Machine learning identification of microhabitat features associated with occupancy of artificial nestboxes by hazel dormice (Muscardinus avellanarius) in a UK woodland site

WILDLIFE BIOLOGY

Research article

Machine learning identification of microhabitat features associated with occupancy of artificial nestboxes by hazel dormice *Muscardinus avellanarius* in a UK woodland site

Joe Malyan[1,2], Amanda J. Lloyd[3] and Manuela González-Suárez[2]

[1]Lead Ranger, Parks and Countryside Department, Bracknell Forest Council, Bracknell, UK
[2]Ecology and Evolutionary Biology, School of Biological Sciences, University of Reading, Reading, UK
[3]Ecological Consultant, Wantage, UK

Correspondence: Manuela González-Suárez (manuela.gonzalez@reading.ac.uk)

Wildlife Biology
2024: e01185
doi: 10.1002/wlb3.01185

Subject Editor: Lucas Wauters
Editor-in-Chief: Ilse Storch
Accepted 9 February 2024

Hazel dormice *Muscardinus avellanarius* have severely declined since 2000 leading to increased legislative protection in the UK and Europe. Artificial nestboxes are widely used for its conservation and monitoring. Previous research has focused on how to identify suitable areas for nestboxes, but where to place individual boxes to promote occupancy is less well understood. Here, we demonstrate the use of machine learning Random Forest regression to predict nestbox occupancy from a wide range of microhabitat variables using a UK woodland as a case study. Random forest models are powerful predictive tools that allow simultaneous testing of many predictors with relatively few observations.

Field data included observed nestbox occupancy (2017–2021) and measurements of 76 microhabitat variables collected in the summer of 2021 from 45 occupied and unused nestboxes located in a deciduous woodland in Berkshire, UK. We applied Random Forest regression to identify important variables and predict nestbox occupancy demonstrating robust approaches to tune model hyperparameters and evaluate importance metrics.

In our study area, nestboxes were more likely to be occupied in sites with more hazel *Corylus avellana*, greater overall tree abundance but not fully closed canopies (optimal 80–85%), more honeysuckle *Lolium periclymenum* and hawthorn *Crataegus monogyna*, and when located further from footpaths and woodland margins. Occupancy over the study period was well predicted using microhabitat variables (13.3% OOB error) but future occupancy was more uncertain (33.3% error for 2021–2023 records).

Modelling approaches that allow consideration of numerous variables from few locations or observations can be help identify relevant features and predict desirable outcomes of conservation actions. Here we demonstrate this approach identifying microhabitat variables that influence artificial nestbox occupancy by hazel dormice in a UK woodland. Findings offer some recommendations for local management that could promote nestbox occupancy and improve monitoring and conservation efforts.

Keywords: conservation, dormouse, habitat selection, mitigation, modelling, *Muscardinus*, RandomForest

NORDIC SOCIETY OIKOS

www.wildlifebiology.org

Introduction

Over the past 50 years, the UK has seen a severe decline amongst many of its native mammal species (Coomber et al. 2021), including hazel dormice *Muscardinus avellanarius*, harvest mice *Micromys minutus* and hedgehogs *Erinaceus europaeus*. Even populations thought to have been stable, and widespread, such as those of stoats *Mustela erminea* and weasels *Mustela nivalis*, are being shown, through new research, to be decreasing at alarming rates (Coomber et al. 2021). Habitat loss through urban expansion and changes in agricultural practices are cited as key drivers of population decline, as well as changes in forestry management. These changes decrease structurally complex and spatially heterogeneous woodlands (Hopkins and Kirby 2007) affecting vulnerable mammal species such as red squirrels *Sciurus vulgaris* (de Raad et al. 2021) and pine martens *Martes martes* (Caryl 2021), as well as hazel dormice (Bright and Morris 1995a). Many of our small UK mammal species are vitally important contributors to biodiversity directly through interactions with various plant and invertebrate species, and also indirectly as prey for other species such as birds of prey and larger mammals (Occhiuto et al. 2021). It is therefore important that we understand the specific habitat requirements of these species to provide suitable mitigation and enhancement, and inform new protective legislation as required.

Hazel dormice have severely declined with a reported 70% reduction in population size since 2000 across the UK (Wembridge et al. 2023). This decline is primarily attributed to a reduction of traditional woodland management techniques, specifically coppicing (Bright and Morris 1995b) and habitat fragmentation (Bright and Morris 1994, Capizzi et al. 2002). Hazel dormice have slow reproduction rates and live at low densities (Bright and Morris 2008) and have been shown to be particularly sensitive to environmental changes and habitat fragmentation (Capizzi et al. 2002), which makes this species especially vulnerable to local declines and extirpation. The decline of hazel dormice has prompted several conservation initiatives designed to protect the species and enhance or create suitable habitats (Bright and Morris 1994, 1995b, Ramakers et al. 2014, Phillips et al. 2022). For example, more active woodland management has led to local recoveries (Goodwin et al. 2018a). Management includes coppicing hazel to maintain a successional status, which slows the progression from an unshaded and productive shrub layer to a high forest with an overshaded understorey (Bright and Morris 1990). This provides hazel dormice with habitat rich in foraging material that can support healthy populations of invertebrate species, which are a vital part of the diet of hazel dormice over the summer months (Bright et al. 2006). Conservation initiatives have also involved reintroduction, with the 1000th individual reintroduced in Lancashire, northwest England in 2021 (People's Trust for Endangered Species 2021). Success has been linked to adequate habitat management (Bright and Morris 2002) and improved connectivity to allow population expansion (Mitchell-Jones and White 2009).

The installation of nestboxes is widely employed in the UK to help reverse hazel dormice declines (Morris et al. 1990) and to monitor population trends (Wembridge et al. 2023). Nestboxes can improve local densities (Morris et al. 1990), potentially due to enhanced survival of young in dry, secure boxes, or because boxes offer greater nesting opportunities than might be naturally available in some habitats. With entrances holes approximately 24–28 mm in diameter, dormice nestboxes are less likely to be used by other larger woodland species, such as squirrels and woodpeckers, which reduces competition (Madikiza et al. 2010). While nestboxes can be important, their uptake by hazel dormice depends on the surrounding environment which needs to be carefully considered when placing boxes (Juškaitis et al. 2013, Mortensen et al. 2022). Placement near food and nesting material sources is likely beneficial because while hazel dormice can travel up to 50 m to collect materials, when resources are available regular travel is generally limited to within 10 m of the nest site (Bracewell and Downs 2017). In fact, one study found that over 70% of nests in nestboxes were made from the plant on which the nestbox was attached (Bracewell and Downs 2017). A shorter journey when encumbered by heavy nest materials reduces the risk of predation, and conserves energy, especially for lactating females (Prentice and Prentice 1988, Juškaitis 2014). Accessibility to these resources is also likely important. As a predominantly arboreal species, hazel dormice are often associated with well-developed tree canopies, and/or understorey layers with abundant horizontal branches (Bright and Morris 2009, Goodwin et al. 2018b). However, an extensive tree canopy and/or understorey layer could limit the amount of sunlight reaching field layer plants affecting flowering and fruiting, and encouraging vertical growth which is less useful for travel (Bright and Morris 1990), so probably intermediate to high tree canopy and/or understorey layer cover is optimal (Juškaitis and Augutė 2008).

Previous research has described general habitat preferences of hazel dormice and explores how overall local conditions affect nestbox occupancy (Bright and Morris 1990, Panchetti et al. 2007, Cartledge et al. 2021, Fedyń et al. 2021), using traditional regression models (linear and generalised) and more recently Ecological Niche Factor Analysis – ENFA (Dietz et al. 2018, Cartledge et al. 2021). However, information is more limited on how microhabitat within a suitable local site influences nestbox use (Mortensen et al. 2022, Phillips et al. 2022). After an area is identified as suitable for installing nestboxes it is still important to determine the optimal locations for nestboxes within the site. Here, we showcase the application of machine learning Random Forest regression to address this knowledge gap. This modelling approach is particularly well-suited for this question due to its strong classification accuracy and applicability to high variable to observation ratios which pose challenges for traditional regression approaches and may be common in local studies where sample sizes may be limited (Cutler et al. 2007). Understanding how microhabitat features influence occupancy can offer recommendations for nestbox placement and site management to aid in the protection and conservation of the hazel dormouse.

Material and methods

Study area and dormice surveys

Our study area was within the grounds of Basildon Park House (BPH), a National Trust property with extensive woodland, located in Berkshire, UK (Fig. 1). The woodland covers approximately 63 ha around the perimeter of BPH and is connected (western boundary) to a further 40 ha woodland (with a small road in between). Management at BPH includes seasonal coppicing over the winter months, and the creation of 'wigwam' structures for trees as protection from deer browsing that improve habitat quality for hazel dormice (Reid et al. 2021).

In 2013, 144 nestboxes were erected at two different woodland areas within BPH: 78 were located in clusters within the northern site, and 66 were positioned along three distinct lines in the southern site (Fig. 1). Within the northern site, mature beech *Fagus sylvatica*, ash *Fraxinus excelsior* and oak *Quercus* spp. are the most frequent tree species with sweet chestnut *Castanea sativa* and field elm *Ulmus minor* occasional. Hawthorn and coppiced hazel are dominant in the understorey layer, with holly *Ilex aquifolium* and young sycamore *Acer pseudoplatanus* frequent, and other species such as field maple *Acer campestre*, whitebeam *Sorbus aria* and rowan *Sorbus aucuparia* occasional. Bramble *Rubus fruticosus* agg. and honeysuckle are abundant in both the understorey and field layer. In the southern site, nestboxes in the densest lines are closest to the boundary between woodland and arable land (in 2021 the crop was rapeseed *Brassica napus*), amongst what used to be a hedge, but is now a line of mature scrub. The main tree canopy species here are oak, lime *Tilia x europaea* and beech, which are frequent, although there are some sections where there are no canopy tree species present. In the understorey cherry plum *Prunus cerasifera* is dominant, with occasional European spindle *Euonymus europaeus* and young oaks. The other two, more sparse lines of boxes within the southern site extend northwards into more mature woodland, where the dominant tree canopy species are beech and oak, with hazel dominant in the understorey, and hawthorn and young sycamore frequent. Bramble is dominant in the understorey and field layer species.

All nestboxes at BPH have been regularly monitored by the Berkshire Mammal Group since 2017, up to four times a year as part of the National Dormouse Monitoring Programme (NDMP). No monitoring occurred between 2013 and 2016. During each survey, nestboxes are recorded as occupied by hazel dormice if a dormouse is present, or if there is a nest with evidence of recent occupation. Otherwise, they are recorded as empty. Using the full survey records between 2018 and July 2021 we classified nestboxes as historically occupied (occupied at least once since 2018) or unoccupied (not occupied since 2018). We did not include existing occupancy records from 2017 to reduce the time lag from recorded occupancy to habitat recorded; such that we focus on nestboxes occupied in the four years prior to recording habitat. We focused on 13 occupied nestboxes at the northern site and all six occupied nestboxes at the southern site (total 19 occupied nestboxes). In the northern site we focused on nestboxes occupied frequently (in more than one survey) and/or recently (occupied at least once in the last three years). If two nestboxes in close proximity (< 10 m) met these criteria, we randomly selected one of them to avoid

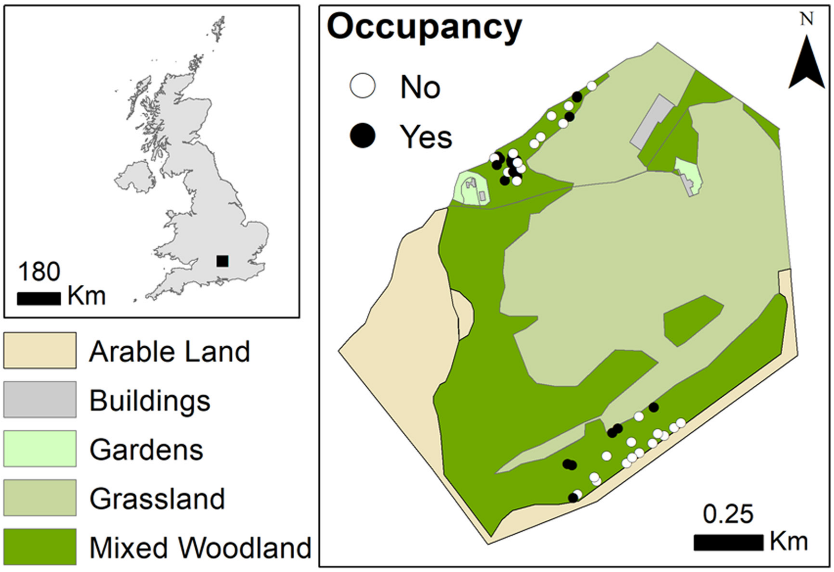

Figure 1. Map of the surveyed nestboxes in Basildon Park House (UK) showing the broad habitat types and the historical occupancy by hazel dormice *Muscardinus avellanarius* of 45 sampled nestboxes. The northern site is located on the northwest boundary and the southern site is located on the southern-southeastern boundary, adjacent to arable land. Top left inset shows the study area location in the UK.

replicating microhabitat data. We also selected 26 historically unoccupied nestboxes (12 in the northern site and 14 in the southern site) using the *SelectRandomByPercent* function in 'ARCGIS' 10.5.1 excluding any nestboxes within 10 m of selected occupied nestboxes. If two selected unoccupied nestboxes were in close proximity (< 10 m) we located an alternative pair in the area at least 10 m apart.

Microhabitat surveys

In March 2021, all selected nestboxes were cleaned and we recorded GPS coordinates (using Handy GPS on iPhone 11, accurate to 3 m), tree species on which the nestbox was installed, height from the base of the nestbox to the ground, and the orientation of the front of the nestbox. Microhabitat data were collected during May and June 2021 at four scales: directly above the nestbox, within a 5 m radius of each box, in four 2 × 5 m quadrats starting 5 m from the nestbox, and using existing GIS layers (Fig. 2, Table 1 for details). Within the 5 m radius cover was visually estimated within four levels: tree canopy, understorey, field layer and ground layer (Eden 2009). Tree canopy reflected trees taller than 4 m, with trees < 4.m classified as part of understorey (Berg and Berg 1998). The four 2 × 5 m quadrats started at the edge of the 5 m radius running with orientations N, E, S and W (Fig. 2). Sampling areas which intersected footpaths or trackways were still assessed as dormice can occasionally cross open ground when foraging or looking for nesting materials (Mortelliti et al. 2013). We acknowledge that the collected microhabitat data represent a particular time of the year and cannot capture seasonal or interannual variability, which may be important for site selection in hazel dormice.

GIS information

Nestbox locations were collected via GPS with a minimum of 3 m accuracy and mapped using ARCGIS. Footpaths around the site, and the woodland margin were walked and recorded using GPS and added as a new layer. These layers were then used to calculate the minimum distance to the woodland margin, the closest footpath, and the nearest neighbouring nestbox.

Statistical analysis

We evaluated the role of microhabitat variables in nestbox occupancy using machine learning Random Forest regression methods (Cutler et al. 2007). This approach ensembles multiple regression or classification trees allowing the estimation of variable importance and conditional effects (Breiman 2001). In our case we predicted occupancy (occupied versus unoccupied) fitting classification models using the function *randomForest* from the R package 'randomForest' (Liaw and Weiner 2002). The randomForest algorithm works as follows: first it draws a user-defined (hyperparameter *ntree*, see below) which describes the number of bootstrap samples obtained from the original dataset. Then, for each bootstrap sample, it grows an unpruned classification or regression tree choosing the best split at each node among a random sample of predictors (the number of predictors to include in that random sample is the user-defined hyperparameter *mtry*, see below). The tree is then used to predict the observations from the dataset that were not in the bootstrap sample.

We fitted two models: a complete model with all measured variables and a simplified model based on variables identified as most important (details below). We fitted a simplified model because measuring 76 variables is

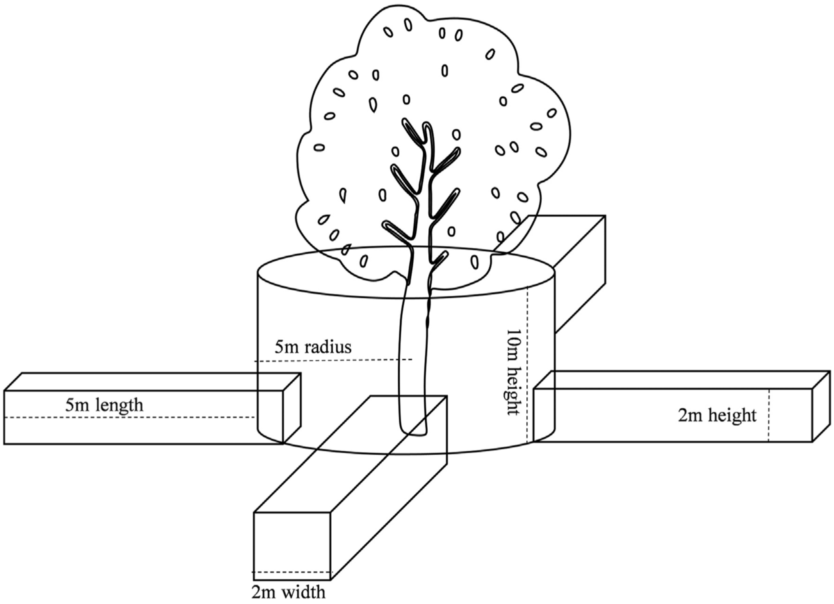

Figure 2. Schematic of the vegetation microhabitat sampling scheme used to study factors influencing nestbox occupancy by hazel dormice *M. avellanarius* in a UK woodland. The central tree is the site of a studied nestbox and we show the radius and quadrat sampling areas. See Table 1 for details on the variables measured at these scales.

Table 1. The 76 microhabitat variables measured at four different scales and used to evaluate the factors influencing nestbox occupancy by hazel dormice *M. avellanarius* in a UK woodland site. Some variable types are described in general but data were collected separately for different plant species, in those the total number of separate predictor variables used in the model is indicated (underlined text) and the individuals plant species are listed under the definition.

Variable/scale	Definition
Above nestbox scale	
Canopy closure	Total percentage of tree canopy and understorey vegetation cover above the focal nestbox estimated as the average cover from two measurements in May and July (to account for seasonal variability in leaf growth). Cover was defined as the percentage of black pixels in processed photos taken with an Apple iPhone 11 levelled-flat on a tripod set at 1 m from the ground and as close to the nestbox as possible without including the box in the photo. Original photos were desaturated to grayscale with the colour curve adjusted to make all pixels either black or white using GIMP 2.10.24
5 m radius	
Tree canopy cover	Percentage of the tree canopy (in 10% increments) occupied by trees of the same species taller than 4 m within a 5 m radius circular area around the focal nestbox. Estimated visually Each identified species described by a separate variable (total 10 variables): ash *Fraxinus excelsior*, beech *Fagus sylvatica*, field elm *Ulmus minor*, field maple *Acer campestre*, holly *Ilex aquifolium*, lime *Tilia x europaea*, oak *Quercus* spp., sweet chestnut *Castanea sativa*, sycamore *Acer pseudoplatanus*, yew *Taxus baccata*
Understorey cover	Percentage of the canopy (in 10% increments) occupied by trees of the same species smaller than 4 m within a 5 m radius circular area around the focal nestbox. Estimated visually Each identified species described by a separate variable (total 19 variables): alder buckthorn *Frangula alnus*, ash *Fraxinus excelsior*, beech *Fagus sylvatica*, blackthorn *Prunus spinosa,* cherry plum *Prunus cerasifera*, elder *Sambucus nigra*, field elm *Ulmus minor*, field maple *Acer campestre*, hawthorn *Crataegus monogyna*, holly *Ilex aquifolium*, lime *Tilia x europaea*, oak *Quercus* spp., rowan *Sorbus aucuparia*, spindle *Euonymus europaeus*, sycamore *Acer pseudoplatanus*, wayfaring tree *Viburnum lantana*, whitebeam *Sorbus aria*, wild cherry *Prunus avium*, hazel *Corylus avellana*
Field layer cover	Percentage of the field layer (in 10% increments) occupied by plants of the same species within a 5 m radius circular area around the focal nestbox. Climbing species cover was estimated up to a height of 10 m. Estimated visually Each identified species or group described by a separate variable (total 16 variables): bluebells *Hyacinthoides non-scripta*, bramble *Rubus fruticosus* agg., cleavers *Galium aparine*, cow parsley *Anthriscus sylvestris*, dogs mercury *Mercurialis perennis*, ground ivy *Glechoma hederaceae*, hedge woundwort *Stachys sylvatica*¸ speedwell *Veronica* spp., herb Robert *Geranium robertianum*, honeysuckle *Lonicera pericylmenum*, lords and ladies *Arum maculatum*, meadow buttercup *Ranunculus acris*, nettles *Urtica dioica*, yellow archangel *Lamiastrum galeobdolon*, fern (group, not identified to species)’, grass (graminoid group, not identified to species)
Ground cover	Percentage of the ground (in 10% increments) occupied by bryophytes within a 5 m radius circular area around the focal nestbox. Estimated visually [Data were also collected for cover of fungi, leaf litter and bare ground too but due to low variability among sites were not considered in the analyses]
Quadrat	
Quadrat cover	Mean percentage cover of individual species over four 2 × 5 m quadrats starting 5 m from the focal nestbox and running North, South, East and West Each identified species described by a separate variable (total two variables): bramble *Rubus fruticosus* agg. and honeysuckle *Lolium periclymenum*
5 m + quadrat	
Tree abundance	Relative abundance of individual tree species within a 10 m radius from the focal nestbox. Obtained by adding the total number of individual trees with a trunk circumference > 40 cm within a 5 m radius circular area and in four 2 × 5 m quadrats starting 5 m from the focal nestbox and running North, South, East and West. Abundance within the 5 m radius area included every individual tree, whilst the four quadrats provided relative abundance within the area 5–10 m from the nestbox based on quadrat totals Each identified species described by a separate variable (total 20 variables): ash *Fraxinus excelsior*, beech *Fagus sylvatica*, blackthorn *Prunus spinosa*., cherry plum *Prunus cerasifera*, elder *Sambucus nigra*, field elm *Ulmus minor*, field maple *Acer campestre*, Guelder rose *Viburnum opulus*, hawthorn *Crataegus monogyna*, hazel *Corylus avellana*, holly *Ilex aquifolium*, lime *Tilia x europaea*, oak *Quercus* spp., plum *Prunus domestica*, spindle *Euonymus europaeus*, sweet chestnut *Castanea sativa*, sycamore *Acer pseudoplatanus*, wayfaring tree *Viburnum lantana*, wild cherry *Prunus avium*, yew *Taxus baccata*
Total trees	Combined relative tree abundance. Sum of ‘Tree abundance’ for all 20 recorded species at each nestbox
Tree richness	Observed tree species richness calculated adding all species with ‘Tree abundance’ > 0 for each nestbox
Local	
Site	Descriptor of the general site where the focal nestbox was located: northern or southern (Fig. 1)
Nestbox height	Straight vertical distance in mm from the base of the nestbox to the ground
Local (GIS)	
Distance to nearest footpath	Distance in metres from focal nestbox to the nearest footpath. Collected via GPS with minimum 3 m accuracy and mapped using ARCGIS
Distance to nearest nestbox	Distance in metres from focal nestbox to the nearest nestbox (all nestboxes, not only those for which habitat data were collected, were considered to measure distance). Collected via GPS with minimum 3 m accuracy and mapped using ARCGIS
Distance to nearest woodland margin	Distance in metres from focal nestbox to the nearest woodland margin. Collected via GPS with minimum 3 m accuracy and mapped using ARCGIS

time-consuming and could be potentially unnecessary. To define the complete model we used all 76 predictor variables (Table 1). We defined 'tune' hyperparameter values by testing performance with 56 combinations of *mtry* (tested all values from 2 to 15) and *ntree* (four values tested: 500, 1000, 5000 and 10 000). We selected the hyperparameter combination that resulted in the lowest OOB (out-of-bag) error.

There are multiple metrics that can be used to assess variable importance in Random Forest models and more robust inferences can be achieved by their simultaneous consideration. We used the package 'randomForestExplainer' (Paluszynska et al. 2020) to assess seven importance metrics: mean minimal depth from top trees, total number of nodes that use the variable to split the data, the total number of trees in which the variable is used, mean decrease in prediction accuracy after the variable is permuted, mean decrease in the Gini index of node impurity by splits based on the variable, total number of trees in which the variable is used for splitting the root node, p-value from a binomial test comparing the number of nodes in which the variable was used compared to the expected number if variables were assigned to nodes at random. To facilitate the identification of the most relevant variables we focused on variables with significant p-values in the binomial test, which were then evaluated in detail using plots representing all metrics and further confirmed via the function *important_variables* from the package 'randomForestExplainer'. Relationships between importance metrics shown in Supporting information.

To define the simplified model we focused on the most important variables identified from the complete model. We tuned hyperparameter values by testing performance with 24 combinations of: mtry (tested all values from 2 to 7), and ntree (four values tested: 500, 1000, 5000 and 10 000). We selected the hyperparameter combination that resulted in the lowest OOB error. From this simplified model we generated dependence plots to show how each variable influences the probability of occupancy using the function *partial* from the R package 'pdp' (Greenwell 2017).

For the complete and simplified models we report OOB overall error, false positive (unoccupied nestboxes predicted to be occupied), and false negatives rates (occupied nestboxes predicted to be unoccupied), and their reciprocals: model accuracy, specificity, and sensitivity. OOB samples represented approximately one-third of the observations drawn with replacement (the default setting). In addition to the OOB validation we explore how well models fitted with habitat data obtained in the summer of 2021 could predict future occupancy records from surveys completed between September 2021 and October 2023 when our surveys identified 16 nestboxes as occupied and 29 as unoccupied (this is an independent validation as this information was not used to define occupancy for model fitting).

The full dataset and R script used for analyses are available in Dryad https://doi.org/10.5061/dryad.1c59zw43q.

Results

We found a diversity of tree and field layer plant species across different areas of the northern and the southern sites (summary in Supporting information). The complete model with all 76 variables had an OOB error rate of 20% (model accuracy = 80%), with 11.5% false positives (specificity = 88.5%), and 31.6% false negatives (sensitivity = 68.4%). Variable importance metrics from this full model revealed 24 variables with significant binomial test p-values, and among those nine variables were identified as most relevant based on the other six importance metrics (Fig. 3; for display purposes variable importance is shown based on the three less correlated metrics, Supporting information). Supporting information shows ranking and values for each importance metrics for all 76 predictors. These eight variables collectively described vegetation measurements at different scales, human impact and nestbox position variables (Fig. 3 for details of each variable).

The simplified model based on the eight most important variables had a OOB error rate of 13.3% (model accuracy = 88.5%) with 15.8% false negatives (sensitivity = 86.7%) and 11.5% false positives (specificity = 84.2%). The model predicted increased probability of nestbox occupancy with more trees within 10 m (total tree), particularly more hazel (tree abundance hazel) and hawthorn (tree abundance hawthorn) and with more understorey cover by hazel (understorey hazel cover %), and at intermediate to high levels of tree and understorey canopy closure above the nestbox (values above ~ 87% cover resulted in lower probability of occupancy. Fig. 3). Occupancy was also more likely in areas located further from footpaths and slightly more likely when away from woodland margins both of which may be sources of disturbance, and for nestboxes located nearer (within 10–15 m) and furthest (> 45 m) from others with a lower occupancy probability at intermediate distances (Fig. 3).

From September 2021 to October 2023 our surveys identified 16 nestboxes as occupied and 29 as unoccupied. The simplified model based on habitat data collected in the summer of 2021, correctly predicted occupancy for 10 of the 16 occupied nestboxes (37.5% false negatives, specificity = 62.5%) and for 20 of the 29 unoccupied nestboxes (31.0% false positives, sensitivity = 69.0%), resulting in an overall 33.3% error rate (model accuracy = 66.7%). Predictions based on the complete model with all variables were identical.

Discussion

Random Forest models allow consideration of multiple predictors even with small sample sizes to explore how fine-scale microhabitat features associate with artificial nestbox occupancy by hazel dormice. For example, in our case study we tested 76 predictors with a sample size of 45 nestboxes. Understanding which features are most relevant and how they affect hazel dormice occupancy can guide the placement of nestboxes for conservation actions within selected sites and

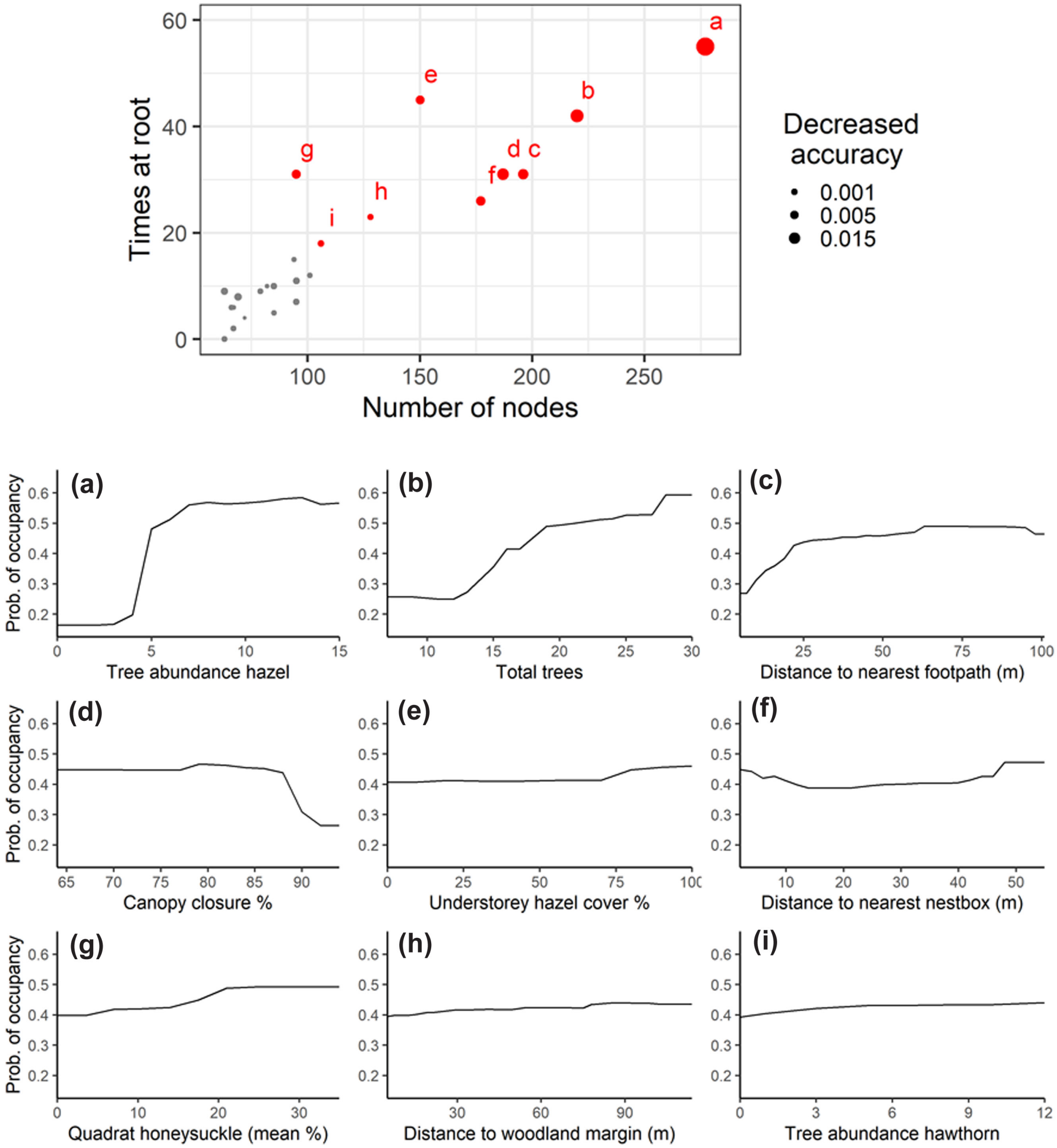

Figure 3. Variable importance and dependence plots for the top selected variables linking microhabitat to nestbox occupancy (probability of occupancy) by hazel dormice *M. avellanarius* in a UK woodland. Top large panel shows the 24 variables with binomial test p-values < 0.05 with values for the three importance metrics that were less correlated in their ranking (chosen to showcase differences in variable importance among metrics. See Supporting information for correlations among importance metrics). Bottom panels (labelled a to i) show changes in predicted probability of nestbox occupancy by hazel dormice for the nine most relevant predictors (red symbols on the top panel) in descending order of variable importance (from left to right, top to bottom). The three displayed metrics are: decreased accuracy (mean decrease in prediction accuracy after the variable is permuted), number of trees (total number of trees in which the variable is used), times at root (total number of trees in which the variable is used for splitting the root node). These three metrics are used for displaying purposes, but all seven metrics (described in the methods) were considered to identify the most important variables shown in red colour and labelled with letters that correspond to the bottom dependence plot panels.

suggest specialised woodland management to promote features encouraging use of already installed nestboxes by hazel dormice. In our application of Random Forest models to a UK woodland, we highlight necessary steps associated with tuning hyperparameter values and the comprehensive evaluation of diverse importance metrics to identify key predictors. We also define a simplified model, in a step that could be particularly valuable for future data collection as this model would require measuring few key variables that could allow evaluation and prediction for a large sample size of nestboxes.

While Random Forest models can be useful, consideration of some steps and potential limitations is necessary. Improved inferences can be achieved by tuning hyperparameters and considering diverse variable importance metrics, but these steps require additional functions and coding that may be challenging. We provide the R script for our analyses which could be adapted to other areas and datasets. In addition, the data needed to describe the system and to test predictions needs consideration. For example, in our case study we had low OOB error rates that suggest good predictive power. For the simplified model the error was equivalent to being able to correctly predict occupancy for nearly 9 out of 10 (13.3% error, 86.7% accuracy). These low errors reflect how well habitat data collected in 2021 could predict occupancy in the current and previous three years (2018–July 2021). However, error rates were much higher (equivalent to about 3 out of 10 wrong) when we tried to predict occupancy into the future based on habitat variables measured in 2021. This additional error may reflect temporal changes in microhabitat conditions, dormice population abundance fluctuations, and/or variation in abundance of natural nesting sites that can encourage or reduce the use of artificial nestboxes. Future work should explore this variability. In addition, it would be interesting to test how well our model predicts occupancy in other sites. This would provide an evaluation on whether the identified microhabitat variables, discussed below, are important generally, or if there is variation across sites that would require site-specific models and management strategies.

In our UK woodland case study, nestboxes were more likely to be occupied by hazel dormice in sites with higher abundance of key vegetation resources (hazel, hawthorn, and honeysuckle), near more trees, where canopies were not fully closed, and when located further from disturbances (footpaths and woodland margins). Although hazel dormice have a flexible diet (Eden 2009), proximity to preferred resources is likely beneficial. Hazel, honeysuckle, and hawthorn have all been previously recognised as important food and nesting resources for hazel dormice (Richards and Hurrell 1984, Prentice and Prentice 1988, Eden 2009, Tooke and Battey 2010, Bracewell and Downs 2017). Tree structure is also likely important because hazel dormice are primarily arboreal, travelling up to 152 m in search of food (Bright and Morris 2009). Our model predicted more than double probability of occupancy when the relative abundance of trees goes from 10 to 30 within 10 m of a nestbox. Previous research has also reported positive effects of tree diversity and abundance on nest box occupancy (Bright and Morris 1990, Juškaitis and Augutė 2008, Mortensen et al. 2022). However, in our study area very high tree canopy and understorey cover (> 85%) were associated with reduced occupancy, with an apparent optimal around 80–85%. This previously unreported effect might have been missed in studies using coarser density indices (Mortensen et al. 2022). Very closed canopies may prevent understorey plant growth and result in lower temperatures that can affect dormice (Goodwin et al. 2018a). Finally, in our study site, higher occupancy was associated with lower disturbance. Although hazel dormice do not completely avoid disturbed sites (Schulz et al. 2012), our study site is a well-visited location, especially at weekends and in the Spring and Summer when hazel dormice are most active. Nestboxes located closer to the footpaths and woodland margins are likely exposed to higher noise levels and potential human disturbance (e.g. people trying to look inside nestboxes, agricultural machinery).

Active management of vegetation already occurs in our study area including hazel coppicing, which combined with the relocation of some nestboxes to areas further away from existing paths and/or some consideration of managing visitor access to paths is likely to be beneficial. For relocations, consideration of distances among nestboxes could also be important. In our study area, nestbox occupancy was greater within clusters of nestboxes (within 10–15 m) and when located further away from each other (> 45 m). Future work is needed to determine if lower occupancy at intermediate distances occurs in other sites and habitats, and how it relates to individual use. This work could also test whether availability of several close-by nestboxes is beneficial for dormice because it allows individuals to remain within a suitable home range area even if competing for nesting sites with other species (Lang et al. 2022).

In conclusion, our study shows how machine learning methods can help address the knowledge gap of how microhabitat features affect nestbox occupancy by hazel dormice. We tested the approach in one woodland location in the UK. Future work in other areas and habitats is still needed. Moreover, additional information is needed in our study site to facilitate the recovery of the hazel dormouse. For example, despite collecting data on dozens of plant species during our vegetation surveys, dormice occupancy seems to be influenced by few key plants. Measuring microhabitat features requires working closely to nestboxes, and thus, to minimize disturbance we completed these during the scheduled monthly monitoring. More frequent surveys may identify rarer but potentially important plants or seasonal changes we were unable to monitor. In addition, our occupancy time-series did not allow analysis of temporal patterns in detail, but it would be interesting to study potential lag effects and temporal changes in habitat and occupancy. Research on variation among individual dormice in their preferences will also be valuable. Marking individual dormice using pit-tags and placement of camera traps near boxes could be used to understand temporal and individual patterns of nestbox use. While we wait for this additional understanding, our results offer some insight into suitable statistical methods that may be applied. We also identify some of the microhabitat variables that influence hazel dormice occupancy of nestboxes in a UK woodland. This information can be useful to guide placement and local scale management to promote conservation of this little mammal.

Acknowledgements – We thank the members and volunteers of the Berkshire Mammal Group for their efforts to monitor and maintain dormice nestboxes in Basildon Park House. Ian White and two anonymous reviewers provided valuable feedback to the manuscript. A previous version of this manuscript was submitted as a thesis in fulfilment of a University of Reading MSc requirement.
Funding – No funding was received to support this project. This research was conducted as part of a MSc programme.

Permits – In compliance with UK regulations, nestboxes and vegetation surveys were conducted in the presence of a member of the BMG with a Natural England dormouse class licence (Dr Amanda Lloyd 2016-21177-CLS-CLS or Ms Debbie Cousins 2016-21346-CLS-CLS).

Author contributions

Joe Malyan: Conceptualization (equal); Data curation (lead); Formal analysis (equal); Investigation (lead); Methodology (lead); Visualization (equal); Writing – original draft (lead). **Amanda J. Lloyd:** Conceptualization (equal); Investigation (supporting); Methodology (equal); Project administration (equal); Supervision (supporting); Writing – review and editing (supporting). **Manuela González-Suárez:** Conceptualization (equal); Data curation (supporting); Formal analysis (equal); Investigation (supporting); Methodology (equal); Project administration (equal); Supervision (lead); Visualization (equal); Writing – review and editing (lead).

Transparent peer review

The peer review history for this article is available at https://publons.com/publon/10.1002/wlb3.01185.

Data availability statement

Data are available from Dryad: https://doi.org/10.5061/dryad.1c59zw43q (Malyan et al. 2024).

Supporting information

The Supporting information associated with this article is available with the online version.

References

Berg, L. and Berg, Å. 1998. Nest site selection by the dormouse *Muscardinus avellanarius* in two different landscapes. – Ann. Zool. Fenn. 35: 115–122.

Bracewell, M. and Downs, N. 2017. Hazel dormouse (*Muscardinus avellanarius*) nest material preferences and collection distances, in southern England. – Mammal Com. 3: 1–10.

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

Bright, P. and Morris, P. 1990. Habitat requirements of dormice (*Muscardinus avellanarius*) in relation to woodland management in southwest England. – Biol. Conserv. 54: 307–326.

Bright, P. and Morris, P. 1994. Dormouse distribution: survey techniques, insular ecology and selection of sites for conservation. – J. Appl. Ecol. 31: 329–339.

Bright, P. and Morris, P. 1995a. A review of the dormouse (*Muscardinus avellanarius*) in England and a conservation programme to safeguard its future. – Hystrix Ital. J. Mammal. 6: 295–302.

Bright, P. and Morris, P. 1995b. A review of the dormouse (*Muscardinus avellanarius*) in England and a conservation programme to safeguard its future. – Hystrix Ital. J. Mammal. 6: 295–302.

Bright, P. and Morris, P. 2002. Putting dormice back on the map. – Br. Wildl. 14: 91–100.

Bright, P. and Morris, P. 2008. Why are dormice rare? A case study in conservation biology. – Mamm. Rev. 26: 157–187.

Bright, P. and Morris, P. 2009. Ranging and nesting behaviour of the dormouse, *Muscardinus avellanarius*, in diverse low-growing woodland. – J. Zool. 224: 177–190.

Bright, P., Morris, P. and Mitchell-Jones, T. 2006. The dormouse conservation handbook. – English Nature.

Capizzi, D., Battistini, M. and Amori, G. 2002. Analysis of the hazel dormouse, *Muscardinus avellanarius*, distribution in a Mediterranean fragmented woodland. – Ital. J. Zool. 69: 25–31.

Cartledge, E. L., Baker, M., White, I., Powell, A., Gregory, B., Varley, M., Hurst, J. L. and Stockley, P. 2021. Applying remotely sensed habitat descriptors to assist reintroduction programs: a case study in the hazel dormouse. – Conserv. Sci. Pract. 3: e544.

Caryl, F. 2021. Pine marten diet and habitat use within a managed coniferous forest. – PhD thesis, Univ. of Stirling, UK.

Coomber, F. G., Smith, B. R., August, T. A., Harrower, C. A., Powney, G. D. and Mathews, F. 2021. Using biological records to infer long-term occupancy trends of mammals in the UK. – Biol. Conserv. 264: 109362.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. and Lawler, J. J. 2007. Random forests for classification in ecology. – Ecology 88: 2783–2792.

de Raad, L., Lurz, P. and Kortland, K. 2021. Managing forests for the future: balancing timber production with the conservation of Eurasian red squirrel (*Sciurus vulgaris*). – For. Ecol. Manage. 493: 119164.

Dietz, M., Büchner, S., Hillen, J. and Schulz, B. 2018. A small mammal's map: identifying and improving the large-scale and cross-border habitat connectivity for the hazel dormouse *Muscardinus avellanarius* in a fragmented agricultural landscape. – Biodivers. Conserv. 27: 1891–1904.

Eden, S. 2009. Living with dormice: the common dormouse: real rodent or phantom of the ancient world. – Papadakis.

Fedyń, I., Figarski, T. and Kajtoch, Ł. 2021. Overview of the impact of forest habitats quality and landscape disturbances on the ecology and conservation of dormice species. – Eur. J. For. Res. 140: 511–526.

Goodwin, C., Suggitt, A., Bennie, J., Silk, M., Duffy, J., Al-Fulaij, N., Bailey, S., Hodgson, D. J. and McDonald, R. A. 2018a. Climate, landscape, habitat, and woodland management associations with hazel dormouse *Muscardinus avellanarius* population status. – Mamm. Rev. 48: 209–223.

Goodwin, C. E. D., Hodgson, D. J., Bailey, S., Bennie, J. and McDonald, R. A. 2018b. Habitat preferences of hazel dormice *Muscardinus avellanarius* and the effects of tree-felling on their movement. – For. Ecol. Manage. 427: 190–199.

Greenwell, B. M. 2017. pdp: an R package for constructing partial dependence plots. – R J. 9: 421–436.

Hopkins, J. J. and Kirby, K. J. 2007. Ecological change in British broadleaved woodland since 1947. – Ibis 149: 29–40.

Juškaitis, R. 2014. Summer mortality in the hazel dormouse (*Muscardinus avellanarius*) and its effect on population dynamics. – Acta Theriol. 59: 311–316.

Juškaitis, R. and Augutė, V. 2008. Habitat requirements of the common dormouse (*Muscardinus avellanarius*) and the fat dormouse (*Glis glis*) in mature mixed forest in Lithuania. – Ekol. Bratisl. 27: 143–151.

Juškaitis, R., Balčiauskas, L. and Šiožinytė, V. 2013. Nest site selection by the hazel dormouse *Muscardinus avellanarius*: is safety more important than food? – Zool. Stud. 52: 53.

Lang, J., Bräsel, N., Beer, S. M., Lanz, J. D., Leonhardt, I. and Büchner, S. 2022. The battle about the box: competition as the main factor behind the choice for resting sites of hazel dormice. – Mammalia 86: 351–354.

Liaw, A. and Weiner, M. 2002. Classification and regression by randomForest. – R News 2: 18–22.

Madikiza, K., Bertolino, S., Baxter, R. and Do Linh San, E. 2010. Nest box use by woodland dormice (*Graphiurus murinus*): the influence of life cycle and nest box placement. – Eur. J. Wildl. Res. 56: 735–743.

Malyan, J., Lloyd, A. J. and González-Suárez, M. 2024. Data from: Machine learning identification of microhabitat features associated with occupancy of artificial nestboxes by hazel dormice (*Muscardinus avellanarius*) in a UK woodland site. – Dryad Digital Repository, https://doi.org/10.5061/dryad.1c59zw43q.

Mitchell-Jones, T. and White, I. 2009. Using reintroductions to reclaim the lost range of the dormouse, *Muscardinus avellanarius*, in England. – Folia Zool. 58: 341–348.

Morris, P., Bright, P. and Woods, D. 1990. Use of nestboxes by the dormouse *Muscardinus avellanarius*. – Biol. Conserv. 51: 1–13.

Mortelliti, A., Santarelli, L., Sozio, G., Fagiani, S. and Boitani, L. 2013. Long distance field crossings by hazel dormice (*Muscardinus avellanarius*) in fragmented landscapes. – Mamm. Biol. – Zeitschrift fur Saugetierkunde 78: 309–312.

Mortensen, R. M., Fuller, M. F., Dalby, L., Berg, T. B. and Sunde, P. 2022. Hazel dormouse in managed woodland select for young, dense, and species-rich tree stands. – For. Ecol. Manage. 519: 120348.

Occhiuto, F., Mohallal, E., Gilfillan, G. D., Lowe, A. and Reader, T. 2021. Seasonal patterns in habitat use by the harvest mouse (*Micromys minutus*) and other small mammals. – Mammalia 85: 325–335.

Paluszynska, A., Biecek, P. and Jiang, Y. 2020. randomForestExplainer: explaining and visualizing random forests in terms of variable importance, ver. 0.10.1. – https://github.com/ModelOriented/randomForestExplainer.

Panchetti, F., Sorace, A., Amori, G. and Carpaneto, G. M. 2007. Nest site preference of common dormouse (*Muscardinus avellanarius*) in two different habitat types of central Italy. – Ital. J. Zool. 74: 363–369.

People's Trust for Endangered Species. 2021. 1000th dormouse released in Britain. – https://ptes.org/1000th-dormouse-released-in-britain/.

Phillips, B. B., Crowley, S. L., Bell, O. and McDonald, R. A. 2022. Harnessing practitioner knowledge to inform the conservation of a protected species, the hazel dormouse *Muscardinus avellanarius*. – Ecol. Solut. Evid. 3: e12198.

Prentice, A. M. and Prentice, A. 1988. Energy costs of lactation. – Annu. Rev. Nutr. 8: 63–79.

Ramakers, J., Dorenbosch, M. and Foppen, R. 2014. Surviving on the edge: a conservation-oriented habitat analysis and forest edge manipulation for the hazel dormouse in the Netherlands. – Eur. J. Wildl. Res. 60: 927–931.

Reid, C., Hornigold, K., McHenry, E., Nichols, C., Townsend, M., Lewthwaite, K., Elliot, M., Pullinger, R., Hotckiss, A., Gilmartin, E., White, I., Chesshire, H., Whittle, L., Garforth, J., Gosling, R., Reed, T., Hugi, M. and Downey, H. 2021. State of the UK's Woods and trees 2021. – Woodland Trust.

Richards, C. G. J. and Hurrell, E. 1984. The food of the common dormouse (*Muscardinus avellanarius*), in South Devon. – Mamm. Rev. 14: 19–28.

Schulz, B., Ehlers, S., Lang, J. and Büchner, S. 2012. Hazel dormice in roadside habitats. – Peckiana 8: 49–55.

Tooke, F. and Battey, N. H. 2010. Temperate flowering phenology. – J. Exp. Bot. 61: 2853–2862.

Wembridge, D., White, I., Freegard, K., Al-Fulaij, N. and Langton, S. 2023. The state of Britain's dormice 2023. – People's Trust for Endangered Species.