H.R., Ye, A., (2014). A mid-term analysis of progress toward international biodiversity targets. *Science*. 346.

Thiele, J., Kellner, S., Buchholz, S. & Schirmel, J. (2018) Connectivity or area: what drives plant species richness in habitat corridors? *Landscape Ecology*, 33 (2), 173–181.

Tjørve, E., Tjørve, K.M. (2017) Species-Area Relationship. *eLS*, 1–9.

Watts, K., Quine, C.P., Eycott, A.E., Moseley, D., Humphrey, J.W., Ray, D., (2008) Conserving forest biodiversity: recent approaches in UK forest planning and management in: Patterns and Processes in Forest Landscapes, Springer Verlag.

Williams, C. & Donald, C. (1957) Changes in organic matter and pH in a podzolic soil as influenced by subterranean clover and superphosphate. *Australian Journal of Agricultural Research*, 8 (2), 179.

Zak, D.R., Freedman, Z.B., Upchurch, R.A., Steffens, M., et al. (2016) Anthropogenic N deposition increases soil organic matter accumulation without altering its biochemical composition. *Global Change Biology*, 23 (2), pp 933–944.

Supplementary material

ZETA DIVERSITY

To calculate zeta diversity, consider four plots with species S_1, S_2, S_3, S_4. Let S_i represent the set and number of species in plot 1. Four orders of zeta can be calculated in the following way.

$$\zeta_1 = \frac{S_1 + S_2 + S_3 + S_4}{\binom{4}{1}} \qquad 1$$

$$\zeta_2 = \frac{S_1 \cap S_2 + S_1 \cap S_3 + S_1 \cap S_4 + S_2 \cap S_3 + S_2 \cap S_4 + S_3 \cap S_4}{\binom{4}{2}} \qquad 2$$

$$\zeta_3 = \frac{S_1 \cap S_2 \cap S_3 + S_1 \cap S_2 \cap S_4 + S_1 \cap S_3 \cap S_4 + S_2 \cap S_3 \cap S_4}{\binom{4}{3}} \qquad 3$$

$$\zeta_4 = \frac{S_1 \cap S_2 \cap S_3 \cap S_4}{\binom{4}{4}} \qquad 4$$

$S_i \cap S_j$ is the number of species shared between plots i and j. For a woodland with 16 plots, zeta can be calculated to ζ_{16} as shown in figure 1.

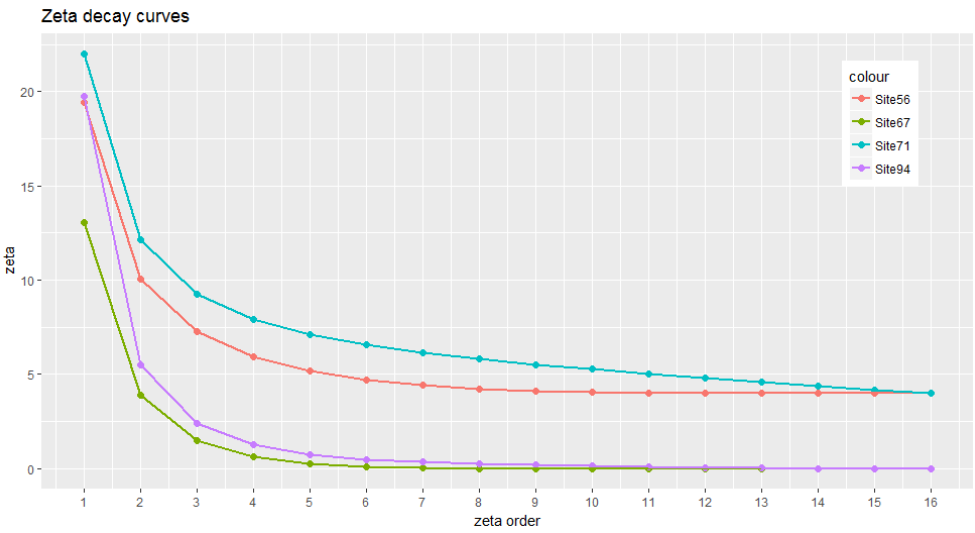

Figure 1. Zeta decay curves. Calculation of the zeta values to ζ_{16}, as shown in equations 1 to 4 above, will give zeta decay curves as shown above.

The slope of the zeta decay curve varies with the probability of finding new species in new plots [Hui & McGeogh, 2014] and the number of common and rare species [Latombe et al., 2017]. If plots contained the same species, then $\zeta_1 - \zeta_2 = 0$ and the initial gradient of the curve = 0. If all plots contained different species then $\zeta_2 = 0$, and the initial gradient of the curve would equal ζ_1.

DECISION TREES

The decision tree algorithms are ensemble methods using multiple trees to improve predictive accuracy. A RF further improves performance by selecting a subset of the predictor variables at each split. In a GBM each new tree is tested and values that have poor predictive power are weighted such that they are more likely to be selected in the next random subset of the observations. In both methods the data is split into train and test sets using a 75/25 ratio. These methods then also use random subsets of training data and so automatically provide a validation set; the values that are not selected can be used to give an "out of bag" (OOB) prediction error for the model.

Hyper parameters define the structure of the decision tree model and must be chosen to optimise the model performance, this was achieved through a hyper parameter tuning grid; a matrix containing a set of permutations of hyper parameters. The model was run across the hyper parameter tuning grid and the hyper parameters which minimized the OOB error were used in the final model.

Over fitting was examined by considering the rmse of the train and test sets. If the error on the train set is less than that of the test set the model is overfit. This was addressed in the random forest through hyper-parameter tuning and in the gradient boosted machine by the ntrees function. This function looks at the prediction error on the OOB set, the number of trees is set as the value when this error stops decreasing and starts to increase.

It was seen in this analysis that both decision tree methods were unstable. That is, they did not consistently predict the same variables as important on repeated runs of the algorithm. Although this was addressed by averaging the algorithms over 100 repeats, the validity of the models in the face of this instability could be investigated. For example, spurious random variables could be included to compare their response to the actual variables.

INTERACTIONS

Interactions were not included in the average model, however, when optimizing the decision trees an interaction depth of 6 was required. This suggest that interactions are important for predicting richness in this data. To improve the average model, the interactions within the decision trees could be examined and the interactions included in the average model

ALTERNATIVE METHODS OF CONSTRUCTING SPECIES AREA CURVES.

At the larger scale, an exact method for finding the expected number of species, $\overline{S_i}$, as area is incremented over the i plots, is given by,

$$\overline{S_i} = S_n - \binom{n}{i}^{-1} \sum_k \binom{n - n_k}{i}$$

S_n is the total number of species in the woodland, n is the number of plots and k is the number of plots in which a species occurs, [Chiarucci et al., 2008]. These values can be calculated using the "exact" option to the specacc function of the vegan R package, [Oksana et al., 2018]. The z values can again be extracted from a log/log fit the the calculated expected species number and area. When these z values were used to form an average model, the results were as shown below.

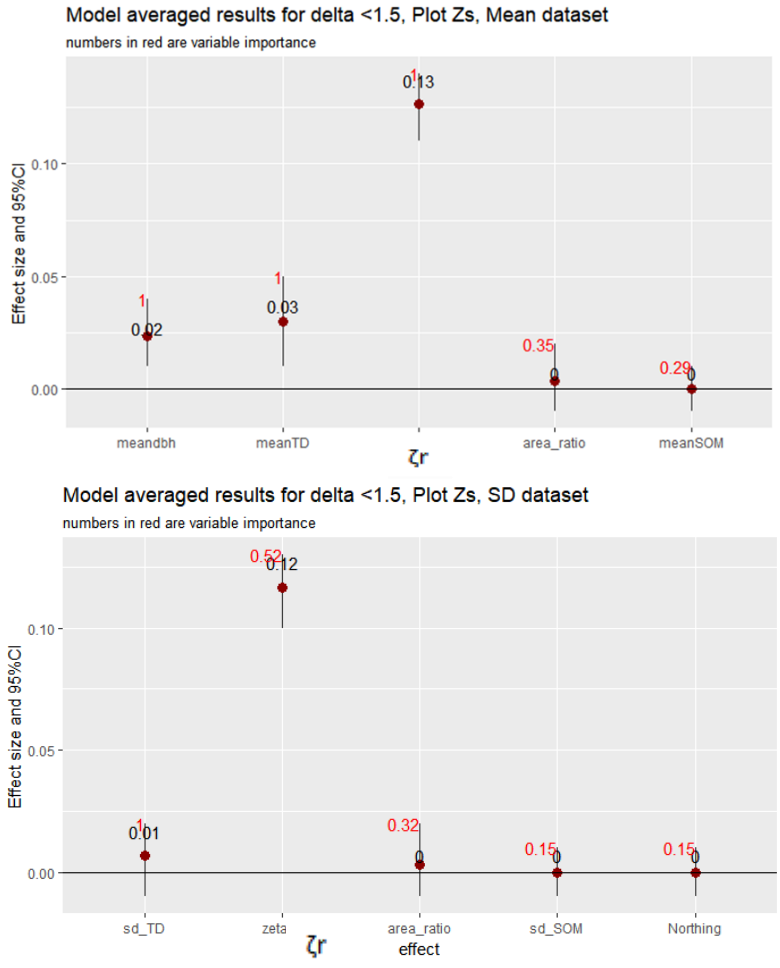

Figure 2. Average model parameter effect sizes using z values derived from the exact method accumulate species

Although ζ_r is still a significant variable with a positive effect on z, mean tree density and mean diameter at breast height also have a positive effect. This seems counter intuitive as both these variables have a negative effect on species richness.

The spearman correlation coefficient between the z values obtained by the exact method and the spatially explicit method is 0.55, demonstrating a difference in the two techniques.

Chiarucci et al., [2009] point out that spatial autocorrelation needs to be accounted for. The z values obtained from the spatially explicit method I used to calcuate the species accumulation curves give more biologically meaningful results, perhaps because they are taking spatial autocorrelation into account. The difference in the curves obtained requires further analysis, as does the behaviour or ζ_r. For example, Baselga et al (2007) show that multisite dissimilarity metrics may be unable to differentiate between turnover due to new assemblages or due to nested assemblages. Therefore, modelling of ζ_r with realistic, simulated species assemblages would be beneficial.