Predicting the distribution of ancient and other noteworthy trees across the UK

A thesis submitted for the degree of

Doctor of Philosophy

The Faculty of Life Sciences at University of Nottingham, Nottingham, UK

in association with the Woodland Trust, Grantham, UK

by

Victoria Jayne Nolan

March 2021

Supervisors

Prof. Tom Reader, University of Nottingham, UK

Prof. Francis Gilbert, University of Nottingham, UK

Dr. Nick Atkinson, Woodland Trust, Grantham, UK

Contents

Common abbreviations 3

Acknowledgements 4

Chapter contributions and publication status 5

List of tables 8

List of figures 9

Abstract 11

Chapter 1: Introduction 12

Chapter 2: An initial analysis of the Ancient Tree Inventory (ATI) and an introduction to the other datasets used in the thesis**.** 22

Chapter 3: Historical maps confirm the accuracy of zero-inflated model predictions of ancient tree abundance in English wood-pastures**.** 51

Chapter 4: Identifying predictors of sampling bias in the Ancient Tree Inventory (ATI) in England**.** 73

Chapter 5: Solving sampling bias problems in presence-absence or presence-only species data using zero-inflated models**.** 86

Chapter 6: Distribution models calibrated with independent field data predict two million ancient and veteran trees in England**.** 123

Chapter 7: Assessing the use of landscape metrics in Species Distribution Modelling: a case study using the UK Ancient Tree Inventory (ATI). 158

Chapter 8: Discussion and Conclusion. 178

References 186

Appendices.....207

A2: Appendix 2 – Additional tables and figures from Chapter 2 207

A3: Appendix 3 – Additional tables and figures from Chapter 3 221

A5.1: Appendix 5.1 - Additional tables and figures from Chapter 5.....224

A5.2: Appendix 5.2 - Simulation methods and results using average temperature as an alternative biological predictor.....230

A5.3: Appendix 5.3 - Derivation of D: 'Deviation from the best model'.....233

A6.1: Appendix 6.1 - Species distribution model parameter tuning and evaluation of alternative methods of splitting of training and test data. 234

A6.2: Appendix 6.2 – Example survey form and instructions for field volunteers 239

A6.3: Appendix 6.3 – Additional figures from Chapter 6 244

Common abbreviations

ATF	Ancient Tree Forum
ATI	Ancient Tree Inventory
DBH	Diameter at Breast Height
ENM	Ecological Niche Modelling
GLM	Generalised Linear Model
NB	Negative binomial
NPPF	National Planning Policy Framework
NTM	National Tree Map
SDM	Species Distribution Modelling
TROBI	Tree Register of the British Isles
TPO	Tree Preservation Order
WT	Woodland Trust
ZI	Zero-Inflated

Acknowledgements

Firstly I would like to thank my supervisors, Tom Reader and Francis Gilbert, and all my other Life Sciences colleagues at the University of Nottingham, for their invaluable advice, help and support throughout my PhD.

Thanks also to my two internal assessors, Markus Eichhorn and Richard Field who provided many helpful comments on drafts of manuscripts and reports. Thanks to Richard especially for his help with Chapter 5 of this thesis and his advice and comments on parts of this section, and also Tim Newbold (UCL) who provided a helpful review of the material and theory behind this chapter. I would finally like to thank Rose Saikayasit and Keith Evans (University of Nottingham) for their support and help with getting set up on and using the university HPC system.

I am deeply grateful for the support provided by the Woodland Trust throughout my PhD, not least to Nick Atkinson, Tom Reed, Kylie Harrison-Mellor, Jill Butler, Karen Hornigold, Kate Lewthwaite, Emily James, Emma Gilmartin, Christine Tansey and all current and former Woodland Trust staff who have assisted with the project in some way. In particular I would like to thank Karen Hornigold for her help organising and completing the historical desk verification for Chapter 3 and the two volunteers, Victoria Willets and Olwyn Spencer, who digitised the historic maps. I would also like to thank Tom Reed for his amazing help over the past year with establishing and completing the field surveys.

I would finally like to acknowledge the amazing efforts of all of the volunteers who carried out the field surveys in 2020 and greatly assisted with the collection of the new data.

This research was financially supported jointly by the University of Nottingham and the Woodland Trust.

Chapter Contributions and Publication Status

Sections of Chapter 1 and 2 are adapted from: Nolan, V., Reader, T., Gilbert, F. et al. The Ancient Tree Inventory: a summary of the results of a 15 year citizen science project recording ancient, veteran and notable trees across the UK. Biodiversity and Conservation, 29, 3103–3129 (2020). https://doi.org/10.1007/s10531-020-02033-2. The main text of this paper (and all text in Chapters 1 and 2), as well as all statistical analysis was written and carried out by V. Nolan with contributions and minor edits from other authors (T. Reader, F. Gilbert and N. Atkinson). Material presented in Chapters 1 and 2 of the thesis has undergone little change from the manuscript other than slight rewording, with the exception of the addition of the 'Aims of Thesis' section in Chapter 1 and 'Other datasets used in thesis' section in Chapter 2. Other differences include page layout and the use of 'I' rather than 'we'.

Chapter 3 is adapted from a manuscript, currently under peer review at the Journal of Applied Ecology (first submitted Nov 2020, major revision submitted April 2021). The conceptual idea, modelling and analysis and initial draft were prepared by V. Nolan, with help and minor edits from other authors (T. Reader, F. Gilbert and N. Atkinson). N. Atkinson also helped with the conceptual design of the historic desk verification work. Material presented in the thesis differs significantly from the most recent manuscript, including omission of the analysis of genus and much of the results from the Poisson model, significant text rewording, model simplification additions and analysis of coefficient magnitudes. Other differences include page layout and the use of 'I' rather than 'we'.

Chapter 4 and 7 were conceptualised by V. Nolan, who also carried out the analysis and initial chapter draft. Edits to the draft and assistance with some statistical concepts was provided by T. Reader and F. Gilbert.

Chapter 5 is adapted from a manuscript currently undergoing the third review process at the Journal of Biogeography, following recommendations for a final minor revision (submitted Feb, 2021, minor revisions submitted May 2021). The idea for this manuscript was designed by V. Nolan with assistance

from T. Reader and F. Gilbert. All analysis was carried out by V. Nolan as well as the preparation of the initial draft. Revisions to the draft and assistance with the major revision were provided by T. Reader and F. Gilbert. Differences between Chapter 5 and the most recently submitted manuscript are small, with only minor edits to several figures, some rewording and a couple of typographical errors. The only other differences include page layout and the use of 'I' rather than 'we'.

Chapter 6 is adapted from a manuscript currently submitted to Ecological Applications (Feb, 2021). Statistical analysis and modelling, and preparation of the initial draft was carried out by V. Nolan. Minor edits were made by T. Reader, F. Gilbert and T. Reed (Woodland Trust) to subsequent versions of the manuscript. All material in Chapter 6 is identical to the submitted manuscript except for page layout and the use of 'I' rather than 'we'.

Chapter 8 was written by V. Nolan, with edits to the draft provided by T. Reader and F. Gilbert.


about
excess
more
habitat
occurrence
study
distribution
records
verification
GLM
model
using
across
Trust
Although
Fig
often
true
km
predictions
both
bias
being
grid
zero
only
occurrences
Therefore
ZJ
trees
most
ATJ
modelling
likely
veteran
predictor
count
many
Table
location
two
components
maps
other
survey
use
each
used
one
ancient
species
type
high
England
method
SDM
all
notable
Chapter
common
zero-inflation
National
land
biased
different
recorded
UK
number
within
data
tree
altitude
large
distance
wood-pastures
Numeric
area
methods
areas
nearest
models
between
alet
sampling
some
information
based
conservation
spatial
predicted
environmental
presence-absence
woodland
record

List of tables

Table 1.1 Definitions and distinctions between different terms used when discussing ancient and other noteworthy trees.

Table 2.1 Information collected about each Ancient Tree Inventory (ATI) record.

Table 2.3 Wald Chi-Squared ANOVA results to test for parameter significance from a multinomial logistic regression model of category (ancient, veteran and notable tree) in relation to Ancient Tree Inventory (ATI) characteristics.

Table 2.4 The ATI star rating system and the number of records within each group.

Table 2.5 Additional data-sets used throughout this thesis selected based on their potential as a predictor of ancient, veteran and notable tree distributions across the UK.

Table 3.1 The 21 variables describing wood-pasture characteristics used as predictors in statistical models of ancient tree abundance.

Table 3.2 Map series used for the historical desk verification of model predictions.

Table 3.3 Estimates of the abundance of ancient trees from a) Zero-Inflated (ZI) model predictions and b) based on the historical desk verification estimates.

Table 3.4 Spearman's rank correlations (r_s) between the predicted ancient tree abundance from the zero-inflated Poisson (ZIP) or negative binomial model (ZINB), and the verification estimates for 60 selected wood-pastures in England.

Table 4.1 Potential predictors of bias in the Ancient Tree Inventory (ATI).

Table 4.2 Spearman's Rank correlation coefficients (r_s) between the value of each sampling bias predictor and record abundance per 1-km^2 grid square.

Table 5.1 Pearson's correlation coefficient (r) between altitude or altitude_randomised (biological predictors) and distance from the nearest town (bias predictor) across the 10 maps with randomly generated sets of 'town centre' locations.

Table 5.2 Sources of zero-inflation in the simulated species occurrence data.

Table 5.3 Ten predictor combinations considered when modelling the simulated species distributions.

Table 6.1 Information from 20 datasets (see Table 2.5) was collected for each 1-km^2 grid cell, and converted into a useable quantitative model predictor. There are 16 continuous predictors and 4 categoric predictors.

Table 6.2 Types of bias correction method applied to the Ancient Tree Inventory (ATI) records when modelling the distribution of ancient and veteran trees across England.

Table 6.3 Model coefficients (± standard error), Z value and p value of significance are shown for the negative binomial ZI model for both the count and zero components.

Table 6.4 Independent field evaluation of model predictions.

Table 6.5 Permutation importance of each of the Maximum Entropy distribution model predictors shown for the model with no bias correction compared to the overall best performing bias corrected model using systematic sampling (SS) at a 2-km resolution.

Table 7.1 The original 16 landscape metrics calculated for each grid square in England based on the National Tree Map.

Table 7.2 Independent field evaluation of model predictions.

List of figures

Figure 1.1 The three main phases of the ageing process of trees.

Figure 2.1 Distribution maps of ancient, veteran, notable and other records in the Ancient Tree Inventory.

Figure 2.2 The percentage contribution of the 12 most common genera to the total number of records in the Ancient Tree Inventory (ATI).

Figure 2.3 The relative proportion of ATI records between three tree categories (ancient, veteran and notable) shown across a) country of record, b) tree form and c) both country and tree form.

Figure 2.4 The relative proportion of ATI records between three tree categories (ancient, veteran and notable) shown across a) the 12 most common genera and b) the 12 most common genera and tree form

Figure 2.5 Mean measured girth (m) of trees recorded in the Ancient Tree Inventory (ATI) shown for three tree categories (ancient, veteran and notable) across country and tree form of record.

Figure 2.6 Standardised Pearson residuals (r) from the Chi-square test of association between habitat and the 12 most common genera (left) and ancient, veteran or notable category (right) in the ATI.

Figure 2.7 Standardised Pearson residuals (r) from the Chi-square test of association between threat and ancient, veteran or notable category (left) and the 12 most common genera (right) in the ATI.

Figure 3.1 Distribution of all wood-pasture in England (as mapped by Natural England in Wood Pasture and Parkland BAP Priority Habitat Inventory for England).

Figure 3.2 Hanging rootograms to visualise the fit of the zero-inflated Poisson (ZIP) and negative binomial (ZINB) models to the ancient tree abundance data in English wood-pastures.

Figure 3.3 Evaluation of abundance predictions from the zero-inflated Poisson (ZIP) and negative binomial model (ZINB).

Figure 3.4 Mean number of ancient trees per wood-pasture across each categorical predictor.

Figure 4.1 Left: Ancient and veteran tree records across England from the Ancient Tree Inventory (ATI). Right: Ancient and veteran tree record abundance (counts of records) per 1-km grid square.

Figure 4.2 Left: Centroid locations (red dots) and kernel density plots of all ATI records uploaded by each individual recorder or organisation. Right: centroid locations and kernel density plots of all the records uploaded by each of the 10 most active individual recorders.

Figure 4.3 Scatterplots of the relationships between ancient and veteran tree abundance from the Ancient Tree Inventory (ATI) and each of the 12 numeric sampling bias predictors for each 1-km grid cell across England.

Figure 4.4 Number of ancient or veteran tree records in the Ancient Tree Inventory (ATI) per km^2 of each land type.

Figure 5.1 Simulation study area consisting of a group of 100 x 100 grid squares of 1 km^2 size randomly placed within England covering a total area of 10,000 km2 (outlined in red) (left), with the biological predictors: altitude (m) and altitude_randomised (m) (randomised altitude layer with no spatial autocorrelation) shown for the study area (right).

Figure 5.2 A simulated species with 5,000 occurrence points showing a) no preference for altitude (random species), b) a preference for high altitudes based on a logarithmic scaler of altitude (altitude species), and c) a preference for high altitudes based on a logarithmic scaler of altitude_randomised (altitude_randomised species).

Figure 5.3 Evaluation of abundance predictions (based on D = 'deviation from the best model') for a hypothetical organism with occurrences simulated based on a preference for high altitudes (altitude species).

Figure 5.4 Example maps of abundance for a hypothetical species ('altitude species') whose occurrence is positively influenced by altitude, produced from two generalised linear models (GLMs) and two Zero-Inflated (ZI) models.

Figure 5.5 Example maps of abundance for a hypothetical species ('altitude_randomised species') whose occurrence is positively influenced by a randomised altitude layer, produced from two generalised linear models (GLMs) and two Zero-Inflated (ZI) models.

Figure 5.6 Comparisons of model predictive power of a) count abundance (top panels) or b) sampling abundance (bottom panels) between a generalised linear model (GLM) and two zero-inflated (ZI) models across varying levels of biological and sampling bias zero-inflation.

Figure 5.7 Spearman's Rank correlation coefficients (rs) between the model predictors (altitude and distance from nearest town) and model predictions for altitude species across three modelling resolutions: 1-km, 2-km and 5-km.

Figure 5.8 Evaluation of MaxEnt, generalised linear model (GLM) and zero-inflated (ZI) model predictions of altitude species presence-absence across the study area based on mean Area under the Curve (AUC) across three scales of data aggregation: 1-km, 2-km and 5-km.

Figure 6.1 Centroid locations of each of the 90 1-km grid squares selected for field verification.

Figure 6.2 a. Corrected Akaike's Information Criterion (AICc), b. Training Area Under the Curve (AUC) and c. Testing Area Under the Curve (AUC) for each species distribution model of ancient and veteran tree distribution across England using four main types of bias correction method (spatial filtering, background restriction, bias files and ZI models).

Figure 6.3 Histogram of the estimated percentage coverage of each grid square during the field surveys.

Figure 6.4 Predicted distribution maps of habitat suitability for ancient and veteran trees across England.

Figure 6.5 Predicted maps of the abundance of ancient and veteran trees across England from the Poisson and negative binomial (NB) zero-inflated (ZI) models.

Figure 7.1 Two 1-km^2 grid squares from Suffolk, England with the overlaid National Tree Map canopies.

Figure 7.2 Percentage of explained variance of each principal component dimension shown from the Principal Component Analysis (PCA) of all 16 landscape metrics.

Figure 7.3 a. Direction and contribution of each landscape metric to Principal Components (PCs) 1 and 2 from the Principal Component Analysis (PCA) of all 16 original landscape metrics. b. Plot of each individual grid square coloured by ancient or veteran tree presence on the axes of PC1 and PC2 from the PCA of all 16 original landscape metrics.

Figure 7.4 Scatterplots of ancient and veteran tree abundance in relation to the seven final landscape metrics.

Figure 7.5 Evaluation of the performance and predictive power of Maximum Entropy models of ancient and veteran tree distributions using environmental predictors (Environment), Principal Components (PCs) of landscape metrics, and seven landscape metrics (Metrics), as well as two combinations of these predictor sets.

Figure 7.6 Distribution maps of ancient and veteran trees across England produced from Maximum Entropy models using five combinations of predictors: a) environmental predictors, b) seven original landscape metrics, c) Principal Components (PCs) of the landscape metrics, d) environmental + landscape metrics and e) environmental + landscape-metric PCs.

Abstract

Ancient, veteran and notable trees are ecologically important keystone organisms and have tangible connections to folklore, history and sociocultural practices. Although found worldwide, few countries have such a rich history of recording and treasuring these trees as the UK, which has resulted in the formation over the past 15 years of a large, comprehensive database of ancient and other noteworthy trees, the Ancient Tree Inventory (ATI). Although the ATI contains over 200,000 recorded trees, there are still thought to be many more that are undiscovered across the UK, and information about their status, condition and distribution is lacking. The primary aim of this thesis is to use the ATI to gain novel and detailed insights into the true distribution of ancient and veteran trees across the UK, important predictors of their presence, and key habitat types in which they are found. The ATI suffers many of the problems of large species databases, including sampling bias, which is a major focus of this thesis. To address this problem, sampling bias is first identified and quantified, and then established and novel bias correction methods are employed to improve predictions of ancient and veteran tree distributions. By combining mathematical models at various scales, from specific habitats to the whole of England, with additional independent data from desk and field surveys, robust accurate distribution maps of ancient and other noteworthy trees are produced and verified. The models suggest that wood-pasture is a particularly important habitat for ancient and veteran trees, and that their distributions are highly influenced by historical features of the environment and human factors. A key result emerging from multiple chapters of this thesis is the potentially large number of undiscovered ancient and veteran trees predicted across England: diverse alternative models produced similar and impressive total estimates of around two million trees. These results can be used to inform the conservation and protection of ancient trees, and highlight the need for more targeted surveying, tree planting and implementation of policy measures to ensure their persistence and survival into the future.

Chapter 1: Introduction

"It is not so much for its beauty that the forest makes a claim upon men's hearts, as for that subtle something, that quality of air that emanation from old trees, that so wonderfully changes and renews a weary spirit"

Robert Louis Stevenson (1875-6)

1.1 What are ancient trees?

Trees are thought to grow and age in three phases (White, 1998; Read, 2000; ATF, 2008a) (Fig. 1.1). First is formative growth occurring from seedling establishment until maturity, when there are rapid increases in crown spread, girth, height and leaf area. The mature phase is reached once the crown is at maximum size; this is generally after 40 – 100 years depending on the tree species (White, 1998). Finally the tree enters the ancient (or senescent) phase, where the characteristics associated with ancient or veteran trees emerge, including a hollowing trunk, holes and cavities, deadwood in the canopy, bark loss and the presence of fungi, invertebrates and other saproxylic organisms (Read, 2000; Rust and Roloff, 2002; ATF, 2008a; Owen & Alderman, 2008). Each phase length differs depending on environmental conditions, management techniques and tree species (Woodland Trust, 2001; Fay, 2002; ATF, 2008a; Owen & Alderman, 2008).

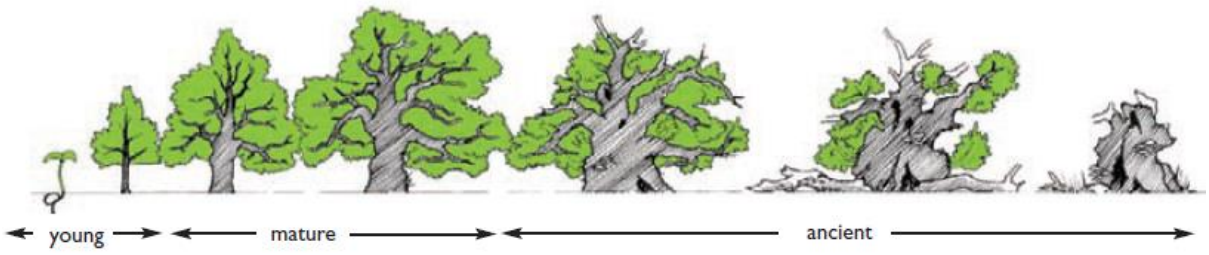


Fig. 1.1 *The three main phases of the ageing process of trees; young (formative), mature and ancient (senescent). The ancient phase can often be the longest phase, and even if a tree appears to be dead, it may have many years of life left, (ATF, 2008a).*

In the literature, the terms ‘veteran’, ‘notable’, ‘champion’, ‘large old’ and ‘heritage’ are often used interchangeably with ‘ancient’ (Read, 2000; Fay, 2002; Pautasso & Chiarucci, 2008; Lindenmayer et al., 2012), which has led to confusion about why a tree is of particular interest (ATF, 2008a). In the UK, the Woodland Trust (WT), one of the largest woodland and ancient tree conservation charities, recognised the need to separate and define these terms to provide clarity when classifying trees in relation to age, size or other characteristics (Woodland Trust, 2001; ATF, 2008a; Lonsdale, 2013) (Table 1.1). Subsequent uses of these terms in this thesis will follow the WT definitions. As there is overlap between the terms ‘ancient’ and ‘veteran’ (i.e. all ancient trees are also veteran trees), any use of the term ‘veteran’ in this thesis refers to only trees that are ‘non-ancient veterans’.

1.2 Value and importance of ancient and other noteworthy trees

Like all trees, trees showing ‘veteran characteristics’ contribute to ecosystem services such as carbon storage, water and microclimate regulation (Rubino & McCarthy, 2003; Lachat et al., 2013; Sist et al., 2014). They are also an important source of decaying and dead wood, a rare and declining habitat throughout Europe (Siitonen, 2001; Butler et al., 2002). Fungi are the main dead wood decomposers (Cooke, 1984; Boddy, 2001), and influence the creation of the hollowing trunk, crevices and water-filled pools that support a diverse range of saproxylic invertebrates, especially beetles (Speight, 1989; Seibold et al., 2018). It is estimated that 6% of British invertebrate species rely solely on decaying wood ecosystems (Alexander, 1999). Ancient trees also support a diverse range of epiphytes, including mosses, lichens and liverworts (Read, 2000; Butler et al., 2002; Ranius et al., 2008).

The cracks and crevices within the decaying branches and stumps of ancient trees are ideal for bats, and all 16 UK species are associated with ancient trees in some way (Rasey, 2004). Similarly, birds roost, nest and feed in the hollows and crevices of ancient and veteran trees. Reptiles and amphibians, in particular grass snakes (*Natrix natrix*), and mammals such as red squirrels (*Sciurus vulgaris*), hedgehogs (Erinaceinae sp.), bank voles *(Myodes glareolus*), wood mice (*Apodemus sylvaticus),* harvest mice (*Micromys minutus*), common dormouse (*Muscardinus avellanarius*) and even wildcat (*Felis silvestris*) make use of ancient and veteran tree habitats (Schmeller et al., 2009; Humphrey, 2005).

Many ancient and other noteworthy trees have famous cultural and historical ties, which present valuable recreational and tourism opportunities (Rackham, 1994; Lonsdale, 2013). One of the oldest UK trees, often reported to be around 2000-2500 years old, is the Ankerwycke Yew (*Taxus baccata*) in Berkshire, where King John is rumoured to have signed the Magna Carta in 1215 (Bevan-Jones, 2016). Other well-known trees include the Major Oak (*Quercus robur*) in Sherwood Forest, the most visited tree in the UK (Everett & Parakoottathil, 2018) which is associated with the story of Robin Hood; the most reliable estimates date it around 800 - 900 years old (Farjon, 2017). Outside England, the 300-year-old Scottish Birnam Oak (*Quercus petraea)* is thought to be a relic of Birnam Wood, famously mentioned in Shakespeare's Macbeth (Woodland Trust, undated).

Table 1.1 *Definitions and distinctions between different terms used when discussing ancient and other noteworthy trees according to the Woodland Trust position statement (2001) and the 'Ancient tree guide 4: What are ancient, veteran and other trees of special interest?' (ATF, 2008a).*

Term	Description
Veteran	Any tree showing 'veteran' characteristics (e.g. hollow trunk, crown retrenchment, crevices and the presence of saproxylic organisms). All ancient trees are veteran trees, but there are some younger trees also classed as veterans that show 'veteran' characteristics due to damage or disease. Veteran trees might also be classed as champion or heritage trees. Throughout this review, all references to a 'veteran tree' are in relation to only trees that are 'non-ancient veterans'.
Ancient	Any tree showing 'veteran' characteristics and that is older than most individuals of the same species. Age is estimated based primarily on girth (as in White, 1998), but also considering the environment and growing conditions of the tree. Approximate age-girth relationships are available for the most common UK tree species (ATF, 2008a). All ancient trees are veteran and heritage trees, and may or may not be champion trees.
Notable	The largest or tallest tree per species in a defined local area e.g. a park or garden. A notable tree has no obvious 'veteran' characteristics.
Champion	The tallest tree or the tree with the largest girth per species in the UK (or other region e.g. England). These trees may or may not be ancient, veteran or heritage trees.
Heritage	Trees with connections to historical or cultural events or trees that provide high aesthetic value. These trees may or may not also be ancient, veteran, notable or champion trees.

Ancient and ageing trees also offer insights into historical and cultural vegetation and land management techniques used in different areas, such as coppicing or pollarding. These techniques involve the periodic cutting of the trunk to just above ground level (coppicing) or breast height (pollarding), from which regrowth is harvested at intervals (Rackham, 1967; Rackham, 1994; Fuller & Warren, 1993; Petit & Watkins, 2003). Both methods can produce stools (coppices) or trunks (pollards) of extreme ages (Lewington, 2012). A tree that has never undergone either of these procedures is usually classed instead as a 'maiden tree' (single-stemmed tree), and often is not able to obtain the same longevity (Petit & Watkins, 2003). The use of these techniques varies spatially and temporally in the UK, and so can inform us about changes in management and landscaping practice (Read, 2000; Barnes et al., 2017). Ancient trees can also be historical relics of boundaries, hedgerows, commons, ancient woodlands and forests, avenues and ancient burial grounds (Stahle, 1996; Lonsdale, 2013, Farjon, 2017).

Additionally, although many ancient or veteran trees are hollow so that dendrochronological analysis of the trunk about tree age and condition is difficult, dead stumps or fallen branches can be used to show evidence of changes in temperature, water availability, disease outbreaks and mechanical damage over time (Kelly et al., 1992; Briffa, 2000; Cherubini et al., 2002; Ballesteros et al., 2010). Finally, ancient and ageing trees are not only relics of the past, but also important genetic resources for the future (Read, 2000; Lonsdale, 2013). Ancient trees display an unusual degree of phenotypic plasticity and have clearly demonstrated their ability to survive disease outbreaks and environmental stress by virtue of their age. These trees may harbour genes for pathogen resistance or stress tolerance (Major, 1967), which we might consider exploiting when planting the next generation of veteran and ancient trees.

1.3 Threats to ancient and ageing trees and associated organisms

Ancient and veteran trees are in global decline, with losses reported in Australia (Fischer et al., 2010), America (Gibbons et al., 2008), South America (Laurance et al., 2000) and Europe (Linder & Östlund, 1998; Jönsson et al., 2009). The key threats to the persistence and future of ancient tree populations are the lack of appropriate tree planting (Read, 2000) and elevated mortality (Gibbons et al., 2008; Le Roux et al., 2014) resulting from poor management e.g. the end of traditional techniques such as coppicing

and pollarding (Lonsdale, 2013), urbanisation, and the intensification of agricultural practices (Read, 2000; Fay, 2004; ATF, 2005). Increasing field sizes, soil compaction, over-grazing and fertiliser applications are all particularly detrimental to ancient trees and associated organisms (Read, 2000; Fay, 2004; ATF, 2005).

There is uncertainty around how ancient and veteran trees and their dependent species will be affected by climate change (Ranius, 2002; Jonsson et al., 2005; Ranius, 2006; ATF, 2008b). The dispersal abilities of saproxylic species in the face of climate change are uncertain (Jonsson et al., 2005; Ranius, 2006) and we may be at risk of losing these dead wood specialists (Sebek et al., 2013). Although ancient trees have shown their ability to survive over many past centuries, they may be less able to cope with rapid environmental and climate changes predicted in the future (Butler et al., 2002; ATF, 2008b). A further impact of climate change and globalisation is the spread of tree-associated diseases and pests (Brasier, 1996; Holdenrieder et al., 2004; Morin et al., 2007). Diseases such as Ash dieback (*Hymenoscyphus fraxineus)* have had devastating impacts on UK trees since 2000 (Mitchell et al., 2014).

1.4 Ancient and veteran trees in the UK

The UK ancient and veteran tree population is of global renown and interest, and there is a large amount of information about certain aspects of the trees including their management and associated arboriculture practices (Read, 2000; Fay, 2002, 2004; Lonsdale, 2013), particular sites with high numbers of ancient and ageing trees (Mountford & Peterken, 2003; Read et al., 2010; Hall & Bunce, 2011), particular genera such as Oak (*Quercus)* or Yew (*Taxus*) (Moir, 2013; Farjon, 2017) and their historic context in the UK landscape (Rackham, 1986, 1994; Fulford, 1995; Butler et al., 2002; Farjon, 2017). Yet despite all this, there is still a lack of consensus and discussion about the large-scale abundance and distribution of ancient, veteran and other noteworthy trees in the UK.

Particular sites that are most well-known for harbouring ancient and veteran trees are wood-pastures and historic parklands (Rackham, 1986, 1994; Hartel et al., 2013; Farjon, 2017). Wood-pastures are

generally characterised as an open, productive land-use type that combines livestock grazing with scattered, actively managed trees (Rackham, 1994; Quelch, 2002). The UK in particular is thought to have some of the highest concentrations of wood pastures in Europe (Rackham, 1994), possibly due to the continuity of land ownership (Butler et al., 2002): it is a recognised UKBAP priority habitat (BRIG, 2011). Another habitat that might include large numbers of ancient trees is ancient woodland (woodland that has existed since at least the 16th century and therefore unlikely to be of plantation origin: Peterken, 1977), yet this has undergone extensive conversion to plantation or other land uses across England and Wales since 1930, and was reported to cover a mere 2.6% of land in 1992 (Spencer & Kirby, 1992). Ancient trees are also found within farmland, in urban areas, as landscape boundaries, in tree avenues, on church grounds, in hedgerows or orchards and on private land or gardens (Rackham, 1994; Read, 2000; Woodland Trust, 2017), yet little is known about the distribution or state of these trees.

Although there is still uncertainty about the overall distribution and condition of ancient and veteran trees, the UK has substantially more information than other countries due to the long-term collation of tree records from a citizen-science project, the Ancient Tree Inventory (ATI) (https://ati.woodlandtrust.org.uk/). Other ancient tree inventories do exist covering a variety of geographical areas around the world. These range from global databases such as 'Monumental trees' (https://www.monumentaltrees.com) containing ~40,000 large, tall, old or notable trees across the world, to more localised regional databases such as the Remarkable Trees of the Brussels-Capital Region (http://bomen-inventaris.irisnet.be) which contains around ~5,800 records. Nevertheless, none of these datasets come close to the size or detail of the ATI. With over 200,000 trees recorded to date, the ATI provides an opportunity to extensively examine our current understanding of UK ancient and ageing tree distributions and condition, from which wider inferences about global ancient tree ecology can be made.

1.5 Aims of thesis

Citizen science data (i.e. data collected by members of the public) are usually stored in online databases, museums and herbariums, and are valuable resources of species records spanning large scales and long

time periods. In ecology, conservation and biogeography, this type of data is often difficult to collect due to financial, geographical and time constraints, so public databases such as the ATI are useful sources of species data for scientific research (Schmeller et al., 2009; Devictor et al., 2010; Tulloch et al., 2013). The ATI provides substantial information about the distribution, condition and attrition of ancient, veteran and notable trees across the UK.

Unlike other UK citizen science projects such as the British Trust for Ornithology's (BTO) Big Garden Bird Watch, or the UK Butterfly Monitoring Scheme, the ATI remains largely under-used and under-appreciated in the scientific community, despite its longevity and number of records. This is likely due to uncertainty regarding the reliability, usefulness and limitations of the ATI: issues that this thesis will evaluate and address. In Chapter 2 of this thesis I introduce the ATI database in detail and provide a descriptive and statistical summary of different components of the dataset in order to outline the current known status and distribution of UK ancient and other noteworthy trees. I also summarise some of the issues and problems that are potentially encountered when using the ATI for research, and attempting to infer about the current distribution of trees. I finally introduce several other common datasets that I use in subsequent chapters of the thesis.

A common method to investigate the true distribution of a species is Species Distribution Modelling (SDM), also called Ecological Niche Modelling (ENM). SDM has been successfully used to predict range shifts in relation to climate change (Beaumont et al., 2007; Chen et al., 2011), the spread of invasive species (Václavík & Meentemeyer, 2012) and has been useful in the development and deployment of many conservation projects (Clement et al., 2014; Mota-Vargas & Rojas-Soto, 2016). SDM is performed through the assessment of known species presence (and sometimes absence) records in relation to environmental or climatic variables. The suitability of locations for this species - their fundamental niche and geographic range - can then be predicted based on climate/ environmental characteristics of other locations (Araújo & Guisan, 2006; Hijmans & Graham, 2006; Mateo et al., 2011). There are a variety of modelling techniques available, with Maximum Entropy (MaxEnt)

modelling being by far the most widely used due to its ability to use presence-only data and to cope with small datasets (Hernandez et al., 2006; Phillips et al., 2006; Elith et al., 2006).

Distribution models of ancient, veteran and notable tree distributions across the UK using the ATI could provide insight into the true distribution of the trees and important environmental determinants of tree presence, as well as highlighting areas with high conservation potential, for example with high suitability for planting trees to become future ancient trees, or current hot-spots of ancient trees that need protecting. The next parts of this thesis use different approaches to SDM in order to produce the most robust, accurate distribution maps of ancient trees. In Chapter 3, I firstly take a targeted SDM approach by concentrating on a particular habitat with known connections to ancient trees: wood-pasture. By modelling ancient tree abundance in wood-pastures across England in relation to environmental, historic and anthropogenic predictors, wood-pastures with high numbers of undiscovered ancient trees are identified. In this chapter I also introduce my first novel method of model verification: using a series of historic Ordnance Survey (OS) maps over time to estimate tree abundance in randomly selected wood-pastures and verify model predictions.

One of the main problems with using citizen-science data such as the ATI in SDM is sampling bias (also called sample selection or survey bias), where certain temporal periods, geographical areas or taxa are sampled more intensively or frequently than others (Phillips et al., 2009; Dickinson et al., 2010; Bird et al., 2014). Sampling bias in SDM can lead to over- or under-estimation of important species-environment relationships (Syfert et al., 2013), meaning that predicted distribution maps may partly represent survey effort rather than species niche requirements (Phillips et al., 2009). Therefore, if sampling bias is not accounted for, predictive SDM maps of ATI records may not reflect the true distribution of the trees. Chapter 4 presents a detailed literature-focused discussion of this issue in SDM for scientific research and the advantages and disadvantages of alternative methods of bias correction. In Chapter 4, I also introduce the problem of sampling bias in the ATI and carry out a short statistical investigation where I quantify potential sampling bias predictors, so that the optimal bias correction

methods can be applied in subsequent chapters in order to produce the most accurate distribution maps of ancient and veteran trees.

Based on findings from Chapter 3 and detailed examination of the literature about sampling bias correction, in Chapter 5 I present an alternative to the common occurrence-based SDM methods such as MaxEnt. I show through a series of simulations how aggregating presence-only or presence-absence species occurrence data into counts of abundance, and then fitting zero-inflated models, can produce robust predictive species distribution maps free from sampling bias. This allows not only removal of bias, but also study of the causes of bias, something which most modelling methods currently are unable to do. In Chapter 6 I then apply this method to the ATI, along with several other common bias correction methods (as outlined in Chapter 4) in order to evaluate the effectiveness of each one in relation to the ATI. This approach builds on Chapter 3 and expands the scale of the research from one habitat to all habitats across England, with modelling occurring at a 1-km resolution. In this chapter, I also introduce my second method of model verification, additional randomised field surveys using trained volunteers that were carried out over autumn and winter of 2020, providing independent data to evaluate model predictions. Based on this independent field validation, I calculate estimates of the potential total number of ancient and veteran trees in England.

My final research chapter (Chapter 7) investigates the use of alternate model predictors based on habitat and landscape structure in SDM, also at a 1-km resolution across England. Quantifying landscape structure involves the use of landscape metrics, which mathematically describe aspects of the landscape at different scales and complexities (Li & Wu, 2004), and their use in SDM has been shown to improve model performance by adding functional ecological information. In this chapter I compare the use of landscape metrics in SDM with models using only environmental predictors, and models using combinations of the two. Landscape metrics were calculated based on an alternative data-set, the National Canopy Map (also called National Tree Map^{TM}) (Bluesky, 2015), which is a map of all canopy higher than 3 m across England and Wales constructed from stereo aerial photography and digital elevation models. I also use the data collected from the field verification to evaluate model predictions

from the landscape metric distribution models, and to create similar estimates of the total number of ancient and veteran trees across England as in Chapter 6.

I finally review my main findings and conclusions about the most accurate distribution of ancient, veteran and notable trees in Chapter 8, where based on all my research using the ATI, I summarise the key environmental determinants of the trees, the most likely true predictive distribution maps and estimates of the total numbers of ancient and veteran trees in England.

Chapter 2: An initial analysis of the Ancient Tree Inventory (ATI) and an introduction to the other datasets used in the thesis.

Adapted from: Nolan, V., Reader, T., Gilbert, F. et al. The Ancient Tree Inventory: a summary of the results of a 15 year citizen science project recording ancient, veteran and notable trees across the UK. Biodiversity and Conservation, 29, 3103–3129 (2020). https://doi.org/10.1007/s10531-020-02033-2.

2.1 Abstract

In this chapter, I firstly present a descriptive and statistical outline of the ATI, including summaries of the current UK ancient, veteran and notable tree distributions, the status and condition of the trees, and key information about the recording process and maintenance of the database. I also outline areas of the ATI that are lacking in knowledge or robust surveying methodology, and that have the potential for improvement or for further study. Examining or correcting these issues with the ATI will become the focus of the remaining chapters of this thesis. Secondly, I then introduce other environmental, topographical and anthropogenic datasets that have the potential to be predictors of ancient tree distributions across the UK that I use in subsequent chapters of this thesis. My initial analysis of the ATI dataset (first carried out in 2018-2019) suggests there are significant differences in the threats, size, form and location of different types of trees, especially in relation to taxonomic identity and tree age. These findings will be used alongside the other datasets presented here to investigate the true distribution of ancient trees across the UK in the subsequent chapters in this thesis.

2.2 Introduction

This chapter describes data from the Ancient Tree Inventory (ATI) provided by the Woodland Trust in late 2018 (accessed 17/12/18). The ATI began as the Ancient Tree Hunt in 2004 and was originally envisaged as a five-year citizen science project between the Ancient Tree Forum (ATF), Tree Register of British Isles (TROBI) and the WT, that encouraged the public to record and map ancient, veteran and notable trees. The success of the original project has resulted in its continuation to the present day (at the time of writing in July 2020) as the ATI and over 200,000 trees have been mapped with many more still being recorded each year. The project was intended to cover the UK, but a small number of records have also been collected across Ireland. The ATI encourages not only the location of trees to be recorded, but also information about their condition, accessibility, survey information and several other characteristics (Table 2.1).

Any member of the public can upload a record to the ATI via an online database system, with the minimum requirement of information added about each tree being location, girth, species (if possible to identify) and access information. Currently 87% of records have completed information for all categories. The WT recording protocol requires all uploaded records to undergo a second verification step, whereby trained WT verifiers revisit each tree to confirm the record, location and associated information. Additionally, a tree can only be classified as ancient, veteran or notable by a verifier. The ATI is actively managed as an online database by the WT, and a record can only be viewed by all members of the public once verified. The verification process is ongoing, so although not all trees in the ATI have currently been verified, they will be in the near future.

For guidance on how to distinguish between ancient, veteran and notable trees, verifiers are encouraged to refer to the WT's Ancient Tree Guide No. 4 (ATF, 2008a) or to the WT website (https://ati.woodlandtrust.org.uk/what-we-record-and-why/what-we-record/). These sources describe in detail the features of each type of tree, as well as providing species-specific estimates of girth measurements for trees in each category. Verifiers are also required to attend an additional training day where they receive further guidance and assistance in distinguishing between the three categories.

In addition to members of the public, many organisations contribute to the ATI and also support, provide advice and campaign on behalf of ancient trees in the UK including the WT, Natural England, Scottish Natural Heritage, the National Trust, the ATF and the Royal Society for the Protection of Birds. Many of these organisations own and manage land containing ancient trees and all are vocal advocates of the importance of ancient trees.

In this chapter I aim to present a descriptive and statistical overview of aspects of the ATI, including the distribution of the trees, differences between ancient, veteran and notable categories, variation across taxa and the status, condition and threats experienced by the trees. I first outline the ATI recording process and structure of the data, and then I present a combined results and discussion about the statistical analyses of the data in three sections, 1) current distribution and characteristics, 2) condition, threats and attrition and 3) survey and recording information and limitations. I also include a final section where I introduce other datasets used in subsequent chapters of this thesis that I deem to have the potential to be predictors of the distribution of trees in the ATI across the UK.

Table 2.1 *Information collected about each Ancient Tree Inventory (ATI) record.*

Field	Description
Tree ID	A unique record ATI ID
Location	Grid reference (6 – 10 significant figures)
Country	Country of tree
County	County of tree
Tree Site	Site name of record
Public Accessibility	Information about tree access
Location access comments	Information about accessibility and site
Woodland Trust Wood	Whether the tree falls within Woodland Trust owned wood
Category	Ancient, veteran or notable
Veteran characteristics	Additional information about veteran characteristics of the tree
Local historic name	Name of tree in local or national context
Tree Form	Tree form and management status e.g. maiden, pollard etc.
Standing status	Whether the tree is standing or fallen
Living status	Whether the tree is dead or alive
Measured girth (m)	Measured girth of tree at breast height (~1.5m)
Height of girth measurement (m)	Height that girth was measured from the ground
Taxon	Taxonomic identity
Image	Possibility to upload an image of the tree
Date of Survey	Date the record was uploaded to the ATI
Organisation	Organisation or individual who has uploaded the record
Verification	Whether the tree has been verified by a Woodland Trust verifier
Rating	Star rating of record
Additional Notes	Additional notes about location, status and access

2.3 Statistical methods used to analyse the ATI

Multinomial logistic regression models were fitted to the ATI data to compare between ancient, veteran and notable tree categories in relation to three predictors (country, tree form and girth) including all second-level interactions. Models were fitted using the R package 'nnet' (Venables & Ripley, 2002). The most parsimonious model was selected based on multidirectional stepwise regression using Akaike's Information Criteria (AIC) and parameter significance was assessed using Analysis of Variance (ANOVA) Wald Chi-Squared tests. Two models were fitted, one using all ATI records, and another using only records from the 12 most common tree genera. The latter model also included genus as an extra predictor. In addition, to describe patterns of variation in the relative frequency of trees in other categories where the information is not necessarily recorded for all trees (habitat and threats for each tree), I used independent chi-square tests of association based on the absolute numbers of records. Finally, Pearson correlation coefficient tests were used to describe trends in data recording and girth measurements over time. Since some of these tests involved repeated analysis of the same variables, it is important to note that their results are not statistically independent; they provide a descriptive analysis of the data and should not be viewed as definitive tests of particular hypotheses about the association between variables in the ATI. All statistical analyses and modelling were carried out in R (R Core Team, 2018).

2.4 Results and discussion

2.4.1 Current distribution and characteristics

Location, category, taxonomy, tree form and girth

There are 169,967 trees across the UK recorded in the version of the ATI used for this analysis (Figure 2.1). The majority (83.1%) are in England, with smaller numbers of records in Scotland (8.4%), Wales (5.3%), Northern Ireland (3.2%) and Republic of Ireland (0.02%). There are 15 records from Jersey, Guernsey and the Isle of Man. Records show a strong geographical bias towards southern English counties, with Berkshire (15,187 records), Herefordshire (10,934 records) and Wiltshire (9,077 records) contributing a combined total of 20.7% of all records (Fig. A2.1). However, this is influenced by the

high number of veteran and notable tree records in these areas; when considering only ancient trees, North Yorkshire and Cumbria become the second and third highest contributors, highlighting the importance of distinguishing between true ancient trees and those in other categories (as defined in Chapter 1, Table 1.1). Like many of the patterns in the ATI data, these geographical biases may reflect recorder bias, as well as biological and historical processes which influence the distribution of ancient trees. The possible nature of the bias is discussed further in Chapter 4 and methods for quantifying and removing it are actively explored in Chapters 4, 5 and 6.

The majority of the trees in the ATI are veteran (103,648 records) or notable (45,618 records), with relatively few recorded as ancient (13,476 records) (Fig. 2.1). 6,867 records have no category i.e. have not yet been verified or the category is unknown, the majority of which fall within N. Ireland. Eighty two of the trees are also listed as heritage trees and 31 as champions. Many trees have saproxylic organisms present on them (noted in their records), including ferns (3,389 records), lichens (36,240 records), moss (31,253 records) and several fungi species (Table A2.1). Although interesting, these observations are not necessarily the most informative, as quite young and small trees may have some moss or lichen. There is also likely to be inevitable bias in the recording and noting of these, depending on the recorders expertise, accessibility, habitat type etc. However, having the option to record locations and information about rare or endangered saproxylic species if found, in the ATI, could be valuable for other conservation purposes and projects.

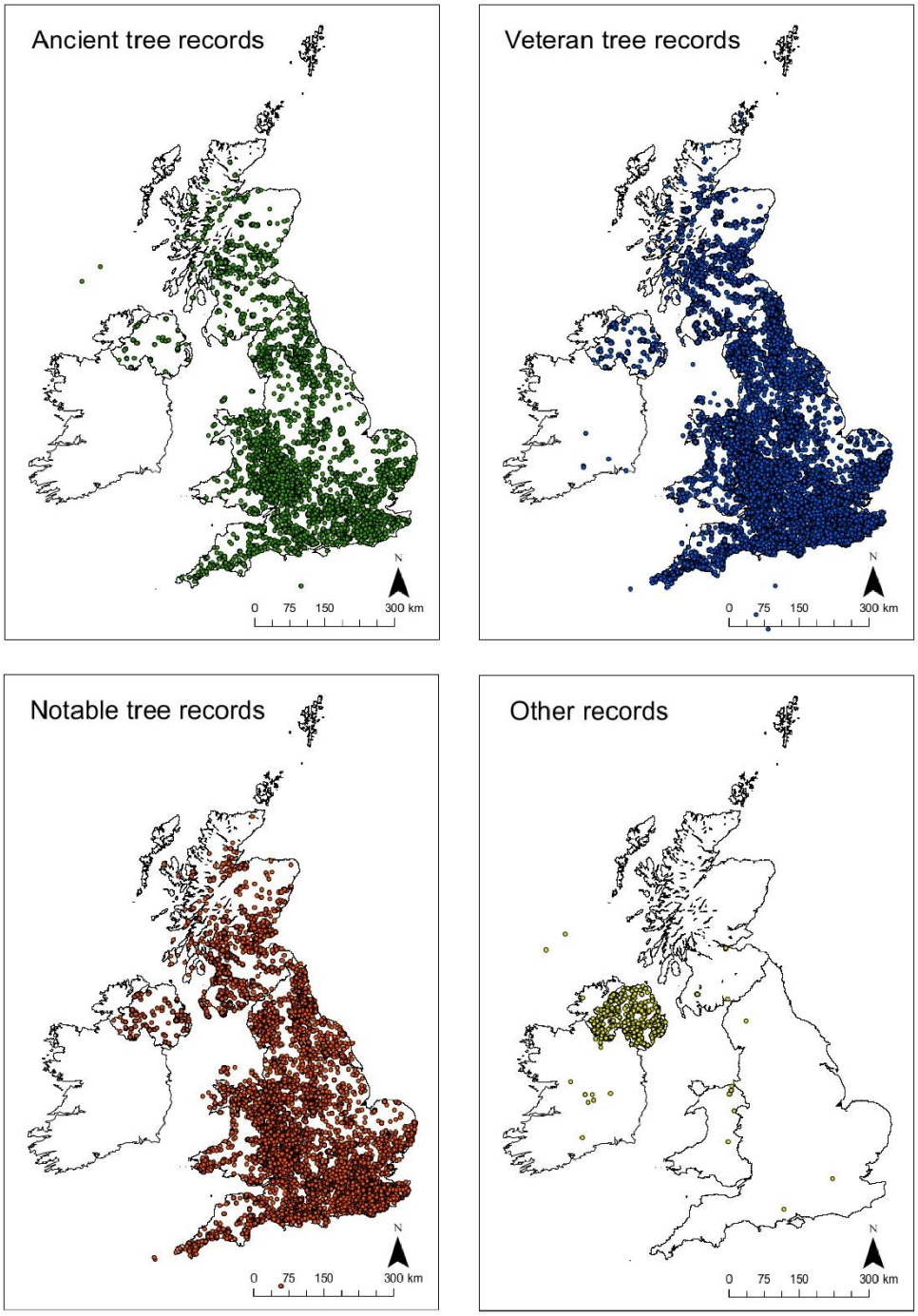


Fig. 2.1 *Distribution maps of ancient, veteran, notable and other records in the Ancient Tree Inventory (ATI) across the UK and Eire. Several records have incorrect grid references in the ATI and therefore do not display in the correct location e.g. outside the UK boundary.*

Two hundred and eleven different taxa have been recorded in the ATI, ranging from family to species level including sub-species, cultivars and hybrid species (Table A2.2). The most common level of identification is genus (81,255 records), so further analysis in this report will focus only on this taxonomic rank. *Quercus* (Oak) is by far the most common genus recorded across the UK, representing almost half of all records (44.2%), followed by *Fagus* (Beech) with 12.4% and *Fraxinus* (Ash) with 6.9%. The 12 most common genera contain 86.4% of all ATI records between them (Fig. 2.2).

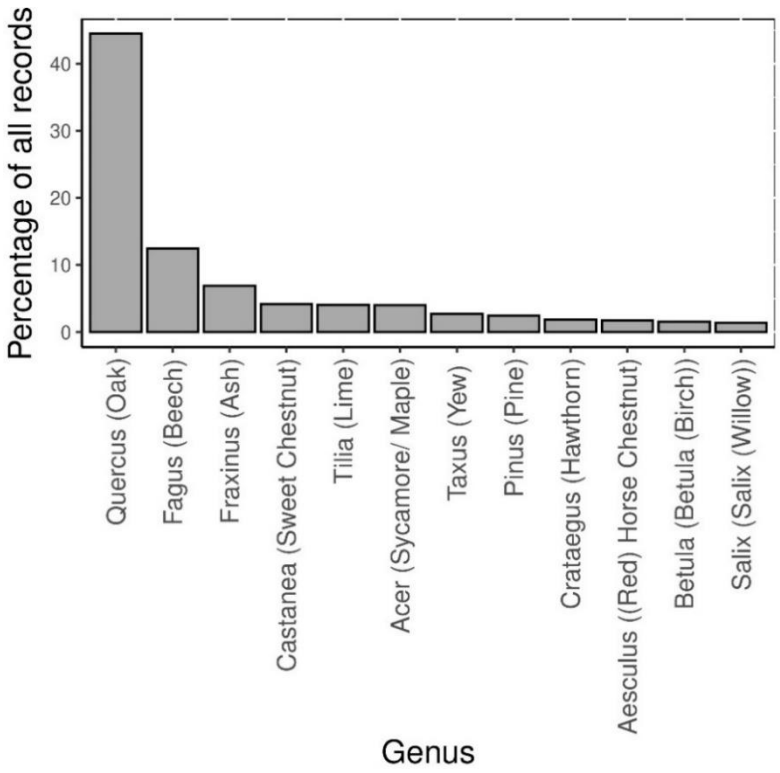


Fig. 2.2 *The percentage contribution of the 12 most common genera to the total number of records in the Ancient Tree Inventory (ATI). The common name(s) of the species present in the ATI that fall within each genus is shown in brackets.*

Strong significant associations were found using multinomial logistic regression models between country, tree form, measured girth, genus and all second level interactions, across the three categories of trees (ancient, veteran and notable) (Table 2.3). When comparing across countries, there are proportionally more ancient tree records in Scotland and Wales than veteran or notable, and

proportionally more notable trees records in Ireland (Fig. 2.3a). Tree form can be a key method in determining whether a tree will survive into its ancient phase. The aim of traditional management tools such as pollarding or coppicing was to extend a tree's life to exploit its resources, and consequently these techniques often produced trees that are many centuries old. Unsurprisingly, therefore, there is a significant association between tree form and category (Table 2.3), with strong links between ancient and veteran trees and pollard form (Fig. 2.3b & 2.3c). However, only 6% of all pollards in the ATI are recorded as being actively managed; this raises concerns about the future survival of the high number of lapsed pollards throughout the landscape.

Ancient trees are also proportionally present more frequently as hedgerow trees or coppards (where the tree is cut at a height intermediate of a coppice and pollard), whereas veteran trees are proportionally more frequently found in 'other' tree forms (such as trees found on cliffs, phoenix trees (fallen trees that are able to re-root and regenerate) or trees of unknown form), and notable trees as maidens or coppices (Fig. 2.3b). However, it is important to note that the definition of a coppard has been revised by the Woodland Trust since this analysis was undertaken; due to the rarity of finding a true coppard, coppards are most likely now recorded as coppices or another tree form. Therefore, inferences about this finding should be taken with caution, as many of these coppard trees may in fact be lapsed coppices rather than true, actively managed coppards.

There were also significant differences between category across tree form and country (Table 2.3), with ancient trees proportionally more common than veteran and notable trees as pollards, hedgerow trees and coppards in England, but more common as pollards or 'other' tree forms in Wales, and coppices, pollards or stumps in Scotland (Fig. 2.3c). Notable trees were most frequently found as maidens in all countries compared to ancient and veteran trees, but presented stronger associations with coppice and multi-stem tree form in England and Scotland than the other categories. Veteran trees showed proportionally stronger associations with 'other' tree forms in both England and Scotland than the other categories, and in general were found in intermediate proportions between ancient and notable category across all tree forms and countries (Fig. 2.3c).

Tree form, country and girth also differed significantly between categories in relation to genus (Table 2.3). The most notable associations were between ancient tree category and *Taxus* (Yew), *Castanea* (Sweet Chestnut) and *Fraxinus* (Fig. 2.4a). *Crataegus* (Hawthorn) had the strongest association with veteran tree form, and *Aseculus* (Horse Chestnut) with notable tree form. The trees most likely to be recorded as coppices belong to *Fraxinus* or *Acer* (Maple) (Fig. 2.4b), particularly in relation to notable trees (Fig. A2.5), and pollards to *Fraxinus*, *Quercus* or *Salix* (Willow) (Fig. 2.4b), especially when in ancient form (Fig. A2.5). Other notable associations include ancient *Fagus* trees and hedgerow or stump form, *Taxus* or *Crataegus* with multi-stem form, and *Pinus*, *Castanea* and *Aesculus* with maiden form (Fig. 2.4b, Fig. A2.5).

Table 2.3 *Wald Chi-Squared Analysis of Variance (ANOVA) results to test for parameter significance from a multinomial logistic regression model of category (ancient, veteran and notable tree) in relation to Ancient Tree Inventory (ATI) characteristics. Two models are fitted, one using all ATI records, and one using only records from the 12 most common genera of tree. The latter model also includes genus as a predictor and second level interactions of genus with the other predictors. Chi-Squared values (d. f.) and parameter significance are shown (* < 0.05, ** < 0.01, *** < 0.001).*

Predictor	**All ATI records**	**Records from the 12 most common genera**
Country	1194.6 (8) ***	897.7 (8) ***
Girth	14104.9 (2) ***	20926.5 (2) ***
Tree Form	12930.4 (14) ***	11768.3 (14) ***
Country: Girth	1704.7 (8) ***	383.9 (8) ***
Country: Tree Form	1276.2 (56) ***	567.8 (56) ***
Girth: Tree Form	1006.1 (14) ***	1334.3 (14) ***
Genus	-	6735.0 (24) ***
Genus: Girth	-	766.1 (24) ***
Genus: Tree Form	-	2186.9 (168) ***
Genus: Country	-	1880.0 (96) ***

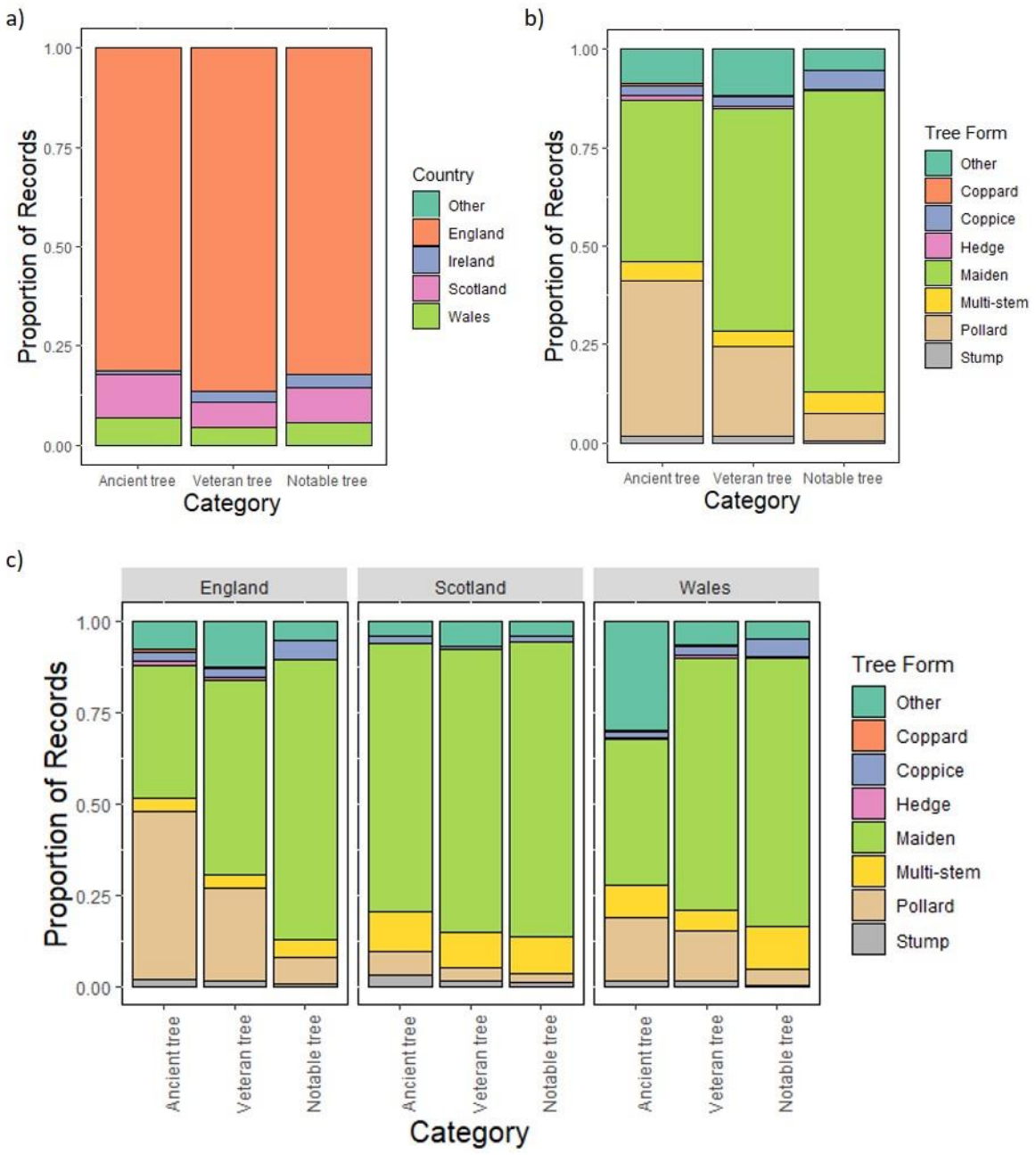


Fig. 2.3 *The relative proportion of Ancient Tree Inventory (ATI) records between three tree categories (ancient, veteran and notable) shown across a) country of record, b) tree form and c) both country (England, Scotland and Wales) and tree form. The country category 'other' refers to trees situated in either Jersey or Guernsey, and 'other' tree forms include cliff trees, phoenix trees and trees of unknown form.*

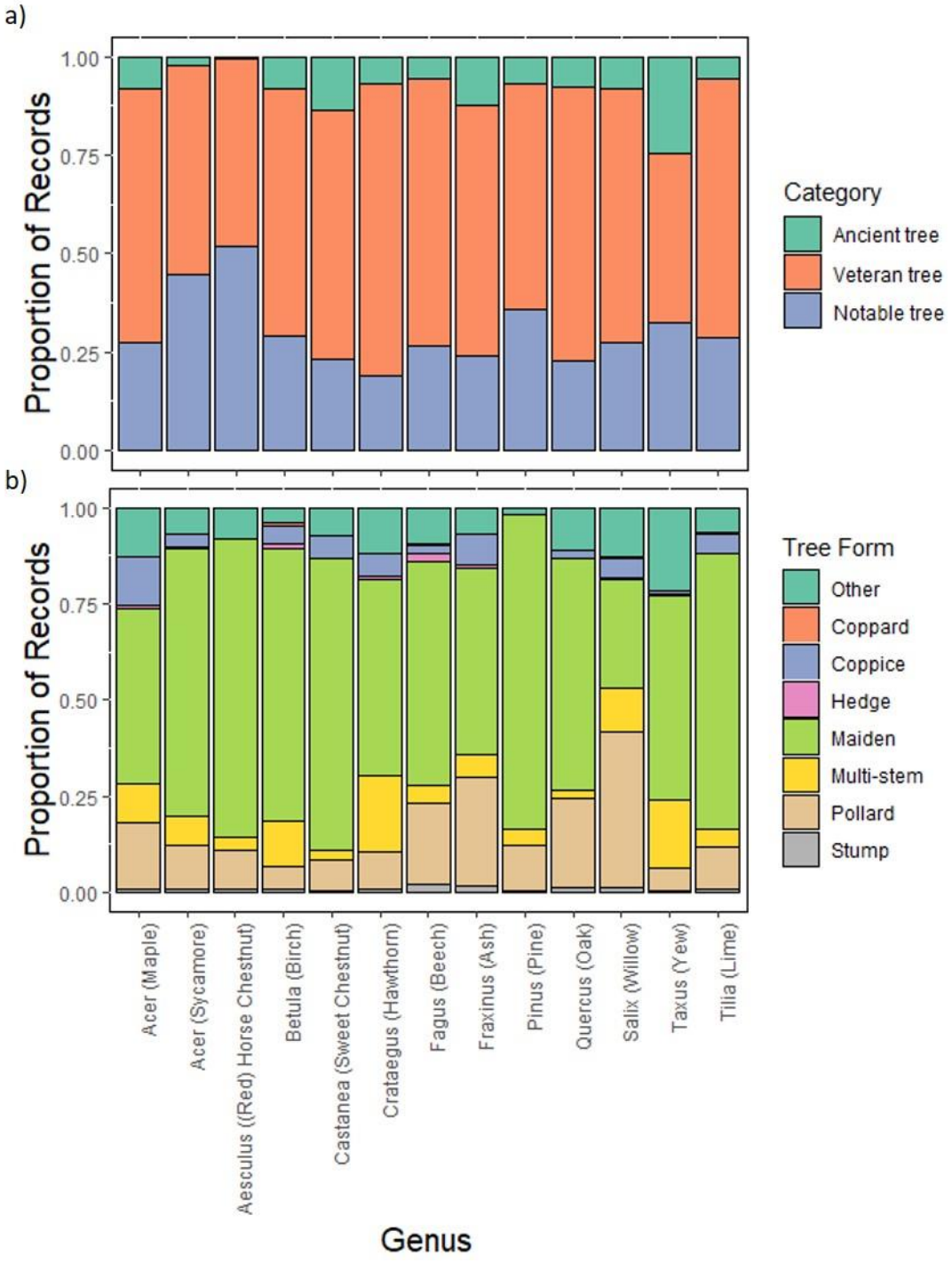


Fig. 2.4 *The relative proportion of Ancient Tree Inventory (ATI) records between three tree categories (ancient, veteran and notable) shown across a) the 12 most common genera and b) the 12 most common genera and tree form. Although Sycamore and Maple belong to the same genus, they are shown separately here to identify any unique associations that may be present. The category 'other' tree forms includes cliff trees, phoenix trees and trees of unknown form.*

There is much discussion about the accuracy and usefulness of the relationship between tree girth and age, and without dendrochronological sampling a tree's age is often over or under-estimated (Hartesveldt et al., 1975; White, 1998; Moir, 2013). The age-diameter relationship has also been shown to vary depending on environmental parameters such as temperature and water runoff (Rohner et al., 2013) and across species (Yunyun et al., 2009). Nevertheless, it is an important characteristic to record and can provide some general idea of the rough age of a tree (White, 1998). Tree girth is usually measured at breast height (~1.5 m above the ground) where possible, although this can be difficult for trees in pollard, coppice or multi-stem form. The mean height at which girth is measured for all ATI records is 1.573 m.

There are 22 trees with measured girths greater than 20 m in circumference (6.4 m dbh), with the largest (a maiden Pedunculate Oak), recorded as 54.18 m (17.2 m dbh) in girth. Most of these can be attributed to recording errors by the volunteers or verifiers i.e. omission of a decimal place; the largest Oak to ever be recorded is thought to be the Marton Oak in Cheshire (13.38 m girth, 4.26 m dbh) (Farjon, 2017). Additional errors also occur when a recorder or verifier incorrectly identifies a cluster of trees or coppices as one multi-stem tree, therefore introducing erroneous inflated girth measurements into the ATI. To reduce the influence of these potentially biased records, only trees with girths below 15 m in circumference (4.8 m dbh) were included in these analyses in this chapter. An interesting initial observation is that there is a weak significant positive correlation between measured girth and date of record ($r = 0.042$, $p < 0.001$), suggesting that there are many large, and potentially old, trees still being discovered.

Mean measured girth differed significantly across category, tree form and country (Table 2.3). As might be expected, ancient trees have larger girths in general than veterans, which in turn are larger than notable trees (Fig. 2.5). The largest mean girth measurements belonged to Welsh ancient trees in the form of coppards, pollards or 'other' tree forms, English ancient trees in the form of hedgerow trees or coppices and Irish ancient coppice trees (Fig. 2.5). The largest veteran trees were Irish coppards or

multi-stem trees, English veteran hedgerow trees or maiden trees in other locations (Jersey or Guernsey). In general, mean measured girth was smaller across Scotland than any other country.

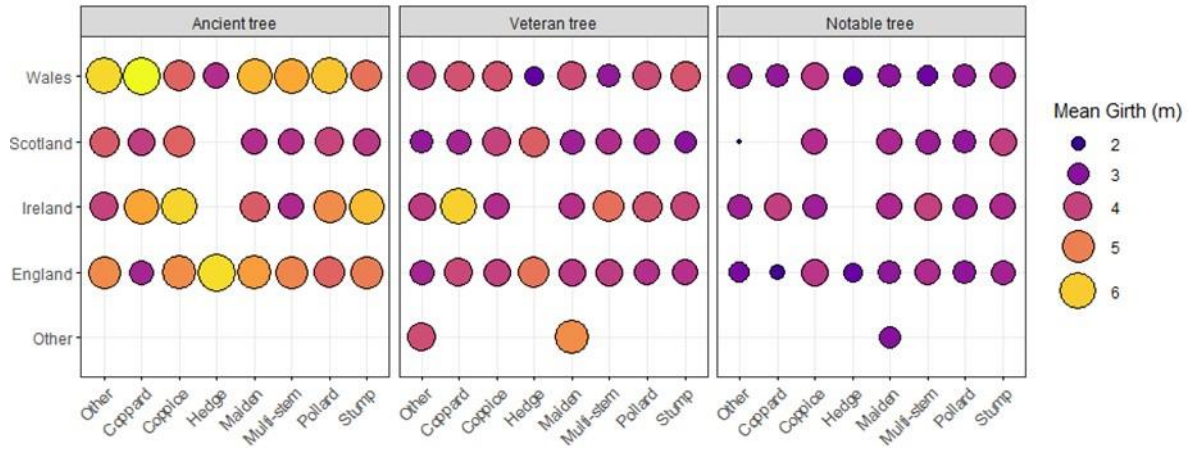


Fig. 2.5 *Mean measured girth (m) of trees recorded in the Ancient Tree Inventory (ATI) shown for three tree categories (ancient, veteran and notable) across country and tree form of record. The larger the circle and the lighter the colour, the larger the mean measured girth. The category 'other' tree forms includes cliff trees, phoenix trees and trees of unknown form.*

Mean measured girth also differed significantly among the 12 most frequent genera in the ATI (Table 2.3; Fig. A2.4). The genus *Castanea* has by far the largest mean girth (4.87 m) followed by *Taxus* (4.10 m) and *Quercus* (3.95 m). *Pinus*, *Betula* and *Crataegus* all have relatively smaller girths. *Quercus* is often thought of as the typical 'ancient tree', especially in England (Farjon, 2017), but surprisingly, *Quercus spp.* in the ATI had relatively smaller girths compared to other species than might be expected, which may be explained by the strong association of Oak with veteran rather than ancient form; as expected, veteran trees have significantly smaller girths than ancient trees. Oak was traditionally the preferred timber tree and its prevalence across the landscape is more due to economic factors than ecology (Barnes et al., 2017). When managed as a maiden tree, Oak was often harvested before reaching its mature phase, so was unlikely to reach great ages (Barnes et al., 2017). Most ancient Oaks remain today in either pollard or coppice form, or as maidens within parkland or wood-pasture (Farjon, 2017). This may also be the case with *Castanea* (Sweet Chestnut) and *Fraxinus*, both of which have strong historical association with coppicing or pollarding practices (Barnes et al., 2017).

Local habitat and site associations

An interesting component of the ATI is the optional recording of habitat information for trees. Although only 69,308 (40.8 %) records have this information recorded, it offers insight into local habitat associations of trees. The number of records with habitat information is higher for ancient trees (47.9 %), compared to veteran (44.2 %) and notable trees (36.3 %). It also varies across country, ranging from 24.5% of records in Scotland to 61.9% of records in N. Ireland. The distribution of records is unequal across habitat types with 29.3% of records associated with woodland, 16.5% associated with parkland and 10.1% associated with field habitat. All other records fall within other habitat types (Fig. A2.2).

Habitat associations depend significantly on genus ($\chi^2 = 29{,}998$, *d.f.* $= 132$, $p < 0.001$) (Fig. 2.6), with combinations of *Tilia* (Lime) in avenue habitat, *Taxus* in churchyards or cemeteries, *Betula* and *Pinus* in upland or moorland habitat, *Fagus* in woodland and *Fraxinus* in field habitat all appearing more frequently than expected. As with *Taxus spp.* in churchyards, *Tilia spp.* (Lime) are familiar elements of avenues, especially on historic estates and parkland (Pigott, 1992; Couch, 2012). Although abundant across much of the UK, Lime trees were favoured for avenues and parkland due to their aesthetically pleasing, tall and long-lived characteristics (Helliwell, 1989). Both *Betula* (Birch) and *Pinus* (Pine) are common upland tree genera, especially in parts of Scotland (Fenton, 1984) and Birch was heavily coppiced in these areas (Barnes et al., 2017). There are fewer strong negative associations, but *Quercus spp.* are present less frequently as avenue trees or in upland/ moorland areas, *Tilia* spp. are less frequently present in woodland and *Fagus spp.* are less frequently present within field habitat than expected.

Ancient, veteran and notable trees also have significantly different habitat associations ($\chi^2 = 2163.7$, *d.f.* $= 22$, $p < 0.001$) (Fig. 2.6). Key habitats for ancient trees include cemeteries or churchyards, wood-pasture, fields and moorland or upland. Ancient trees are less likely to be present as avenue trees, in gardens, alongside roads, railways or other public rights of way, or in woodland. Veteran trees are also found less frequently than expected in parkland, gardens and avenues, but do have strong positive associations with woodland habitat and hedgerows. In contrast, notable trees follow opposite patterns

to ancient or veterans and are more likely to be associated with avenues, gardens, parkland and public rights of way, and less likely to be found in hedgerows, fields, moorland or upland habitat, wood-pasture or woodland.

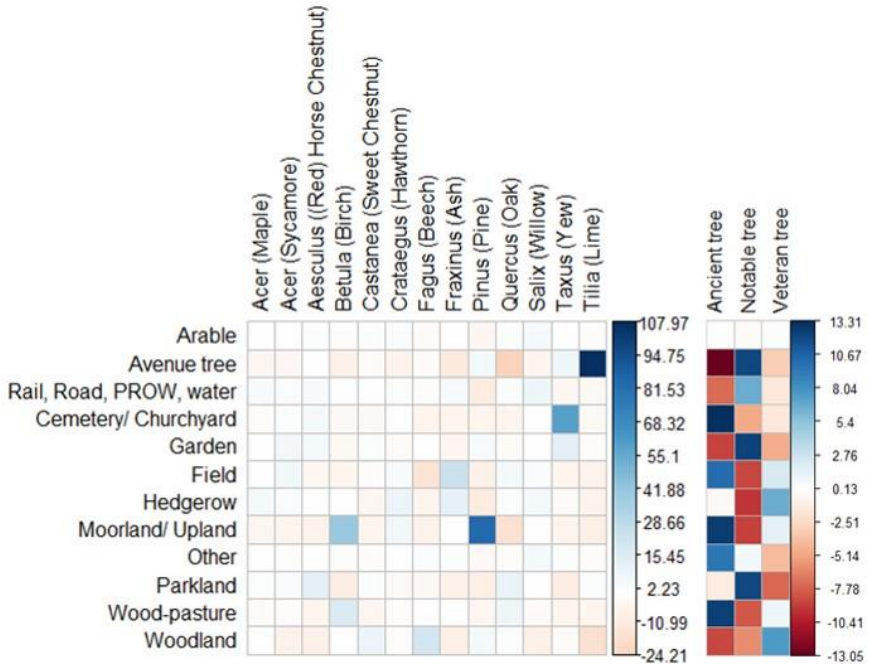


Fig. 2.6 *Standardised Pearson residuals (r) from the Chi-square test of association between habitat and the 12 most common genera (left) and ancient, veteran or notable category (right) in the Ancient Tree Inventory (ATI). The higher the absolute residual value, the more that association contributes to the Chi-square statistic. + represents positive associations, whereas - represents negative associations, and the darker the square the stronger the relationship. Although Sycamore and Maple belong to the same genus, they are shown separately in to identify any unique habitat associations that may be present.*

Where possible, recorders are encouraged to name the site on which a tree is found, and as a result, 69,308 trees can be located to 1,466 specific named land areas. As with habitat, records appear biased towards public parks, large estates and historic forests, with the top 20 named sites (most of which fall into one of these three categories) containing 21.9% of all records between them (Fig. A2.3). Additionally, certain land owners have contributed heavily to the ATI, with 2,925 records appearing on WT owned land across the UK. Separate analysis using publicly accessible National Trust databases

across England (National Trust open data: 'limited access land' and 'always open land', accessed 08/01/19) shows that approximately 11.5% of all English ATI records fall within National Trust land. The NT is an environmental and heritage conservation charity and has the largest number of subscribing members of the public of any organisation across England, Wales and Northern Ireland. Since its foundation in 1895, the NT has acquired ownership of over 350 properties and 2470 km2 of land. Churchyards and cemeteries also feature heavily, and contain 38.5% of all trees found on a named site. As before, these patterns probably reflect a combination of recorder bias and biological and historical processes, so further analysis might reveal interesting details for the understanding and conservation of ancient trees across the landscape.

By assessing the distribution of records across different scales (from country to individual site or habitat), it is possible to gain insight into suitable locations for the persistence and survival of ancient trees to inform conservation and management action. Additionally, areas with few records, through either a lack of ancient trees or lack of surveys, can be targeted for future surveys, verification work or tree planting. The current distribution maps show records heavily clustered in southern English counties around London. These counties have strong associations with historic Royal forests, hunting grounds and private parks such as Richmond Park or Epping Forest (Farjon, 2017). Similarly, Savernake forest, Windsor Great Park, Ashridge Estate and the New Forest (the four sites with the highest record abundance) currently are or have been owned at some point by the monarch. The continuity of the monarchy and aristocracy in the UK, unlike other European countries such as France and Germany, is likely to be one of the main influences on the high abundance of UK ancient trees (Rackham, 1976; Butler et al., 2002; Farjon, 2017).

2.4.2 Condition, threats and attrition

Standing status and threats

The standing and living status of each tree provides valuable information about the current condition, threats and attrition of ancient trees in the UK. In this version of the ATI (December, 2018) most trees

(93.4% of ancient trees, 95.7% of veteran trees, 98.3% of notable trees) are recorded as alive and standing. Only one tree is suffering from suspected Ash dieback and 15 trees are suffering from acute Oak decline or chronic Oak decline. However, this is very likely linked to observer bias; diseases such as Ash dieback are relatively difficult to spot in ancient trees, and so the total affected numbers are likely to be much higher. Additionally, there is likely to be a bias towards the recording of living (and therefore healthy) trees, as opposed to those that are fallen or dead, which may explain the low prevalence of diseased and dead trees in the ATI. Furthermore, most records have not been revisited since their initial recording, which for some trees is almost 16 years ago, so it is likely that some trees have subsequently been lost. Inferences about threats and disease should therefore be considered as speculative and a likely underestimation of the true, current status and attrition rate of ancient trees.

There is an additional option in the ATI to add information about apparent threats to trees, although this is highly likely to be biased by expertise in this area e.g. confident tree recorders such as arboriculturalists are much more likely to notice and record threats than the average ATI recorder. Therefore, any inferences about threats to particular trees should take into consideration potential recorder biases. 17,499 specific instances of a threat have been recorded that are tree-specific and include 'Compaction of root area' (31% of threats), 'Grazing damage' (27% of threats), 'Over shading' (15% of threats), 'Major tree surgery' (8% of threats), 'Cultivation close to tree' (7% of threats) , 'Vandalism' (4% of threats), 'Development or building' (3% of threats), 'Vehicle damage' (3% of threats) and 'Fire damage' (2% of threats).

Nevertheless, relatively few records have associated threats recorded (less than 10% of records are recorded as threatened) and there is no way to assess the completeness of this field, so it is likely that many more trees are threatened in some way. For example many other threats may be less observable to recorders such as nitrogen deposition and drought (Lindenmayer et al., 2012; Lonsdale, 2013). By understanding the individual age and species-specific threats, conservation work can be targeted in these areas to better protect the most vulnerable trees and ensure our current mature phase trees will reach their ancient phase and the continuity of deadwood habitats across the landscape. One obvious example

is to promote continuous, appropriate management of coppices and pollards as part of any conservation plan for ancient trees.

There is a significant association between the 12 most frequent genera in the ATI and the type of threat (χ^2 *test = 581.2, d.f. = 96, p < 0.001)* and also between the tree category and type of threat (χ^2 *= 158.58, d.f. = 16, p < 0.001*) (Fig. 2.7). Ancient trees are the most threatened category relative to veteran and notable trees, showing positive associations with six out of nine threats, the most prominent of which are grazing and fire damage, but also include cultivation, vandalism, tree surgery and development. Large, hollow trees are known to be vulnerable to fire. (Lanner, 2002; Becker & Freeman, 2009; Crane et al., 2017) especially in hot, arid places such as California or Australia (Lindenmayer et al., 2012). Although wildfires are infrequent and localised in the UK, fire damage through vandalism is a more common cause of ancient tree loss (Kirby et al., 1995). Over-grazing around ancient trees is also a common threat, due to the browsing of the bark, leaves or suckers from the tree, trampling of roots and high levels of excreted nutrients around the base of the tree (Manning et al., 2006; Hartel & Plieninger, 2014). Veteran trees are very strongly impacted by over shading, and notable trees by cultivation and root compaction compared to other categories.

The greatest associations recorded between particular threats and genera are those between *Fagus* and over-shading, *Taxus* and tree surgery, *Tilia*, *Aesculus* and *Pinus* and grazing, and *Pinus* and root compaction. *Quercus* experiences the most threats relative to any other genera, showing strong associations with seven out of nine including vehicle damage, cultivation and root compaction, which is worrying as records of this genus comprise almost 50% of the ATI. The strong anthropogenic interest in old trees, especially old Oaks can sometimes be counter-productive: excessive, inappropriately managed visitation to sites such as National Trust parklands or historic houses can result in pressure on exposed, scattered old trees. Conservation measures such as fencing can help to protect ancient trees against threats from livestock, human influences and cultivation (Fischer et al., 2009).

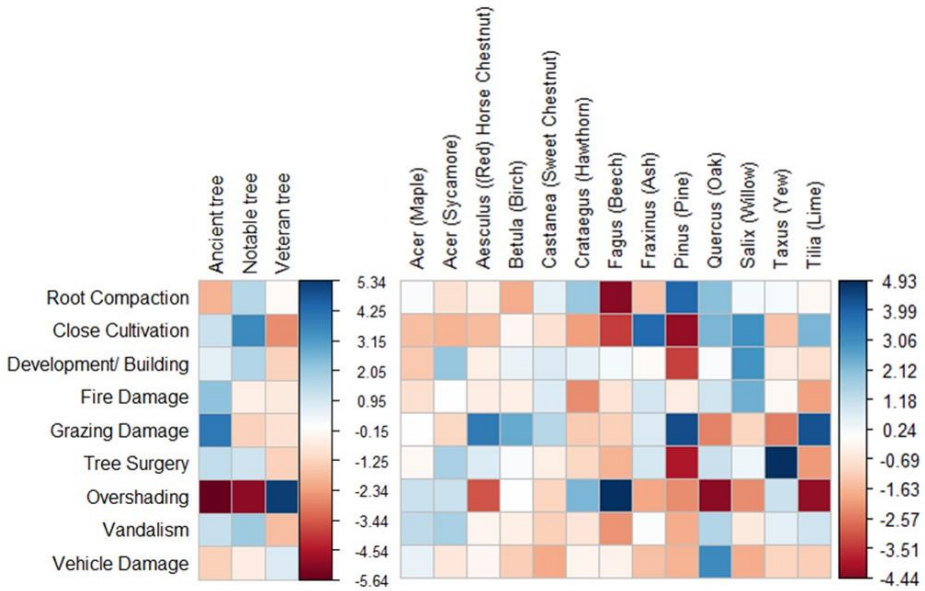


Fig. 2.7 *Standardised Pearson residuals (r) from the Chi-square test of association between threat type and ancient, veteran or notable category (left) and the 12 most common genera (right) in the Ancient Tree Inventory (ATI). The higher the absolute residual value, the more that association contributes to the Chi-square statistic. + represents positive associations, whereas - represents negative associations, and the darker the square the stronger the relationship. Although Sycamore and Maple belong to the same genus, they are shown separately in to identify any unique habitat associations that may be present.*

Legal protection for trees with veteran characteristics in the UK has improved significantly in the past few years, and the recently published 2019 National Planning Policy Framework (NPPF) now recognises both ancient trees and ancient woodlands as 'irreplaceable habitat' (NFFP, 2019). Other protective measures include Tree Protection Orders (TPOs) or legislation protecting other species or habitats such as bat roosts, designated sites, hedgerows, or scheduled ancient monuments (Read, 2000). Nevertheless, all of these measures can be overridden for 'exceptional reasons' such as health and safety concerns or national infrastructure projects (Read, 2000; NPPF, 2019).

Attrition rate over time

The global decline of decaying and dead wood habitat is a growing issue (Gibbons et al., 2008; Fischer et al., 2010; Lindenmayer et al., 2012), and it appears that UK ancient, veteran and notable trees are no

exception. Dead wood in different forms within an ecosystem is an important resource for many organisms (Hjältén et al., 2007; Lõhmus et al., 2010; Svensson et al., 2016), so its removal is likely to have cascading impacts across rural landscapes. Ancient and veteran tree loss is also hugely detrimental for wildlife and biodiversity in urban environments (Stagoll et al., 2012). Le Roux et al. (2014) predicted declines in urban ancient and veteran tree populations of 87% over the next 300 years under current management strategies. These declines were not halted by increasing the recruitment rate of ancient trees, and under the worst management scenarios, urban ancient and veteran tree populations were predicted to disappear within 115 years (Le Roux et al., 2014). Development or building works were found to present significant threats to *Acer* (Sycamore) and *Salix,* two of the UK's most abundant urban tree genera. As information about urban tree populations is sparse, it is important to increase quickly our understanding of their abundance and distribution to prevent their decline and the loss of ecological functions.

Since 2004, 47 ancient, 227 veteran and 84 notable trees are recorded as lost. Although some trees are only discovered once they are lost i.e. as a stump or fallen tree, some of the lost trees were previously recorded as alive, and have been revisited over time and have had their records updated. There does not appear to be any geographical pattern to these lost trees, and their distribution reflects that of ancient, veteran and notable trees across the UK. There is no significant correlation between year and proportion of lost ancient tree records collected ($r = -0.0004$, $p = 0.999$), or proportion of lost notable tree records collected ($r = -0.725$, $p = 0.814$). However, there is a moderate positive correlation between year and proportion of lost veteran tree records collected ($r = 0.582$, $p = 0.023$). The data suggest that the proportion of veteran trees recorded as lost has increased over the past 20 years, either through increased surveying of lost trees or through an actual increase in the number of trees that have been removed or damaged.

Although this information gives some insight into attrition rates, the interpretation of this is problematic because we don't know the total standing ancient, veteran or notable tree population sizes in the UK, or the rate at which new recruits are entering or leaving each population i.e. how many notable trees are

becoming veteran, and how many veterans are becoming ancient. Therefore, it is hard to know whether rates of loss are typical, and hence whether they reflect a population-level decline. Nevertheless, the establishment of the ATI means that we now have the ability to monitor long-term changes in the size and demographic constitution of ancient, veteran and notable trees, so future investigation into attrition rates should be more informative and reliable.

2.4.3 Recording and survey information

A feature of the ATI that increases its usefulness beyond being a simple database of species occurrences, is the extra information collected about the recording and verification process including accessibility of sites, recorder or organisation identity, date, verification status and rating. These factors allow more accurate assessment of record reliability before being added to the ATI, as well as more detailed assessment of recorder biases, which in theory should result in highly accurate distribution maps. I recommend both current and future citizen science projects to implement similar features in their recording process, as it should firstly allow assessment of the record reliability and secondly could help to establish a framework for a longer-term monitoring program of a particular species or ecosystem.

Accessibility and recorder

The majority of records are found on open access land where members of the public have unrestricted access (Fig. A2.6). Nevertheless, a remarkably high number of records (37.5%) are found on private land, where there are no public footpaths, which is another important cause of spatial bias in the ATI (see Chapter 4 for more information). These records result mostly from pre-arranged site visits with consent from the landowner or site manager, and although more challenging to organise, they are useful in both obtaining ATI records and raising awareness of the importance of ancient trees on the site and within the wider landscape. The largest contribution of records (13.1%) to the ATI comes from the collective input of individual members of the public, but an additional 101 charities, consultancies and conservation organisations have contributed records, the top 10 of which have recorded 55% of all records. Although most are national organisations and charities, several county-specific ancient tree-recording groups have made significant contributions.

Date of tree record

The ATI began in 2004 as the Ancient Tree Hunt, which is reflected in the sudden increase in the number of records in 2005. Records added before this date have come from original Tree Register of the British Isles records that had been collated over many years and provided the initial inspiration for the ATI. The earliest record is of an ancient Oak tree recorded in 1900 at Croft Castle, Herefordshire with a 7.63 m girth. The years following 2004 saw many more records added, even after the original project ended and the Ancient Tree Hunt became the ATI. However, in recent years (2011 onwards) there has been a decline in the number of records added each year, although 2017 proved a better year than the past five.

The number added in 2018 is very low due to the fact that the ATI online recording platform was being updated throughout the spring and summer period and was unavailable during this time. Records from 2018 were retrospectively added to the ATI by the WT throughout 2019/ 2020 and over this period around 9,000 more records have been uploaded. Currently, the date associated with each record is its date of upload to the ATI website, so may not necessarily be the date a tree is recorded. The delay between the two processes can be lengthy (potentially several years difference), but due to the longevity of the trees this is unlikely to make a substantial difference to the overall distribution map, providing that the tree is not felled or damaged.

Verification steps in the ATI

Best practice is for each ATI record to be revisited by a WT verifier after it is uploaded, to confirm the tree's location and status and to maintain credibility of the ATI. Reassuringly, a high proportion of ATI records have been verified at least once, with some trees being revisited many times to assess their status and persistence. However, this proportion is slightly less for ancient trees (97.7%) compared to veterans (98.3%) or notables (99.2%). Verification work in the first instance should be targeted at N. Ireland, where the highest percentage (26.2%) of records are unclassified as ancient, veteran or notable,

compared to 2.9% in England, 3.2% in Scotland and 8.5% in Wales. Although most records have undergone this verification, there is concern that some have not and may therefore be incorrect.

As an extra step, each record has been provided with a star rating to reflect its validity and reliability (Table 2.4). This was determined by the Woodland Trust ATI managers and relates predominantly to the level of verification for each record. Citizen science programs can introduce more errors or bias than traditional scientific recording methods (Dickinson et al., 2010; Crall et al., 2011), but record verification by volunteers has been shown to be more cost-effective than traditional data collection by professional scientists and less error-prone than using unverified records (Gardiner et al., 2012). These extra steps should help identify and eliminate biased or false records in the ATI.

Table 2.4 *The Ancient Tree Inventory (ATI) star rating system and the number of records within each group.*

Rating	No. of records	% of Total	Reason for rating
5	77,767	45.75	Recorded and verified by WT verifiers on site
4	46,087	27.12	Verified by a WT verifier or Quality Assured records but not verified on site
3	30,932	18.20	Verified by volunteers of another organisation
2	5,905	3.47	Data that has proved unreliable and unverified
1	9,276	5.46	Unverified but with potential of being 5 star

A potential improvement to the recording process could be the introduction of remote, online verification using the uploaded photos, which would increase the quantity of records that could be verified to a high standard providing that the photos are of a good quality. Throughout the project, there has been periodic background screening and revision of parts of the dataset carried out remotely by an expert head verifier and other trained volunteers, in order to increase the accuracy and robustness of the

data. This usually includes running queries in the data to target records with missing information or fields with suspicious values. Nevertheless, not all records have received this extra attention due to the time consuming nature of this process and large quantities of records. Therefore, increasing the capacity and efficiency of this process, perhaps through the additional 'lead' verifiers, or by encouraging the public to upload additional photos of the trees, could increase the robustness of the data and the data-capturing process.

Limitations of the ATI recording process

Firstly, the confusion regarding the terminology of ancient and other noteworthy trees presents problems in understanding the true nature of each record. The classification of an ancient or veteran tree based on the presence-absence of 'veteran' characteristics is a reliable distinguisher from notable trees. However, the difference between an ancient and veteran tree is much more arbitrary, and likely to vary across recorder or verifier based on experience or geographical region. Additionally, comparisons between different studies on ancient trees, whether in the UK or between other international ancient tree populations, will be difficult if definitions are not standardised. Although the WT provide guidelines for tree age based on girth measurements per species (ATF, 2008a), as mentioned previously, tree age is thought to vary hugely across different environmental conditions. Therefore, it is likely there are a number of ATI records that have been subjectively misclassified, and awareness of this issue is important when separating ancient trees from veteran trees for conservation purposes. Nevertheless, concentrating purely on trees displaying 'veteran' characteristics i.e. the characteristics most ecologically important for saproxylic organisms, mitigates this issue to a large extent until clearer definitions of ancient and veteran trees can be established.

Although the ATI is the most comprehensive database of ancient and ageing trees to date, it suffers many of the drawbacks of a citizen science recording program, including sampling bias. Sampling bias results from the 'ad-hoc', non-representative recording method of public recording schemes and is often present in online, museum or herbarium datasets (Boakes et al., 2010; Rocchini et al., 2011). ATI recording and the subsequent distribution maps are likely to be strongly influenced by the home location

of the most active recorders and WT verifiers, and accessibility to sites such as private estates and parks. For example, the majority of the top 20 named sites with the most trees are all large, well-known, accessible parks, so the high abundances in these areas may be from large numbers of visitors who simply enjoy visiting here. Similarly, the comparatively low number of trees in Northern Ireland and Ireland could be a result of a lack of recorders, or low levels of interest or awareness of the ATI. Recording is also likely to be biased towards areas with good coverage of public rights of way, footpaths and roads, so working with farmers and landowners, and raising awareness of the ecological benefits of ancient trees in the landscape could help gain access to sites for recording purposes.

There are currently a variety of statistical methods that are able to cope with large, biased species datasets which is discussed in more detail in Chapter 4, including spatial filtering of occurrence records, producing bias layers to capture the anthropogenic influence of recording or using statistical models to account for bias (Phillips et al., 2009; Fourcade et al., 2014; Boria et al., 2014; Bird et al., 2014). The ATI is unique in its abundance of unusually good information about recorder location and identity, so presents a brilliant opportunity for bias correction methods to be applied, which is the focus of Chapter 6. Many citizen-science projects require recorders to provide an estimate of their level of expertise or education (Kosmala et al., 2016), which can greatly benefit scientific research based on these data (Johnston et al., 2018). Therefore, collecting more information about sampling effort, time spent in the field, number of volunteers and level of expertise in identification for each record would help to address the issues caused by these biases. In the meantime, identified patterns and conclusions drawn from information in the ATI should be considered in parallel to potential patterns of bias, and caution should be applied when using the ATI for ecological or conservation research without prior consideration of the data limitations.

The reliability of the grid references should be also questioned, as even though the majority are recorded with 1 m precision (10-figure grid references), or at the very least 100 m precision (6-figure grid reference), several records display outside of the UK boundary and are certainly incorrect. It is likely therefore, that there are other records within the UK boundary that are also incorrect. Increasing the

number of trees that are revisited over the next few years will increase the accuracy of the ATI and ideally reduce the amount of false information, yet this may be difficult if an incorrect grid reference has been provided and there are many recorded trees on a site. WT verifiers are often familiar with the problem and the inclusion of an image of the tree mitigates it somewhat, nevertheless there are several thousand duplicated records in the ATI that have been recorded multiple times. Additional work to remove these records is being currently undertaken by WT staff members and other experienced individuals.

A possible remedy for the incorrect grid references would be the development of a smartphone app to collect high quality GPS location data, as well as provide a more accurate estimate of the time the tree is recorded. Many large citizen-science projects such as 'eBird', 'Project Noah or 'What's Invasive!' currently benefit from mobile-phone record collection methods (Newman et al., 2012; Teacher et al., 2013; Luna et al., 2018) and although not currently available for the ATI, a mobile app could be a valuable asset to the project for data acquisition. However, as mentioned previously, the ATI website was redeveloped in the summer of 2018, providing a more user-friendly interface and additional recording features such as the ability to update record information for previously recorded trees. Since this redevelopment, there has also been an ongoing boost in marketing and awareness about the ATI so we should hopefully expect to see an increase in the annual number of records added in the future as the website becomes more popular and awareness of the scheme grows.

2.5 Additional data-sets used in this thesis

Information from a variety of data-sets was collected in addition to the ATI, following careful consideration of the potential influence of each one, on ancient, veteran and notable tree distributions across the UK (Table 2.5). These included a variety of predictors considered to have either ecological influence on the distribution of the trees, or that might be a predictor of the sampling bias or have an influence on the record collection procedures. The data-sets can be grouped as being of historical importance, having environmental influences (either topographical or land classifications) or

anthropogenic factors. Further explanations of the categories from each of the categoric data-sets (Land Class, Agricultural Class, Soil type and Countryside type) can be found in Tables A2.3 to A2.6. All of the historical predictors were digitised manually using the georeferenced maps from literature using ArcGIS version 10.3 (ESRI, 2011).

Table 2.5 *Additional data-sets used throughout this thesis selected based on their potential as a predictor of ancient, veteran and notable tree distributions across the UK, along with the source the data were accessed from, the original format of the data obtained (as a point, polygon, line or raster format) and the reason for their general inclusion as a potential predictor of ancient and other noteworthy tree distributions across the UK.*

Predictor	Format	Source (date accessed)	Justification for inclusion as a predictor
Historical predictors			
Historic forests (1327 – 1336)	Polygon	Neilson, 1940 in The English Government at Work (Willard and Morris, 1940) - The Forests: 1327 – 1336 (02/07/18)	In the UK certain types of historic sites such as these are thought to be less likely to have been deforested, and their ancient trees are more likely to have been protected than in the wider countryside (Rackham, 1976; Farjon, 2017), particularly due to their continuous Royal or aristocratic ownership across the centuries (Butler et al., 2002).
Medieval moated sites	Point	Aberg, 1978 - Medieval moated sites (05/07/18)	
Medieval Deer parks	Point	Rackham, 1976 - Trees and Woodland in the British Landscape (05/07/18)	
Tudor Deer parks	Point	The Counties of Britain: A Tudor Atlas by John Speed (Nicolson and Hawkyard, 1989) – (03/07/18)	
Countryside type	Polygon	Rackham, 1976 - Trees and Woodland in the British Landscape (05/07/18)	The divisions in the historical landscape are likely to highly influence the management and persistence of tree populations (Rackham, 1976). See Table A2.4 for more information.
Topographical predictors			
Watercourses	Line	OS Open Rivers V.10/2018 (Vector) (07/01/19)	Environmental characteristics such as these shape the micro-climate experienced by the trees throughout their whole lives, and are likely to influence the species composition, dispersal, decay and other dynamics of ancient and veteran tree populations (Hall & Bunce, 2011; Barnes et al., 2017; Hartel et al., 2018)
Altitude (1-km)	Raster	Altitude (elevation above sea level (m)) - WorldClim DEM (10/05/18)	
Soil type (1-km)	Raster	EU Soil Database – World Reference Base (WRB) for Soil Resources full soil code (WRBFU) (24/09/18)	

Anthropogenic predictors			
Town centre	Point	Government Open Data – English Town Centres 2004 (19/03/2018)	The presence of ancient and veteran trees across the UK landscape has experienced strong human influences across many centuries (Rackham, 1976; Farjon, 2017; Barnes et al., 2017). Therefore it is likely that proximity to towns, cities and roads will have shaped the planting and management of ancient and veteran trees. Additionally, many of these characteristics also are likely to influence ancient and veteran tree sampling due to issues around accessibility, favouring certain sites etc. (Reddy & Dávalos, 2003; Mair & Ruete, 2016).
Major city	Point	Office of National Statistics (ONS) - Major Towns and Cities 2015 (29/11/2017)	
Commons	Point	Government Open Data – Commons register 2015 (18/12/18)	
Major road	Line	Government Open Data - Major Road Network 2016 (05/11/2017)	
Minor road	Line	OS Open Map Local V.10/2018 (Vector) - Road (07/01/19)	
Buildings	Polygon	OS Open Map Local V.10/2018 (Vector) – Building (07/01/19)	
Land classification predictors			
Ancient woodland	Polygon	Natural England - Ancient Woodlands (England) inventory (08/01/2018)	Ancient and veteran trees can sometimes be found in woodland (especially ancient woodland (Peterken, 1977)) so could be an important habitat in which they are present (Lonsdale, 2013).
National Forest	Polygon	Government Open Data - National Forest Inventory (NFI) 2016 (04/12/17)	
Traditional orchard	Polygon	Natural England - Traditional Orchards HAP England (10/01/18)	Ancient and veteran trees have strong connections to wood-pasture habitat (Hartel et al., 2013; 2018; Chapter 3) and traditional orchards (Barnes et al., 2017).
Wood-pasture	Polygon	Natural England - Wood Pasture and Parkland BAP Priority Habitat Inventory (4/12/17)	
National Trust land	Polygon	National Trust – Open data: limited access land and always open land (08/01/19)	The National Trust is a large organisation in the UK that holds vast areas of land with historic or natural interest and therefore have strong links to ancient and veteran trees (Nolan et al., 2020).
Agricultural Class (1-km)	Raster	Natural England - Provisional Agricultural Land Classification England 2013 (13/04/18)	Land use change, agricultural intensification and urbanisation are strong influences on ancient and veteran tree decline around the world (Read, 2000; Fay, 2002; ATF, 2005, 2011; Lonsdale, 2013). In addition, tree populations have specialised niche requirements to grow and survive, and are likely to be adapted to particular environmental conditions relating to specific land types (Barnes et al., 2017).
Land Class (1-km)	Raster	Centre for Ecology and Hydrology (CEH) - Land Cover Map 2015 (LCM2015, 1km dominant target class) (29/03/17)	
Special Areas of Conservation (SAC)	Polygon	Natural England – Special Areas of Conservation England (01/04/20)	Special Areas of Conservation (SAC) are protected sites that significantly contribute to the conservation of particular habitats or species. Many SAC designations relate to the importance of a site for saproxylic organisms and therefore potentially ancient trees.

Chapter 3: Historical maps confirm the accuracy of zero-inflated model predictions of ancient tree abundance in English wood-pastures.

3.1 Abstract

Ancient trees have important ecological, historical and social connections, and are a key source of dead and decaying wood, a globally declining resource. Wood-pastures, which combine livestock grazing, open spaces and scattered trees, are significant reservoirs of ancient trees, yet information about their true abundance within wood-pastures is limited. England has extensive databases of both ancient trees and wood-pasture habitat, providing a unique opportunity for a large-scale case study to address this knowledge gap. In this chapter, I investigate the relationship between the abundance of ancient trees in a large sample of English wood-pastures and various environmental predictors, in order to identify wood-pastures with high numbers of undiscovered ancient trees. Twenty-one digitised environmental, topographical and anthropogenic variables were collected for 5,571 wood-pastures across England, and using the UK Ancient Tree Inventory (ATI), I predicted the abundance of ancient trees within each wood-pasture. I also introduce a novel model verification step using series of historic maps with detailed records of trees to validate my model predictions; this allows verification using completely independent data, often a challenging hurdle in many modelling scenarios. Important predictors of ancient tree abundance included wood-pasture area, distance to several features including cities, commons, historic Royal forests and Tudor deer parks, and different types of soil and land classes. Model predictions of tree abundance correlate well with historic map verification estimates. They suggest there are ~101,400 undiscovered ancient trees in all wood-pastures in England, an increase of around 10 fold in the total current number of ancient tree records. Historical maps and statistical models can be used in combination to produce accurate predictions of ancient tree abundance in wood-pastures, and inform future targeted surveys of wood-pasture habitat, with a focus on those deemed to have undiscovered ancient trees.

3.2 Introduction

Ancient trees (often referred to as 'veteran trees' or 'large, old trees') are found worldwide and are important ecological structures, in particular as a source of dead and decaying wood, in many ecosystems (Read, 2000; Siitonen, 2001; Butler et al., 2002). The characteristics that define an ancient tree, such as a hollowing trunk and branches, crevices and water-filled pools, enable them to act as 'keystone species', supporting a wide range of saproxylic and non-saproxylic species, including fungi (Boddy, 2001), invertebrates (Speight, 1989), epiphytes (Read, 2000; Ranius et al., 2008) and larger vertebrates (Rasey, 2004; Ruczynski & Bogdanowicz, 2008). At a landscape scale, ancient trees provide ecosystem functions and have strong regulatory influences on local nutrient cycles and microclimate (Rubino & McCarthy, 2003; Lonsdale, 2013). Additionally, ancient trees are known for their cultural and historical ties, and can inform us of past land management and use, historical climate and changing social behaviours (Rackham, 1976, 1980; Read, 2000), as well as providing valuable tourism opportunities (Rackham, 1994; Lonsdale, 2013).

Wood-pastures, royal forests and historic parklands are habitats which often contain an abundance of ancient trees (Rackham, 1994; Hartel et al., 2013; 2018; Farjon, 2017). These also include deer parks, commons (land owned collectively by many people with traditional shared grazing or harvesting rights), and chases (common land in the UK used by many for hunting without prosecution). These habitats, referred to here collectively as 'wood-pasture', usually combine livestock grazing with scattered trees either in maiden form or actively managed as pollards, where the tree is periodically cut at breast height and the trunk and branches are removed for use as animal fodder, or for particular industrial purposes (Petit & Watkins, 2003). The resulting landscape is productive, open and relatively undisturbed, providing an ideal environment for the development and persistence of ancient trees (Quelch, 2002; Hartel et al., 2018). Wood-pastures also more generally support high densities of rare flora and fauna (Rosenthal et al., 2012), and their conservation value is recognised throughout Europe (Dorresteijn et al., 2013; Hartel et al., 2018). Several studies have mapped European wood-pasture (Hartel et al., 2013; Plieninger et al., 2015), and it is estimated that it covers an area of ~203,000 km^2 (Plieninger et al., 2015).

Despite their importance, ancient trees are in global decline (Gibbons et al., 2008; Fischer et al., 2010), particularly due to the spread of disease and pests, urbanisation, and agricultural expansion (Read, 2000, ATF, 2005, 2011; Lindenmayer et al., 2012). In addition, there is a lack of tree planting and appropriate management to ensure the continuity and replacement of ancient tree populations and dead-wood habitats (Read, 2000). To add to this, wood-pasture is also considered an increasingly threatened habitat, particularly across Europe, (Hartel & Plieninger, 2014; Forejt et al., 2017), where overgrazing, the decline of old trees, and land-use intensification and conversion are having major impacts (Kirby, 2015). Additionally, although the connection between wood-pasture and ancient trees is generally agreed upon, few studies, with the exception of Hartel et al. (2013; 2018) and Moga et al., (2016) in Romania, have investigated the true abundance or distribution of ancient trees within wood-pastures at an international or even a national scale. Further investigation and quantification of the links between ancient trees and wood-pasture at larger scales i.e. across other regions, countries or continents, would enable more effective conservation and protection of ancient trees.

Compared to Europe and the rest of the world, both the number of ancient trees and the concentration of wood-pastures in the UK, and particularly in England, is extremely high (Rackham, 1994; Fay, 2004; Lonsdale, 2013). This is often attributed to the long history of continuous Royal and aristocratic land ownership and management of forests and parkland (Butler et al., 2002). Additionally, the UK has the most comprehensive ancient tree database in the world; the Ancient Tree Inventory (ATI). The ATI began as a citizen-science collaboration project in 2004 between the Woodland Trust (WT), the Ancient Tree Forum (ATF) and The Tree Register of the British Isles (TROBI), and over 200,000 ancient and other notable trees have been mapped since its beginning (Butler, 2014; Nolan et al., 2020). The extraordinary number of ancient trees recorded in the ATI presents a unique opportunity to investigate quantitatively the large-scale determinants of ancient tree abundance in wood-pastures, with the aim of identifying sites likely to contain undiscovered ancient trees across England.

The non-random, ‘ad-hoc’ recording method of the ATI means that the inventory is thought to be far from complete, and many more ancient trees in the UK, including those at risk from the many factors

that threaten their survival, are likely to have gone unrecorded. This also means the ATI is likely to suffer from high levels of sampling bias, because certain geographical locations or time periods have been more extensively surveyed than others (Phillips et al., 2009; Mair & Ruete, 2016). It is suspected that there are many partially or completely un-surveyed sites, including wood-pasture that actually contain ancient trees; currently ~ 44 % of all ATI ancient trees are located in a wood-pasture, yet these wood-pastures represent only ~ 9 % of the total number of wood-pastures across England. The patchy recorded occurrence of ancient trees means that the data display a high level of zero-inflation, which presents a problem when trying to model tree abundance using conventional methods. Hence, in the present study I use zero-inflated (ZI) models to model the data and create abundance predictions.

The accuracy of large-scale spatial models of the distribution and abundance of organisms is best assessed by comparison with independent data collected in the field (Chatfield, 1995). However, such data are seldom available and model verification typically involves retaining one or more subsets of the original data as pseudo-independent 'test' data sets. In this chapter, I take advantage of the uniquely detailed mapping of trees in England over the past 200 years to perform a novel form of model verification using completely independent data on the location of the organisms I am attempting to model. I use of a series of historical Ordnance Survey (OS) maps with detailed records of trees across England, together with the National Tree Map (NTM) (Bluesky National Tree Map, 2015) which depicts the current location, extent and height of all trees above 3 m across England. By overlaying these maps across time, abundance estimates were obtained for a randomly selected sample of wood-pastures to verify model accuracy and predictive power.

This chapter provides quantitative evidence for the drivers of the important relationship between ancient trees and wood-pastures in England, and highlights the international need to establish and expand ancient tree inventories such as the ATI. I hope these findings will assist with conservation efforts to locate and protect our ancient tree populations, and to ensure their survival into the future.

3.3 Methods

3.3.1 Study area and ancient tree records

Data describing the distribution of 5571 mapped wood-pastures in England were obtained from Natural England (Wood Pasture and Parkland BAP Priority Habitat Inventory for England, accessed 04/12/17) (Fig. 3.1). The digitised wood-pasture polygons cover an area of ~2780 km^2. The wood pasture polygon dataset originally comprises 9815 sites of wood pasture collected from national and local record sources dated between 1976 to 2011 and verified using aerial photography and UK Ordnance Survey (OS) County Series maps (Epoch 1-4) that span 1846 - 1969 (for more information see https://data.gov.uk/dataset/bac6feb6-8222-4665-8abe-8774829ea623/wood-pasture-and-parkland-england). All rivers and metalled roads were excluded by Natural England during digitisation in accordance with their mapping rules, so many wood pastures are artificially divided into multiple polygons. For this analysis all polygons sharing a common place name and within 100 m of each other were dissolved into a single feature, as were all polygons, named or unnamed, within 50 m. Finally, wood pasture polygons with an area less than 1-m^2 were removed, leaving 5571 wood pasture polygons ranging in area from 8.6 m^2 to 270 km^2. The dataset is not exhaustive as, for example, churchyards and other burial grounds that may be considered as wood-pasture or parkland habitat are excluded unless completely within another wood-pasture, but it covers a significant land area and is the most comprehensive and up-to-date inventory of this habitat in England available to date.

Ancient tree records in England were obtained from the ATI (Woodland Trust, accessed 17/12/18). I excluded all unverified (one or two star) records (see Chapter 2 for more information), and 185 records with incorrect or missing grid references. 10,450 records of ancient trees in England were retained, 4,582 (43.8%) of which fall within a wood-pasture polygon. Ancient tree abundance (number of ancient trees per wood-pasture) was subsequently calculated. Abundance ranged from 0 to 392, but was right-skewed with 91.4% of wood-pastures containing no ancient tree records (Fig. A3.1). Thus, the data showed severe zero-inflation i.e. there were significantly more zeroes than expected when compared to a standard Poisson distribution (Van den Broek test 1995: $\chi^2=14,356.69, df = 1, p < 0.001$).

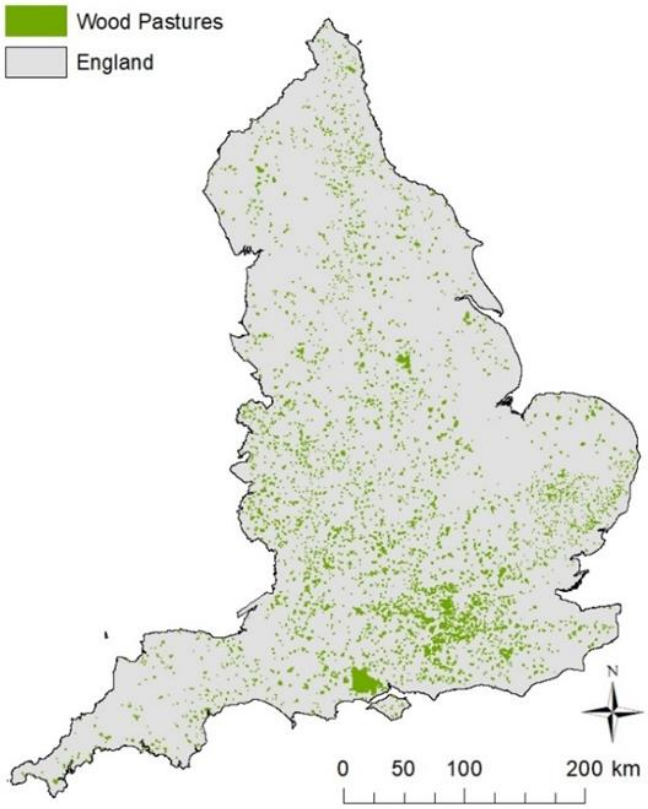


Fig. 3.1 *Distribution of all wood-pasture in England (as mapped by Natural England in Wood Pasture and Parkland BAP Priority Habitat Inventory for England). Although shown in the figure, wood-pastures on the Isle of Wight were not included in the analysis.*

3.3.2 Predictor variables

A variety of sources was used to collect data on 21 characteristics for each wood-pasture (Table 2.5, Table 3.1). Justification for the inclusion of each predictor in this analysis are the same as those detailed in Table 2.5, and all are deemed relevant to ancient trees in wood-pastures for the same reasons outlined previously. Wood-pasture area (km^2) was square-root transformed due to the large range of values. All 16 numeric predictors were centred (the mean of each predictor is subtracted from each value of the predictor) and standardised (the centred values were divided by the standard deviation of the predictor values). Under-represented categories of the three categorical predictors (land classification, countryside type and soil type) were combined to aid model fitting resulting in five categories of land class ('Broadleaved', 'Arable', 'Grassland', 'Urban' and 'Other') (Table A2.3), five categories of soil

type (‘Limited root growth’, ‘Fe/Al’, ‘Clay’, ‘No Profile’ and ‘Other’) (Table A2.5) and four types of countryside (‘Ancient’, ‘Planned’, ‘Highland’ and ‘Cornwall’) (Table A2.4).

Two binomial predictors were used: whether the wood-pasture covered agricultural land or not (4,653 wood-pastures are on agricultural land: defined as all agricultural land ranging from Grade 1-5 (Table A2.6)), and whether the wood-pasture covers land owned by the National Trust (NT); there are 244 wood-pastures on NT land. The minimum resolution possible at which to obtain the categoric predictors (including agricultural land) was 1-km^2, so the value (or average/ most common value if a wood-pasture covered multiple 1-km^2 grid squares) was extracted for each wood-pasture. As a result, many wood-pastures, which are recorded at a smaller resolution than the categoric predictors, fell within squares not necessarily designated as specific wood-pasture or parkland type habitat: some wood-pastures were assigned categories of land use based on squares whose primary designation was agricultural, urban or woodland. Nevertheless, including these land use predictors provides key information about the local environment and surroundings of the wood-pastures, which I believe could be important determinants of ancient tree distributions. All data processing was carried out in ArcGIS (ESRI, 2011) and R (R Core Team, 2018).

Table 3.1 *The 21 variables describing wood-pasture characteristics used as predictors in statistical models of ancient tree abundance. There are 16 continuous predictors, 2 binomial predictors and 3 categoric predictors.* *Reasons for the inclusion of each predictor in the analysis in this chapter are the same as those detailed in table 2.5, and equally apply to wood-pastures.*

Dataset	Predictor (after processing)	Format
Wood-pasture	Wood-pasture area (km^2)	Numeric
Town centres	Distance from nearest town center (km)	Numeric
Major city	Distance from nearest major city (km)	Numeric
Historic forest	Distance from a royal forest (km)	Numeric
Medieval moated site	Distance from a moated site (km)	Numeric
Medieval Deer park	Distance from a medieval deer park (km)	Numeric
Tudor Deer park	Distance from a Tudor deer park (km)	Numeric
Commons	Distance from a commons (km)	Numeric
Ancient woodland	Cover of ancient woodland (%)	Numeric
Traditional orchard	Cover of traditional orchard (%)	Numeric
National Forest	Cover of forest or woodland (%)	Numeric
Buildings	Cover of buildings (%)	Numeric
Major road	Distance from a major road (km)	Numeric
Minor road	Length of minor roads per km^2 of wood-pasture (km)	Numeric
Altitude	Mean altitude across wood-pasture (m)	Numeric
Watercourse	Distance from a water course (km)	Numeric
National Trust land	National Trust owned land	Binomial
Agricultural class	Agricultural Land	Binomial
Countryside type	Type of countryside	Categoric
Soil type	Most common soil type across wood-pasture	Categoric
Land class	Most common land classification	Categoric

3.3.3 Statistical modelling

Zero-inflated (ZI) models (Lambert, 1992) have been used effectively in ecology to model species data with excess zeroes and have been shown to be superior to equivalent Generalised Linear Models (GLM) (Potts & Elith, 2006). They have two parts producing two sets of coefficients; a 'zero' logistic component modelling the probability of an observation being an excess zero, and a 'count' component

generating the count estimates (see Lambert, 1992 or Welsh et al., 1996 for more information), and thus two different types of model predictions can be produced (Zeileis et al., 2008) (see Chapter 5 for more information). If all excess zeros are 'true absences' (arising from either unsuitability of the habitat or ecological stochastic processes) then the 'zero component' is modelling causes of biological aggregation. If some or all excess zeroes arise from 'false absences' (arising from sampling, detection or misclassification errors), abundance predictions from the whole ZI model (hereafter known as 'model abundance' predictions) reflect the abundance that would be observed in the presence of the sampling error in the data. In this case, predictions produced purely from the 'count' component of the ZI model (hereafter known as 'count abundance' predictions), will typically be a better reflection of the true ecological or environmental processes that determine species abundance. As I suspect the excess zeroes arise primarily from the lack of sampling of wood-pastures, I assume that the ZI 'zero' component will predominantly model the processes determining the likelihood that a wood-pasture has been sampled, whereas the 'count' component will model the ecological processes determining the suitability of the wood-pastures for ancient trees.

Ancient tree abundance data were modelled using two ZI models with different distributions: a zero-inflated Poisson model (ZIP) and a zero-inflated negative binomial (NB) model (ZINB), using the 'pscl' package in R (Zeileis et al., 2008). Discrete count data are most commonly modelled using a Poisson distribution and log link function (Zuur et al., 2007; Bolker et al., 2009; Cameron & Trivedi, 2013), assuming that the mean and variance are equal. This assumption is incorrect when dealing with overdispersed or aggregated data where the variance is greater than the mean, which occurs often in ecological data. In these cases, a NB distribution is more suitable, as it allows adjustment of the variance independently from the mean using an extra model parameter, Θ (also referred to as $1/\alpha$) (Gardner et al., 1995). If excess zeroes under a Poisson distribution are the result of biological aggregation, an NB model can be used to account for these extra zeroes. However, if the data are still zero-inflated with respect to a NB distribution, then a ZINB model can be more appropriate. Comparative model fit to the data was assessed using Vuong's (1989) closeness test for non-nested models, likelihood ratio tests (package: 'lmtest': Zeileis & Hothorn, 2002), the significance of the Θ parameter, and visual analysis

of hanging rootograms (package: 'countreg', Kleiber & Zeileis, 2016). ZI models were also fitted separately for the two most common genera present across all wood-pastures (*Quercus* and *Fraxinus*) to assess differences in the environmental determinants between these taxa. No other genera were modelled owing to their low prevalence (comprising < 1% of records of all ancient trees in wood-pastures).

Model predictions were created using 10-fold cross validation; the data were split into 10 equal parts, with each subsample sequentially used as test data, and the other nine subsamples as the training data. Both 'count abundance' and 'model abundance' predictions were considered, as well as the predicted probabilities that each observation is an excess zero (i.e. the probability predictions from the 'zero' component only). Abundance predictions were evaluated against observed ancient tree abundance using Spearman's rank correlation coefficient (r_s) and root mean square log error (RMSLE). In addition, the probability of observing the data based on the predictions was calculated for each model; for every wood-pasture, a Poisson or NB probability distribution function was simulated based on the mean predicted count from the ZIP or ZINB model respectively. The natural log probability of obtaining the observed abundance under this simulated distribution was summed for all wood-pastures to produce an overall probability of obtaining the observed results.

Spatial autocorrelation and collinearity between predictors violate model assumptions and can result in inaccurate parameter and standard error estimates, inflated type I and II errors and biased inferences (Dormann et al., 2013; Thompson et al., 2017). No collinearity between the raw numeric predictors was detected using the following tests and thresholds: VIF (vif = 2), Leamer's method (leamer = 0.1), Pearson correlation coefficients (r = 0. 5) and determinant of the correlation matrix (detr = 0.01). In addition, no collinearity was detected in the fitted model residuals using adjusted generalised VIF (gVIF) (corrected for degrees of freedom) (gvif = 5) (R packages: 'mctest', Imdad Ullah et al., 2016; 'car', Fox and Weisberg, 2011). No spatial autocorrelation was detected in ancient tree abundance across all wood pasture sites using Moran's I (package: 'ape', Paradis & Schliep, 2019) (Observed = -0.0004, Expected = -0.0002, p = 0.339) and correlations between model residuals and wood pasture

midpoint northing and easting coordinates were weak (ZIP Spearman correlations: northing: 0.140, easting: -0.033; ZINB Spearman correlations: northing: 0.115, easting: -0.002).

3.3.4 Model verification

The ideal method for ecological model verification is the evaluation of predictions using an independent dataset, yet it is often time-consuming and costly to collect extra data from the field; here I propose a more efficient, novel method of verification using historic maps. Three map series were selected (Table 3.2), the first two of which are country-wide historic OS maps with detailed records of mature free-standing trees, designated as having a 'very high' or 'high' UK coverage respectively according to the EDINA Historic Digimap Service. The last map is the National Tree Map (NTM) (Bluesky, 2015), a digitised polygon-based dataset of the location, extent and height of all tree canopies over 3 m in height across England and Wales recorded as present in 2015, which is between 116-169 years after the date of the earliest map series I used. By overlaying all three map series (between 1846 – 2015) the persistence of individual trees can be traced over a time to provide an estimate of current ancient tree abundance within wood-pastures.

Table 3.2 *Map series used for the historical desk verification of model predictions. The first two series consist of historic maps (Edina Historic Digimap Service), and the last one of recent (2015) digitised tree canopies of all trees and shrubs above 3 m in height across England.*

	Map Series	**Date**	**Source**
1	County Series First Edition Survey Map (Epoch 1)	1846-1899	Ordnance Survey County Series 1st Edition [TIFF geospatial data], Scale 1:10,560. Using: EDINA Historic Digimap Service, http://edina.ac.uk/digimap, Downloaded: June 2019
2	National Grid Imperial Map First Edition (Epoch i5)	1948-1977	Ordnance Survey National Grid Imperial Map 1st Edition [TIFF geospatial data], Scale 1:10,560. Using: EDINA Historic Digimap Service, http://edina.ac.uk/digimap, Downloaded: July 2019
3	National Tree Map™ (NTM)	2015	Bluesky National Tree Map 2015, http://www.bluesky-world.com/#!national-tree-map/c1pqz. Accessed via Woodland Trust, 2018 as GIS vector layer

All wood pastures were categorised into one of four groups based on the observed presence-absence of ancient trees and the predicted probability of being an excess ('false') zero. These probability predictions were then converted into binary presence-absence. Fixed thresholds (usually 0.5) are often used for classification into binary groups in these circumstances in ecology but have been shown to perform poorly in comparison to more objective, variable methods based on prevalence, mean probability or sensitivity-specificity approaches (Liu et al., 2005; Freeman & Moisen, 2008). Therefore, I chose to use a variable threshold which was the mean predicted probability of a wood-pasture being an excess zero (i.e. the mean zero component probability prediction across all wood-pastures) for each of the 10 different sets of training data i.e. a different threshold therefore was used for each training data set. The four groups therefore comprised a) wood-pastures with ATI records predicted to contain trees, b) wood-pastures with ATI records predicted not to contain trees, c) wood-pastures with no ATI records predicted to contain trees and d) wood-pastures with no ATI records predicted not to contain trees. Fifteen wood-pastures from each group were randomly selected resulting in 60 wood-pastures overall that underwent verification.

Two volunteers from the Woodland Trust digitised all freestanding (i.e. non-woodland) trees within the wood-pasture polygon boundary for the first two map series by placing a single point in the middle of each OS tree symbol. Each of these symbols is taken to mean a mature, free-standing tree (~75-100 years old) at the time of mapping (see https://maps.nls.uk/view/128076885). Only freestanding trees were selected rather than those in woodland patches as these usually were documented using a generic woodland 'symbol' instead. The volunteers had no knowledge of the observed or predicted abundance of ancient trees for each wood-pasture.

NTM Canopy polygons containing a digitised tree from both the first and second OS map series were retained and considered to be ancient as they represented free-standing trees in 2015 which were probably already mature 116-169 years previously, meaning that they were at least 166 years old, and likely to be over 200 years old; the majority of trees reach the mature stage (prior to becoming ancient) by 100 years old (White, 1998). The abundance per wood-pasture of probable ancient trees was thus

obtained. I aimed to account for discrepancies and errors between the map series that may have occurred from either the original mapping methods or the digitising of the paper maps, by allowing an area of uncertainty around each historic tree. The verification process was therefore carried out for three different levels of accuracy using 1) the digitised tree point itself, 2) a 5-m buffer around the digitised tree and 3) a 10-m buffer around the digitised tree.

Verification abundance estimates were assessed against both 'count abundance' and 'model abundance' predictions using Spearman's Rank correlation coefficient (r_s). Linear regression models were fitted in R using the 'stats' package, modelling the ZIP and ZINB model predictions in relation to the verification estimates for the 60 wood-pastures across the three different levels of accuracy (no buffer, 5-km and 10-km). These models were then used to predict total ancient tree abundance across a) all wood-pastures, b) wood-pastures currently containing ancient tree records and c) wood-pastures with no records.

3.4 Results

3.4.1 Model performance, parameter estimates and predictions

Abundance of ancient trees in wood-pastures in England was best modelled with a zero-inflated negative binomial (ZINB) model, which accounts for biological overdispersion as well as additional zero inflation. The ZINB model provided a more appropriate fit to the training data than an equivalent zero-inflated Poisson (ZIP) model, based on the Vuong AIC-corrected test ($z = -5.974$, $p < 0.001$) and the likelihood ratio test ($\chi^2 = 6{,}089.3$, $p < 0.001$). Additionally, the significant Θ parameter in the ZINB model suggests overdispersion is present in the data, meaning the ZIP model is not appropriate to use with this dataset (Table A3.1). Visual analysis of hanging rootograms for each model suggest the ZIP model is highly under-predicting wood-pastures with zero records and over-predicting wood-pastures with small numbers of records (less than 10) (Fig. 3.2).

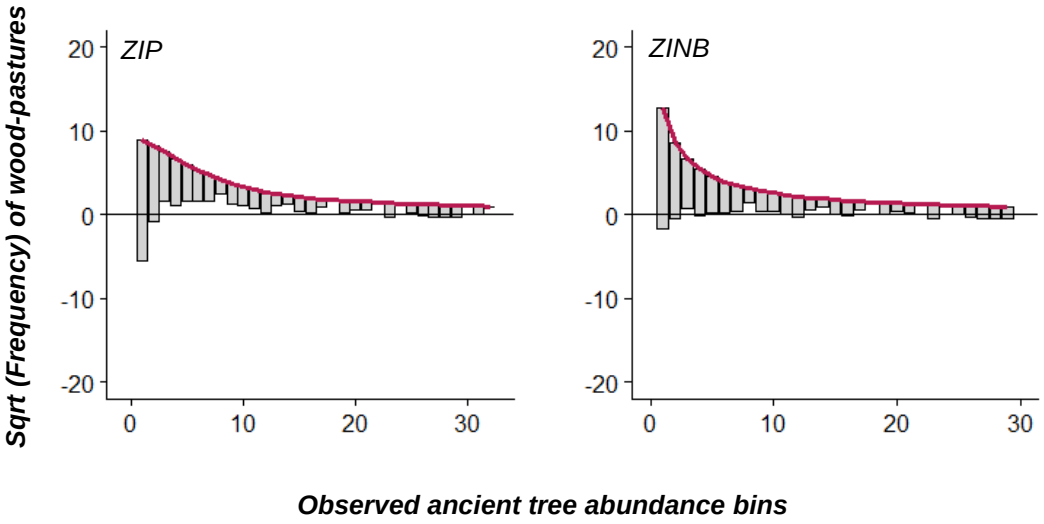


Fig. 3.2 *Hanging rootograms to visualise the fit of the zero-inflated Poisson (ZIP) and negative binomial (ZINB) models to the ancient tree abundance data in English wood-pastures. The (square root) expected number of wood-pastures containing a certain ancient tree abundance is represented by the red line, and the observed number of wood-pastures by the grey bars. Therefore, bars that fall below a count frequency of zero are being under-predicted in a particular count bin, and bars that do not reach a count frequency of zero are being over-predicted by the model.*

The ZINB 'count abundance' (from the count component) predictive performance based on the cross-validation test data was significantly better than that of the ZIP for all three evaluation metrics (predicted probability of obtaining results, r_s and RMSLE) (Fig. 3.3). There was no difference in predictive power of 'model abundance' (from the whole ZI model) for two of the metrics (predicted probability of obtaining results and RMSLE) but ZINB 'model abundance' predictions correlated more strongly with original ancient tree abundance per wood-pasture than those from ZIP. 'Count abundance' predictions suggest that there are 50,784 (ZIP) or 13,848 (ZINB) ancient trees across all wood-pastures in England, which is between 3 and 9 times more than the total number already known (Table 3.3a). ZIP 'model abundance' predictions are quite poor, and actually suggest there are fewer trees in total than those already known about.

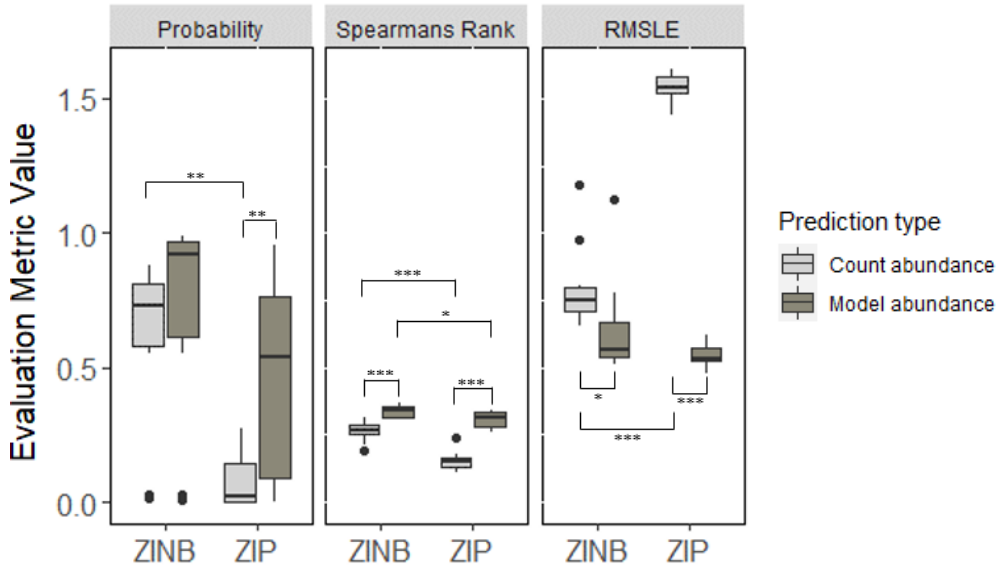


Fig. 3.3 *Evaluation of abundance predictions from the zero-inflated Poisson (ZIP) and negative binomial model (ZINB). Two types of abundance predictions are evaluated: 'count abundance' predictions from the 'count' component of the ZI models and 'model abundance' predictions from the whole ZI model. Values shown represent the median, quartiles and range for each evaluation metric across all 10 cross-validation folds. Three metrics were considered: Spearman's Rank Correlation Coefficient between the model predictions and observed abundance (Spearmans Rank), Root Mean Squared Log Error (RMSLE), and the probability of obtaining the results based on the observed and predicted values (Probability). The best performing models will therefore have the highest Spearmans Rank values, probabilities closest to one and the lowest RMSLE values. Significance levels are represented by $p < 0.05$ = *, $p < 0.01$ = **, $p < 0.001$ = *** and were calculated using the Welch two sample t-test.*

Parameter estimates of the best-performing model (ZINB) suggest ancient tree abundance is positively influenced by increasing wood-pasture area, increasing distance to the nearest city and nearest Royal forest, decreasing distance to the nearest Tudor deer park, distance to a common, and length of minor roads per km^2 of wood-pasture (Table A3.1). Ancient tree abundance is also predicted to differ significantly across certain land classifications and soil types (Fig. 3.4), and is higher on NT and non-agricultural land (Table A3.1). The logistic parameter estimates from the ZINB model provide insight into the factors that influence the odds of a wood-pasture being an excess ('false') zero, which is most

likely to arise because a wood-pasture has not been sampled and has undiscovered ancient trees. Such wood-pastures are more likely to be large, have a low coverage of forest or woodland and are on agricultural land. Soil type and land class also influence the probability a wood-pasture is an excess ('false') zero.

Table 3.3a *Estimates of the abundance of ancient trees for the zero-inflated Poisson (ZIP) and negative binomial model (ZINB) based on predictions from either the 'count' component of the ZI model ('count abundance') or the whole model ('model abundance). Three wood-pastures deemed to be outliers due to extreme predictions (all* 10^{11} *times greater than the next highest predicted abundances) were removed.* ***3.3b*** *Estimates of abundance of ancient trees for the zero-inflated Poisson (ZIP) and negative binomial model (ZINB) based on the historical verification estimates. Estimates were obtained across the three levels of accuracy (no buffer, 5-m buffer and 10-m buffer).*

		Count abundance predictions *('count' component)*			**Model abundance predictions** *('count' and 'zero' component)*		
Model Estimates (a)		All wood-pastures	Wood-pastures with records	Wood-pastures without records	All wood-pastures	Wood-pastures with records	Wood-pastures without records
ZIP		50,784	4,122	46,662	4,376	1,869	2,506
ZINB		13,848	7,118	6,729	11,306	6,909	4,397
Verification Estimates (b)							
ZIP	No buffer	127,489	7,940	119,559	51,468	22,666	28,811
	5-m	1,022,218	63,607	958,633	221,062	97,360	123,752
	10-m	2,526,901	157,206	2,369,717	437,695	192,777	245,032
ZINB	No buffer	101,402	29,900	71,516	70,284	43,120	27,177
	5-m	368,411	108,649	259,836	266,208	163,330	102,949
	10-m	701,925	207,021	495,067	511,783	314,008	197,931

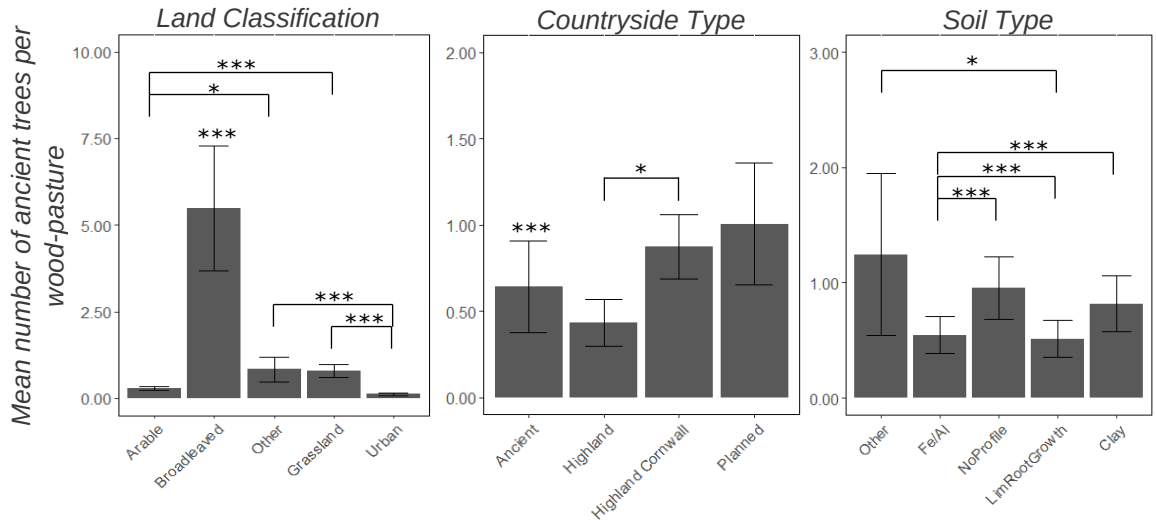


Fig. 3.4 *Mean number of ancient trees per wood-pasture across each categorical predictor. Error bars = ± 1 SE. Dunn's Test of Multiple Comparisons using Rank Sums significance tests were used to assess which categories differ significantly from each other. Significance levels are shown using stars: *** = p < 0.001, ** = p < 0.01, * = p < 0.05, and brackets are used to indicate which categories actually significantly differ from each other. Categories associated with stars but no brackets are significantly different from all other categories.*

There were 321 wood-pastures containg *Quercus* records and 66 containing *Fraxinus* records. The abundances of both genera were positively influenced by wood-pasture area and being on National Trust land, and were negatively influenced by being on 'Limited Root Growth' soil and increasing distance from a watercourse (Table A3.2). *Fraxinus* abundance was also higher at high altitudes, away from major cities, in planned countryside, on grassland or on land with traditional orchards, and lower when nearer to medieval deer parks, there was a large coverage of national forest and on agricultural land. *Quercus* abundance was also positively influenced by being on 'other' soil types and being on broadleaved habitat land, and negatively influenced by being on urban land, at lower altitudes and with greater coverage of minor roads (Table A3.2).

3.4.2 Model verification using historic maps

Verification estimates of ancient tree abundance across 60 selected wood-pastures ranged from 0 to 2,108 across the three levels of spatial accuracy, with mean values ranging from 20 (standard error = ±

4) (no buffer) to 202 (standard error = ± 43) (10-m buffer). 'Count abundance' predictions from the ZINB model correlated remarkably well with the historic map verification estimates, whereas ZIP predictions correlated less strongly (Table 3.4). ZINB predictions performed better as I allowed for greater levels of inaccuracy in the precise location of trees in the historic maps (i.e. as buffer size increased), whereas this had little effect on the ZIP predictions.

Additionally, 100 % of wood-pastures categorised as true positives (predicted to contain records when they actually do) and 13 out of 15 wood-pastures (87 %) categorised as false negatives (predicted to contain records but currently there are none) were verified as having ancient trees using the historic maps. The other two categories are slightly less clear-cut, with 8 out of 15 (53 %) 'true negative' wood-pastures (correctly predicted by the model to contain no records) and 9 out of 15 (60 %) 'false positive' wood-pastures (predicted to not contain records when there are some) having evidence of ancient trees based on historical maps.

Table 3.4 *Spearman's rank correlations (r_s) between the predicted ancient tree abundance from the zero-inflated Poisson (ZIP) or negative binomial model (ZINB), and the historic map verification estimates for 60 selected wood-pastures in England. Predictions considered include the 'count abundance' predictions from the 'count' component of the ZI models and the 'model abundance' predictions from the whole ZI model. Coefficients are shown for each of the three levels of assumed accuracy of the historic maps from highest accuracy (no buffer) to lowest accuracy (10-m buffer) along with p values representing test significance (p < 0.05:*, p < 0.01:**, p < 0.001:***).*

Model	Prediction	Spearman's Rank Coefficient (r_s)		
		No buffer	5 m	10 m
ZIP	Count abundance	0.365**	0.432***	0.432***
	Model abundance	0.663***	0.681***	0.678***
ZINB	Count abundance	0.553***	0.582***	0.594***
	Model abundance	0.701***	0.710***	0.720***

Based on the linear regression models fitted using the ZI 'count abundance' predictions (from the ZI count component) and historic map verification estimates, the total estimates of ancient trees in English wood-pastures range from 101,402 (ZINB with no buffer) to 2,526,901 (ZIP with 10-m buffer) (Table 3.3b). It is most likely the true number falls closer to the lower, more conservative estimate from the ZINB model, which consistently outperformed the equivalent ZIP model in a variety of evaluation metrics. This estimate is 22 times the number of ancient tree records currently in English wood-pastures, and almost 10 times the total number of ancient tree records in England.

3.5 Discussion

Ancient trees are keystone organisms in the landscape, and it is important to understand where they are and how they might best be protected and managed for long-term conservation. This research identified important environmental and anthropogenic factors that positively and negatively influence ancient tree abundance in English wood-pastures, both for all trees, and for *Quercus* and *Fraxinus* genera. As seen in previous studies (Moga et al., 2016; Hartel et al., 2018), wood-pasture area is a strong predictor of ancient tree abundance. This is to be expected, since larger areas by definition can contain more trees, but it may also be the result of historical management and land-ownership: many of the larger wood-pastures are either royal forests or former aristocratic estates, which have actively managed trees over the centuries in ways to continuously sustain and benefit from them (Quelch, 2002). Wood-pasture habitat is an important resource for the development and persistence of ancient tree populations, yet is not considered to be self-sustaining (Quelch, 2013). Constant, active management of both land and trees is needed in the form of sustainable grazing and continuation of traditional pollarding techniques (ATF, 2009; Lonsdale, 2013).

Abundance was also influenced by three human factors, distance to a city, length of minor roads and agricultural land. In all cases, true ancient tree abundance is higher when away from high anthropogenic pressures. There are many threats to the future survival of ancient trees, especially agricultural intensification (Read, 2000; Fay, 2004; ATF, 2005) and urbanisation (Le Roux et al., 2014). It is

important to mitigate these threats, and implement protection measures such as Tree Preservation Orders (TPOs) or scrub planting (Read, 2000; ATF, 2009) and policy changes (Lindenmayer et al., 2014).

Sampling bias is a common artefact in many large species databases (Phillips et al., 2009) and is thought to be present also in the ATI. Verification of the abundance estimates confirmed that the majority (almost 90%) of wood-pastures predicted to have ancient trees, but having no ATI records, did in fact contain at least one ancient tree. Model coefficients from the 'zero' component of a ZI model provide insight into the factors that influence the probability of an excess zero (Lambert, 1992), and thus inform us about predictors of sampling bias in the ATI. One such factor is the occurrence of wood-pastures on agricultural land, or land not covered by ancient woodland or forests. Citizen-science recorders are known to favour interesting areas or species (Kramer-Schadt et al., 2013); for example I found ancient tree abundance to be much higher on NT land. Agricultural land is generally less appealing for ancient tree surveys, and is also is likely to be less accessible and have fewer public rights of way. As ancient trees on agricultural land are likely to be at increased risk of mortality from increasing field sizes, soil compaction, over-grazing and fertiliser applications (Read, 2000; Fay, 2004; ATF, 2005), these areas should be a priority for future surveys which aim to identify ancient trees in need of conservation intervention.

Historic maps are an incredibly useful source of information about past land use, management and socio-cultural factors, yet they are often undervalued in scientific research (Roper, 2003). In the UK the extensive collection of Ordnance Survey maps dating as far back as 1801 provides a unique, unrivalled source of historical landscape characterisation, and have been used successfully in geographical and ecological studies (Cowley et al., 1999; Sutherland, 2012; Visser, 2014). The high level of detail included in these maps, such as the specific locations of individual trees and different types of woodland patches, present a rare opportunity to address ecological research questions such as ours, where environmental, historical and anthropological factors are all being considered to define the niche of organisms that can live to be over 1000 years old.

Abundance estimates from the historic map verification work correlated highly with the model predictions, providing strong support for a) the predictive power of the model, b) the hypothesis that many wood-pastures are 'false absences' and actually do contain ancient trees and c) the benefits of historic maps for addressing landscape-scale scientific questions. The most conservative estimate of ancient tree abundance in English wood-pastures (based on predictions from verification work and ZINB model with no area of uncertainty) was 101,402. Although at first glance this may seem an overestimate, as it represents a 2112 % increase on the known number of ancient trees in wood-pastures, it is not implausible. Because only 9% of wood-pastures contain 10,450 (43%) ATI ancient tree records, a figure close to 100,000 ancient trees (i.e. a 10-fold increase) is possible, depending on the completeness of sampling across all wood-pastures. Other estimates of ancient tree totals have suggested figures close to nine million ancient or veteran trees (trees that are becoming ancient trees or show ancient characteristics) across the whole UK (Fay, 2004). Therefore, my value of ~100,000 in wood-pastures seems if anything conservative. Either way, my predictions highlight the fact that, even in the UK, where sampling is relatively good, most ancient trees in the landscape are yet to be recorded.

It is important to consider the accuracy of the OS maps used to verify my model predictions, especially as the early historic maps are thought to have the most inconsistencies (Harley, 1968; Visser, 2014) and there are likely to be a variety of caveats with using the historic maps, resulting in both under- and overestimation of ancient tree abundance. My decision to map only free-standing ancient trees and exclude woodland patches is likely to have contributed to under-estimation of true abundance: although frequently less common, ancient trees can be found in woodland (Rackham, 1980). Additionally, inconsistencies and the misplacement of the historic tree symbols would also result in underestimation if the tree is still around today but did not fall within an NTM canopy polygon. This risk could be relatively high, particularly as there was no standardised key for the tree symbols in the first OS map. Alternatively, overestimation of abundance may have occurred where the locations of trees recorded during verification actually reflected places in which more than one individual had been recorded over time. For example, a mature tree recorded on an early map may have been felled and another immediately planted in its place. Although I deemed this unlikely to happen, given that the interval

between any two map series was around 50-100 years, barely sufficient time for many species, especially free-standing Oaks, to reach maturity (White, 1998), it could have resulted in some immature or mature trees being labelled as ancient.

Finally, both under- and overestimation of abundance could have occurred owing to the interspecific differences in the age at which a tree reaches maturity and then becomes ancient (White, 1998; ATF, 2008a; Lonsdale, 2013). By assuming that a mature or ancient tree, minimally 40 years old (White, 1989) in the first County series map, will now be at least 200 years old, this time period may be too long for the shorter-lived species to survive until the present day. Many fruit trees such as plum or pear, for example, will never reach 100 years old. Conversely, for some species such as Yew, which is generally only ancient after 800 years, this time period may not be long enough to classify it now as ancient. However, the majority of records were *Quercus* and *Fraxinus*, both of which often survive beyond 200 years, but are very likely to show ancient characteristics by this age or soon thereafter.

Nevertheless, despite these suspected errors, it is likely the under- and overestimation of abundance largely cancel each other out, a view supported by the strong correlations with the model predictions. I believe the potential use and benefits of historical maps for ecological studies is high, and aim to draw attention to the possibilities that these often underused resources offer for research at a landscape scale. I also hope these findings could allow targeted surveys of wood-pastures with high predicted suitability for ancient trees to assist with the conservation and protection of valuable UK ancient trees and wood-pasture habitats.

Chapter 4: Identifying predictors of sampling bias in the Ancient Tree Inventory (ATI) in England.

4.1 Abstract

Sampling bias in large species datasets is problematic and can result in inaccurate, error-prone models and prediction maps when used in Species Distribution Modelling (SDM). There are a variety of proposed methods to correct for sampling bias in SDM, yet there is little consensus as to which is most appropriate. Nevertheless, before attempting to correct for sampling bias it is useful to first gain insight into the potential sources and levels of sampling bias so the most optimum correction methods can be applied. Although the long-running UK Ancient Tree Inventory (ATI) has collected an impressive number of ancient, veteran and notable tree records over the past 15 years through the efforts of citizen-scientists, it is thought to suffer heavily from sampling bias; recording is likely to have been focused in areas with high population densities, easy accessibility and with recreational or aesthetic interest to survey. Therefore, in this chapter, I aim to firstly discuss the problem of sampling bias in species data and potential methods to correct for this, and secondly I investigate and quantify sampling bias in the ATI using a variety of statistical approaches and descriptive methods.

4.2 Introduction

The fields of ecology and conservation rely heavily on understanding links between species distributions and the environment. Species Distribution Modelling (SDM) is widely used to explore these links, and has a broad range of applications including predicting species distributions under climate change (Beaumont et al., 2007; Dormann et al., 2007), modelling the spread of invasive species (Václavík & Meentemeyer, 2012) and conservation planning (Wilson et al., 2006; Linkie et al., 2009). Species occurrence or abundance data from large, observational datasets are increasingly being used in SDM (Pearce & Boyce, 2006; Schmeller et al., 2009; Tiago et al., 2017a). The extensive spatial and temporal coverage of the data, as well as the growing ease of online access, provides numerous benefits over often costly and labour-intensive sampling methods employed in more focused scientific studies of distribution (Dickinson et al., 2010; Dwyer et al., 2016; Gouraguine et al., 2019). However, the increasing frequency of use of such datasets within scientific research has received much comment and criticism (Dickinson et al., 2010; Tulloch et al., 2013; Bird et al., 2014; Tiago et al., 2017b).

Although large species record collections can be generated using hypothesis-led, systematic sampling protocols (Schmeller et al., 2009; Pocock & Evans, 2014), much of the available data for SDM comprise presence-only occurrence records originating from citizen-science projects, museum or herbarium collections, record lists and online databases (Pearce & Boyce, 2006). There is often little information about the source or survey effort accompanying the records, so the assumption of random sampling needed for SDM is infrequently met (Boakes et al., 2010; Rocchini et al., 2011). As a result, sampling bias (also called sample selection or survey bias) is often present in such data: certain temporal periods, geographical areas or taxa are sampled more intensively or frequently than others (Phillips et al., 2009; Dickinson et al., 2010; Bird et al., 2014). This can result from a variety of causes including variation in accessibility e.g. land-use, distance to roads or paths, or elevation (Reddy & Dávalos, 2003; Kadmon et al., 2004; Schulman et al., 2007; Mair & Ruete, 2016), or proximity to a recorder's home or base location (Fourcade et al., 2014; Mair & Ruete, 2016), and a tendency to focus on interesting features such as endangered species or conservation areas (Kramer-Schadt et al., 2013).

Sampling bias in SDM can lead to over- or under-exaggeration of important species-environment relationships (Syfert et al., 2013), so predicted distribution maps may partly represent survey effort rather than species niche requirements (Phillips et al., 2009). Although SDM has been widely used over the last two decades, the influence of sampling bias in SDM has received relatively little attention (Kramer-Schadt et al., 2013; Boria et al., 2014; Mair & Ruete, 2016). Several authors have questioned studies using species datasets where the reliability of the records has not been evaluated (Loiselle et al., 2008; Yackulic et al., 2013; Fourcade et al., 2014). Ideally, the best strategy to mitigate it is to design a sufficient, accurate recording scheme with systematic protocols and recording of survey effort, rather than having to deal with bias retrospectively (Tweddle et al., 2012). Unfortunately, many of the well-known, established citizen science and data recording projects, which have already collected numerous records that span large spatial areas and temporal periods, do not have these ideal characteristics, and yet they are the often the best available source of data for species of conservation significance.

Proposed methods to correct for sampling bias generally rely on either spatial filtering of occurrence records, or the manipulation of background data ('pseudoabsences') (Phillips et al., 2009; Kramer-Schadt et al., 2013; Fourcade et al., 2014, Boria et al., 2014). Spatial filtering techniques such as a systematic re-sampling of occurrence records are one of the best methods to remove sample bias for many models and bias types (Fourcade et al., 2014; Beck et al., 2014). Other filtering techniques involve randomly removing occurrences within a certain distance of each other (Veloz, 2009; Boria et al., 2014; Varela et al., 2014), or sampling one point per cluster (Fourcade et al., 2014). These techniques produce more accurate models with reduced overfitting and autocorrelation, but are limited by sample size, as reducing the number of occurrence records can result in poor model predictions (Wisz et al., 2008). There is also the risk of reducing clustering in areas that truly represent high ecological value for a species (Fourcade et al., 2014).

The second method involves manipulating the background data (i.e. the pseudo-absences) or environmental model predictors so that they mimic the bias in the occurrence data; SDM predictions made using the manipulated background data should then reflect the actual species niche rather than

sampling effort (Ponder et al., 2001; Phillips et al., 2009; Fourcade et al., 2014). Techniques for manipulating background data include restricting the area from which the selection of background pseudo-absence points are taken (Phillips, 2008; Fourcade et al., 2014), splitting the data into different areas (Gonzalez et al., 2011; Fourcade et al., 2013; 2014) and using weighted bias variables in the model representing a direct measure of sampling effort or a proxy, such as a map of road networks (Dudík et al., 2005; Elith et al., 2010). Another method is target group sampling (TGS): locations with occurrence records of a similar species that was surveyed in a comparable way, and where the species/ taxa being modelled is absent, can be classed as 'true' absences (Phillips et al., 2009; Syfert et al., 2013; Hertzog et al., 2014). These approaches usually require some prior knowledge of the source of the bias or an appropriate TGS species, which can limit their application (Dudík et al., 2005; Phillips, 2008).

A third option for tackling the problems caused by sampling bias is the use of statistical models that can account for some of the causes of bias (Bird et al., 2014; Isaac et al., 2014). These include Geographically Weighted Regression (GWR) (Brunsdon et al., 1998), Maximum Entropy (MaxEnt) with a bias layer, autoregressive and spatially-explicit models (Dormann et al., 2007) and mixed effect models (Bird et al., 2014), although again, most of these require prior knowledge of the source of the bias.

Several studies have attempted to compare alternative methods of sampling bias correction (Fourcade et al., 2014; Boria et al., 2014) but generally have compared only one or two methods, and tend to focus on one case study, taxon or species (Fourcade et al., 2014). Therefore, there is not one widely accepted method of correction, and the optimal choice is likely to depend somewhat on the study in question and type of data used. Nevertheless, when using any large species database collected through non-strategic, uncontrolled methods, it is essential to consider sampling bias in the data and apply the most appropriate correction methods in order to produce the most accurate models and predictions.

Before attempting to remove sampling bias from a dataset, it may be useful to first determine the type, strength and possible predictors of the bias (Boakes et al., 2010). For example, Reddy and Dávalos

(2003) confirmed sample bias in citizen science data of African passerines by testing statistically for significant effects of accessibility and conservation status of the areas compared to randomly generated data points. Kadmon et al. (2004) similarly found woody plants in Israel were significantly more sampled along roadsides. In this chapter, I aim to investigate sampling bias in the ATI by using similar methods to identify and quantify potential sampling bias predictors in the ATI. Some of these factors such as distance from roads, distance from towns and elevation are likely to be predictors of the true ecological distribution of trees as well, but some, such as recorder location and accessibility of sites, will only be predictors of sampling bias. By confirming the cause and extent of sampling bias in the ATI, the process of bias correction in the subsequent chapters can be more targeted and relevant to the ATI dataset, and can result in the production of more accurate predictive distribution maps.

4.3 Methods

A grid consisting of 130,754 cells of 1-km x 1-km resolution was created within the boundary of England: this was the maximum total number of grid cells possible that fell completely within the boundary. Ancient and veteran tree records were obtained from the ATI (accessed 17/12/18). I chose to exclude all records with a rating below 3 stars and are therefore un-verified and potentially unreliable (see Chapter 2). This left 93,404 ancient and veteran tree records within my generated grid in England (Fig. 4.1). Ancient and veteran tree abundance was subsequently calculated for each 1-km grid cell by summing the number of occurrence records per cell.

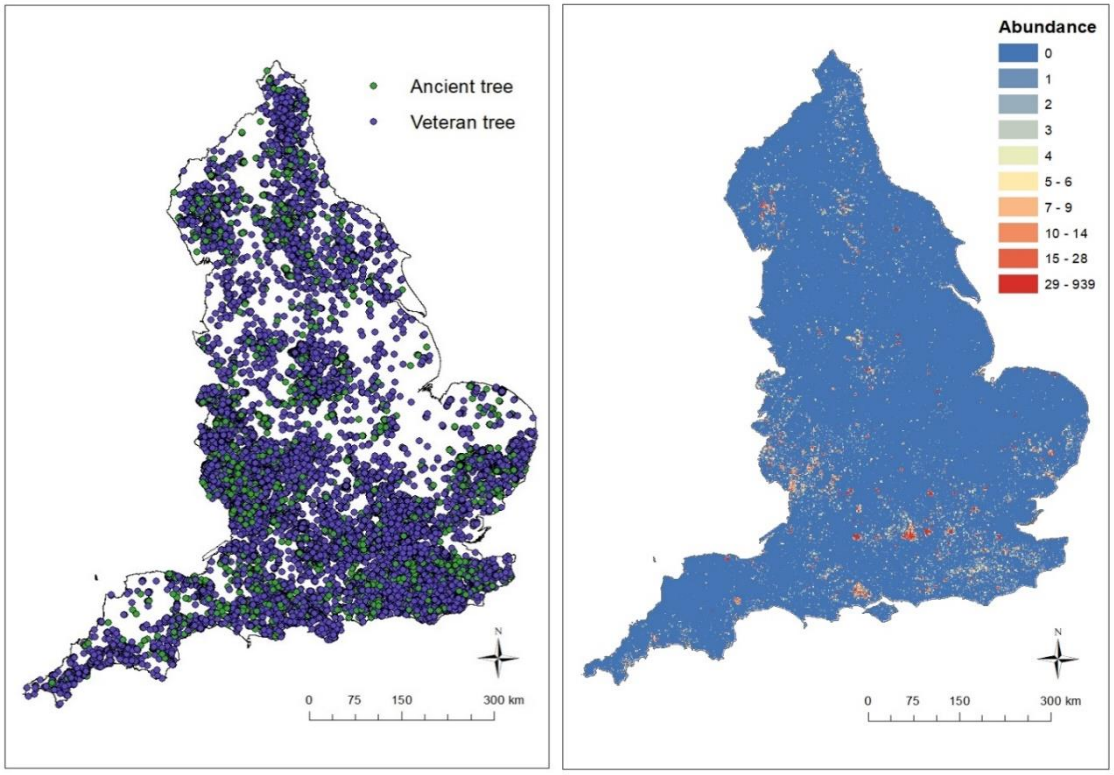


Fig. 4.1 *Left: Ancient and veteran tree records across England from the Ancient Tree Inventory (ATI). There are 94,024 records in total (10,450 ancient and 83,574 veteran). Right: Ancient and veteran tree record abundance (counts of records) per 1-km grid square. Abundance ranges from 0 (blue) to 939 (red).*

Potential predictors of sampling bias in the ATI were hypothesised based on similar studies, literature about the trees, and prior knowledge of the ATI (Table 4.1). These included a mixture of environmental and anthropogenic factors, as well as spatial biases. Raster layers of each predictor were created at a 1-

km resolution across England with the same extent as the 1-km grid cells. In addition, for each original ATI occurrence record, the value of each numeric raster bias predictor was extracted at that precise location. Due to confidentiality issues, the exact location of each recorder's home was unable to be obtained, so recorder home base location was determined as the centroid of all records collected by an individual recorder. There are 1,610 independent recorders of records, consisting of either organisations, projects or individuals. The top source of records is from the Woodland Trust (Woodland Trust batch upload or Ancient Tree Hunt), contributing 48,361 (52.5%) of ancient and veteran tree records. A significant number of records (21,659: 23.5%) have been uploaded by the 10 most active recorders (Fig. 4.2): 912 recorders have only ever uploaded one single record. Visual analysis of potential spatial biases relating to recorder location was carried out through the production of kernel density plots of the centroids for either all recorders or the top 10 recorders. Kernel density plots at a 1-km resolution were created using ArcGIS version 10.3 (ESRI, 2018) based on planar distances between each centroid location.

Spatial autocorrelation of record density was tested using Moran's I, along with collinearity between all numeric bias predictors, tested using Pearson's product correlation coefficient (r). Due to the non-linear relationships between ancient and veteran tree abundance and many of the sampling bias predictors (Fig. 4.2), Spearman's rank correlation coefficient (r_s) significance tests were used to assess the relationship between abundance and each numeric sampling bias predictor per 1-km grid square. Two-sample Kolmogorov-Smirnov (K-S) tests were used to compare statistically the frequency distributions for the extracted raster values of seven of the numeric bias predictors (altitude and distance to nearest town, city, major road, minor road, watercourse or recorder base) at each tree location, to the frequency distributions of the bias predictor values extracted from an equal number of randomly simulated point locations across England.

Finally, significant differences in the frequencies of record densities within each category of land use, agricultural type, National Trust land, Special Area of Conservation (SAC) and countryside type (Table 4.1) were assessed using Chi Squared (χ^2) tests. These compared the observed number of records per

category weighted by the proportion of area covered by each land category. All raster processing was carried out in ArcGIS version 10.3 (ESRI, 2018) and all statistical analysis in R (R Core Team, 2018).

Table 4.1 *Potential predictors of bias in the Ancient Tree Inventory (ATI) based on similar studies, literature about the trees and prior knowledge of the dataset. Predictor values were calculated for each 1-km grid cell and converted to raster format. They include 12 numerical predictors and five categorical predictors.*

Potential bias predictor	Raster predictor (for a 1-km grid square)	Type of predictor	Potential reason/ support from literature for inclusion as a bias predictor
Altitude	Mean altitude (m)	Numeric	High altitude sites are likely to be more difficult to access and survey because of lower road densities, rugged terrain and limited by the physical fitness of recorders.
Distance from a town	Distance from nearest town (km)	Numeric	The greater the distance from a town or city, the lower the population density (i.e. fewer recorders) and the more difficult/ costly/ time-consuming and less desirable it is for recorders to travel to more remote recording sites. (Reddy & Dávalos, 2003; Parnell et al., 2003)
Distance from a city	Distance from nearest city (km)	Numeric	
Distance from a major road	Distance from nearest major road (km)	Numeric	The density of roads across the landscape greatly influences the ability of recorders to access survey sites and also reduces the likelihood or chance encounters with survey species (Freitag et al., 1998; Reddy & Dávalos, 2003; Parnell et al., 2003; Kadmon et al., 2004).
Distance from a minor road	Distance from nearest minor road (km)	Numeric	
Distance from a watercourse (river, stream etc.)	Distance from nearest watercourse (km)	Numeric	Similarly to roads, the density of watercourses also is likely to influence accessibility of sites, as many watercourses are banked by public rights of way or small roads and are more desirable to visit and likely to be surveyed (Reddy & Dávalos, 2003).
Location of recorders	Distance from nearest recorder's base (km)	Numeric	The location of recorders is likely to be one of the most influential factors of sampling bias as recorders are more likely to survey closer to their homes (or in favourite visiting/ holiday spots). Therefore, often species distributions maps reflect recorder density rather than true species distributions (Dennis & Thomas, 2000; Fourcade et al., 2014).
Latitude	Latitude of centre	Numeric	As with recorder location, recording, accessibility and interest in ancient and veteran trees is likely to vary spatially and will therefore influence sampling bias. There is likely to be spatial autocorrelation in the ATI, both from ecological clustering of records and sampling biases.
Longitude	Longitude of centre	Numeric	
Wood-pasture coverage	Cover of wood-pasture (%)	Numeric	As well as wood-pasture, ancient woodland and other types of forest having ecological associations with ancient and veteran trees (Farjon, 2017; Hartel et al., 2018), many wood-pastures and forests are desirable places to visit, and have easy access (foot-paths, roads etc) and tourist attractions (cafes, public toilets etc.), so are likely to be more visited and thus surveyed sites.
Ancient woodland coverage	Cover of ancient woodland (%)	Numeric	
National Forest coverage	Cover of National Forest (%)	Numeric	
Type of agricultural land use	Most common agricultural class	Categoric	Different types of land use e.g. urban or broadleaved type land are likely to have variable interest in their recording and different levels of accessibility. Agricultural land in particular is often difficult to survey unless there are public rights of way or roads across the land (Freitag et al., 1998; Parnell et al., 2003).
Type of general land use	Most common land class	Categoric	
Historic countryside type	Most common historic countryside type	Categoric	Different types of countryside (ancient, planned, highland or highland Cornwall and likely to have different levels of survey interest and accessibility influencing recording (see Table A2.4 for more information).
Whether land is owned by the National Trust	National Trust (NT) Land present in square (Y/N)	Categoric	The National Trust is a charity organisation that owns and manages property and land with historic connections or natural interest, and therefore has strong connections to ancient and veteran trees. National Trust properties are also highly popular tourist attractions, and their accessibility and recreational focus are likely be a major influence on the surveying and large numbers of ancient and veteran trees on these sites (Freitag et al., 1998; Reddy & Dávalos, 2003).
Whether land is a conservation area	Special Area of Conservation (SAC) present in square (Y/N)	Categoric	Similarly to National Trust land, conservation areas are visited by the public for recreation, and are likely to be more surveyed as a result of their desirability and ease of access (Freitag et al., 1998; Reddy & Dávalos, 2003).

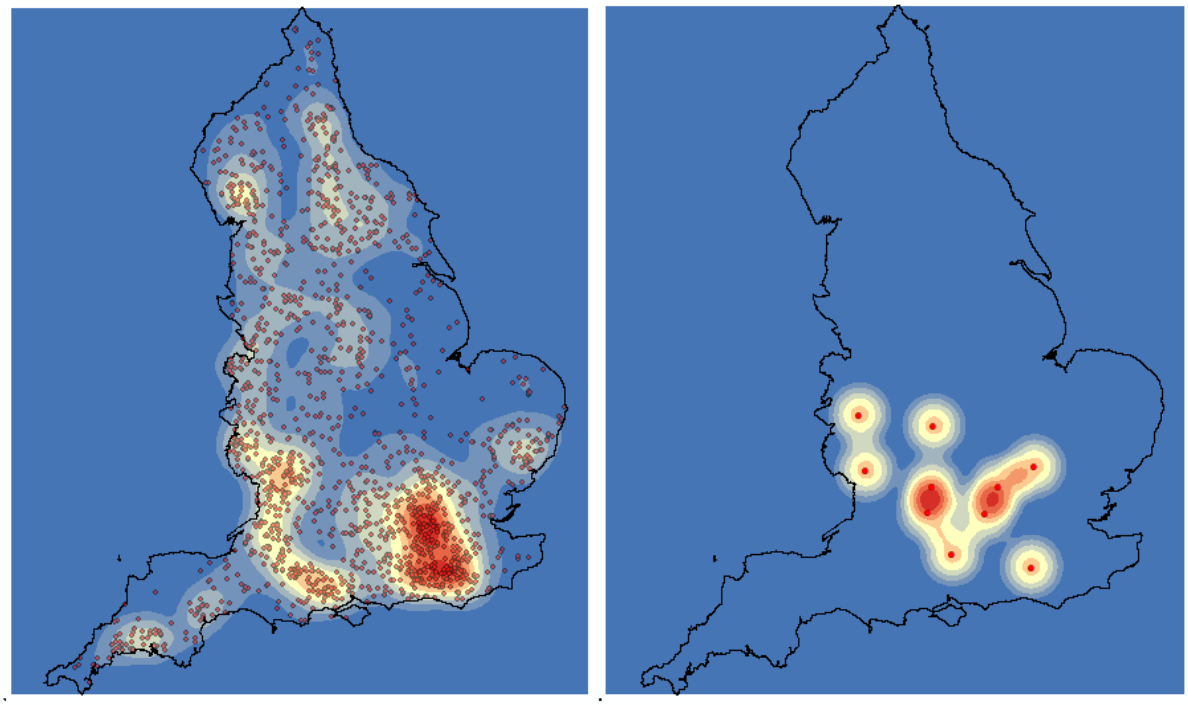


Fig. 4.2 *Left: Centroid locations (red dots) and kernel density plots of all the records uploaded by each individual recorder or organisation of ancient and veteran trees to the Ancient Tree Inventory (ATI) and the kernel-density plots of the centroids. Right: centroid locations of all the records uploaded by each of the 10 most active individual recorders and the kernel-density plots of the centroids.*

4.4 Results

The abundance of ancient and veteran tree records per grid square shows significant spatial clustering (Moran's I: $z = 281.9$, $p < 0.001$) and significant correlations with latitude and longitude (Table 4.2); abundance increases with longitude but decreases with latitude i.e. abundance is highest in the south east of England. Pearson correlation coefficients suggest there is low collinearity between potential numerical bias predictors in Table 4.1 (all r values fall below an absolute value of 0.65) (Fig. 4.3).

There are significant differences between ATI ancient and veteran tree records and randomly generated points in the frequency distributions of altitude, distance from nearest town, city, major road, minor road, watercourse and recorder base (Table 4.2). Additionally, record density per 1-km grid square correlated significantly with all bias predictors (although the coefficients were low for many - see Table 4.2); density was higher in cells significantly closer to towns, cities, roads and rivers and closer to the

nearest recorder's home base. Density was also higher in cells with greater coverage of wood-pasture, ancient woodland and national forest. Interestingly, density was also higher at greater altitudes, which suggests perhaps that this predictor is a greater influence on the real distribution of ancient and veteran trees than on sampling effort.

Ancient and veteran tree record density differed significantly across all five types of land use: agricultural class ($\chi^2 = 89{,}969$, $d.f. = 8$, $p < 0.001$), countryside type ($\chi^2 = 12{,}301$, $d.f. = 3$, $p < 0.001$), land class ($\chi^2 = 126{,}106$, $d.f. = 10$, $p < 0.001$), NT land ($\chi^2 = 19{,}883$, $d.f. = 1$, $p < 0.001$), and Special Area of Conservation (SAC) ($\chi^2 = 226.35$, $d.f. = 1$, $p < 0.001$) (Fig. 4.4). Record density was particularly high on any land considered non-agricultural, broadleaved or other land class, ancient countryside, National Trust land or on an SAC (Fig. 4.4).

Table 4.2 *Spearman's Rank correlation coefficients (r_s) between the value of each sampling bias predictor and record abundance per 1-km grid square. Results from a two-sample Kolmogorov-Smirnov (K-S) test comparing the frequency distributions of potential sampling bias predictor values at record locations and the frequency distributions of values at an equal number of simulated random locations are shown.*

Bias source	**r_s**	**K-S test**
Altitude (m)	0.0187***	D = 0.137, p < 0.001***
Distance from nearest town (km)	-0.0740***	D = 0.172, p < 0.001***
Distance from nearest city (km)	-0.0298***	D = 0.203, p < 0.001***
Distance from nearest major road (km)	-0.0509***	D = 0.124, p < 0.001***
Distance from nearest minor road (km)	-0.0602***	D = 0.113, p < 0.001***
Distance from nearest watercourse (km)	-0.0415***	D = 0.183, p < 0.001***
Distance from nearest recorders base (km)	-0.2086***	D = 0.074, p < 0.001***
Cover of wood-pasture	0.2551***	-
Cover of ancient woodland	0.1873***	-
Cover of national forest	0.1892***	-
Latitude	-0.1000***	-
Longitude	0.0092***	-

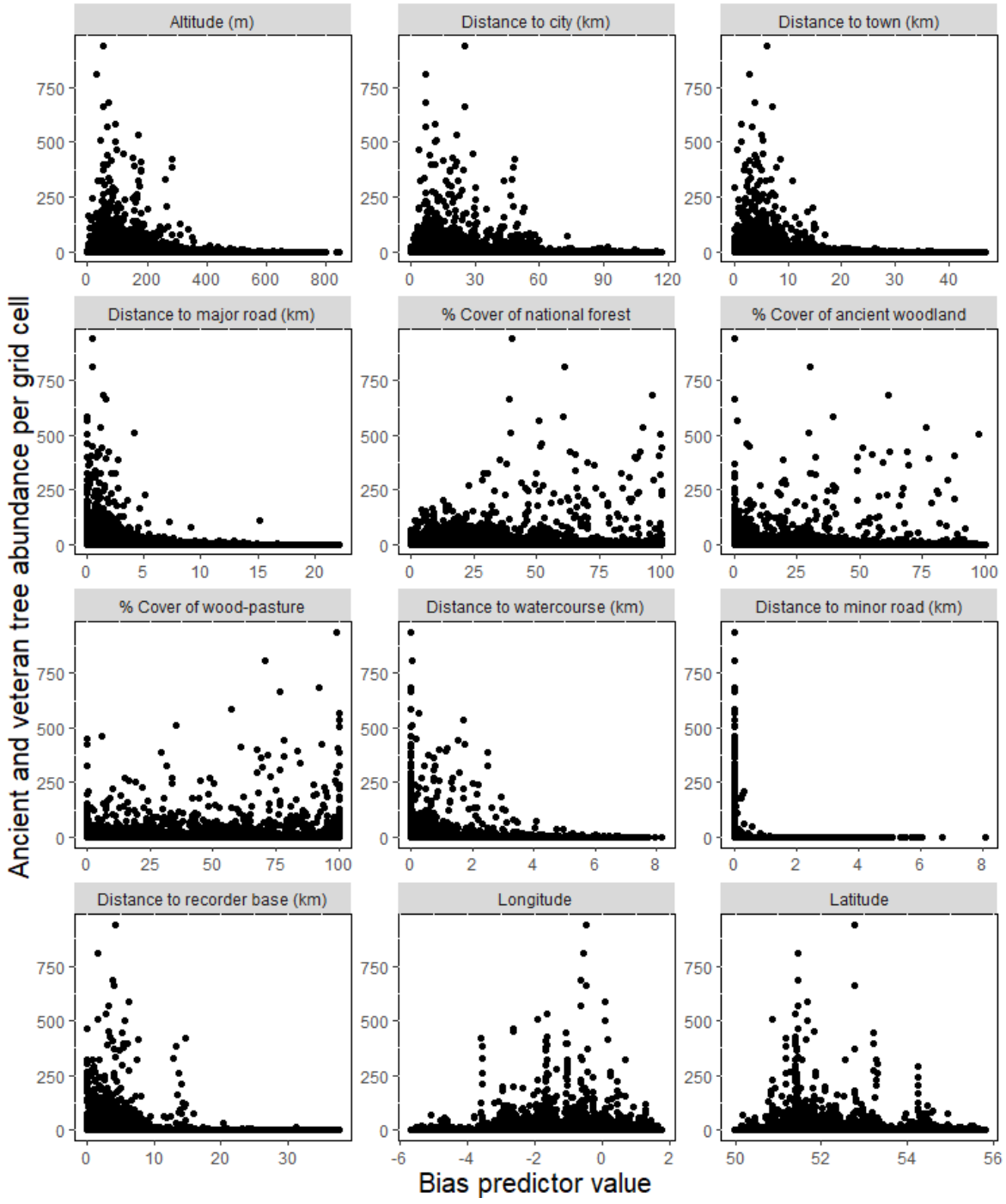


Fig. 4.3 *Scatterplots of the relationships between ancient and veteran tree abundance from the Ancient Tree Inventory (ATI) and each of the 12 numerical sampling bias predictors for each 1-km grid cell across England.*

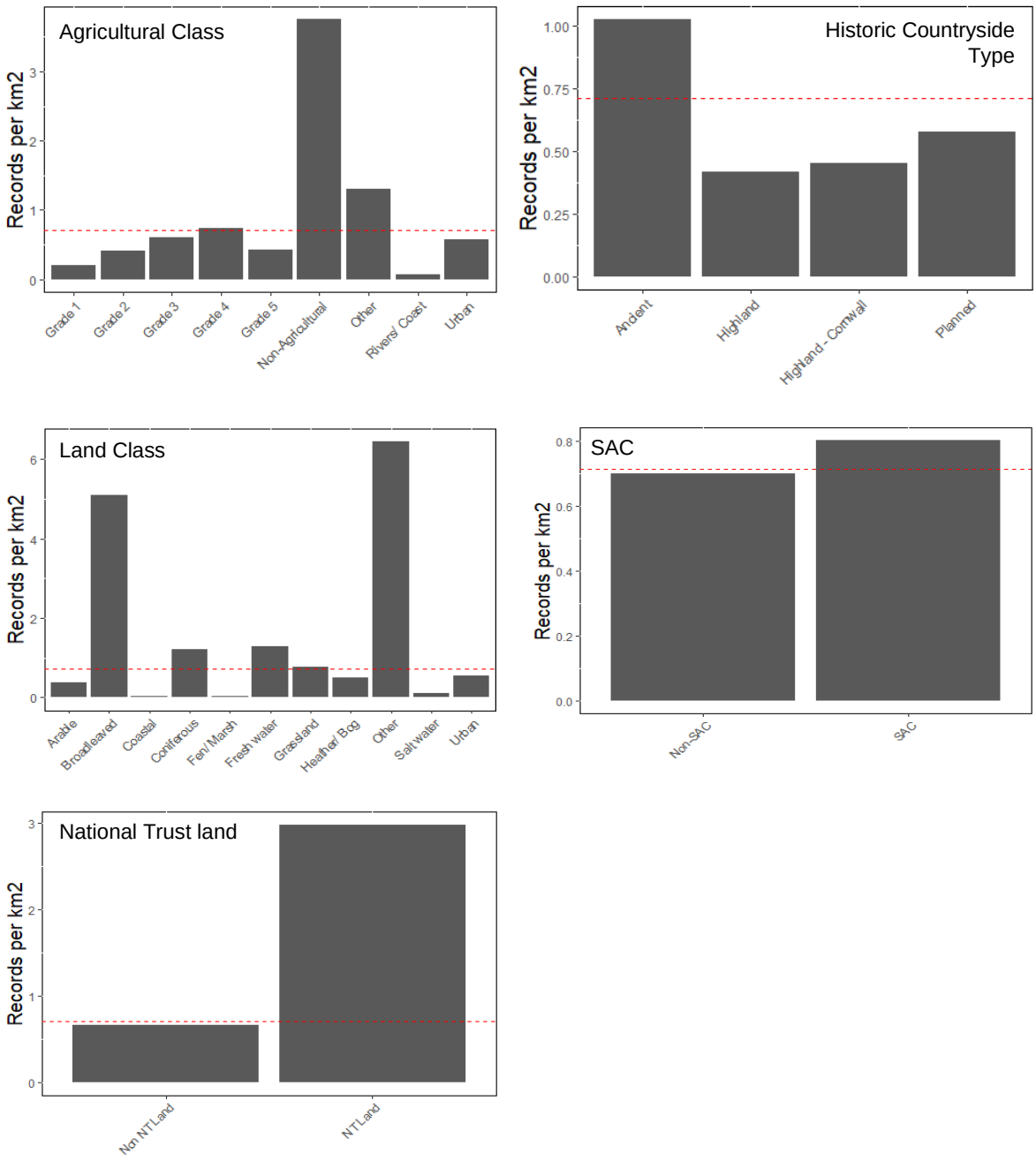


Fig. 4.4 *Number of ancient or veteran tree records in the Ancient Tree Inventory (ATI) per km^2 of each land type (Agricultural Class, Historic Countryside Type, Land Class, Special Area of Conservation (SAC) or National Trust land). Dotted red lines represent the expected record density of records if land type has no influence on abundance.*

4.5 Discussion

Analysis of potential sampling bias predictors in the ATI identified a variety of environmental, anthropogenic and land use variables that are likely to have some influence on the recording of ancient and veteran trees across England. To summarise, recording is likely to be biased by all anthropogenic factors including recorder location, and distance to features such as cities, towns, rivers and roads. There is also spatial clustering in the ATI with significant correlations between abundance and latitude and longitude, although these relationships appeared to be non-linear so are likely to have an even greater influence than suggested from the correlation analysis. Therefore, bias correction measures might focus on the inclusion of some or all of these factors as either predictors or weights for the selection of absence points in species distribution models (Dudík et al., 2005; Elith et al., 2010).

The cover of wood-pasture, ancient woodland and national forest also all positively influenced the record density, which may be a result of ecological influences, but is possibly an artefact of sampling bias through recorders selecting places they deem more likely to contain such trees, as well selecting more desirable and aesthetic places to visit and survey. There is also likely to be a similar process happening with the other land uses, especially National Trust land and SACs, as in fact many National Trust sites and SACs are also wood-pastures or woodland (Harvey, 1987; Nolan et al., 2020). Bias corrections methods that might be applicable in these cases could include spatial filtering of occurrence records to remove emphasis on highly surveyed areas (Fourcade et al., 2014).

The only factor identified as unlikely to be influencing sampling bias was altitude: abundance was greater with higher altitudes, the opposite of my initial hypothesis that accessibility and terrain would limit surveying at higher altitudes. Nevertheless, the correlation between altitude and abundance is weak (0.019), so there may be some underlying bias not picked up, for example sampling bias at low altitudes e.g. coastal areas where accessibility may be low. All of these bias predictors have the potential to influence the SDM process and to reduce the accuracy and robustness of the predictions and resulting distribution maps. Therefore, sampling bias correction with regard to these potential bias sources will be the focus of Chapters 5 and 6.

Chapter 5: Solving sampling bias problems in presence-absence or presence-only species data using zero-inflated models.

5.1 Abstract

Large species record databases such as those generated through citizen science projects, archives or museum collections, are being used with increasing frequency in Species Distribution Modelling (SDM) for conservation and land management. Although the broad spatial and temporal coverage of the data is advantageous to ecological researchers, such data often suffer from sampling bias and consequently, zero-inflation; there are more zeroes (which are potentially 'false absences') in the data than expected. In this chapter, I demonstrate how pooling presence-absence or presence-only data into a 'pseudo-abundance' count, can allow identification and removal of sampling bias through the use of zero-inflated (ZI) models, and thus solves a common SDM problem. I present the results of a series of simulations based on hypothetical ecological scenarios of data collection using random and non-random sampling strategies. My simulations assume that the locations of occurrence records are known at a high spatial resolution, but that the absence of occurrence records may reflect under-sampling. To simulate pooling of presence-absence or presence-only data, I count occurrence records at intermediate and coarse spatial resolutions, and use ZI models to predict the counts (species abundance per grid cell) from environmental layers. The results show that ZI models can successfully identify predictors of bias in species data, and produce abundance prediction maps that are free from that bias. This phenomenon holds across multiple spatial scales, thereby presenting an advantage over presence-only SDM methods such as binomial GLMs or MaxEnt, where information about species density is lost, and model performance declines at coarser scales. My results highlight the value of converting presence-absence or presence-only species data to 'pseudo-abundance' and using ZI models to address the problem of sampling bias. This method has huge potential for ecological researchers when using large species datasets for research and conservation.

5.2 Introduction

Species Distribution Modelling (SDM) is widely used to address important ecological questions about species distributions and the environment (Dormann et al., 2007; Phillips et al., 2009; Elith et al., 2011). Species occurrence or abundance data from large, observational datasets are increasingly being used in SDM (Pearce & Boyce, 2006; Schmeller et al., 2009; Tiago et al., 2017b). The extensive spatial and temporal coverage of the data, as well as the growing ease of online access, provides numerous benefits over often costly and labour-intensive sampling methods employed in more focused scientific studies of distribution (Dickinson et al., 2010; Dwyer et al., 2016; Gouraguine et al., 2019). Although large species record collections can be generated using hypothesis-led, systematic sampling protocols (Schmeller et al, 2009; Pocock & Evans, 2014), much of the available data for SDM comprise presence-only occurrence records originating from citizen-science projects, museum or herbarium collections, record lists and online databases (Pearce & Boyce, 2006). There is often little information about the source or survey effort accompanying the records (Boakes et al., 2010; Rocchini et al., 2011), and as a result sampling bias is often present in such data: certain temporal periods, geographical areas or taxa are sampled more intensively or frequently than others (Phillips et al., 2009; Dickinson et al., 2010; Bird et al., 2014).

Sampling bias in SDM can lead to over- or under-estimation of important species-environment relationships (Syfert et al., 2013), and predicted distribution maps may partly represent survey effort rather than species niche requirements (Phillips et al., 2009). Proposed methods to correct for sampling bias generally rely on either spatial filtering of occurrence records, or on the manipulation of background data (‘pseudoabsences’) (Phillips et al. 2009; Kramer-Schadt et al., 2013; Fourcade et al., 2014, Boria et al., 2014). Both of these techniques have limitations: the former results in a dataset of reduced sample size and statistical power (Wisz et al., 2008), whereas the latter usually requires some prior knowledge of the source of the bias (Dudík et al., 2005; Phillips, 2008). A third option is the use of statistical models that can account for some of the causes of sampling bias (Bird et al., 2014; Isaac et al., 2014). These include Geographically Weighted Regression (GWR) (Brunsdon et al., 1998), Maximum Entropy (MaxEnt) with a bias layer, autoregressive and spatially-explicit models (Dormann

et al., 2007) and mixed effect models (Bird et al., 2014), although again, most of these require prior knowledge of the source of the bias.

One specific problem with many large species databases, which is partly caused by sampling bias, and which is especially noticeable in databases that record species abundances, is zero-inflation: the presence of more recorded zeroes or locations where data are absent than expected under standard distributions (binomial, Poisson, negative binomial etc.) (Martin et al., 2005). These excess zeros can arise from multiple processes. Some are considered to be 'true zeros', which result from either ecological processes that render a site unsuitable for occupancy by a given species, or stochastic processes, such as a sudden random extinction event in an otherwise suitable location (Cunningham & Lindenmayer, 2005; Martin et al., 2005). In contrast, 'false zeros' are locations where a species occurs but was not recorded because of errors or omissions in the sampling method (Dénes et al., 2015). These errors are either systematic and occur repeatedly throughout the survey process (for example through a lack of detection or poor survey design), or are owing to sampling bias, because some geographical areas have not been sampled at all (Bird et al., 2014).

Generalised Linear Models (GLMs) are a common method for analysing relationships between species occurrences or abundance and environmental variables, but excess zeros are problematic for GLMs. If excess zeros are not accounted for, GLMs may suffer from biased parameter estimates and poor predictive power (Lambert, 1992). As a possible solution to this problem, zero-inflated (ZI) models and their components (extensions of GLMs) have been widely discussed in the literature (Lambert, 1992; Welsh et al., 1996; Zuur et al., 2009). ZI models consist of two parts: a logistic component that models the probability of an observation being an excess zero (hereafter called the "zero component"), and a "count component" that models a count (e.g. species abundance) under an assumed distribution (Lambert, 1992). Both components of ZI models are capable of producing zeros, and a key feature is the ability to include different predictor combinations in each component. In other words, they can model the different sources of zeros independently (Wenger & Freeman, 2008; Zuur et al., 2009).

ZI models, which require counts of occurrences (i.e. abundance), are rarely considered in SDM, because most large datasets record only the presence of a species at a site, not the abundance. SDM methods that can use presence-only data, such as MaxEnt, are therefore most commonly applied (Phillips & Dudík, 2008; Fitzpatrick et al., 2013; Fourcade et al., 2014). Furthermore, where abundance is recorded, sampling effort is often not standardised across sites, and hence variation in abundance may simply reflect variation in sampling effort. Thus, SDM typically attempts to predict species presence, rather than abundance. However, the ability of ZI models to model separately the two processes underlying the generation of zeroes in a species dataset could provide an alternative method to model and account for sampling bias. In addition, ZI models can be used with any species database that records abundance directly, or by aggregating presence-only or presence-absence data into counts of occurrence that can be modelled using common count distributions. In this chapter, I therefore propose ZI models as a new, alternative method to address problems of sampling bias in SDM. I present here the results of a series of simulations based on hypothetical ecological scenarios representing the large-scale collection of species occurrence data that aim to address three particular research questions.

My first research question is whether undersampling and sampling bias (resulting in excess 'false' zeroes) can be modelled and accounted for using ZI models, in order to improve species distribution predictions. ZI models have been used effectively to model true and false zeros in ecological count data, such as when modelling the abundance of rare species (Welsh et al., 1996; Cunningham & Lindenmayer, 2005; Martin et al., 2005). They are particularly prevalent in the field of occupancy-abundance modelling (Sileshi et al., 2009; Smith et al., 2012), especially when there are false zeros in the data owing to systematic sampling errors from imperfect detection (Wenger & Freeman, 2008; Sólymos et al., 2012; Williams et al., 2016). Such occupancy-abundance models can account for detection errors without the need for repeated sampling (Sólymos al., 2012; Dénes et al., 2015). However, research into zero-inflation caused by spatio-temporal sampling bias in species occurrence data is scarce. A few studies have used ZI models to identify and quantify sources of bias in species data (Dwyer et al., 2016; Williams et al., 2016; Tiago et al., 2017a), yet none have tested the ability of the models to produce accurate predictions of species occurrence or abundance from biased data. I

outline through my simulations how accurate species distribution maps can be produced using ZI models to fit ZI data suffering from sampling bias, and I describe the required criteria during model fitting and prediction for this to occur. In particular, my simulations also address my second research question: under what levels of zero-inflation is my ZI model method most appropriate?

My final research question considers the benefits of pooling fine-scale occurrence data to model occurrence density across coarser spatial scales. Species presence is normally modelled at the smallest spatial scale (grid cell size) possible, given the resolution of the records and environmental layers used to build the model. Counting or aggregating presences across grid cells at a larger spatial scale to generate "abundance" data intuitively seems to be a bad idea, because it throws away information about the precise location of the records. However, this process of aggregating occurrences may be inevitable if predictor layers have lower spatial resolution than occurrence location data, and I propose here that it may actually present considerable advantages. Aggregated counts of occurrences are commonly not a direct measure of true abundance (the total number of individuals of the target species), since each raw occurrence often represents a locality which is home to several or many individuals. Regardless, modelling 'abundance', and any zero-inflation therein, may give important clues to sources of bias in the data which are not obvious in the raw occurrences, and the benefits of being able to identify and eliminate bias could outweigh the costs of any loss of spatial resolution caused by aggregation. Therefore, counting occurrence records at larger spatial scales in order to model "occurrence density' across the study area may be a better alternative to traditional presence-only SDM methods. Indeed, abundance models have been shown to perform better than presence-absence models fitted using the same data across multiple spatial scales (Howard et al., 2014; Johnston et al., 2015).

Other methods do exist that propose aggregating occurrences into counts of 'abundance' that may also provide advantages when using spatially biased species data, including Poisson point models (Renner et al., 2015; Komori et al., 2020). These models can incorporate bias predictors when modelling intensity rather than occurrences across the study area. Nevertheless, they still require a-priori

knowledge about potential bias predictors, whereas I show here that ZI models are able to provide an indication of potential sources of sampling bias in the data when the exact sources are unknown.

I do not attempt to provide a detailed statistical summary of ZI models and theory (there is much associated literature already available), but aim to draw attention to the main modelling methods and usefulness of ZI models for ecological researchers and species distribution modellers dealing with large, biased databases. This method benefits from being applicable to both presence-only and presence-absence data: ZI models can be built using counts of occurrence made at a coarser resolution than the original data. I argue that ZI models can provide insight into, and correction methods for, the bias in large species databases, and that they can be powerful and effective SDM tools.

5.3 Methods

My general approach was to use ZI models to predict the observed number of species occurrences per grid cell for a series of simulated species using predictors of either the biology of the species and/or sampling bias in the data. I envisaged a large species for which it is theoretically possible to survey all individuals in a landscape (e.g. trees, large animals). The true distribution of all individuals was simulated for each species, and this distribution was then sampled incompletely, with or without spatial sampling bias. Before sampling, the true abundance of the species could be calculated by summing occurrences per grid square. But with incomplete sampling, the observed or "sampling abundance" per grid cell is an underestimate. An alternative way to view my simulations, which is more realistic for species which are small or hard to enumerate (e.g. smaller plants, most insects), is to consider each occurrence in the raw data to represent a recorded encounter with the species at a local site which may contain many individuals. In such cases, the models do not strictly predict abundance, but instead they predict what I might call "occurrence density".

As a result of the two-part nature of ZI models, two types of abundance predictions can be produced. Assuming that all excess zeros arise from incomplete sampling, the first type of prediction is of true,

biological abundance (or occurrence density) across the study area, created only from the count component of the model, which I call here the 'count abundance prediction'. This is likely to be the desired modelling outcome, especially for conservation and land-management planning. The second type of prediction, which I here call the 'sampling abundance prediction', comes from the whole model (combining both the count and zero component) and therefore represents the predicted abundance (or occurrence density) that would be recorded if sampling were carried out in the same way as when collecting the data that were used to fit the model. Bias in sampling will be reflected in this second prediction. However, if some excess zeros arise also from biological zero-inflation, for example if a species is clustered, the zero component will reflect some of the underlying biological processes as well as the sampling bias. In this case, the count abundance prediction will only partially reflect the true species abundance. The best type of prediction to use will therefore depend on the estimated strength of biological zero-inflation versus the bias in the data.

5.3.1 Simulation study area and predictor variables

I simulated the occurrence of a hypothetical species in a study area that consisted of a 100 x 100 cell grid at 1-km^2 resolution placed randomly within the boundary of England (Fig. 5.1). The total area covered by the grid is therefore 10,000 km^2 and there are 10,000 individual grid cells. Two predictor variables were selected across this area. The first was a 'biological predictor' that I chose to be 'altitude', which I used to define the relationship between the simulated species occurrences and environment (Meynard et al., 2019). Real values for altitude (m) across the study area were obtained from WorldClim DEM (accessed 10/05/18) at a 1-km^2 resolution and ranged from 0 to 284 m above sea level (Fig. 5.1). The choice of biological predictor for a simulation study of this sort is necessarily somewhat arbitrary, but I chose altitude because it is both a plausible predictor of occurrence for a range of organisms, and it is quite strongly spatially auto-correlated, an important possible source of biological zero inflation in the abundance data formed when occurrences are counted across grid cells at intermediate spatial scales. The actual biological mechanism underlying the relationship between altitude and species occurrences is not important for this study, but altitude is a good proxy for a suite of environmental variables such

as temperature or precipitation commonly used in SDM which have direct effects on species distributions.

Because altitude is spatially autocorrelated, and so is the sampling bias I wanted to investigate (see below), there was a risk that biological and sampling bias predictors in my simulations could correlate: depending on the positions of the simulated towns on my map, there could be a strong correlation between real altitude and sampling effort. Thus, in order to allow us to investigate the impact of sampling bias completely independently of the biological predictor, I also generated an alternative 'biological predictor' with no autocorrelation: a spatially random control variable. This control variable (henceforth labelled 'altitude_randomised') was created by randomising the real altitude values across the study area at a 1-km^2 resolution (Fig. 5.1), and hence removed any correlation between altitude and distance from town. Pearson's correlation coefficients are also shown for this predictor across each replicate simulation (Table 5.1).

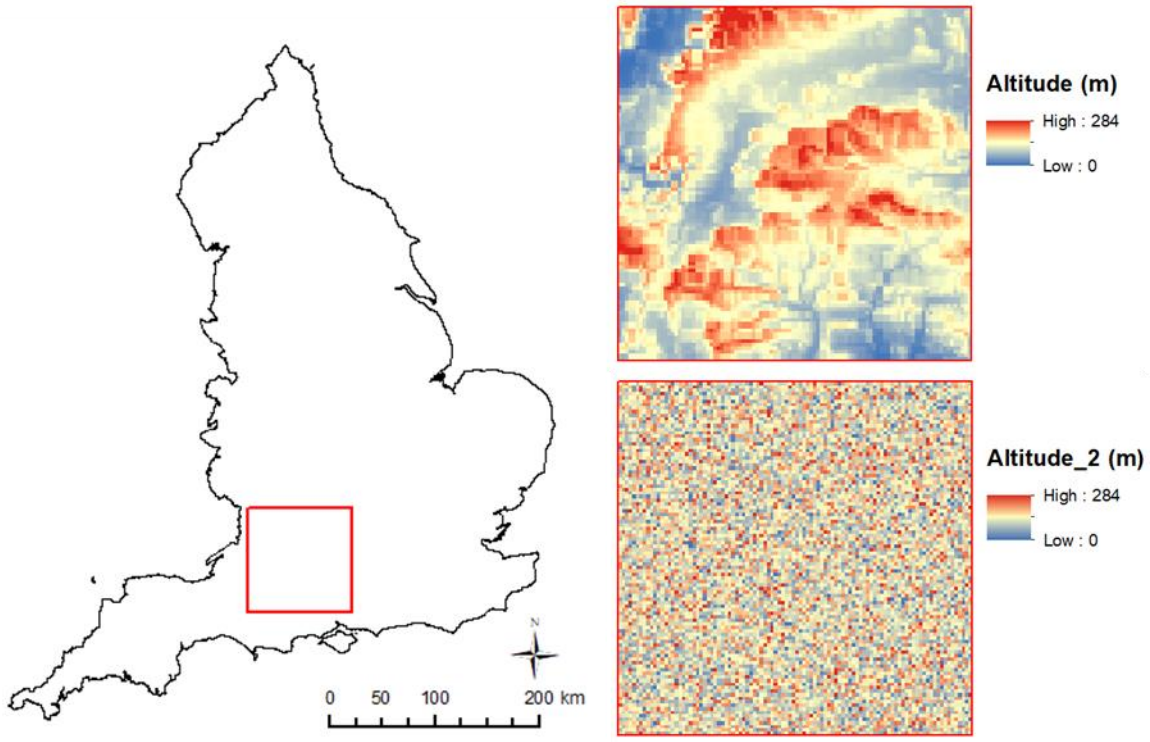


Fig. 5.1 *Simulation study area consisting of a group of 100 x 100 grid squares of 1 km^2 size randomly placed within England covering a total area of 10,000 km^2 (outlined in red) (left), with the biological predictors: altitude (m) and altitude_randomised (m) (randomised altitude layer with no spatial autocorrelation) shown for the study area (right).*

The second predictor of observed species occurrence was a 'bias predictor' ('distance from nearest town') which affected the virtual sampling of the simulated species. I assumed that the greater the distance from a town, the lower the feasibility and likelihood of sampling occurring, as has previously been seen in ecological studies (Reddy & Dávalos, 2003; Parnell et al., 2003). Unlike with altitude, I chose to simulate a hypothetical bias layer rather than use values based on the locations of real towns, in order to ensure the lowest possible correlation between the two predictors, although some correlation between them was likely because of spatial autocorrelation in both. Within the study area, 10 points representing 'town centres' were randomly placed, and the distance from the nearest town (m) was calculated for each grid cell, creating a continuous predictor layer at 1-km^2 resolution across the study area. The process of generating the 'town centres' was repeated 10 times, creating 10 sets of randomly placed 'town centres' (Fig. A5.1.1): by both randomising town locations and repeating this process 10 times, impacts of spatial autocorrelation between the bias and biological (altitude or altitude_randomised) predictors can be limited as much as possible. Pearson's correlation coefficients between predictors are shown for each repeat (Table 5.1).

Table 5.1 *Pearson's correlation coefficient (r) between altitude or altitude_randomised (biological predictors) and distance from the nearest town (bias predictor) across the 10 maps with randomly generated sets of 'town centre' locations.*

	Pearson's correlation coefficient (r)	
Repetition	**Altitude**	**Altitude_randomised**
1	0.277	-0.022
2	0.053	0.011
3	-0.114	-0.009
4	-0.197	-0.014
5	0.331	-0.018
6	-0.067	0.006
7	-0.397	-0.002
8	-0.245	0.010
9	0.031	-0.007
10	-0.171	-0.000

To summarise, I had three variables in total across the simulation study area: two biological predictors ('altitude' and 'altitude_randomised'), and one bias predictor ('distance from nearest town'). All predictors were centred (the mean of each predictor was subtracted from each value of the predictor) and scaled (the centred values were divided by the standard deviation of the predictor values) so that the differences in units of the predictors was removed.

5.3.2 Simulating the virtual species

To obtain counts of 'abundance' to use in ZI models, I first simulated species occurrences across the study area and then aggregated them into counts of "abundance" (alternatively interpreted as occurrence density – see above). Because I assumed that the simulated distribution of occurrences was the complete true distribution, all other locations are assumed to be 'true absences'. Therefore, when aggregating the raw occurrence points into 'abundance' counts, a value of 0 represented a true absence and any value greater than 0 a true presence.

The recommended first step in a simulation study is to define the relationship between the environment and occurrence points (Meynard et al., 2019). I modelled the distributions of three simulated species each with 5,000 occurrence points (Fig. 5.2). The occurrence points of the first species ('random species') were simulated randomly across the study area, and show no preference for any environmental condition. The second and third species were simulated based on the two biological predictors ('altitude' and 'altitude_randomised') and were assumed to favour high altitudes; these species were named 'altitude species' and 'altitude_randomised species' respectively. I chose these three scenarios in order to create datasets in which different kinds of zero-inflation occur. For the random species, zero-inflation can only occur as a result of sampling (where sites which are not sampled might be incorrectly recorded as zeros), whilst for the altitude species and altitude_randomised species, zero-inflation can result both from sampling and from the fact that grid cells are potentially not suitable for the species because of environmental conditions.

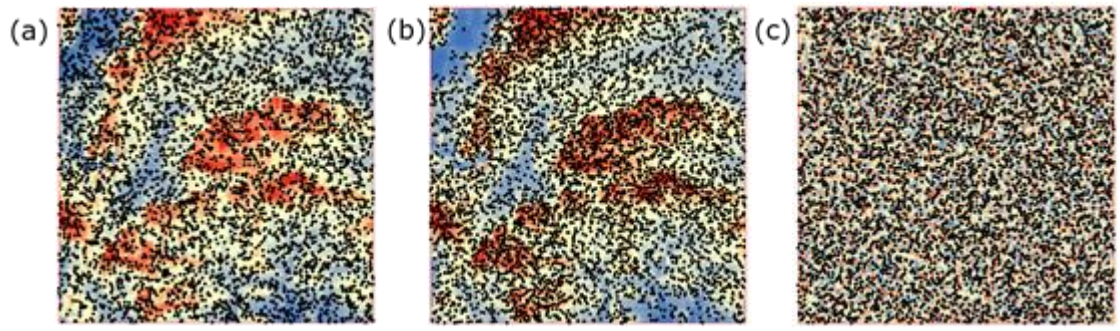


Fig. 5.2 *A simulated species with 5,000 occurrence points showing a) no preference for altitude (random species), b) a preference for high altitudes based on a logarithmic scaler of altitude (altitude species), and c) a preference for high altitudes based on a logarithmic scaler of altitude_randomised (altitude_randomised species).*

I then simulated the effect of the relationship between the biological predictors and species occurrences by creating layers of the probability of occurrence which varied according to altitude or altitude_randomised (see Meynard and Kaplan, 2013; Meynard et al., 2019). Initially I tried using a linear relationship between the altitude predictor layers and probability of occurrence, but this introduced relatively little zero-inflation in the data. For the purposes of investigating sampling bias and zero-inflation I therefore chose to use a logarithmic relationship, whereby probability of occurrence rapidly increases initially with small increases in altitude, but gradually tapers off at higher altitudes. This heavily disfavours low altitude values, and the majority of these will be assigned low probability values close to zero. Hence, biological aggregation of the occurrence points was effectively increased, yielding greater zero-inflation. Each biological predictor was resampled to a 100 m x 100 m resolution across the study area, and were then rescaled using the 'rescale by function' tool in ARCGIS version 10.3.1 (ESRI, 2018), such that the new probability of occurrence layers (ranging between 0 and 1) were logarithmically related to the biological predictors.

Five thousand occurrence points were placed across the study area (using the ArcGIS tool: 'Create Spatially Balanced Points') based on these altitude and altitude_randomised occurrence probability layers. Due to computation limitations of the 'Create Spatially Balanced Points' tool, only one occurrence point can be placed within a single raster cell. Therefore a resolution of 100 m x 100 m was

chosen for the probability layers so that up to 100 species occurrences could be placed in each 1-km^2 grid cell. Although visually the altitude_randomised species appears to be randomly distributed across the study area, it is actually the underlying altitude grid square values that are randomised: occurrences of the altitude_randomised species still occur at higher densities in grid squares with higher altitude values. As I used a logarithmic species response to the altitude_randomised layer, significant (biological) zero-inflation still occurs in the raw data: occurrences are unlikely in low altitude grid cells, generating lots of true zeros when occurrences were counted per grid cell (Table 5.2). Only the random species distribution is completely random across the study area.

Finally, true (raw) species abundance (total number of occurrence points) was calculated for each 1-km^2 grid cell. I felt the chosen grid scale was appropriate because, although the maximum abundance per grid cell is strictly 100, no grid cells reached this value (the maximum was six occurrences per 1-km grid cell), and I therefore assumed that it was unlikely that the shape of the distribution of abundances would be significantly affected by the upper bound (i.e. unbounded distributions such as Poisson or negative binomial were likely to be appropriate). In addition, using this grid scale sets up a situation where location data are available at a higher resolution than the environmental predictors. Hence, I am simulating a situation in which modellers must make a decision about how to aggregate high resolution data across grid cells to create models which predict species distributions based on lower resolution environmental predictors.

Table 5.2 *Sources of zero-inflation in the simulated species occurrence data.*

	Source of zero-inflation		
	True abundance (before sampling)	**Random sampling**	**Biased sampling**
Species			
Random	No zero-inflation	Sampling	Sampling
Altitude	Biological	Biological and sampling	Biological and sampling
Altitude 2	Biological	Biological and sampling	Biological and sampling

5.3.3 Simulating the sampling strategies

I considered two sampling strategies across the study area to represent alternative scenarios of ecological data collection. The first is random sampling, where every 1-km grid cell has an equal chance of being visited and sampled. If visited, I assume all species occurrences in the cell are recorded (i.e. there is no detection error) and the result is the true (raw) abundance (count of all occurrences) for each visited grid cell. The second sampling strategy is affected by spatial sampling bias and relates to the 'bias predictor', where the probability of a grid cell being sampled decreases as distance from the nearest 'town centre' increases. The grid cells selected for this strategy were chosen based on a probability layer created using a logarithmic scaler of the 'distance from nearest town' predictor, again using the 'rescale by function' ArcGIS tool. This time high probability values close to one were assigned to cells with small numerical values i.e. cells closer to towns and more likely to be sampled, whereas low probability values close to zero were assigned to cells with large 'distance from nearest town' values. For each strategy, 2,000 grid cells (20% of the total) were sampled and species abundance was noted for each one. All other (unsampled) squares were assigned an observed abundance of zero, creating a zero-inflated dataset. All sources of zero-inflation in the simulated species abundance data before and after sampling are shown in Table 5.2.

5.3.4 Simulation 1: Investigating the accuracy of species distribution maps from ZI models

To address my first question regarding the accuracy of ZI model predictions of abundance, I focused initially on the performance of ZI Poisson models, and how this compared with equivalent conventional Poisson GLMs. I include comparisons between a) ZI and GLM models, b) count and sampling abundance predictions from ZI models, and c) alternative ZI models fitted using different combinations of biological and bias predictors.

I chose to fit four GLMs and six ZI models for each of the three sets of species abundances per 1-km^2 (random, altitude and altitude_randomised), all fitted with a Poisson distribution but with different combinations of the biological or bias predictors (Table 5.3). These included combinations where different predictors were tested in the count and zero components of the ZI models. Where the biological

predictor was included, models for the "altitude species" were fitted using altitude as a predictor, and models for the altitude_randomised species were fitted using altitude_randomised. Model fitting was repeated 10 times, each time using a different set of simulated 'town centres' (Fig. A5.1.1). Thus, there are three species (random, altitude, altitude_randomised), two sampling strategies (random and biased) and 10 repeats, resulting in 60 total simulation runs. Model performance was evaluated using AIC, averaged across the 10 repetitions. All ZI and GLM models were fitted in R version 3.6.3 (R Core Team, 2019) using packages 'stats' (R Core Team, 2019) and 'pscl' (Zeileis et al., 2008).

Table 5.3 *Ten predictor combinations were considered when modelling the simulated species distributions. Four Generalised Linear Model (GLM) and six Zero-Inflated (ZI) model structures were considered using combinations of the biological predictors (either altitude or altitude_randomised) and the bias predictor (distance from nearest town), including different combinations in the count and zero components of the ZI models.*

Model	Predictors (GLM/ ZI Count component)	Predictors (ZI Zero component)
GLM1	Null (No predictors)	N/A
GLM2	Biased	N/A
GLM3	Biological	N/A
GLM4	Biological + bias	N/A
ZI1	Null (No predictors)	Null
ZI2	Biological + bias	Biological
ZI3	Biological	Biological + bias
ZI4	Biological	Biological
ZI5	Bias	Bias
ZI6	Biological + bias	Biological + bias

Abundance predictions from each model were created using 10-fold cross-validation, where the data were split into 10 subsets and each subset was used iteratively as the test data for which predictions were created and the other nine subsets as training data. For the ZI models both count abundance and sampling abundance predictions were evaluated. Model predictions were evaluated using a novel metric based on the probability of obtaining the model predictions, that I named 'deviation from the best model' (*D*) (See Appendix A5.3 for more information). I used this metric, rather than conventional

measures of performance (e.g. root mean square) typically employed in presence-only or presence-absence modelling, because it produces a measure of fit for count or abundance predictions which is independent of the mean. *D* ranges from a minimum of one for a perfect model where model predictions are equal to the true raw abundance data, and increases without limit as model predictive performance decreases. Spearman's rank correlation coefficients (r_s) were also used to compare model abundance predictions to the original model covariates.

To check that my results were not overly sensitive to the choice of predictor, simulations using average temperature (^oC) (WorldClim, accessed 10/05/18) at a 1-km^2 resolution, as an alternative biological predictor, were also carried following the same methodology (see Appendix A5.2): the results parallel those of altitude, and so were omitted from the main results and discussion.

5.3.5 Simulation 2: Examining the impact of the extent of zero-inflation in the data

To address my second question, about the effect of varying the extent of zero-inflation in the data (both as a result of biological processes and sampling bias) on the effectiveness of the ZI models, I carried out a second simulation. In my first simulation, I assumed 20% of grid cells were sampled, but in Simulation 2 zero-inflation resulting from sampling bias was adjusted by varying the number of cells sampled from the grid, ranging from 1000 (10%) to 10,000 (100%) at 10% increments. Therefore, the highest level of zero-inflation occurred when 1000 cells were sampled, and thus 9000 cells were assigned an abundance of zero simply because they were not sampled, and the lowest level of zero-inflation occurred when 10,000 cells were sampled and none were assigned an abundance of zero for this reason. At the same time zero-inflation resulting from biological processes was adjusted by adding a threshold below which the altitude species can no longer survive, but keeping constant the number of true occurrence points generated each time. With higher altitude thresholds, the species occurrences were increasingly aggregated, and more cells were classified as true zeros. Altitude across the study area ranged from 0 to 284 m, so I tested threshold values of 0 m, 50 m, 100 m, 125 m, 150 m, 175 m and 200m (see Table A5.1.1 for number of cells above each threshold). Above these thresholds, species occurrences were placed in a similar way based on weighted probability calculated from a logarithmic

scaler of the original altitude predictor as described previously. Both the random species and altitude species were examined in scenarios with varying sample sizes, but obviously only the latter was tested using the altitude threshold method.

Based on the results of Simulation 1, I selected three predictor combinations to fit the models and create predictions. These included the GLM with both the bias and biological predictor (GLM4) and two of the ZI models which differ only in the inclusion (ZI6) or exclusion (ZI2) of the bias predictor from the zero component (Table 5.3). Although theoretically a ZI model that has only the biological predictor in the count component, but both the biological and bias predictor in the zero component (as with ZI3), would be the most obvious choice, in the real world the bias predictor may also have some biological influence on the species distribution, and the researcher may not be sure whether it is a better predictor of bias or biology. I therefore chose to use ZI6 rather than ZI3, to simulate better a real world modelling scenario in which the causes of bias are unknown.

Model performance (*D*) was calculated for each simulation run with a particular combination of sample size and altitude threshold. Finally, in order to evaluate the improvement in model performance created by adding predictors of zero inflation, the difference in '*D*' was calculated between each model (GLM4 and ZI2, GLM4 and ZI6, and ZI2 and ZI6). This was repeated using both count abundance and sampling abundance predictions for the ZI models. Again, model fitting was repeated 10 times each with two sampling strategies (random and biased). Therefore, there were 200 simulation runs for the random species (10 repeats, two sampling strategies and 10 levels of sampling zero-inflation), and 1,400 simulation runs for the altitude species (10 repeats, two sampling strategies, 10 levels of sampling zero-inflation and seven altitude thresholds (levels of biological zero-inflation)).

5.3.6 Simulation 3: Comparing abundance versus presence-absence when aggregating spatial data

Often when fitting distribution models the only data available are presence-only, and multiple species occurrences within a grid cell are usually classified as a single presence. Often the predictors are only available at a coarser spatial scale than the species occurrence data, forcing the modeller to aggregate

occurrences into coarser scale presence-only or presence-absence estimates. The coarser the resolution at which the distribution is modelled, the more information is lost about both the precise location of species occurrences, and species abundance (or occurrence density). However, if occurrences are instead aggregated into count data, information about abundance or occurrence density is retained at all scales, which may be more beneficial for conservation purposes. Therefore, even if only presence-only data are available, ZI models fitted at a larger spatial scale using the summed counts of occurrence may provide a better modelling method than traditional presence-only SDM that aggregate multiple occurrences into presence-absence data. This effect is likely to be more pronounced when the species data are biased, because ZI models attempt to model the excess zeroes from sampling bias, whereas other methods, unless they explicitly incorporate bias correction, make no attempt to model or remove the bias.

My final simulation study addressed this question by comparing the performance of Poisson GLM and ZI models predicting abundance of the altitude species (as was carried out in Simulation 1) with two commonly used modelling methods that predict presence-absence: presence-absence binomial GLMs, and presence-only MaxEnt models. This represents a scenario where the raw species occurrences (simulated at a 100m resolution) are available at a greater resolution than the predictors (at a 1-km resolution), so the modeller is required to make a decision on how to aggregate the data.

To fit the binomial GLM presence-absence models, the source data for which need to be in the form of presence-absence rather than abundance, simulated 1-km cells that received an abundance count of zero based on either the random or biased sampling strategy for the ZI models in Simulation 1 (i.e. 80% of cells that were not considered to have been sampled) were classified automatically as an absence, and any cell with species occurrences that was sampled was classified as a presence. All binomial GLMs were fitted using the package 'stats' in R. As with Simulation 1, two GLMs were fitted, one with only the biological predictor ('Binomia-GLM1' equivalent to GLM3) and one with the biological and bias predictor ('Binomial-GLM2' equivalent to GLM4). Binomial occurrence predictions (i.e. predicted

probability of presence) were estimated across the study area from each model using 10-fold cross-validation.

Two MaxEnt presence-only models were also fitted to the altitude species occurrence data, one with altitude as the only predictor ('Maxent1'), and one with both altitude and distance from nearest town as predictors ('Maxent2'). To produce presence-only data collected under a random or biased sampling strategy, only occurrence points at a 100m resolution that fell within a 1-km cell that had been sampled for the ZI models in Simulation 1 were retained; only these cells would be classified by MaxEnt as a presence. Each model was fitted using the 'dismo' package (Hijmans et al., 2017) in R, at a 1-km resolution with 10,000 randomly selected background 'pseudo-absences' and 10 repetitions across each set of town centres. All other MaxEnt parameter settings were set to the default options, including 1,000 iterations, regularization multiplier = 1, and specifying a logistic output (Naimi & Araújo, 2016).

Comparing the performance of count/abundance models (Poisson GLM and ZI models) and presence/presence-absence models (MaxEnt and binomial GLMs) required evaluation metrics which could work with both types of model. As it is less feasible to convert presence-absence predictions to abundance to use 'D', two other evaluation metrics were selected: Area Under the Curve (AUC) and the Spearman's Rank correlation coefficient (r_s) between the model predictors ('altitude' and/ or 'distance from town') and each of the model predictions of count/abundance (GLM/ ZI) or habitat suitability (MaxEnt/ binomial GLM). In order to calculate AUC for the ZI and GLM models, abundance predictions were converted to binary presence-absence predictions, using an abundance threshold above which the species was considered to be predicted to be present. As I outline in Chapter 3, variable thresholds like 'mean probability' are shown to be more robust and a better classification method than fixed thresholds (e.g. if I categorised abundance using a threshold of below or above one). In addition, some models produced predicted abundances that all fell below one. Therefore, the threshold I chose for conversion was the mean abundance prediction across all grid cells for each individual model i.e. the threshold varied across each GLM or ZI model. Mean AUC was calculated across the 10 repetitions for each model based on the presence-absence predictions for all models compared to the true presence-

absence based on all occurrence locations across the study area. It should be noted that neither of these metrics offer a perfect measure of model performance. AUC causes a loss of information from the Poisson GLMs and ZI models, which are designed to predict abundance, while Spearman's rank retains more of the information in the predictions of both types of model, but is necessarily relatively crude.

Finally, in order to assess the impact of the scale of data aggregation on the performance of abundance and presence-absence models, additional models were fitted and compared across two other scales of increasing coarseness: 2-km and 5-km. The larger the grid cell, the larger the mean count of occurrences per cell, and hence the more data potentially lost by converting to presence-absence. ZI count abundance predictions at a 2-km and 5-km scale were obtained following the methodology of Simulation 1 using the ZI6 model structure and again converted to presence-absence predictions. MaxEnt and binomial GLM presence-absence predictions at a 2-km and 5-km scale were obtained following the methodology outlined previously in Simulation 3. Model predictors (altitude and distance from town) were converted to coarser scales by calculating the mean values of each predictor at a 1-km resolution for each 2-km or 5-km cell. As before, all predictions were evaluated using AUC and Spearman's Rank correlation coefficient (r_s).

5.4 Results

5.4.1 Simulation 1: Investigating the accuracy of species distribution maps from ZI models

The results from Simulation 1 confirm that count abundance predictions from the ZI models provide the most accurate estimates (according to the metric *D*) of true species abundance (Fig. 5.3, Fig. A5.1.3). Estimating true abundance based purely on the biology of the species rather than sampling processes is usually the aim of ecological research, and these results suggest the count abundance predictions are most likely able to fulfil these aims. In contrast, all GLMs are poor at predicting true abundance because they do not separately model the excess (false) zeros generated by grid cells that have not been sampled. The problem is exaggerated when sampling is not just incomplete, but is also biased; if the GLM includes a predictor which is correlated with sampling effort (distance from nearest town), the model performs even less well (compare pink and blue bars for GLM3 (without bias predictor) and GLM4

(with bias predictor) in Fig. 5.3) because it detects a spurious negative association between this predictor and abundance (top panels, Fig. A5.1.4). Similarly, ZI sampling abundance predictions (predictions from the whole model that potentially include the influence of sampling bias) perform poorly; rather than estimating true abundance, reflecting the species niche, they predict abundance as it would appear to observers employing each sampling strategy (Fig. 5.3, Fig A5.1.3). Again, these predictions are particularly poor when sampling is biased (compare pink and blue bars for ZI2 and ZI6 in Fig. 5.3). These findings hold true for all three species (altitude, altitude_randomised and random) (Fig A5.1.3).

The ability to model excess zeros separately led to dramatically improved predictive power of true abundance for all ZI models (see count abundance predictions in Fig. 5.3 and Fig A5.1.3), although one (ZI2) performed relatively less well than the others when sampling was biased (Fig. 5.3, Fig. A5.1.3). In ZI2, the bias predictor was included in the count component but not the zero component, meaning that like the GLMs it detected a spurious negative association between abundance and distance from the nearest town (middle panels, Fig. A5.1.4); if they included the bias predictor, the other ZI models (e.g. ZI3 or ZI6) correctly detected that it was positively associated with the probability of an excess zero being recorded (lower panels, Fig. A5.1.4).

Predicted distribution maps based on both the count abundance predictions and sampling abundance predictions also support these findings (Fig. 5.4 & 5.5). Maps produced using ZI count abundance predictions that account for bias where necessary (i.e. including predictors of bias in the zero component when sampling is biased), correlate strongly with the biological predictor layer (altitude) ($r_s > 0.9$) and show little influence of bias (distance from towns) (Fig. A5.1.5). When sampling is biased, both neglecting to account for the bias in the zero component, or using the sampling abundance predictions, results in low accuracy distribution maps that correlate more strongly with the bias predictor (r_s value between -0.64 to -0.71) and less strongly with the biological predictor (r_s values between 0.60 to 0.74) (Fig. A5.1.5). Distribution maps produced by the GLMs were also less accurate when sampling was biased and predictors correlating with bias were included (Fig. 5.4 & 5.5). Maps from the GLMs which include the bias predictor (GLM4) show a strong influence of sampling bias similar to that seen in the

ZI sampling abundance predictions. These maps show relatively weak correlations to the altitude predictor (r_s = 0.60) compared to their counterpart GLMs that do not include the bias predictor (GLM3) (r_s = 0.99) (Fig. A5.1.5). The prediction map from the GLM including both the biological and bias predictors (GLM4) with biased sampling also shows a strong correlation to the bias predictor (r_s = -0.72).

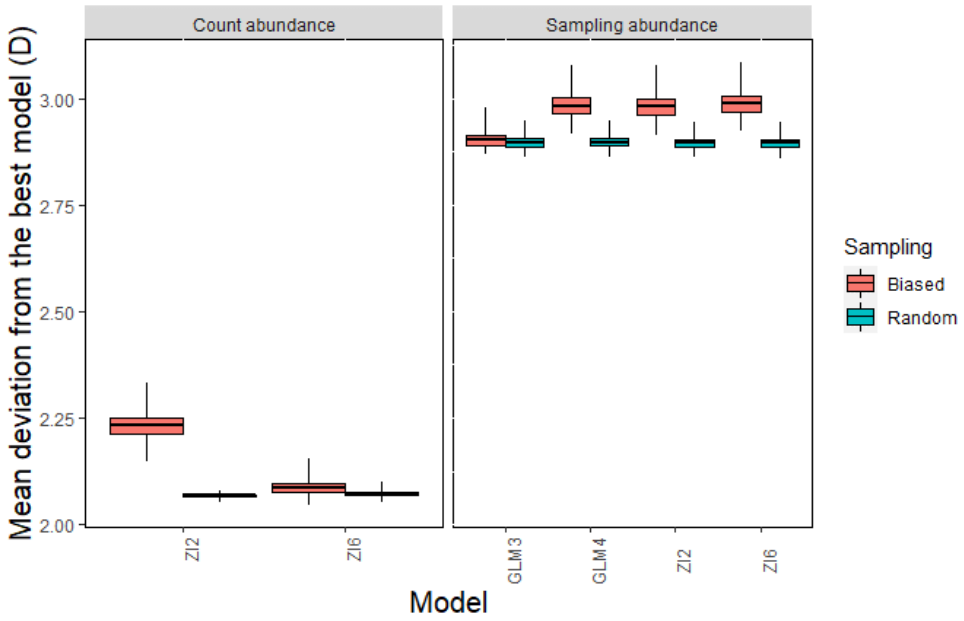


Fig. 5.3 *Evaluation of abundance predictions (based on D = 'deviation from the best model') for a hypothetical organism with occurrences simulated based on a preference for high altitudes (altitude species). Mean D values (± SE and data range) are shown for each sampling strategy (random or biased) across the 10 model repetitions for four models: two non-zero-inflated generalised linear models (GLM3 including only the biological predictor and GLM4 including the biological and bias predictor) and two zero-inflated (ZI) models (ZI2, which does not account for bias in the zero component and ZI6, which does). Both sampling abundance (abundance from the whole model including the potential impact of sampling bias) and count abundance (abundance from the ZI count component only) are evaluated separately for the ZI models. Only sampling abundance can be obtained from the GLMs.*

Additional maps that depict the probability of each grid cell being an excess zero (i.e. predictions from the zero component of a ZI model) further highlight the ability of ZI models to model separately the biological and sampling processes, as well as providing insight into the nature of bias in the species

data (Fig. 5.4 & 5.5). This means that in real studies in which the sources of sampling bias are unknown, inclusion of predictors that may correlate with sampling bias (e.g. distance to towns or roads, accessibility, land-use etc.) in both the count and zero components of ZI models can help to both model and identify likely causes of bias. This is a unique feature of the ZI models, and is something which the GLMs are unable to reproduce; these models cannot provide insight into the bias or prediction maps that eliminate sampling effects within the data.

Conventional measures of model performance (AIC) were consistent with the results from model testing in that the ZI models performed better than the GLMs, and ZI models that included the bias predictor in the zero component (ZI3 and ZI6) performed better than those that did not (ZI2) (Fig. A5.1.2). However, in contrast to D, the majority of models fitted using biased data that included the bias predictor produced lower AIC values than their corresponding random sampling models. This is possibly because AIC measures the model fit to the data and there is more likely to be overfitting due to the inclusion of the bias predictor. With D, the models are being tested against a separate 'unseen' test data set (each of the 10 CV folds not used to train the model) and therefore is a better measure of predictive power of the model; D is less likely to be influenced by model overfitting.

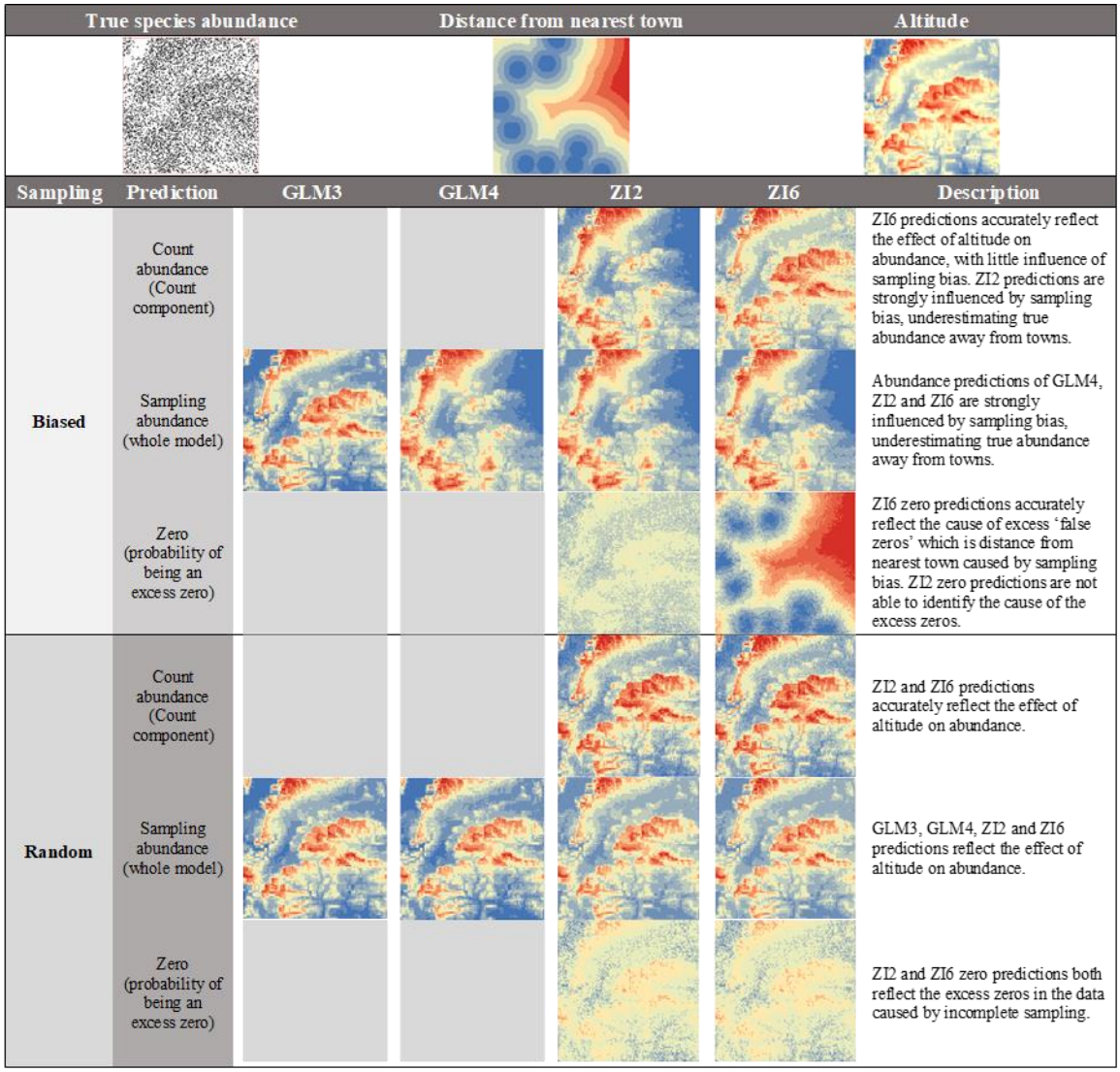


Fig. 5.4 *Example maps of abundance for a hypothetical species ('altitude species') whose occurrence is positively influenced by altitude, produced from two generalised linear models (GLMs) and two Zero-Inflated (ZI) models. Models were built with either data collected by randomly sampling grid cells (random) or with sampling bias (biased). Abundance maps from GLM3 (including the biological predictor only) and GLM4 (including both the biological and bias predictor) are produced using sampling abundance predictions (i.e. from the whole model). Both count abundance and sampling abundance predictions can be produced from the ZI models along with a map of the probability a cell is an excess zeros (zero). Both ZI models include a biological predictor (altitude) of both abundance and excess zeros, and bias predictor (distance from the nearest town) of abundance. ZI6 also includes 'distance from the nearest town' as a predictor of excess zeros. Individual cells are colour-coded based on abundance for the abundance predictions or on probability of being an excess zero for the zero predictions (high = red, low = blue).*

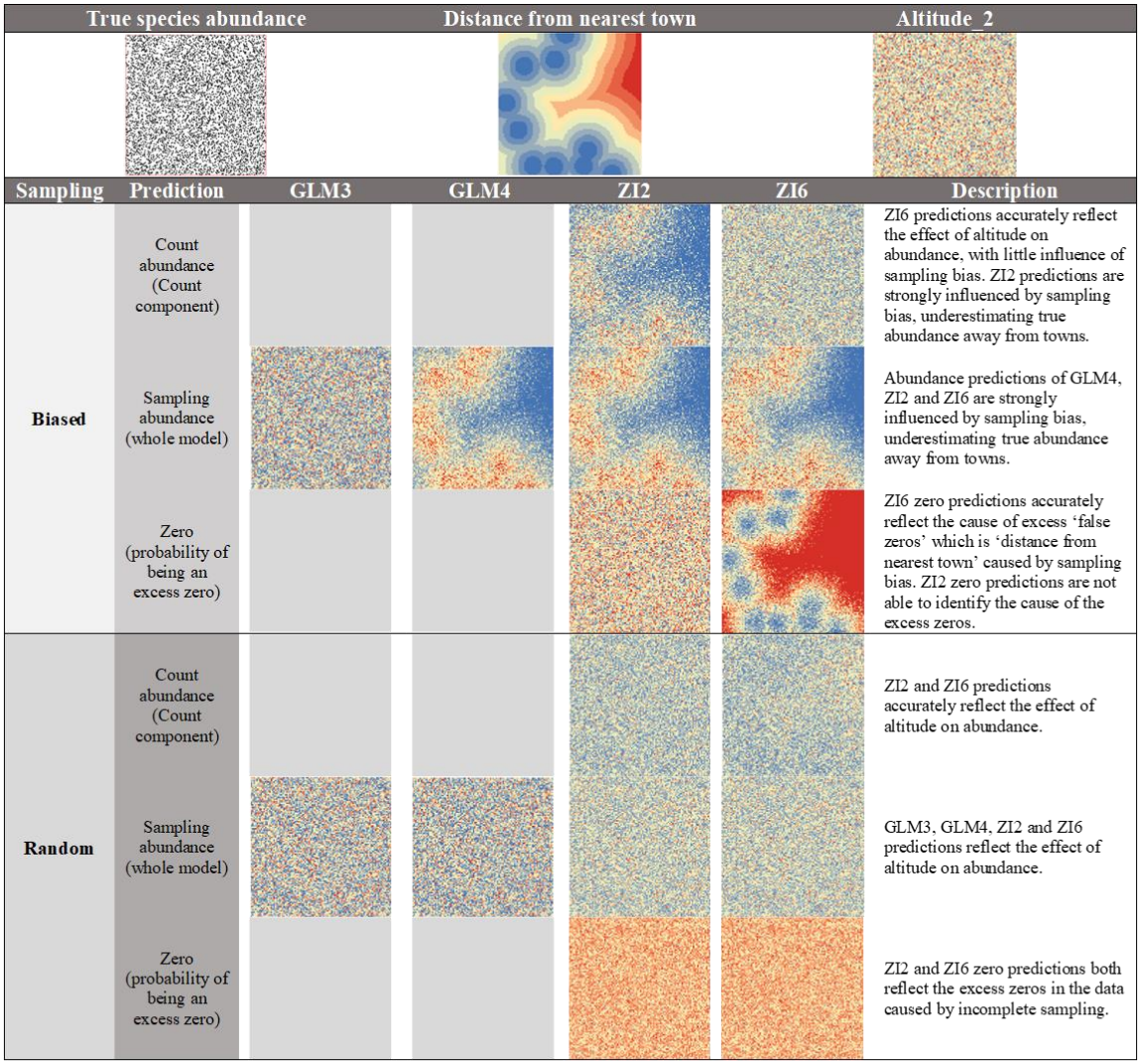


Fig. 5.5 *Example maps of abundance for a hypothetical species ('altitude_randomised species') whose occurrence is positively influenced by a randomised altitude layer, produced from two generalised linear models (GLMs) and two Zero-Inflated (ZI) models. Models were built with either data collected by randomly sampling grid cells (random) or with sampling bias (biased). Abundance maps from GLM3 (including the biological predictor only) and GLM4 (including both the biological and bias predictor) are produced using sampling abundance predictions (i.e. from the whole model). Both count abundance and sampling abundance predictions can be produced from the ZI models along with a map of the probability a cell is an excess zeros (zero). Both ZI models include a biological predictor (altitude) of both abundance and excess zeros, and bias predictor (distance from the nearest town) of abundance. ZI6 also includes 'distance from the nearest town' as a predictor of excess zeros. Individual cells are colour-coded based on abundance for the abundance predictions or on probability of being an excess zero for the zero predictions (high = red, low = blue).*

5.4.2 Simulation 2: Examining the impact of the extent of zero-inflation in the data

Real species occurrence or abundance data will suffer from variable levels of zero inflation resulting from both biological and sampling processes. Therefore, the better performance of ZI models compared with GLMs described in Simulation 1 may not occur in all circumstances. In Simulation 2, I explored which is generally the better choice of model under different levels of biological and sampling bias zero-inflation. As anticipated, ZI count abundance predictions and GLM abundance predictions have similar accuracy when the data are not zero-inflated: when the whole study area is surveyed, all absences are 'true absences', the species is randomly distributed with no biological zero-inflation, and the difference in performance is zero (Fig. 5.6a, see random species (R) in left and middle panels). When considering the random species only (i.e. with no biological zero-inflation), as less of the study area is surveyed, zero-inflation as a result of sampling increases, and therefore the effectiveness of ZI model count abundance predictions improves in comparison to GLMs. Although this phenomenon occurs under both sampling strategies, it is most noticeable when both sampling is biased and that bias is accounted for in the model (by including the bias predictors in the ZI zero component as in ZI6 for example).

As with the random species, when there are high levels of incomplete sampling for the altitude species (e.g. ~20% or fewer cells are sampled), ZI model count abundance predictions are consistently better than GLM predictions, regardless of biological zero-inflation (Fig. 5.6a, left and middle panels). However as more of the area is surveyed (> 20%), the difference in performance decreases. At low levels of biological zero-inflation, this difference tends towards zero. However, at higher levels of biological zero-inflation, GLM predictions are actually more accurate than the ZI model count abundance predictions under both random and biased sampling scenarios. This can best be understood by looking at Fig. 8b showing the results based on sampling abundance predictions from the ZI model, rather than count abundance predictions: in contrast to the count abundance predictions, as biological zero-inflation increases, ZI sampling abundance predictions increasingly outperform those of the GLM. This is because the zero component of the ZI model, which is combined with the count component to create the sampling abundance prediction, is able to predict the excess zeroes caused by the biological

driver, while the GLM cannot. Therefore, if high levels of biological zero-inflation are suspected in the data, both the count and sampling abundance predictions should be considered and evaluated before choosing the best predictions of species abundance.

Reiterating the results from Simulation 1, when sampling is random there is no benefit of including the bias predictor in the zero component under any levels of sampling or biological zero-inflation (Fig. 5.6a & b, top right panels). Under biased sampling scenarios, models accounting for bias (by including the bias predictor in the zero component as in ZI6 for example) are most effective when there are high levels of sampling-related zero-inflation and low levels of biological zero-inflation. As either the area surveyed or biological zero-inflation increases, the effectiveness of these models reduces compared to models that fail to account for bias (Fig. 5.6a, bottom right panel). Nevertheless, the majority of differences seen between ZI models are relatively small compared to those between the ZI models and GLMs.

a) Count abundance

Random

Number of sampled grid squares	Difference between GLM4 and ZI6								Difference between GLM4 and ZI2								Difference between ZI2 and ZI6							
	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200
1000	-1.63	-1.47	-1.63	-1.94	-2.23	-2.52	-3.56	-7.38	-1.64	-1.06	-1.19	-1.42	-1.55	-1.63	-1.99	-3.63	0.00	0.00	0.00	0.01	0.01	0.02	0.02	0.02
2000	-0.96	-0.83	-0.91	-1.02	-1.04	-1.11	-1.12	-1.34	-0.96	-0.83	-0.91	-1.02	-1.04	-1.12	-1.12	-1.34	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00
3000	-0.59	-0.47	-0.52	-0.63	-0.50	-0.33	0.03	0.83	-0.59	-0.47	-0.53	-0.64	-0.51	-0.34	0.02	0.83	0.00	0.00	0.01	0.00	0.01	0.01	0.01	0.00
4000	-0.37	-0.24	-0.17	-0.25	-0.07	0.15	0.78	2.40	-0.37	-0.25	-0.18	-0.26	-0.08	0.15	0.78	2.39	0.00	0.00	0.00	0.01	0.01	0.01	0.01	0.00
5000	-0.22	-0.10	-0.06	0.10	0.16	0.50	1.27	3.60	-0.22	-0.09	-0.06	0.10	0.16	0.49	1.26	3.59	0.00	-0.01	0.00	0.00	0.01	0.01	0.01	0.00
6000	-0.13	-0.02	-0.03	0.10	0.44	0.78	1.63	4.43	-0.13	-0.02	-0.03	0.10	0.44	0.78	1.63	4.43	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
7000	-0.07	-0.01	-0.01	0.15	0.43	0.96	1.91	4.71	-0.07	-0.01	-0.01	0.15	0.43	0.96	1.90	4.71	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8000	-0.03	0.00	0.00	0.19	0.48	1.08	2.10	5.02	-0.03	0.00	0.00	0.19	0.48	1.08	2.10	5.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9000	-0.01	0.00	0.02	0.24	0.59	1.14	2.15	4.97	-0.01	0.00	0.02	0.24	0.59	1.14	2.15	4.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10000	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Biased

Number of sampled grid squares	Difference between GLM4 and ZI6								Difference between GLM4 and ZI2								Difference between ZI2 and ZI6							
	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200
1000	-1.71	-1.58	-1.70	-2.02	-2.14	-2.39	-3.00	-4.80	-1.64	-1.44	-1.58	-2.00	-2.13	-2.48	-2.86	-4.69	-0.07	-0.13	-0.11	-0.02	0.06	0.09	0.08	0.05
2000	-1.04	-0.90	-1.00	-1.15	-1.16	-1.17	-1.44	-2.55	-0.95	-0.75	-0.86	-1.14	-1.23	-1.23	-1.49	-2.58	-0.09	-0.14	-0.13	0.00	0.00	0.07	0.05	0.02
3000	-0.68	-0.53	-0.57	-0.65	-0.55	-0.38	-0.07	0.70	-0.60	-0.40	-0.43	-0.64	-0.61	-0.45	-0.13	0.66	-0.09	-0.13	-0.14	-0.02	0.06	0.07	0.06	0.04
4000	-0.44	-0.32	-0.27	-0.30	-0.16	0.05	0.64	1.96	-0.37	-0.19	-0.14	-0.14	-0.22	-0.01	0.59	1.92	-0.07	-0.13	-0.13	-0.16	0.06	0.06	0.05	0.04
5000	-0.29	-0.16	-0.11	0.18	0.48	0.42	1.16	2.68	-0.23	-0.06	-0.06	0.04	0.05	0.35	1.11	2.65	-0.06	-0.10	-0.05	-0.06	0.04	0.06	0.05	0.03
6000	-0.18	-0.07	-0.05	0.10	0.30	0.61	1.45	3.18	-0.13	-0.02	-0.03	0.10	0.33	0.55	1.40	3.15	-0.05	-0.05	-0.01	0.00	-0.03	0.06	0.05	0.03
7000	-0.10	-0.03	-0.02	0.13	0.42	0.95	1.74	4.43	-0.07	-0.01	-0.02	0.13	0.41	0.90	1.71	4.42	-0.03	-0.03	0.00	0.00	0.01	0.04	0.03	0.01
8000	-0.05	-0.02	0.00	0.19	0.47	0.98	2.04	4.76	-0.03	0.00	0.00	0.18	0.47	0.94	2.01	4.75	-0.02	-0.01	0.00	0.00	0.00	0.04	0.03	0.01
9000	-0.02	-0.01	0.01	0.23	0.57	1.11	2.11	4.81	-0.01	0.00	0.01	0.23	0.57	1.10	2.06	4.80	-0.01	0.00	0.00	0.00	0.00	0.00	0.05	0.00
10000	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.03	0.28	0.65	1.22	2.21	5.09	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

-10	-7.5	-5.0	-2.5	0.0	2.5	5	7.5	10

-0.4	-0.3	-0.2	-0.1	0.0	0.1	0.2	0.3

b) Sampling abundance

Random

Number of sampled grid squares	Difference between GLM4 and ZI6								Difference between GLM4 and ZI2								Difference between ZI2 and ZI6							
	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200
1000	0.00	0.00	-0.02	-0.11	-0.21	-0.36	-0.73	-1.73	0.00	0.00	-0.02	-0.11	-0.22	-0.36	-0.75	-1.88	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.15
2000	0.00	0.00	-0.01	-0.09	-0.17	-0.31	-0.61	-1.49	0.00	0.00	-0.01	-0.09	-0.17	-0.31	-0.61	-1.53	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04
3000	0.00	0.00	-0.02	-0.08	-0.17	-0.32	-0.60	-1.49	0.00	0.00	-0.01	-0.08	-0.17	-0.32	-0.60	-1.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
4000	0.00	0.00	-0.05	-0.09	-0.19	-0.34	-0.63	-1.51	0.00	0.00	-0.05	-0.09	-0.19	-0.34	-0.63	-1.52	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
5000	0.00	0.00	-0.06	-0.25	-0.21	-0.38	-0.68	-1.59	0.00	0.00	-0.06	-0.25	-0.21	-0.38	-0.68	-1.61	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
6000	0.00	-0.01	-0.06	-0.26	-0.25	-0.42	-0.74	-1.74	0.00	-0.01	-0.06	-0.26	-0.25	-0.42	-0.74	-1.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
7000	0.00	-0.01	-0.06	-0.26	-0.42	-0.48	-0.81	-1.87	0.00	-0.01	-0.06	-0.26	-0.42	-0.48	-0.81	-1.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8000	0.00	-0.01	-0.06	-0.26	-0.42	-0.61	-0.88	-2.00	0.00	-0.01	-0.06	-0.26	-0.42	-0.61	-0.88	-2.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9000	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-0.94	-2.08	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-0.94	-2.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10000	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Biased

Number of sampled grid squares	Difference between GLM4 and ZI6								Difference between GLM4 and ZI2								Difference between ZI2 and ZI6							
	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200	R	0	50	100	125	150	175	200
1000	0.00	0.00	-0.01	-0.08	-0.15	-0.26	-0.49	-1.35	0.00	0.00	-0.01	-0.09	-0.27	-0.36	-0.68	-1.77	0.00	0.00	0.00	0.02	0.06	0.10	0.16	0.36
2000	0.00	0.00	-0.01	-0.09	-0.18	-0.32	-0.62	-1.56	0.00	0.00	-0.01	-0.11	-0.23	-0.42	-0.78	-1.92	0.01	0.01	0.00	0.00	0.00	0.10	0.16	0.36
3000	0.00	0.01	-0.01	-0.09	-0.18	-0.33	-0.61	-1.46	0.00	0.00	-0.03	-0.10	-0.22	-0.41	-0.74	-1.82	0.01	0.01	0.01	0.02	0.04	0.08	0.14	0.37
4000	0.01	0.01	-0.03	-0.10	-0.20	-0.36	-0.65	-1.55	0.00	0.00	-0.05	-0.16	-0.24	-0.42	-0.77	-1.90	0.01	0.01	0.02	0.06	0.04	0.06	0.12	0.35
5000	0.01	0.01	-0.05	-0.26	-0.39	-0.39	-0.70	-1.66	0.00	-0.01	-0.06	-0.26	-0.25	-0.45	-0.80	-1.88	0.01	0.02	0.02	0.04	0.03	0.06	0.09	0.22
6000	0.01	0.01	-0.05	-0.26	-0.30	-0.43	-0.76	-1.74	0.00	-0.01	-0.06	-0.26	-0.38	-0.48	-0.86	-2.00	0.01	0.02	0.01	0.00	0.08	0.06	0.09	0.26
7000	0.01	0.01	-0.06	-0.26	-0.42	-0.48	-0.82	-1.88	0.00	-0.01	-0.06	-0.26	-0.42	-0.52	-0.86	-1.99	0.01	0.01	0.00	0.00	0.00	0.04	0.05	0.11
8000	0.01	0.00	-0.06	-0.26	-0.42	-0.59	-0.88	-2.00	0.00	-0.01	-0.06	-0.26	-0.42	-0.62	-0.92	-2.08	0.01	0.01	0.00	0.00	0.00	0.03	0.04	0.08
9000	0.01	0.00	-0.06	-0.26	-0.42	-0.63	-0.98	-2.10	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.00	-2.15	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.04
10000	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	-0.01	-0.06	-0.26	-0.42	-0.63	-1.02	-2.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

-2.5	-2	-1	-0.5	0	0.5	1	2	2.5

-0.8	-0.6	-0.4	-0.2	0	0.2	0.4	0.6

Fig. 5.6 *Comparisons of model predictive power of a) count abundance (top panels) or b) sampling abundance (bottom panels) between a generalised linear model (GLM) and two zero-inflated (ZI) models across varying levels of biological and sampling bias zero-inflation. Values represent the mean difference in D ('deviation from the best model') between GLM4 (containing both biological and bias predictor), ZI2 (excludes the bias predictor from the zero component) and ZI6 (includes the bias predictor in the zero component). Biological zero-inflation was increased by introducing a minimum altitude threshold below which the species cannot survive and therefore reducing its environmental niche. Sampling-related zero-inflation was increased by decreasing the number of grid cells sampled across the study area in increments of 10% from 100% to 10%. Negative (red) values show scenarios where the ZI model performs better than the GLM (left and middle panels) or where ZI6 performs better than ZI2 (right panel), whereas positive (blue) values show scenarios where GLM4 outperforms the ZI models or ZI2 outperforms ZI6. 'R' represents the values for the random species whose occurrence is not related to altitude.*

5.4.3 Simulation 3: Comparing abundance versus presence-absence data across multiple spatial scales

The results from Simulation 3 support my hypothesis that, when dealing with biased species data, modelling aggregated count data using ZI models is a better choice than modelling aggregated presence-absence or presence-only data, as is commonly done in traditional SDM studies, using approaches such as binomial GLMs or MaxEnt (Fig 5.7). The only model to perform consistently well across all spatial scales when dealing with the biased species data was the ZI model, which maintained strong correlations to the biological predictor ($r_s > 0.9$) and low correlations to the bias predictor ($-0.12 < r_s < 0.07$) across all scales (Fig. 5.7). Predicted maps of the altitude species distribution also show that the ZI model count abundance predictions provide the most accurate reflection of the true species distribution as the scale of data aggregation increases (Fig A5.1.6). Binomial-GLM2 and MaxEnt2 models, which incorporate the bias predictor, produced predictions that are heavily influenced by sampling bias at a 1-km scale, with strong correlations to the bias predictor ($r_s < -0.75$) (Fig. 5.7, Fig. A5.1.6). These increase in strength as scale increases to 2-km and 5-km, so that both model predictions produce correlations to the bias predictor close to one ($r_s < -0.92$). Both MaxEnt1 and binomial-GLM1 (which do not include the bias predictor) were able to produce accurate predictions with the biased data at a 1-km resolution, although performance declined as the scale became coarser. Even when the species data was collected using a random sampling strategy, the performance of the presence-absence models declined as the scale became coarser and more information was lost with data aggregation (Fig 5.7); this phenomenon was not seen in the ZI models and performance remained high as scale increased.

			Random Sampling						Biased Sampling					
	Model	Predictor	1 km		2 km		5 km		1 km		2 km		5 km	
			r	se	r	se	r	se	r	se	r	se	r	se
Binomial occurrence predictions	Binomial GLM1	Altitude	0.998	0.000	0.980	0.008	0.377	0.232	0.998	0.000	0.640	0.231	0.333	0.281
		Town	-0.037	0.071	-0.037	0.071	-0.066	0.070	-0.037	0.071	-0.142	0.046	-0.166	0.047
	Binomial GLM2	Altitude	0.999	0.000	0.872	0.037	0.364	0.202	0.607	0.052	0.313	0.078	0.195	0.106
		Town	-0.060	0.080	0.003	0.124	0.121	0.179	-0.778	0.020	-0.940	0.017	-0.957	0.018
MaxEnt occurrence predictions	MaxEnt1	Altitude	0.996	0.003	0.725	0.141	0.097	0.292	0.988	0.009	0.466	0.224	0.278	0.277
		Town	-0.039	0.070	-0.025	0.070	-0.069	0.061	-0.043	0.068	-0.155	0.034	-0.193	0.037
	MaxEnt2	Altitude	0.964	0.007	0.704	0.087	0.192	0.188	0.558	0.066	0.297	0.080	0.181	0.105
		Town	-0.076	0.090	-0.029	0.104	-0.332	0.156	-0.781	0.014	-0.896	0.025	-0.930	0.028
ZI count abundance predictions	ZI6	Altitude	0.944	0.013	0.991	0.002	0.994	0.002	0.960	0.012	0.992	0.002	0.985	0.004
		Town	0.069	0.091	-0.074	0.076	-0.071	0.078	0.069	0.086	-0.034	0.081	-0.112	0.090

Fig. 5.7 *Spearman's Rank correlation coefficients (r_s) between the model predictors (altitude and distance from nearest town) and model predictions for altitude species across three modelling resolutions: 1-km, 2-km and 5-km. Three types of model are compared: 1) binomial generalised linear models (GLMs) that predict the probability of occurrence, 2) Maximum Entropy (MaxEnt) models that predict the probability of occurrence and 3) zero-inflated (ZI) models that predict the true (count) abundance of the species. Binomial-GLM1 and MaxEnt1 include only the biological predictor in the model, whereas Binomial-GLM2 and MaxEnt2 include both the biological and bias predictor. ZI6 model includes the bias and biological predictor in both the count and zero component. Values represent the mean coefficients (r_s) and standard error (se) across the 10 simulated sets of 'town centres' using data collected under two sampling strategies (random and biased). Coefficients are colour-coded based on strength: the darker the colour, the stronger the correlation. Red values represent positive correlations, whereas blue represent negative correlations.*

Model evaluation using mean AUC based on the presence-absence predictions also supports these findings (Fig. 5.8). At a 1-km scale, all models fitted using biased data performed worse than the corresponding model using random data, with the exceptions of GLM3 and Binomial-GLM1 (both of which do not include the bias predictor) and ZI6 (which does include the bias predictor). Additionally, all presence-absence models (binomial GLMs and MaxEnt) performed worse than ZI6 regardless of the sampling strategy. As scale increases, the presence-absence models have a much larger variance in performance across the 10 repetitions of the town centre sets than the abundance models, with some repetitions producing AUC values below 0.5 and above 0.9 (Fig. 5.8). Nevertheless, when dealing with biased species data, the ZI6 model performed better on average than both MaxEnt models (with or

without the bias predictor) and Binomial-GLM2 (with the bias predictor) at a 2-km scale, and better than MaxEnt2 (with the bias predictor) and both binomial GLMs at a 5-km scale.

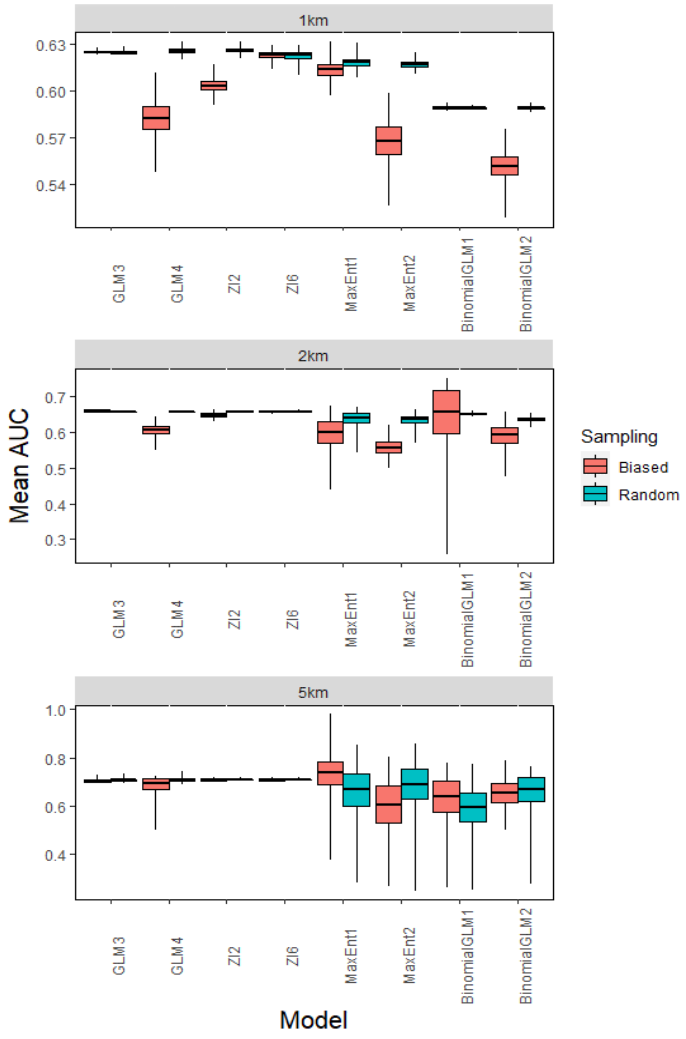


Fig. 5.8 *Evaluation of MaxEnt, generalised linear model (GLM) and zero-inflated (ZI) model predictions of altitude species presence-absence across the study area based on mean Area under the Curve (AUC) across three scales of data aggregation: 1-km, 2-km and 5-km. Mean AUC values (± SE and data range) for each sampling strategy (random or biased) across the 10 model repetitions are shown for two MaxEnt models and two binomial generalised linear models (GLMs): MaxEnt1 and Binomial-GLM1 which includes only altitude as a predictor, and MaxEnt2 and Binomial-GLM2 which includes altitude and distance from town as predictors. Abundance predictions were converted into binary presence-absence predictions for two non-zero-inflated generalised linear models (GLM3 including only the biological predictor and GLM4 including the biological and bias predictor) and two zero-inflated (ZI) models (ZI2 which does not account for bias in the zero component and ZI6 which does account for bias in the zero component) based on a threshold equal to the mean predicted abundance per model type (see Methods for more information).*

5.5 Discussion

Sampling bias in species data is problematic for SDM, and many researchers call for greater awareness and development of correction methods to deal with this issue (Araújo & Guisan, 2006; Bystriakova et al., 2012; Kramer-Schadt et al., 2013). My simulations using ZI models highlight a novel approach for dealing with sampling bias and zero-inflation in SDM, which I believe can be applied to a wide variety of ecological and conservation research questions that use large databases of species records. My results reveal that ZI models have the potential both to reduce the impact of bias on predictions which are used for biological inference, and to provide insights into previously unknown causes or correlates of sampling bias. This method can be used with both raw abundance data, and with abundance data created by summing occurrences from presence-only data at a coarser resolution, and therefore offers an alternative to traditional presence-only SDM methods. As spatial occurrence data is often present at a finer scale than the environmental predictors, decisions about data aggregation have to be made when fitting distribution models. I found that even though information about the precise location of species occurrences is sacrificed, aggregating species occurrences into counts of abundance and fitting ZI models produces better estimates of a species distribution, especially when the species data is biased by sampling methods, than aggregating occurrences into presence-absence form at a coarser spatial scale, as is common with traditional SDM methods such as binomial GLMs or MaxEnt.

Species distribution maps are an important resource for conservation planners (Rodríguez et al., 2007), yet there is often little consideration of inaccuracies or uncertainty in these maps or associated models (Elith et al., 2002; Zuquim et al., 2014). My results show how the biological information value of maps based on GLM, MaxEnt and ZI sampling abundance predictions can be reduced by sampling bias. In contrast, the distribution maps produced from the predictions from the count component of ZI models are accurate reflections of the species niche and true abundance, even when species data are spatially biased, providing that the bias influence is accounted for in the model by included all predictors suspected of capturing or correlating with the bias in both ZI count and zero components. If in doubt about whether a predictor is likely to be a source of bias, inclusion in both parts will not only alleviate the problem of bias, but will also provide insight into whether it actually is a introducing a large number

of excess ('false') zeros. Additionally, ZI model coefficients allowed examination of potential causes of bias; in ZI6 (the model including both the bias and biological predictor in the zero component) from Simulation 1, 'distance from nearest town' was influential only in the zero component, and was not spuriously identified as influencing true abundance. Currently, there are few statistical models that allow post-modelling identification of bias sources. Many SDM techniques rely on prior understanding and some form of quantification of the bias in order to remove it (Phillips, 2008), so ZI models provide an advantage over these traditional bias correction methods in their ability to shed light on potential causes of bias.

If all excess zeros are false zeros, count abundance predictions from ZI models should always reflect the true species niche, and the zero component will be modelling only excess zeros from non-biological, sampling processes. However, this scenario is unlikely in ecological systems. In reality, as in my simulations with the altitude and altitude_randomised species, the excess zeros will result from a combination of biological zero-inflation and sampling zero-inflation. Therefore, the count abundance prediction may not always be predicting true abundance, and the zero component may actually be dominated by biological processes, as I suggest is the case for the results from Simulation 2. In this case, the sampling abundance prediction will actually be a more accurate reflection of true species abundance. Nevertheless, by examining the significance and influence of predictors in both components, their plausibility as causes of bias can be inspected: biological predictors of abundance are likely to be significant in both parts of the ZI model, whereas sampling predictors are unlikely to appear influential in the count component.

After identifying potential bias predictors, modellers can make more informed choices about whether to eliminate these predictors from either ZI component, whether the zero component is more heavily dominated by biological or sampling processes, and if the count abundance or sampling abundance is more likely to reflect true species abundance. A good understanding of the biology of the species being modelled is therefore key. Additionally, despite the post-model-fitting ability of ZI models to distinguish bias, beginning any analysis of a zero-inflated dataset, it is important also to try and identify

the source of excess zeros as either from biology or sampling processes (Martin et al., 2005). Consequently, although one benefit of ZI models is the ability to use different sets of covariates in the count and zero components (Lambert, 1992; Zuur et al., 2009), it is important only to include appropriate, relevant predictors in each part where possible.

The collection of species data varies widely in its scale and standardisation, from single museum specimens collected by natural history experts, to more local, standardised recording schemes (Pocock & Evans, 2014) and to international, opportunistic recording schemes such as eBird (Sullivan et al., 2009). The more standardised and directed the protocols, the lower the likelihood of sampling bias and 'false zeros' in the data. In these cases, a simple Poisson or negative binomial GLM may suffice rather than a ZI model; at very low levels of zero-inflation the performance of the GLMs was shown to be equal to that of the ZI models in Simulation 2. Nevertheless, my findings from Simulation 2 suggest that, regardless of biological zero-inflation, when sampling is suspected to be very incomplete (estimated coverage of total study area < ~20%), ZI models will always be the optimum choice. At low levels of biological zero-inflation, I found ZI models to be more effective than GLMs even when sampling coverage approached levels as high as 90%, as might be the case for species with broad ranges that have been extensively documented, such as important or conspicuous species in countries with long histories of species record keeping.

As well as the Poisson distribution, the negative binomial distribution is also often used for count data, which can also be applied within a zero-inflated modelling framework (Ridout et al., 2001; Minami et al., 2007; Zuur et al., 2009). The negative binomial distribution is able to model an extra proportion of the excess zeros compared to the Poisson distribution through the use of an extra model parameter (Ɵ) (Fisher, 1941) and can therefore account for biological aggregation and overdispersion in ecological data (Lindén & Mäntyniemi, 2011). I chose not to investigate a ZI negative binomial model in these simulations to remove confusion when communicating my main message, although I acknowledge that under high levels of biological zero-inflation (as in Simulation 2), such models may well be more effective that the ZI Poisson models. Therefore, when analysing presence-only species data suffering

from high levels of sampling bias, a ZI Poisson model will usually be effective, but it is valuable to know that there are different ZI model types that can be used to address ecological or statistical issues that may arise in species data.

The majority of SDM research to date has focused on producing presence-absence or presence-only distribution maps of species or communities (Brotons et al., 2004; Phillips et al., 2006; Lyashevska et al., 2016). Species abundance maps are produced more infrequently, often due to the practical difficulty of measuring absolute abundance (Lyashaveska et al., 2016). However, their ability to display extra information about density means they are often more informative and preferred (Pearce & Ferrier, 2000; Barry & Welsh, 2002; Johnston et al., 2015).

Although count data are known commonly to suffer from zero-inflation, ZI models have been used to produce accurate species abundance maps from systematically collected species data in very few studies (Bouyer et al., 2015; Lyashaveska et al, 2016), and none have acknowledged or explored bias in their data. It is also not recommended to use SDM to predict species abundance from presence-only or presence-absence data (Jiménez-Valverde et al., 2020), so ZI models that fit abundance by default should be able to cover this methodology gap in the field of SDM. Additionally, scale is hugely important in SDM. Species distributions are often modelled at coarse resolutions across national or international scales due to the availability of predictors, even though occurrences relate more to localised environmental factors (Guisan et al., 2007; Kuemmerlen et al., 2014). The coarser the grain size used in presence-absence or presence-only SDM, the more the raw occurrences are aggregated into a binary variable and density information is lost. Therefore, it is likely that at coarse resolutions, using abundance rather than occurrence data preserves more information and will produce more accurate maps of habitat suitability.

My findings from Simulation 3 suggest that when having to decide how to aggregate data to match the coarser resolution of the environmental predictors, the best method is to aggregate species occurrences into counts of abundance and fit using a ZI model, rather than aggregate into presence-absence data and

fit using a traditional SDM method such as MaxEnt. This provides two main benefits over presence-absence methods in that a) ZI models are able to identify and account for bias without prior knowledge of the bias sources and b) extra information about species abundance is retained and modelled. However, it should be noted that for some species that may be very rare or have been extremely poorly sampled, aggregating data points using this ZI method for SDM may be inappropriate, as the sample size and abundance counts may be too low (i.e. only have abundance counts of one even at very coarse resolutions). In these rare cases, traditional MaxEnt SDM methods that are more robust to low sample sizes combined with bias correction techniques may be a better option (Phillips et al., 2009).

I also found that as scale became increasingly coarser, only the ZI models retained a high level of predictive power and were an accurate reflection of species niche compared to MaxEnt or binomial GLMs, especially when the data suffered from sampling bias. I believe that ZI models have an advantage over other statistical methods in that they can be used with either presence-absence data or abundance data collected from citizen science projects: presence-absence data can just be aggregated into a count at a particular resolution. Furthermore, scale was shown to have little influence on the predictive power of ZI models providing bias was accounted for. Nevertheless, this was only simulated across relatively small resolutions (up to 5-km) due to the limitations of the study area and requirement for zero-inflated data, whereas many studies map distributions at larger scales (> 10-km) (Thuiller et al., 2006; Luoto et al., 2007). It is uncertain therefore whether this pattern holds true across more coarse scales of analysis.

In this chapter, I have investigated the performance of ZI models under a relatively restricted set of scenarios. For example, I chose to use a simple scenario in which only two predictors, a biological predictor and a bias predictor, generate patterns in the species distribution. The altitude species was assigned a simple preference for high altitudes, when in fact, there are likely several different environmental influences on the species niche. Furthermore, some of these biological predictors of species presence will also predict sampling bias. Therefore, it is important that prior consideration is

given to the possible influences of any predictor included in the model on both ecological processes and sampling behaviour before it is decided whether to include it in either part of the ZI model.

GLMs, and by extension ZI models, have been criticised for their inability to capture the complex, non-linear relationships which may often characterise species responses to the environment, in contrast with more modern methods such as MaxEnt or other machine learning techniques which are more flexible (Austin, 2002). Nevertheless, GLMs and ZIs also have some clear benefits, such as the ease with which they can be applied, and the transparency of their design. Here, I have shown an additional benefit of ZI models not yet available with any other modelling approach: the ability to simultaneously account for bias and to make inferences about it, when predicting distributions from incomplete sampling. I believe that my approach using ZI models has broad applicability to a variety of scenarios when bias is present, and there are suspected predictors of bias available. ZI models should be especially valuable when species abundance is of interest to the modeller, such as when modelling distributions of individual large animals or trees. Although I acknowledge that GLMs and ZI models have limitations, there is a range of options for more complex versions of these models, such as those incorporating polynomial terms, interactions and LASSO variable selection (Hastie et al., 2009; Vollering et al., 2019), which might allow such models to capture non-linear/complex responses to the environment at the same time as modelling the causes of excess zeroes.

In my simulations, I assume that all 'false absences' are due to sampling bias, but it is likely that in many cases, particularly for rare or cryptic species, they are also generated by detection errors (Fitzpatrick et al., 2009; Dickinson et al., 2010; Kosmala et al., 2016). The species range size and the scale of detectability of the individuals is likely to influence the interpretation of the model "abundance" predictions. For example under-estimation of true abundance could occur when modelling small organisms which appear frequently during the survey, and will be more representative of the likelihood of successfully sampling the species. On the other hand, over-estimation could occur when modelling large, mobile organisms that cover multiple sampling locations, so prediction abundance might be a proxy of the probability of encountering one of a small number of individuals. Hence, there may be

three sources of excess zeros: true zeros from unsuitable habitat, false zeros from lack of sampling and false zeros from detection error. When detection errors are significant, ZI models will not be able to distinguish between the different types of false zeros; but by including predictors in both the count and zero components of the model that capture the processes generating all types of zeros, I believe that ZI models will still be able (mostly) to account for these excess 'false' zeros, and combined with expert knowledge can provide some information about their sources.

5.6 Conclusion

Large collections of species data are extremely useful for SDM and conservation, and yet are limited by issues associated with the recording processes, including sampling bias and zero-inflation. My simulations show that ZI models can fit biased data and identify sources of bias. Most importantly for conservation, by using only predictions from the count component of the ZI model (i.e. the count abundance predictions), biased species data can be used to produce distribution maps comparable to those using unbiased data. I also highlight the importance of considering the use of abundance data in SDM, especially at large spatial scales, when valuable ecological information about density is lost if data in each cell are converted to presence-absence or presence-only. ZI models are advantageous compared to other commonly used SDM techniques such as MaxEnt owing to their ability to retain information about abundance and also to identify and remove bias without prior knowledge of the bias sources. I believe ZI models have been largely overlooked in ecological research, even though they have a huge potential to be useful in SDM, and could have great benefits for conservation and our environment.

Chapter 6: Distribution models calibrated with independent field data predict two million ancient and veteran trees in England.

6.1 Abstract

Large, citizen-science species databases are powerful resources for predictive Species Distribution Modelling (SDM) yet are often subject to sampling bias. There are many proposed methods to correct for this, but little consensus as to which is most effective, not least because the true value of model predictions is hard to evaluate without extensive independent field sampling. I present here in this chapter a nationwide, independent field validation of distribution models of ancient and veteran trees, a group of organisms of high conservation importance, built using a large and internationally unique citizen-science database: the Ancient Tree Inventory (ATI). This validation exercise presents an opportunity to test the performance of different methods of correcting for sampling bias, in the search for the best possible prediction of ancient and veteran tree distributions in England. I fitted a variety of distribution models of ancient and veteran tree records in England in relation to environmental predictors, and applied different bias correction methods including spatial filtering, background manipulation, the use of bias files and finally, Zero-Inflated (ZI) regression models. I then collected new independent field data through systematic surveys of 52 randomly selected 1-km^2 grid squares across England to obtain abundance estimates of ancient and veteran trees. Calibrating the distribution models against the field data suggests there are around ten times as many ancient and veteran trees present in England than the records currently suggest, with estimates ranging from 1.7 to 2.1 million trees compared to the 200,000 currently recorded in the ATI. The most successful bias correction method was systematic sampling of occurrence records, although the ZI models also performed well, significantly predicting field observations, and highlighting both likely causes of undersampling and areas of the country in which many unrecorded trees are likely to be found. My findings provide the first robust nationwide estimate of ancient and veteran tree abundance, and demonstrate the enormous potential for distribution modelling based on citizen science data combined with independent field validation to inform conservation planning.

6.2 Introduction

Citizen-science species databases and other large species record collections are becoming increasingly useful in conservation research and planning, and are able to provide a great deal of information about species distributions across large geographical areas and temporal periods (Pearce & Boyce, 2006; Schmeller et al., 2009; Tiago et al., 2017b). Nevertheless, sampling in this sort of species data is a widely acknowledged problem (Phillips et al., 2009; Hijmans, 2012; Syfert et al., 2013). Sampling bias results in certain areas or species being sampling more intensively or frequently, most commonly because of issues relating to accessibility and the location of the recorders, for example travel time from a recorder's home to a survey site (Dennis & Thomas, 2000), distance from roads or the availability of pathways (Reddy & Dávalos, 2003; Kadmon et al., 2004; Schulman et al., 2007), or elevation/ terrain steepness (Mair & Ruete, 2016). The selective surveying of rare, 'special' species or interesting geographic areas also generates sampling bias in species data (Reddy & Dávalos, 2003; Snäll et al., 2011; Kramer-Schadt et al., 2013). Quantifying bias is further complicated by different taxa suffering from different causes of spatial bias (Mair & Ruete, 2016).

Species Distribution Modelling (SDM) is a common and effective tool for understanding and predicting species distributions and distributional shifts (Beaumont et al., 2007; Chen et al., 2011; Clement et al., 2014). SDM works by assessing the known presence (and sometimes absence) records of a species in relation to environmental variables. The suitability of locations for this species, reflecting its fundamental niche and geographic range, can then be predicted based on environmental characteristics (Araújo & Guisan, 2006; Hijmans & Graham, 2006; Mateo et al., 2011). Many modelling techniques are available, with Maximum Entropy (MaxEnt) modelling being by far the most widely used because of its ability to use presence-only data and to cope with small datasets (Hernandez et al., 2006; Phillips et al., 2006; Elith et al., 2006). Sampling bias in species data can greatly influence SDM performance and quality, as it leads to exaggeration of the importance of the environmental conditions for the species in the better surveyed locations (Syfert et al., 2013). Therefore, predicted species distributions from models built with biased records can vary dramatically compared to the actual distribution: the predictions partly represent survey effort rather than species niche requirements (Phillips et al., 2009).

Incorrect model predictions are particularly detrimental in the planning of conservation projects and decision making about which areas should be protected or subject to management (MacKenzie, 2005). Various methods to assess and correct for sampling bias have been developed recently, and issues created by sampling bias in SDM and citizen science recording schemes are now widely recognised (Phillips et al., 2009; Kramer-Schadt et al., 2013; Fourcade et al., 2014; Boria et al., 2014). However, thorough evaluations of these methods using independently collected, unbiased species data are scarce, and the true value of many distribution models built using biased data remains unclear.

Ground-truthing of model verifications using independently collected, unbiased new data is the ideal scenario when testing model performance and predictions, yet distribution models are rarely tested in this way (Greaves et al., 2006; Costa et al., 2010; Fabri-Ruiz et al., 2019). The reasons for this are obvious, as the time and financial cost of large-scale surveys is often prohibitive and difficult. However, the networks of volunteer recorders for many citizen-science projects may lend themselves to planned ground truthing, and with some forward planning, robust, strategic sampling methods could be applied in many of these large projects. In this chapter I use a large, volunteer survey network of a nationwide citizen-science project, the UK Ancient Tree Inventory (ATI), to do just that: by recruiting a sample of enthusiastic volunteers who regularly record for the project, I carried out nationwide, randomised surveys in order to validate model predictions independently using the newly collected unbiased species data, with the aim of selecting the most robust predictive models of species distributions.

Dead and decaying wood ecosystems are highly complex and fragile, and are found world-wide (Hodge & Peterken, 1998; Siitonen, 2001; Butler et al., 2002; Seibold & Thorn, 2018). They provide resources and habitats for numerous threatened and endangered saproxylic species (Jonsson et al., 2005; Seibold et al., 2015). Ancient and veteran trees exhibit ‘veteran characteristics’ such as a retrenched crown, hollowing trunk, holes and cavities (Read, 2000; ATF, 2008a; Nolan et al., 2020), and are essential contributors to the persistence of dead and decaying wood ecosystems in most biomes, supporting a wide range of fungi, epiphytes, invertebrates, birds and mammals (Speight, 1989; Read, 2000; Humphrey, 2005). The strong historic and cultural significance of ancient and veteran trees also

provides insight into past landscape use and management, and important events in human and environmental history, as well as changes and developments in social behaviour and landscape structure over time (Rackham, 1976; Read, 2000; Zhang et al., 2017; Nolan et al., 2020). Nevertheless, ancient and veteran trees and their associated habitats and species are declining around the world (Gibbons et al., 2008; Fischer et al., 2010; Le Roux et al., 2014; Kirby & Watkins, 2015). Factors such as urbanisation and agricultural intensification, alongside a lack of planting, management and awareness of the development of ancient and veteran tree populations, are all contributing to their steady decline (Read, 2000; Fay, 2002; ATF, 2005, 2011; Lindenmayer et al., 2012; Lonsdale, 2013). In addition, relatively few countries have knowledge about, or are actively recording, the locations and condition of ancient and veteran trees sufficiently well for conservation measures to be effective (Nolan et al., 2020).

The UK is unique in having excellent records of ancient and veteran trees. The Ancient Tree Inventory (ATI) (formerly known as the Ancient Tree Hunt), is a national database of over 200,000 ancient, veteran and other noteworthy trees (Nolan et al., 2020). The ATI is a great example of a successful and popular citizen-science project, with hundreds of new tree records uploaded to the online inventory managed by the Woodland Trust each month by members of the public, ecological organisations and specialised ancient tree volunteer recorders. Nevertheless, like many citizen-science projects and online species databases, because of the non-random, unstructured nature of the recording process, there is likely to be a high level of sampling bias in the ATI. Therefore, the current distribution map of ancient and veteran trees based on the ATI may be more reflective of recorder activity in certain locations than it is of the true geographical distribution of trees. It is also likely that there is huge under-recording of trees in many areas, especially those that are less accessible, less interesting to survey or further away from centres of human population (Phillips et al., 2009; Mair & Ruete, 2016). Thus, despite the large number of records collected, there are thought to be many more undiscovered ancient and veteran trees in the UK including those that are at risk of damage or destruction (Nolan et al., 2020). Obtaining insight into the true distribution of ancient and veteran trees, as well as under- or well-surveyed areas (i.e. patterns of sampling bias), is therefore key for the conservation and protection of this important component of biodiversity.

A further problem of using non-randomly sampled species data, as found in the ATI, which is often encountered in SDM is the lack of information about true absences - locations where the species is definitively not present, rather than those that have simply not been surveyed (Hastie & Fithian, 2013). Presence-only SDM overcomes this by generating 'pseudo-absence' points across the study area. These points are usually positioned at random (Stockwell & Peters, 1999), but they can be weighted by geography, environment or target group sampling (Hirzel et al., 2001; Phillips & Dudík, 2008). However, the method of pseudo-absence generation has been shown to influence model outcomes (Wisz & Guisan, 2009; Barbet-Massin et al., 2012) and can result in unreliable models (Liang et al., 2018).

Predictive species distribution maps based on abundance are much less common than those based on presence or presence-absence, because most large species datasets record only species occurrence (Lyashevska et al., 2016). If the spatial predictors in SDM are only available at a greater resolution than the occurrence data, occurrences have to be aggregated to presence-only or presence-absence at the same resolution, which results in loss of vital information about species density across the study area (Johnston et al., 2015; see Chapter 5). An alternative to aggregating occurrences to presence-absence data is to aggregate them into counts of occurrences (i.e. abundance or pseudo-abundance), at the resolution of the spatial predictors, an approach which retains information about species density and can produce better fitting, more accurate predictive maps (Howard et al., 2014; Johnston et al., 2015; see Chapter 5). One problem with this method is that the new aggregated abundance data are highly likely to be zero-inflated compared with the standard distributions which they are typically expected to follow (Martin et al., 2005; Bird et al., 2014), but this can be overcome with the use of Zero-Inflated (ZI) models (Lambert, 1992). ZI models, which have received relatively little attention in the field of SDM, are able to cope with such data and shown in Chapter 5 to be able to both identify causes of sampling bias, and to facilitate its removal in simulated species data. Here, I use the ATI case study to test my recently proposed method of sampling bias correction using ZI models (see Chapter 5).

The aim of this study is to produce the best possible, unbiased prediction of the current distribution of ancient and veteran trees in England using distribution modelling and large-scale field validation.

Collecting additional data also presents an interesting opportunity to evaluate independently the effectiveness of a variety of bias correction methods in relation to my distribution models, which is something that relatively few studies attempt. I fit distribution models with a variety of different bias correction methods, including ZI models, and evaluate their performance and predictive power using both common internal model validation methods and my independently collected, unbiased field estimates of ancient and veteran tree abundance. Thorough, independent evaluation of the most robust, accurate predictive maps of ancient and veteran tree distribution can assist with future targeted surveys and provide estimates of the work needed to find undiscovered trees to add to the ATI for their protection, as well as helping to estimate the landscape-scale biological value of this habitat-rich resource as a whole.

6.3 Methods

6.3.1 Study species and environmental predictors

Methods in this chapter follow those using the same 1-km grid and ATI ancient and veteran tree records across England from Chapter 4 (see Methods, Chapter 4). Twenty environmental, topographical and anthropogenic datasets were then selected for predictive modelling across the study area for each 1-km grid cell (see Chapter 2, Table 2.5, and Table 6.1). Four predictors were categorical (agricultural class, land class, soil type and type of historic countryside) and 16 were numeric. No strong correlations were found between any numeric predictor (Pearson's correlation coefficient threshold +/- 0.6, Variance Inflation Factor (VIF) < 5). Each predictor was converted to raster format at a 1-km resolution. All processing of predictors was carried out in ArcGIS version 10.3.1 (ESRI, 2018).

Table 6.1 *Information from 20 datasets (see Table 2.5) was collected for each 1-km grid cell, and converted into a useable quantitative model predictor. There are 16 continuous predictors and 4 categoric predictors.*

Original Dataset	Predictor (after processing)	Format
Countryside type	Type of countryside	Categoric
Soil type (1-km)	Most common soil type	Categoric
Agricultural class	Most common agricultural classification	Categoric
Land class	Most common land classification	Categoric
Historic forest	Distance from a historic forest (km)	Numeric
Medieval moated sites	Distance from a moated site (km)	Numeric
Medieval Deer Park	Distance from a medieval deer park (km)	Numeric
Tudor Deer Park	Distance from a Tudor deer park (km)	Numeric
Watercourses	Distance from a water course (km)	Numeric
Altitude (1-km)	Mean altitude (m)	Numeric
Town centre	Distance from nearest town center (km)	Numeric
Major city	Distance from nearest major city (km)	Numeric
Commons	Distance from a commons (km)	Numeric
Major roads	Distance from a major road (km)	Numeric
Minor roads	Length of minor roads (km)	Numeric
Ancient woodland	Cover of ancient woodland (%)	Numeric
National Forest	Cover of forest or woodland (%)	Numeric
Traditional orchard	Cover of traditional orchard (%)	Numeric
Wood-pasture	Cover of wood pasture (%)	Numeric
National Trust land	Cover of National Trust owned land (%)	Numeric

6.3.2 Bias correction techniques

Four types of bias correction method were tested, three of which are conventional presence-only or presence-absence SDM techniques that have been used and evaluated previously (Kramer-Schadt et al., 2013; Fourcade et al., 2014; Beck et al., 2014). These were 1) spatial filtering of occurrence records, 2) restriction of the selection of pseudo-absence background data and 3) the use of bias files in the models (Table 6.2). Three methods of spatial filtering were tested, the first of which was 'systematic sampling' (Fourcade et al., 2014; Beck et al., 2014), where grids of 2-km, 5-km and 10-km resolution were created with the same extent as that of the occurrence records. One occurrence record was then randomly

sampled from each 2-km, 5-km and 10-km grid cell, resulting in a filtering of occurrence records from a total of 94,024 to 11,261, 5,504 and 2,495 final occurrence records respectively.

The second method was 'cluster analysis' (Fourcade et al., 2014), whereby all occurrence records within 1-km of each other were grouped together as a single cluster. Thus, some records in the same cluster were greater than 1-km distance from each other but all were < 1-km from at least one other record in the cluster. From each cluster, a single occurrence record was randomly selected and retained. All records that were further than 1-km away from the next record and did not fall within a cluster were also retained, resulting in a final total of 1,583 occurrence records. The final spatial filtering method was 'weighted distances' (Veloz, 2009; Kramer-Schadt et al., 2013; Boria et al., 2014), where the distance of the nearest record was calculated for each occurrence location, and rescaled into a probability of weighted distances between 0 and 1. A total of 20,000 occurrence records were then selected based on these weighted probability distances, so that records with large distances to the nearest other record were more likely to be selected (i.e. had a probability closer to 1). The processing of the occurrence records for each of these three filtering methods was carried out manually in R (R Core Team, 2018) and ArcGIS.

The other two bias correction methods are both types of manipulation of the selection of the pseudo-absences from the background when fitting distribution models, but differ based on their requirements. The first method, background restriction (Table 6.2), requires no prior knowledge of sampling bias, but involves restricting the area within which the pseudo-absence data were selected (Phillips, 2008; Fourcade et al., 2014). This was done by creating buffer areas around each occurrence point at 1-km, 2-km, 5-km and 10-km distances, within which the pseudo-absence selection was confined. The second method employs bias files which are proxies of likely sources of bias across the study area (Dudík et al., 2005; Elith et al., 2010). The bias file is used to influence the weighted selection of pseudo-absence locations. Six different potential bias sources were considered (Table 6.2). Two of these bias files were record density (number of trees per grid square) and recorder density (centroid location of each recorder's specific records). Having access to information about recorder locations allows us to examine

true factors that cause sampling bias, rather than just environmental proxies, which is something that many large databases are unable to do.

The fourth bias correction method was based on the novel approach developed in Chapter 5, whereby the presence-only ATI records were aggregated into a count of occurrences per 1-km grid cell ('abundance'). In some cases it is likely that this abundance measure is more likely 'pseudo-abundance', as in many species databases single occurrences represent the presence of multiple individuals at a single location. With the ATI data it is less likely this is the case, because each tree is recorded as a single individual, so I use the term 'abundance' throughout, although I acknowledge that 'pseudo-abundance' may be more appropriate in other cases. This results in 12,687 cells (9.7%) containing one or more records. Abundance ranged from 0 to 939 trees per 1-km grid cell and shows severe zero-inflation with respect to a Poisson distribution (Chi Squared test: $\chi 2 = 283{,}637.96$, $p < 0.001$). Aggregating to count data allowed ZI models to be fitted to the 'pseudo-abundance' data and used to both identify and correct for sampling bias (see Chapter 5).

Table 6.2 *Types of bias correction method applied to the Ancient Tree Inventory (ATI) records when modelling the distribution of ancient and veteran trees across England.*

Method	Type	Description
Spatial filtering	Systematic sampling	Randomly sampling one occurrence point per grid cell of 2-km, 5-km or 10-km resolution.
	Cluster analysis	Randomly sampling once occurrence point per grouped cluster of records within 1-km distance.
	Weighted distances	Sample 20,000 occurrence points based on weighted probabilities of distance to nearest other occurrence location. Occurrences with greater distances to other occurrence locations were more favoured in the selection process.
Background restriction	Restricting background selection area	Restricting the area within which pseudo-absences are randomly chosen by creating buffers at varying distances (1-km, 2-km, 5-km and 10-km) around each occurrence location. Pseudo-absences generated were then confined solely to these areas.
Bias files	Recorder location	Weighted probability surface for the selection of 10,000 pseudo-absence points based on a kernel density analysis of the locations of recorder home bases (centroid locations of all records uploaded by each recorder).
	Density of towns and cities	Weighted probability surface for the selection of 10,000 pseudo-absence points based on a kernel density analysis of the locations of all town and city centroids across England.
	Density of roads (major and minor)	Weighted probability surface for the selection of 10,000 pseudo-absence points based on a kernel density analysis of all major and minor roads across England.
	Altitude	Weighted probability surface created by rescaling altitude values at a 1-km resolution across England for the selection of 10,000 pseudo-absence points.
	Distance to nearest of wood-pasture	Weighted probability surface for the selection of 10,000 pseudo-absence points based on a 1-km resolution raster layer of distance to the nearest wood-pasture across England.
	Record density (abundance of records per 1-km grid cell)	Weighted probability surface for the selection of 10,000 pseudo-absence points based on record density per 1-km grid cell (i.e. abundance of ancient and veteran tree records).
ZI Models	Use of 'pseudo-abundance'	Aggregating presence records to a count of 'pseudo-abundance' at a resolution of 1-km, and fitting ZI models to identify and correct for sampling bias (see Chapter 5). Predictions of abundance for each grid cell can be used to create a distribution map of ancient and veteran trees across England.

6.3.3 Modelling and analysis

MaxEnt presence-only models were fitted to the ancient and veteran tree occurrence records under each of the presence-only bias correction methods (spatial filtering, background manipulation and bias files) at a 1-km resolution using 'ENMeval' package in R (Muscarella et al., 2014). An additional model with no bias correction (i.e. the raw occurrence data) was also fitted for comparison. All models were fitted using 10,000 pseudo-absence background points, which were randomly sampled across the study area unless explicitly different due to the bias correction method. All other MaxEnt parameters were left at their default values (Phillips & Dudík, 2008). Model tuning using combinations of feature classes 'Linear (L)', 'Linear and Quadratic (LQ)', 'Linear, Quadratic and Product (LQP)' or 'Linear, Quadratic, Product, Threshold and Hinge (LQPTH)' and regularisation measures of 0.5, 1, 2, 3, 4, and 5 was undertaken, and the best fitting model for each bias correction method selected based on the corrected Akaike information criterion (AICc) (See Appendix 6.1). All 20 predictors (Table 6.1) were used for each model, but for models using bias files based on one or more of the predictors (towns and cities, roads, altitude or wood-pasture bias files), models were fitted both with and without those particular predictors for comparison.

Model predictions were created for each MaxEnt model and evaluated using 10 fold cross-validation (CV), where the data are randomly split into 10 parts, with each part sequentially acting as the 'test' data during internal model evaluation while the other nine are used to train the model. Initial analysis (not shown) was used to evaluate alternative non-random methods of geographically splitting the data into training and test data, but these proved less effective than CV (see Appendix 6.1). Models were evaluated using AICc and 'Area Under the Curve' (AUC) for the training and test data. AICc is a test of model fitting and performance based on goodness of fit and its ability to avoid overfitting, and can be used to compare between the fit of different models (Akaike, 1973). AUC on the other hand is a measure of a model's predictive power based on the ROC curve and its ability to correctly classify observations across all possible thresholds of classification of the probability of presence (Fielding & Bell, 1997; Lobo et al., 2008). AUC has been critised as an evaluation metric of distribution modelling

(Lobo et al., 2008; Peterson et al., 2008), yet still remains one of the most widely used evaluation methods in SDM.

For the fourth bias correction method (ZI models), ZI models were fitted to the pseudo-abundance data. ZI models are an extension of GLMs and combine two components: 1) a "zero component" which models the probability that an observation is an excess zero, and 2) a "count component", which produces the count estimates (Lambert, 1992; Welsh et al., 1996; Zuur et al., 2009). By having two parts, processes generating true zeroes and excess (potentially false) zeroes can be modelled separately (Zuur et al., 2009). When used for SDM with species abundance data suffering from sampling bias, the zero component can model the probability that an abundance of zero at a particular location is truly zero or not, and the count component can then produce an estimate of true abundance at that location (see Chapter 5 for more information). Therefore, ZI models have great potential to model geographically biased species data, and to allow examination of the sources of bias, if unknown, as well as producing predictive SDM maps free of bias. Several studies have used ZI models to examine sampling bias in species data (Dwyer et al., 2016; Williams et al., 2016; Tiago et al., 2017a), but none have used this method to produce prediction maps from real species data.

ZI models were fitted with either a Poisson or negative binomial (NB) distribution. Both error distributions are commonly used for count data and can be applied within a ZI model framework (Zuur et al., 2009). A NB distribution allows for more overdispersion in the data than the Poisson distribution and can account for some (but often not all) of the excess zeroes in zero-inflated datasets through the use of an extra parameter (Θ) (Fisher, 1941). Therefore, it may be more appropriate to use this distribution if there is biological aggregation in the data (Lindén & Mäntyniemi, 2011). In this case, the pseudo-absence data show huge overdispersion (variance/ mean = 122.7), so it is likely that a NB distribution will be more appropriate, even if there is still zero-inflation. Performance of each model was compared using Vuong's AICc test for non-nested models (Vuong, 1989).

All environmental predictors were included in both components (count and zero) of the ZI models in order to examine both the potential influence of each predictor on both species' ecology and sampling behaviour (Table 6.1). All numeric predictors were centred and scaled. Several categories from the categorical variables soil type, agricultural class and land class were combined to aid model fitting. Therefore, there were three agricultural classes ('Agricultural', 'Non-Agricultural' and 'Other'), 10 land classes ('Arable', 'Grassland', 'Urban', 'Coniferous', 'Coastal', 'Freshwater', 'Saltwater', 'Heather/ Bog', 'Broadleaved' and 'Other'), and 10 soil types ('Luvisol', 'Cambisol', 'Gleysol', 'Fluvisol', 'Podzol', 'Leptosol', 'Arenosol', 'Histosol', 'Urban' and 'Other'). All models were fitted in R using package 'pscl' (Zeileis et al., 2008). No collinearity was found in the model residuals (Generalised VIF (GVIF) <10) and spatial autocorrelation was low, with weak correlations between latitude and longitude and model residuals (+/- 0.015).

A ZI model is capable of producing three types of predictions: 1) a prediction of abundance from the count component, 2) a prediction of abundance from the whole model, taking into account the processes generating the excess zeroes and 3) a probability prediction (known as the 'zero prediction') that an observation is an excess zero (see Chapter 5 for more information). If all zeroes are true zeroes (i.e. there are no false absences), then the most accurate prediction of abundance will be the second of these (abundance from the whole model), because the excess zeroes are the result of some underlying biological process. However, if a proportion of the excess zeroes result from sampling bias, then the count component prediction (hereafter known as the 'count abundance' prediction) may be a more accurate representation of the true species abundance, and the 'model abundance' prediction will partly reflect the processes underlying the sampling bias. Therefore, the whole model prediction of abundance can provide insight into the sources of sampling bias in the model, while the count prediction provides estimates of abundance free from bias. As the level of sampling bias in the ATI is unknown, both types of predictions could be informative and therefore were evaluated separately (see Chapter 5 for more information).

Model cross-validation predictions (both count and whole model predictions of abundance) from the ZI models of ancient and veteran tree abundance for each 1-km grid cell were created using 10 fold cross-validation as described above. Cross-validation predictions were evaluated using Spearman's Rank correlation between predictions and raw abundance, Root Mean Square Log Error (RMSLE) and training and test AUC. Following the same methods and reasoning as Chapter 5, training and test AUC were calculated for each CV fold by converting the abundance predictions from the Poisson and NB models into presence-absence predictions based on a varying threshold of the mean predicted abundance across all grid squares.

6.3.4 Field surveys and model verification

A set of 90 1-km grid cells was selected across England for further independent model verification using field surveys. These squares comprised two groups: 1) 50 of the squares that were selected completely at random so that there would be no additional sampling bias in the results and 2) 40 squares that were selected based on model predictions to ensure that, despite the high proportion of squares which contain no trees, there was good representation in the sample of squares with existing tree records in the ATI and/or predicted tree presences that could be verified (Fig. 6.1). These 40 squares were selected using the ZI NB 'model abundance' predictions; I used the highest performing ZI model and one of the best fitting models overall to generate these predictions. The ZI NB predictions were firstly categorised as being either low or high predicted abundance using the same classification method that I used from Chapter 3: predictions above the mean predicted probability that a square contains zero records (i.e. the mean zero prediction from the ZI model across each grid square) threshold were categorised as high predicted abundance, and all predictions below categorised as low predicted abundance. Then each square was categorised into one of four groups: 1) no ATI records and low predicted abundance, 2) no ATI records and high predicted abundance, 3) ATI records and low predicted abundance and 4) ATI records and high predicted abundance. From each group 10 squares were randomly selected, resulting in the 40 ZI model squares.

Each of the 90 squares was assessed for accessibility using aerial maps and photography. If a square was deemed completely inaccessible (no roads or public rights of way present) then it was discarded and another square selected in the same manner (n = 4 out of 90). A survey form was created for each square containing details about location, what to record (number and location of ancient and veteran trees, date of survey, photographs), how to record any trees found on the form, possible car parking spaces for the recorder during the survey, and all roads and public rights of way (Appendix 6.2). Recorders were also encouraged where possible to record species or genera of the trees found, although this was not included in the analysis in this study due to the relatively low number of individuals of each different species recorded (Fig. A6.3.1). This was likely because of the difficulty in identifying tree taxa when out of leaf during the late autumn/ winter months, as well as problems classifying any trees that were recorded from a distance because of accessibility issues.

The aim of each survey was to cover each 1-km grid as completely and thoroughly as possible, using multiple trips if necessary and binoculars to view areas from afar that were not accessible. In order to maximise the chances that every ancient and veteran tree in the square was found during the surveys, 'areas of interest' were designated on each survey form to help the recorders avoid wasting their time sampling areas with a very low likelihood of ancient trees e.g. industrial parks, new housing estates, open fields, etc., determined using aerial photography and Ordnance Survey Open Street Maps. Only those areas deemed very unlikely to have any trees (or at least any ancient or veteran trees) were not covered under an 'area of interest'. Therefore, I assume that if all areas of interest had been surveyed with 100% coverage, then all ancient and veteran trees had been found. Each survey required the recorder to note the time spent surveying the whole square and each individual area of interest, as well as estimating the percentage from each area of interest that was covered during the survey. Any parts of the whole square that were not surveyed were either the result of not being an area of interest, accessibility issues or due to the lack of time of the recorders.

Recorders comprised a range of volunteers from different sources including the Ancient Tree Forum, Woodland Trust staff members, Woodland Trust ancient tree recorders, Woodland Trust ancient tree

verifiers and other independent volunteer ancient tree enthusiasts. Initially one square was assigned to each recorder, according to geographical proximity to their home, although some recorders completed several squares if no other recorder lived sufficiently close to that square. The recorders had no prior knowledge of any model predictions. Unfortunately, although squares were first assigned from March, due to extensive Covid-based travel restrictions at various points throughout 2020, many recorders assigned to squares were unable to complete them, and 39 out of 90 squares were completed by volunteers (Fig. 6.1). An additional 13 squares were completed by the authors, resulting in a total of 52 squares of the initial 90 (58 %) being completed (Fig. 6.1). Although the authors had prior knowledge of the model predictions, care was taken wherever possible to carry out the surveys impartially. All surveys were carried out throughout the months of August to December, travel restrictions permitting, during daylight hours.

Three metrics were obtained from the field work results: 1) whether ancient or veteran trees were present or absent in each square (presence-absence), 2) raw abundance of ancient and veteran trees found in each square and 3) estimated density of ancient and veteran trees per square in relation to survey effort of the volunteer (number of ancient and veteran trees / estimated total area of the whole grid square surveyed in m^2). Presence-absence verification metrics were analysed using AUC in relation to each of the model predictions of either habitat suitability (MaxEnt models) or abundance (ZI models). For this, the ZI 'model abundance' predictions were converted into binary presence-absence form based on a varying threshold of the 'median' prediction across all 90 grid squares. Although I previously use the 'mean probability' as my threshold in all other parts of this thesis, I chose to use median here instead as several of the abundance predictions were extremely high and would therefore skew the mean resulting in the majority of predictions being classed as absences. The raw abundance and density field work metrics were analysed using Pearson's and Spearman's correlation coefficients: both coefficients were used in order to examine the effect of two potential outliers. AUC was selected based on the necessity to have a metric that could compare predictions of abundance and habitat suitability: it is much more feasible to convert abundance to presence-absence rather than to do the opposite. This metric is not perfect, and is likely to result in a loss of information from the ZI models. Using the

correlations provides an alternative, albeit crude, method of direct assessment of the predictions against field verification results.

In order to calibrate the models and provide total estimates of ancient and veteran tree numbers across England, a linear regression model was fitted for each set of model predictions for the 52 surveyed grid squares in relation to either raw tree abundance or tree density from the field surveys. Each of these linear regression models was then used to calibrate each model's predictions for all of the grid squares across England in order to provide predictions of abundance or tree density in each grid square. These estimates were then summed across all grid squares to predict the total number of ancient and veteran trees across England.

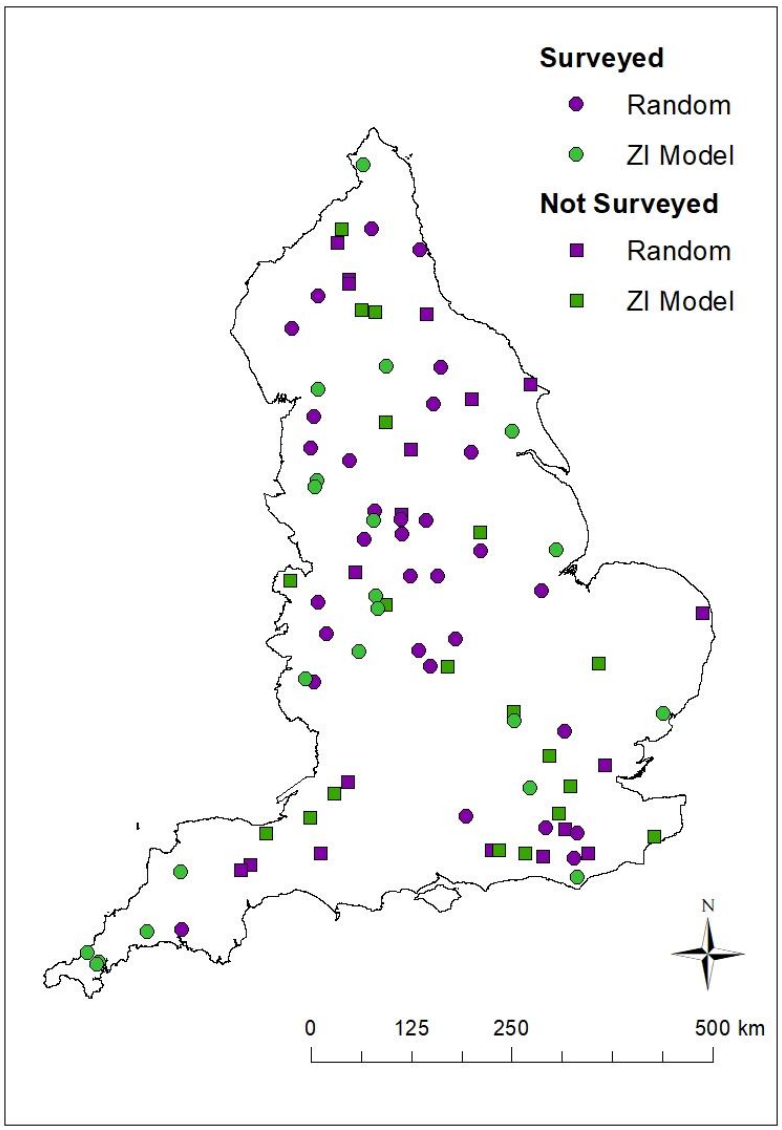


Fig. 6.1 *Centroid locations of each of the 90 1-km grid squares selected for field verification. 50 squares (purple) were selected at random across England and 40 squares (green) were selected based on the model predictions from the zero-inflated (ZI) models. Squares that were actually surveyed for the field verification are indicated as circles, whereas those that were not able to be surveyed due to travel restrictions are indicated as squares.*

6.4 Results

6.4.1 Model fitting and performance using internal model validation

Internal model validation suggests that the highest performing bias correction method based on AICc was the 'cluster analysis' spatial filtering technique, followed by systematic sampling at a 5-km and 10-

km resolution (Fig. 6.2a). All other spatial filtering methods also performed better than the model with no bias correction. Similarly, ZI models performed well compared to other methods, particularly when using a NB distribution based on internal model evaluation. All other bias correction methods showed little difference compared to a model with no bias correction. The most effective bias file was 'record density', followed by altitude and wood-pasture (Fig. 6.2a), with the least effective being towns and cities. Nevertheless the differences among all bias files were relatively small. There was also little difference between the background restriction methods, all of which performed relatively poorly.

When tested against the data used to build the models (cross-validation predictions) using AUC, there appeared to be little improvement in model predictive power when using any bias correction method in relation to the model with no bias correction (Fig. 6.2b & c). Nevertheless, models fitted with bias files provided the best predictions overall based on both training and test AUC, particularly those using 'altitude', 'wood-pasture' and 'roads'. Background restriction using a 10-km buffer was the best background manipulation method and weighted distance was the best spatial filtering method. ZI models performed relatively poorly based on predictive power compared to all other models, although as mentioned in the methods, this is likely owing to a loss of information when converting abundance to presence-absence to calculate AUC.

As suspected, the ZI NB model provided a better fit to the data than the ZI Poisson model (Vuong AICc test: $Z = -22.72$, $p < 0.001$; NB AICc = 128783.0, d.f. =80; Poisson AICc = 290932.5, d.f. = 81). Evaluation of model predictions using internal model validation showed support for the NB model having overall greater predictive power compared to the Poisson model (Fig. 6.2b & c). Additionally, as well as the NB model outperforming the Poisson model, the 'model abundance' predictions showed stronger correlations to the raw data (Poisson $r_s = 0.257$ and NB $r_s = 0.277$), than the 'count abundance' predictions (Poisson $r_s = 0.203$ and NB $r_s = 0.226$), as well as lower error margins (whole model prediction RMSLE: Poisson – 0.566, NB – 0.583; count prediction RMSLE: Poisson – 1.492, NB – 0.706). This is likely because the excess zeroes, as well as being the result of sampling bias, are sometimes caused by ecological processes (e.g. biological aggregation), so excluding the zero

component completely from the model predictions (as in the count abundance prediction), removes important biological information from the abundance prediction.

Model coefficients for the ZI NB model (Table 6.3) provide insight into predictors influencing both the count component and excess zeroes. The count prediction of the abundance of ancient and veteran trees was positively associated with higher altitudes, being closer to Tudor deer parks, commons (land owned collectively by many people with traditional shared grazing or harvesting rights) and National Trust sites, being further away from towns and cities, having a greater coverage of forest and wood-pasture but less coverage of orchard, and being associated with fewer minor roads (Table 6.3). The count prediction of abundance also differed significantly across agricultural class, countryside type, land class and soil type, and was most likely to be highest on non-agricultural, freshwater or broadleaved land classes and fluvisol soil type (Table 6.3).

Many predictors had an influence on the levels of zero-inflation, indicating a potential influence on sampling bias. The likelihood an observation is an excess zero (and is potentially an un-sampled square) increased with increasing coverage of minor roads, wood-pasture, orchard, ancient woodland and forest. Squares that have an observed abundance value of zero were also more likely to be excess (potentially 'false') zeroes if they were further from watercourses, historic forests, moated sites and nearer to commons, National Trust land, medieval and Tudor deer parks and at lower altitudes, as well as on different land types, soil classes and countryside types (Table 6.3). Interestingly, moated sites, historic forests, medieval deer parks, ancient woodland and watercourses had a significant influence only in the zero-component, suggesting they are stronger influences of sampling than of the true underlying ecology determining the tree distribution.

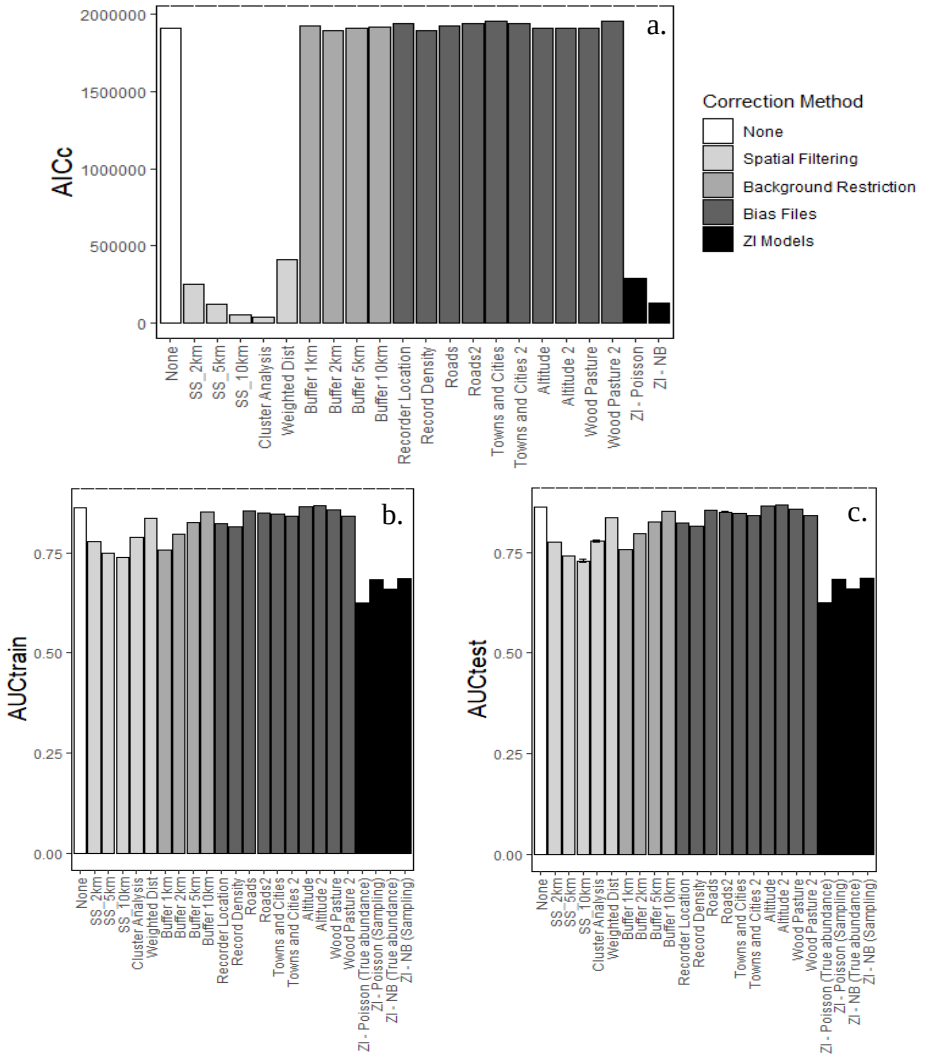


Fig. 6.2a *Corrected Akaike's Information Criterion (AICc),* ***6.2b*** *Training Area Under the Curve (AUC) and* ***6.2c*** *Testing Area Under the Curve (AUC) based on internal model validation for each species distribution model of ancient and veteran tree distribution across England using four main types of bias correction method (spatial filtering, background restriction, bias files and ZI models). Spatial filtering methods include Systematic Sampling (SS) at resolutions of 2-km, 5-km and 10-km, cluster analysis and weighted distance method. Background restriction involved the selection of 'pseudo-absence' points only from buffers of varying distance (1-km, 2-km, 5-km and 10-km) around each occurrence point. Bias files involved weighting the selection of 'pseudo-absence' points according to a proxy for the bias. Where the chosen bias source is also a model predictor, models were fitted with and without the predictor; models missing the predictor are indicated with '-2'. Finally two zero-inflated (ZI) models were fitted using either a Poisson or a negative binomial (NB) distribution. Predictions of both 'count abundance' (from the count component) and 'model abundance' (from the whole model) are shown. Error bars represent ± variance, although in most cases are too small to be visible at this scale.*

Table 6.3 *Model coefficients (± standard error), Z value and p value of significance are shown for the negative binomial ZI model for both the count and zero components.*

Model predictor	***Count component***		***Zero component***	
	Coefficient (±SE)	***Z value***	***Coefficient (±SE)***	***Z value***
Intercept	*-3.276 (0.670)*	***-4.893****	*-3.704 (1.330)*	***-2.785****
Agricultural class - Agricultural	*0.601 (0.279)*	***0.031****	*-0.894 (0.359)*	***-2.490****
Agricultural class – Non-Agricultural	*0.703 (0.283)*	***0.013****	*-0.338 (0.368)*	*-0.918*
Altitude	*0.074 (0.033)*	***0.026****	*0.147 (0.039)*	***3.809****
Type of Countryside - Ancient	*0.036 (0.045)*	*0.431*	*-0.669 (0.060)*	***-11.20****
Type of Countryside - Highland	*-0.328 (0.067)*	***-4.926****	*-0.610 (0.091)*	***-6.672****
Type of Countryside - Cornwall	*0.088 (0.135)*	*0.514*	*1.413 (0.141)*	***10.02****
Landclass - Broadleaved	*2.349 (0.597)*	***3.937****	*0.416 (1.031)*	*0.403*
Landclass – Heather/Bog	*1.529 (0.610)*	***2.509****	*1.031 (1.031)*	*0.999*
Landclass – Saltwater	*2.072 (0.753)*	***2.752****	*1.664 (1.137)*	*1.465*
Landclass – Freshwater	*2.783 (0.637)*	***4.368****	*0.639 (1.066)*	*0.600*
Landclass – Coastal	*1.268 (0.684)*	*1.855*	*1.654 (1.090)*	*1.517*
Landclass – Coniferous	*2.100 (0.604)*	***3.477****	*1.963 (1.031)*	*1.904*
Landclass – Urban	*2.322 (0.596)*	***3.893****	*1.109 (1.024)*	*1.083*
Landclass – Arable	*1.991 (0.594)*	***3.355****	*0.802 (1.019)*	*0.787*
Landclass - Grassland	*2.152 (0.593)*	***3.627****	*0.625 (1.019)*	*0.614*
Soil type - Luvisol	*0.455 (0.123)*	***3.699****	*0.692 (0.213)*	***3.246****
Soil type - Cambisol	*0.227 (0.122)*	*1.857*	*0.702 (0.212)*	***3.303****
Soil type - Gleysol	*0.310 (0.124)*	***2.498****	*0.912 (0.215)*	***4.241****
Soil type - Fluvisol	*0.574 (0.158)*	***3.638****	*1.242 (0.239)*	***5.207****
Soil type - Podzol	*0.360 (0.147)*	***2.449****	*0.852 (0.253)*	***3.364****
Soil type - Leptosol	*0.406 (0.137)*	***2.953****	*0.434 (0.228)*	*1.905*
Soil type - Arenosol	*0.022 (0.157)*	*0.139*	*0.681 (0.262)*	***2.601****
Soil type - Histosol	*-0.573 (0.295)*	*-1.943*	*1.366 (0.350)*	***3.900****
Soil type - Urban	*0.266 (0.140)*	*1.894*	*1.200 (0.249)*	***4.812****
Tudor Deer Park	*-0.130 (0.029)*	***-4.494****	*0.500 (0.036)*	***13.92****
Moated Site	*-0.050 (0.034)*	*-1.477*	*-0.321 (0.037)*	***-8.639****
Historic Forest	*0.003 (0.022)*	*0.146*	*-0.252 (0.029)*	***-8.632****
Medieval Deer Park	*-0.041 (0.021)*	*-1.955*	*0.080 (0.028)*	***2.874****
National Trust	*-0.380 (0.021)*	***-17.71****	*0.275 (0.028)*	***9.644****
Cities	*0.120 (0.025)*	***4.856****	*0.012 (0.032)*	*0.383*
Towns	*0.095 (0.028)*	***3.391****	*-0.016 (0.034)*	*-0.486*
Commons	*-0.096 (0.017)*	***-5.545****	*0.079 (0.024)*	***3.265****
Major Roads	*0.013 (0.023)*	*0.566*	*-0.050 (0.028)*	*-1.755*
Cover of forest	*0.226 (0.028)*	***8.177****	*-0.312 (0.039)*	***-8.057****
Cover of ancient woodland	*-0.009 (0.018)*	*-0.478*	*-0.475 (0.071)*	***-6.696****
Cover of orchard	*-0.020 (0.010)*	***-1.990****	*-0.778 (0.097)*	***-8.014****
Cover of wood-pastures	*0.374 (0.012)*	***31.93****	*-18.99 (3.498)*	***-5.431****
Watercourse	*0.030 (0.016)*	*1.883*	*-0.293 (0.024)*	***-12.08****
Minor Roads	*-0.117 (0.026)*	***-4.473****	*-0.653 (0.050)*	***-13.13****
Log(theta)	*-2.105 (0.019)*	***-113.1****		

6.4.2 Model validation using independent random field surveys

New independent surveys of 52 1-km grid squares resulted in a total of 459 ancient and veteran trees being recorded (94 ancient and 365 veteran), 285 of which had not previously been recorded on the ATI. Before the surveys only 15 out of 52 squares had records of ancient or veteran trees, but this number was increased to 38 out of 52 following the surveys. Seven squares received 100% survey coverage, and 32 squares (62%) had at least 50% of their area surveyed (Fig. 6.3). Accessibility was an issue for some squares, although only three squares received a survey coverage of < 20%.

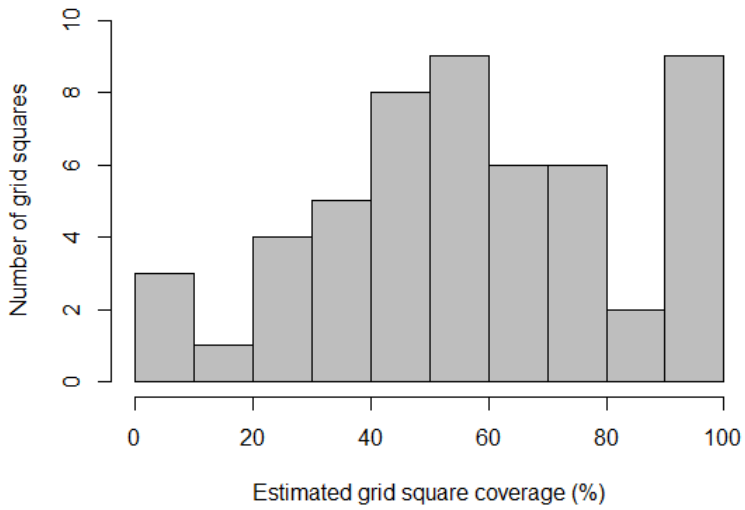


Fig. 6.3 *Histogram of the estimated percentage coverage of each grid square during the field surveys. Percentage coverage was estimated by totalling the area covered from each 'area of interest' and any other areas that the recorders were able to survey.*

Many of the bias corrected models produced predictions that strongly correlated with the field estimates of ancient and veteran tree abundance or tree density, and bias correction substantially improved the predictive power of the distribution models compared to the uncorrected model (Table 6.4). However, the evaluation of the performance of each model when predicting raw abundance or density of ancient and veteran trees depended heavily on whether the raw values (Pearson correlation coefficients) or ranked values (Spearman correlation coefficients) were used. This discrepancy was caused by two outlier squares with extremely high predictions of abundance that were likely inflating the accuracy of the raw predictive power of the models when evaluated with Pearson's correlation (Fig. A6.3.2).

Field estimates of both raw tree abundance and density based on the Spearman ranked correlations provided good support for systematic sampling, and showed significant, strong correlations with model predictions, particularly at a 2-km and 5-km resolution (Table 6.4). The only other methods that increased model predictive power relative to the uncorrected model were the cluster analysis spatial filtering technique, and the wood-pasture bias file (Table 6.4). When evaluated using estimates of survey effort (i.e. against tree density) rather than with the raw abundance of trees per grid square, all these techniques produced predictions with stronger correlations to the field estimates, and the best bias correction was still systematic sampling at either a 2 or 5-km resolution, although using the wood-pasture habitat as a bias file also produced good results (Table 6.4).

As with the ZI models, the most important predictor of ancient and veteran tree habitat suitability across all MaxEnt models was the cover of each square by wood-pasture, which was especially true for the uncorrected model (Table 6.5) where it accounted for over 66% of variable importance. Other important predictors in the uncorrected model included National Trust land, cover of forest or ancient woodland and soil type (Table 6.5). When using the optimum sampling bias correction method (systematic sampling), wood-pasture variable importance dropped significantly by almost 50%, although it was still the most important variable. Other big changes included an increase in permutation importance of the type of countryside and the distance to a Tudor deer park, both by 11% (Table 6.5). The most important predictors of ancient and veteran trees from the systematic sampling model were therefore similar to the ZI models, and included wood-pasture cover, cover of forest, distance to a Tudor deer park, type of countryside and also the presence of minor roads.

Table 6.4 *Independent field evaluation of model predictions. Model predictions were evaluated against a) field verification estimates of the presence-absence (P-A) of ancient and veteran trees per square using Area Under the Curve (AUC), b) field estimates of raw tree abundance (total number of trees recorded per square) using Pearson's (r) and Spearman's (r_s) correlation coefficient tests and c) field estimates of tree density (number of trees in relation to estimated percentage cover of each square) also using Pearson's and Spearman's correlation coefficient tests. See Methods for a detailed description of each bias correction method. Values in bold represent those that are significant. Where indicated, significance levels are: p < 0.05: *, p < 0.01: **, p < 0.001: **. For each model the total predicted abundance of ancient and veteran trees (T) across England was calculated from a linear regression model between the model predictions and field verification data (both raw tree abundance and tree density) for the 52 surveyed squares.*

	P-A	Tree abundance					Tree density				
Model	**AUC**	**r**	**p**	**r_s**	**p**	**T**	**r**	**p**	**r_s**	**p**	**T**
None	0.613	**0.589**	***	**0.339**	*	911,842	**0.490**	***	**0.410**	**	1,867,480
ZI Poiss. (Count)	0.549	**0.804**	***	0.190		873,219	**0.667**	***	0.216		1,830,904
ZI Poiss. (Whole model)	0.549	**0.865**	***	**0.286**	*	831,650	**0.715**	***	**0.329**	*	1,791,807
ZI NB (Count)	0.600	**0.929**	***	0.223		837,739	**0.763**	***	**0.275**	*	1,800,014
ZI NB (Whole model)	0.500	**0.930**	***	0.270		831,096	**0.765**	***	**0.345**	*	1,792,819
SS 2km	**0.658***	0.239		**0.431**	**	1,054,659	**0.354**	**	**0.534**	***	1,930,281
SS 5km	**0.664***	0.222		**0.429**	**	1,045,737	**0.421**	**	**0.528**	***	1,862,409
SS 10km	0.626	0.245		**0.382**	**	1051987	**0.390**	**	**0.476**	***	1,912,788
Cluster Analysis	0.618	**0.642**	***	**0.378**	**	842,714	**0.666**	***	**0.476**	***	1,725,977
Weighted Distance	0.557	**0.442**	***	0.224		1,008,806	**0.366**	**	**0.284**	*	1,961,867
Buffer 1km	0.597	**0.919**	***	0.185		1,091,502	**0.749**	***	0.209		2,042,407
Buffer 2km	0.614	**0.852**	**	**0.335**	*	991,488	**0.712**	***	**0.390**	**	1,943,748
Buffer 5km	0.618	**0.377**	**	**0.309**	*	991,114	**0.310**	*	**0.362**	**	1,945,543
Buffer 10km	0.602	**0.400**	**	0.199		975,943	**0.326**	*	0.259		1,932,274
Recorder Density	0.586	**0.463**	***	0.215		975,943	**0.383**	**	0.251		1,894,780
Record Density	0.443	**0.584**	***	0.251		975,943	**0.529**	***	**0.319**	*	2,010,821
Towns and Cities 2	0.470	0.156		0.080		1,097,494	0.135		0.110		2,045,136
Towns and Cities	0.395	0.090		0.229		1,120,545	0.076		0.258		2,069,004
Roads 2	0.556	**0.464**	***	0.179		1,008,015	**0.382**	**	0.215		1,962,109
Roads	0.446	0.207		0.125		1,001,956	0.169		0.163		1,956,314
Altitude 2	0.591	**0.627**	***	**0.308**	*	1,100,728	**0.532**	***	**0.346**	*	2,049,180
Altitude	0.468	0.102		0.131		964,732	0.082		0.187		1,915,020
Wood-pasture 2	0.621	**0.869**	***	**0.396**	**	1,141,381	**0.755**	***	**0.473**	***	2,088,979
Wood-pasture	**0.656***	**0.708**	***	**0.359**	**	826,052	**0.582**	***	**0.421**	**	1,788,310

Table 6.5 *Permutation importance of each of the Maximum Entropy distribution model predictors shown for the model with no bias correction compared to the overall best performing bias corrected model using systematic sampling (SS) at a 2-km resolution. The percentage change in permutation importance between the two models is also shown, with positive values representing variables that become more important when bias is corrected for and negative values less important.*

Predictor	***Permutation Importance***		
	No Correction	***SS (2-km)***	***% Change***
Agricultural class	0.209	1.230	1.021
Altitude	0.528	1.866	1.338
Type of Countryside	0.193	11.86	11.667
Land class	1.734	1.929	0.195
Soil type	5.473	7.526	2.053
Tudor Deer Park	2.196	13.47	11.274
Moated Site	0.000	1.249	1.249
Historic Forest	0.010	3.893	3.883
Medieval Deer Park	0.703	0.143	-0.56
National Trust	7.947	6.560	-1.387
Cities	0.297	0.927	0.63
Towns	0.000	0.640	0.64
Commons	0.112	0.595	0.483
Major Roads	0.007	0.506	0.499
Cover of forest	6.997	14.88	7.883
Cover of ancient woodland	5.672	1.856	-3.816
Cover of orchard	0.010	0.246	0.236
Cover of wood-pastures	66.20	18.83	-47.37
Watercourse	0.801	3.556	2.755
Minor Roads	0.909	8.247	7.338

When considering the raw Pearson correlation coefficients, the ZI models perform much better in comparison with the uncorrected model with very strong correlations between field estimates of both abundance and density and model predictions (Table 6.4). This is especially true for the ZI NB model, and based purely on this evaluation metric, the ZI NB appeared to be the best method of all to deal with sampling bias. However, because of the outlier grid squares (Fig. A6.3.2), Spearman correlations are likely a better measure of performance, but it is interesting to see the high performance of the ZI models at correctly predicting squares with very high abundances of trees (Fig. A6.3.2).

Prediction maps of ancient and veteran tree distributions from models using bias correction show substantial differences compared to the uncorrected model (Fig. 6.4, see Fig. A6.3.3 and A6.3.4 for maps from all models) with much more variation in habitat suitability across England when using systematic sampling (Fig. 6.4). This model's prediction map suggests there are more areas with high suitability, especially in the south-east of England, the Lake District and in Herefordshire. In contrast, the bias file models using record-density or wood-pasture habitat suggests there are relatively few areas of high suitability, many of which are actually wood-pastures (Fig. 6.4). Prediction maps of abundance from the ZI models are shown in Figure 6.5 and show some areas of high suitability, particularly around London and the New Forest National Park in the south. Maps of the zero predictions from the ZI models provide interesting insight into areas with high numbers of excess zeroes, where trees are likely to have been particularly under-recorded. These maps suggest under-recording in much of Cornwall and Devon, Norfolk and other counties in the East of England and in parts of Northumberland.

Calibrated model predictions of the total number of ancient and veteran trees across England using the field data are very similar across all models, with around 2 million trees (1.7 – 2.1 million) predicted based on the estimated tree density (which accounts for estimated survey effort) from the field validation from all models (Table 6.4). This prediction ranges from 1,725,977 when using the spatial filtering technique, cluster analysis, to 2,088,979 when using the wood-pasture bias file (thus the range across all models is 363,002 trees). Predictions of the total number based on the raw abundance with no correction for survey effort are obviously lower, and range from 826,052 with the wood-pasture bias file to 1,120,545 (towns and cities bias file) (Table 6.4).

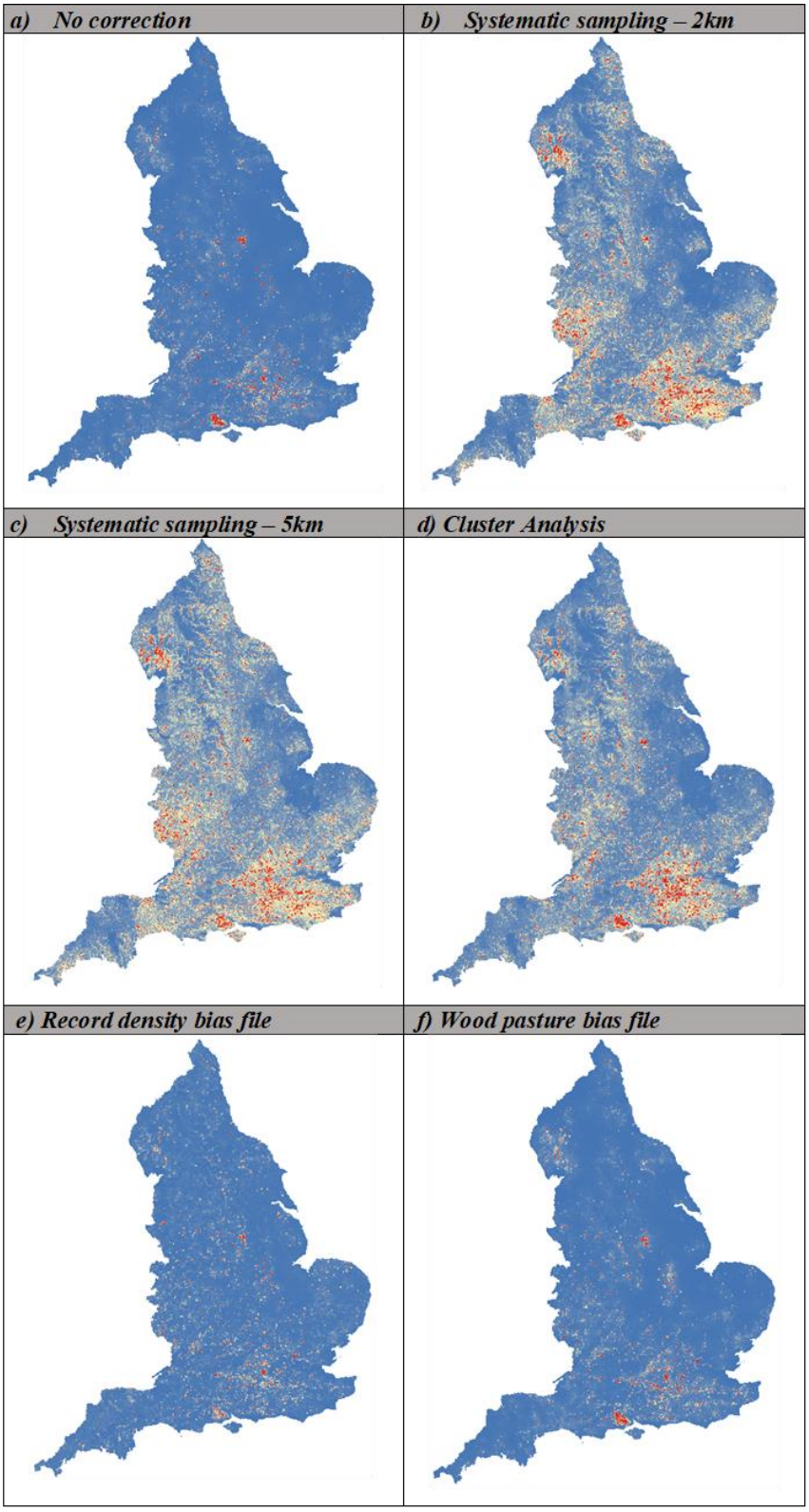


Fig. 6.4 *Predicted distribution maps of habitat suitability for ancient and veteran trees across England from a) a model with no bias correction, and some of the highest performing Maximum Entropy bias correction methods: b, and c) systematic sampling using grids of 2km and 5km resolution, d) cluster analysis, e) record density bias file, and f) wood-pasture bias file. Habitat suitability ranges from low suitability (blue) to high suitability (red).*

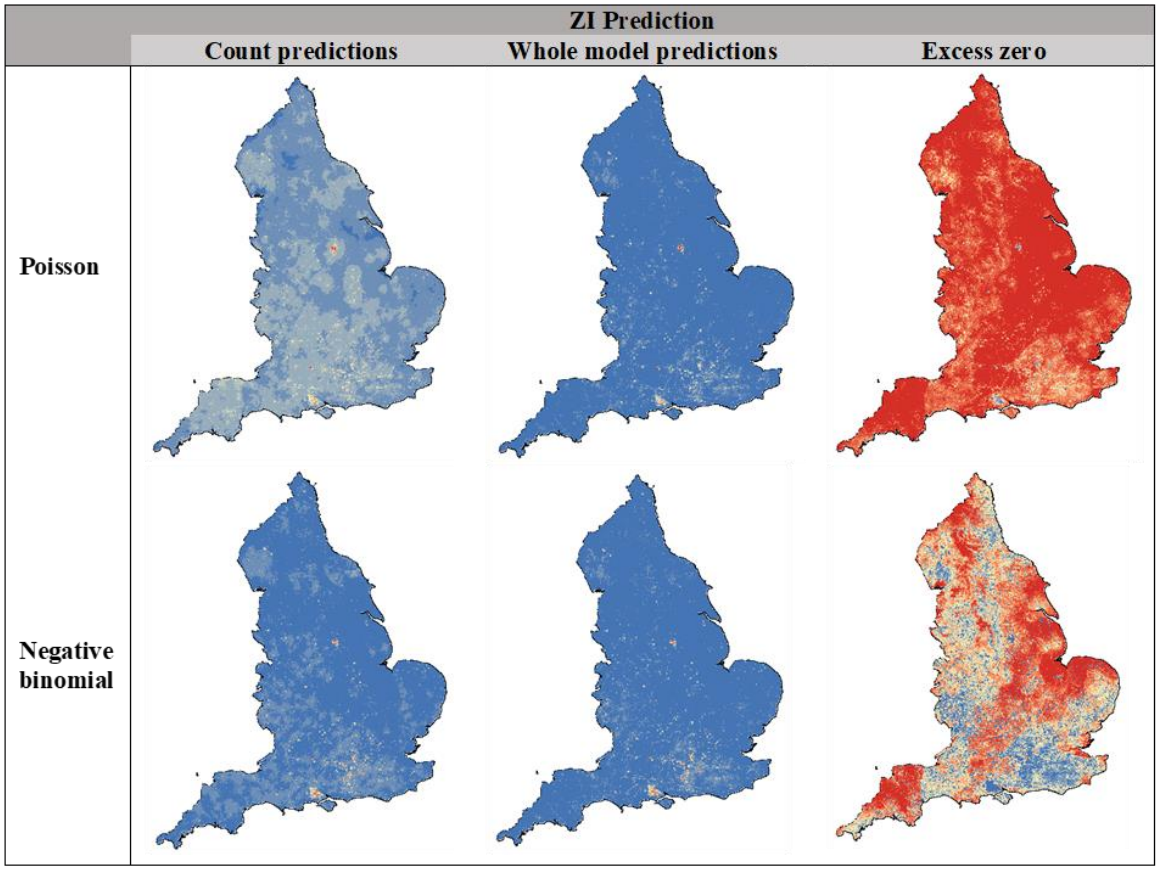


Fig. 6.5 *Predicted maps of the abundance of ancient and veteran trees across England from the Poisson and negative binomial (NB) zero-inflated (ZI) models. Three types of predictions are shown 1) count abundance prediction only from the count component of the ZI model, 2) whole model abundance prediction, from the whole of the ZI model and 3) the excess zero prediction, which represents the probability that an observation is likely to be an excess zero (i.e. a 'false absence'). Red areas in the predicted abundance maps represent areas of high abundance, whereas in the zero probability maps they represent places where it is likely there is under sampling.*

6.5 Discussion

In this chapter, I have presented a rare empirical test of the ability of models fitted using a large citizen-science species database to provide an unbiased prediction of the distribution of ancient and veteran trees across a large geographic area. My results using robust independent field verification methods show that there are indeed many undiscovered ancient and veteran trees across England, and that only a small proportion of the ancient and veteran tree population has been mapped. By evaluating and selecting the best bias correction methods to apply to my distribution models, we can produce accurate predictive maps of the locations of these previously unknown trees, to inform future targeted surveying and conservation plans for these valuable components of terrestrial biodiversity.

It has long been suspected that there are many unrecorded ancient and veteran trees across England with great ecological importance in terms of their dead-wood habitats and associations with saproxylic species (Read, 2000; Butler et al., 2002; Fay, 2004). This study provides strong support for the need to find and record these trees. The field surveys covered a very small percentage of the area of England (0.04%), yet they increased the number of known ancient and veteran trees by a total of 285 records, more than a 100% increase on the number of trees recorded in the ATI in these locations before the surveys. From these surveys alone, there are clear large gaps in our knowledge of the current distributions of these trees, suggesting that many of them may remain unaccounted for in current strategies for protection, ecological monitoring and management. This is true despite the fact that in the UK such trees are much better recorded at the level of the individual than they are in most other parts of the world.

The total number of ancient and veteran trees across England predicted by all the models based on the field estimates of abundance also emphasises the very high number of trees that are still unrecorded. Based purely on the raw abundance of trees recorded during the surveys, estimates totalled around one million trees, more than five times the number of ancient and other noteworthy trees currently in the ATI. However, when estimates of sampling effort for each square were factored in, to account for the parts of each square that were inaccessible in the field survey, the estimated total based on tree density

is around two million trees for several models. Although this is welcome news, as it shows that much more dead and decaying wood habitat is available across the country than was previously known, it is also worrying that so many valuable trees are unrecognised or recorded as ancient or veteran and may lack conservation measures or protection. This is the first study to provide quantitative nationwide estimates of the true number of ancient and veteran trees; my previous research in Chapter 3 was focused purely on wood-pastures in England, which cover an area of ~2,780 km^2: it predicted around 100,000 such trees just in this habitat (see Chapter 3). Other estimates have guessed figures close to nine million ancient or veteran trees across the whole UK (Fay, 2004), so our estimates do not seem wildly inflated. Nevertheless, my results suggest that there is much work to do to find these trees and add them to the ATI.

Field validation with independent, unbiased sampling is the gold standard when evaluating the performance of distribution models and predictive maps, and yet it is rarely used (Getz et al., 2018). Instead, model performance is typically assessed using methods of internal validation: often retaining a portion of the data to test the models, or using a cross-validation approach, are considered sufficient to validate the models and make accurate predictions (Fielding & Bell, 1997) with AUC the most common evaluation statistic used for this. However, measuring model accuracy using AUC and cross-validation has been criticized, because it is likely to inflate perceptions of model performance owing to spatial autocorrelation in the species data (Lobo et al., 2008; Peterson et al., 2008). Additionally, any data retained to test the model from a biased species dataset will suffer the same bias as the data used to fit the model, thereby giving false confidence that significant predictors of species occurrence are predicting the underlying ecology, when they are actually predictors of sampling effort. Therefore, in order to evaluate models fully, and to assess the utility of different sampling bias correction methods, it is important to use unbiased field data where possible. In this study, field validation provided support for the need for bias correction when modelling ancient and veteran tree occurrences, and was able to increase my confidence in the model predictions. As a result, my maps can be relied upon to be biological informative and also robust against the obvious sampling bias in the ATI, something which relatively few studies can attest to. Alongside fine-tuning modelling procedures and understanding

ecological systems, the feasibility of collecting additional data for model validation should always be an important consideration of any ecological study.

Spatial filtering, especially the systematic sampling technique, proved to be one of the most effective bias correction methods overall based on both internal validation using AICc and field validation. This method is known to be particularly useful for wide ranging, heavily sampled species and has been shown to reduce both type I and type II errors (Kramer-Schadt et al., 2013). However, it is often limited by sample size, because reducing the number of occurrence records can result in poor model predictions. Furthermore, the best choice of spatial filter may differ depending on environment; for example Boria et al. (2014) suggest that mountain regions need smaller spatial filters than other areas. There is also the risk of reducing clustering in areas which truly represent high ecological value for a species (Fourcade et al., 2014). Nevertheless, the large number of records in the ATI, as well as the large range of the trees across the UK, means spatial filtering is likely to be highly effective for this database. A similar study using spatial filtering with large species databases also reported good results when comparing to independent field data (Law et al., 2017), and concluded that their models were suitable to be applied to practical management scenarios. I believe that these similarly high-performing, independently validated models are also suitable for management applications, and could provide valuable insight into the areas most suitable for immediate practical ancient and veteran tree conservation measures.

It is notable that field validation often ranked models differently compared to internal model validation; based purely on internal model evaluation, I would have inferred that the best bias correction method was the cluster analysis spatial filtering technique, followed by the ZI models, both of which performed less well when evaluated against the field data using AUC or Spearman rank correlations. The performance of the bias files also differed greatly between internal and field validation, although wood-pasture habitat performed well using both methods. Wood-pastures have strong connections to ancient and veteran trees, and are the most studied of the habitats in which these organisms are found (Rackham, 1994; Farjon, 2017; Hartel et al., 2018). Additionally, many wood-pastures in the UK form part or the whole of a site of interest from a tourism or aesthetic point of view, for example National Trust sites or

public parkland (Rackham, 1994; Lonsdale, 2013). Therefore, it is no surprise that wood-pasture spatial distributions have strong influences on recorded ancient and veteran tree distributions, via effects on both ecology and sampling effort: both the count prediction of abundance and the probability of a grid square being sampled from the ZI models were predicted to be higher in grid squares with greater coverage of wood-pasture. In the bias corrected MaxEnt models, wood-pasture importance as a predictor did decrease significantly compared to the uncorrected model, suggesting it has a large influence on sampling bias in the ATI, yet it still remained the most important predictor overall. This explains why in all the predicted distribution maps, even when sampling bias was corrected for, there are many grid squares containing wood-pastures that have very high suitability for ancient and veteran trees.

Background manipulation methods also performed differently between internal model and field validation. They were relatively good at predicting raw tree abundance found during the field surveys, especially in squares with high numbers of trees, but not so good at predicting tree density (accounting for survey effort estimates), or producing models that fitted well to the original data. Although there has been some success with this method in other studies (Phillips et al., 2009), it has previously been considered to perform worse than other methods (Fourcade et al., 2014), possibly because background points were restricted to too narrow an area, reducing model accuracy (Thuiller et al., 2004). Understanding the optimum background area size, and considering both the species range and the extent of sample bias, are likely to be case specific and should be considered before using this method for bias correction.

The performance of ZI models varied the most across validation methods; internal model evaluation showed that ZI models provided a very good fit to the raw data, but low predictive power, whereas field validation suggested that the models are very suited to predicting raw abundance or density, especially of outlier observations where abundance was high, but poor at predicting presence-absence, and ranked abundance and density. Nevertheless, one benefit of the use of ZI models in comparison to all the other methods is that it is the only one to provide some independent insight into potential causes of bias in

the original data by examining potential causes of excess zeroes (see Chapter 5). Many predictors in my study had some influence on the proportion of excess zeroes in the ATI data, the majority of which are likely to influence both the ecology of the trees and the likelihood of them being sampled, including altitude, type of land or soil, distance to roads and watercourses, historic land use and cover of forests, woods and wood-pasture. Nevertheless, it is likely that predictors which also influenced the count component of the ZI model, for example altitude, have more influence on the ecology, whereas those influencing only the zero component, such as distance from a watercourse, are more likely reliable indicators of sampling effort. The high number of predictors potentially influencing both the tree ecology and sampling processes is likely the reason why the whole model predictions were better overall than the count predictions: a proportion of the excess zeroes in the ZI zero component are probably biological zeroes, rather than being caused by undersampling. Removing the influences of these processes from the model predictions (which is what the count abundance predictions do) would therefore remove meaningful biological information from the overall prediction maps.

A major benefit of the use of zero-inflated models is that they can be used to generate distribution maps of the predicted excess zeros, providing insight into areas which may have been under or oversampled, and thereby helping those planning future sampling and conservation efforts. In my study, Cornwall and Devon counties were, for example, predicted to have high numbers of excess zeroes and are therefore good candidates for extra targeted surveys. Although ZI models have been used to fit distribution models before (Bouyer et al., 2015; Lyashevska et al., 2016), this is the first time that they have been successfully applied to identify causes of, and to correct for, sampling bias, and my results highlight their potential advantages over other more conventional methods of sampling bias correction. I believe ZI models have strong potential in the fields of ecological modelling and practical conservation.

6.6 Conclusion

My results first and foremost provide a robust prediction of ancient and veteran tree distributions across England which can be used for conservation planning and decision making. Until now, there has been

no real measure of the landscape-scale value of this habitat and how it interconnects. My work shows the overall collective value of this irreplaceable natural resource and should frame the debate for further serious discussion about what level of effort will be required to map, monitor and manage ancient and veteran trees in the future. In addition, despite the difficulties presented by a global pandemic, my study demonstrates how citizen scientists can be mobilised to conduct independent field validation of models built from large publicly-accessible databases, increasing confidence in, and the utility of, model predictions. My results also underline the impact of sampling bias in citizen-derived datasets on the effectiveness of ecological models in conservation. Correcting for sampling bias is essential for preventing incorrect inferences from distribution models influencing practical conservation decisions.

Chapter 7: Assessing the use of landscape metrics in Species Distribution Modelling: a case study using the UK Ancient Tree Inventory (ATI).

7.1 Abstract

Understanding how landscape structure and composition influence species distributions and biodiversity is key to conservation and land management, especially as human pressure on the landscape increases. Landscape metrics mathematically quantify aspects of the landscape structure, and their use within Species Distribution Modelling (SDM) can improve model performance and predictions by adding relevant fine-scale ecological information about the species or community in question. In this chapter, I quantify the landscape across England using a unique, large-scale dataset of all tree canopies, the National Tree Map (NTM). By calculating 16 landscape metrics that describe properties of the canopies (size, shape etc.) and their connectivity across the landscape, the landscape structure of England was defined. Maximum Entropy (MaxEnt) models of ancient and veteran tree distributions using different subsets of the landscape metrics as predictors were compared to those fitted using only environmental predictors (see Chapter 6), as well as models fitted using combinations of the two. Models were evaluated using the independent field data from Chapter 6 and total estimates of ancient and veteran tree numbers across England calculated. Quantifying the landscape structure based on the NTM revealed key fine-scale insights into types of landscapes more likely to have ancient and veteran trees including those with a large number of scattered and irregular tree canopies. However, landscape metrics did not improve SDM performance or predictions of ancient and veteran tree distributions, probably because of the coarser scale of fitting the distribution models compared to the fine-scale information captured by the metrics. Predictions of the total number of ancient and veteran trees across England were similar to those of Chapter 6, and again suggest there are around 2 million ancient and veteran trees nationwide, reinforcing the urgent need to find and record these valuable trees for their conservation and protection.

7.2 Introduction

Landscape ecology relates to the spatial interaction of organisms and ecological processes with landscape patterns and structures (Turner, 1989; Haines-Young & Chopping, 1996; Kupfer, 2012). Our landscape has experienced intensive disruption from land use change, fragmentation, pollution and urbanisation, which have impacted many species at all spatial scales (Cozzi et al., 2008; Boyd et al., 2008; Powers & Jetz, 2019). Fragmentation of habitats is of particular concern for conservation, and the reduction of both habitat quality and connectivity has had great influence on species decline and endangerment (Fahrig, 2003; Saltre et al., 2015). Conservation and land management relies on optimising the landscape to benefit and enhance the ecological processes of the target species or ecosystem, so a good understanding of landscape structure, including the composition and configuration of different land types, is essential for protecting biodiversity and ecosystem functions (Pino et al., 2000; Cozzi et al., 2008; Banks-Leite et al., 2011).

The term 'landscape' is highly influenced by human perception, and is often defined in terms of the scale and nature of interactions between humans and the environment (Troll, 1968; Wiens & Milne, 1989). Humans have been highly influential in shaping the past and present landscape structure, with human-related processes such as fire suppression, settlement, land-use change and urbanisation all contributing to changes in landscape structure and consequently shaping local ecology and biodiversity (Baker, 1992; Aubad et al., 2010; Threlfall et al., 2012). The decline of ancient and other trees with veteran characteristics around the world is linked to strong anthropogenic pressure on the landscape from urbanisation, agricultural expansion and development (Laurance et al., 2000; ATF, 2005; 2011; Lonsdale, 2013). Furthermore, it has been suggested that the *original* determinants of the distribution of these trees might also be more linked to human activities in the landscape than natural ecological processes (Rackham, 1994; Barnes et al., 2017). Historical tree management practices such as coppicing or pollarding have increased the longevity of many trees, so spatial geographic variation in these techniques can partially explain the current distribution of ancient and veteran trees (Read, 2000; Barnes et al., 2017). Trees that were deliberately planted by humans as parts of avenues, boundaries, landscaped gardens and in other prominent positions make up a significant proportion of the current known ancient

and veteran tree records (Lonsdale, 2013, Farjon, 2017; Nolan et al., 2020). Analysis of the landscape structure (heavily linked to anthropogenic processes) in relation to ancient and veteran trees could provide insight into areas more likely to contain them and the potential true distribution of ancient and veteran trees across the landscape.

Quantification of landscape structure often involves the use of landscape metrics, mathematical descriptions of aspects of the landscape at different scales and complexities (Li & Wu, 2004). Metrics can be calculated for the whole landscape, for a series of land or habitat classes or for each habitat patch (McGarigal, 1995). The development of metrics to measure landscape ecology accelerated during the late 1980s, and since then, hundreds of metrics have been used to describe all aspects of landscape structure. Many of these are based on O'Neil (1988), who proposed three general measures of landscape structure: dominance (a measure of diversity), contagion (habitat aggregation) and shape (habitat shape complexity). Additions to these measures include patch size, perimeter, patch type proportion, patch perimeter fractal dimension, simple edge contrast and patch type adjacency (Turner, 1989), as well as proximity, patch elongation, linearity, core area and edge area (Gustafson and Parker, 1992), although many of these are thought to be highly correlated and redundant (Baskent and Jordan, 1995; Haines-Young & Chopping, 1996). Landscape metrics have been widely used in ecological research, for example to predict the spread of invasive species (Lustig et al., 2017), to investigate land use change and fragmentation within a Mediterranean ecosystem (Lamine et al., 2018) and to plan conservation of lowland English forests under fragmentation (Baalman & Kirby, 1995).

Describing the landscape structure also requires the identification of particular types of land or habitat that can be used for quantitative calculation of landscape metrics. These could include urban areas (Connors et al., 2013), agricultural land classes (Griffith et al., 2000) or waterbodies (Connors et al., 2013; Yuan et al., 2014), but are most commonly related to vegetation, for example forest patches (Aubad et al., 2010) or specific plant types (Cristofoli et al., 2010). The presence of trees, regardless of their age, within a landscape has a dramatic influence both on abiotic factors defining the landscape including soil erosion, flooding, temperature, rainfall and soil characteristics (Vailshery et al., 2013;

Pardon et al., 2017; Song et al., 2019; Turner-Skoff & Cavender, 2019), and on individual organisms, populations and ecosystems via the provision of habitats, food sources and connectivity (Rasey, 2004; Manning et al., 2006; Rossi et al., 2016). The distribution of trees is also highly linked to the distribution of humans (Rackham, 1994; Polyakov et al., 2008; Cloke and Jones, 2020); trees are integral in many parts of our landscape and play a variety of roles in urban, sub-urban, agricultural and rural landscapes (Barnes et al., 2017; Cloke and Jones, 2020). Different geographic locations have significantly different compositions and configuration of trees (Barnes et al., 2017), and although they represent only one type of habitat, they could be ideal feature to use for landscape structure analysis.

England has an innovative resource to assist with the quantification of landscape structure using trees across a large, national scale: the National Tree Map™ (NTM) (also called the National Canopy Map) produced by the mapping company, Bluesky International Limited. The NTM is a digitised vector map of all canopy higher than 3 m across England and Wales, constructed from stereo aerial photography and digital elevation models. All trees are represented as single polygons that show location, height and canopy extent (for more information see Chapter 3). The NTM is an accurate and useful tool that has been used to model urban vegetation (Casalegno et al., 2017), human health in response to allergenic pollen (McInnes et al., 2017), mental health in relation to urban greeness (Sarkar et al., 2018) and carbon dioxide emissions in central London (Björkegren & Grimmond, 2018). Using the NTM, of which ancient and veteran trees are a small subset, habitat patches can be defined with a high level of accuracy for landscape quantification. Theoretically the NTM should be able to highlight detailed variation in landscape structure to predict the true distribution of ancient and veteran trees in England.

The configuration and spatial connectedness of trees in the NTM, as well as the shape and size of each canopy polygon, could provide a useful method to assess landscape structure: multiple connected canopies are most likely representative of woodland areas or plantation, whereas linear canopy rows could represent hedgerows. As an example, Fig. 7.1 portrays two different landscapes based purely on the canopies from the NTM, shown using two selected 1-km^2 subsets of the NTM from the county of Suffolk in England. Prior knowledge and findings from this thesis and other literature suggest that

ancient and veteran trees are more likely to be found free-standing and in open landscapes, such as wood-pasture, and in places with less anthropogenic pressure (Rackham, 1994; Butler et al., 2002; Farjon, 2017). These habitat characteristics are seemingly represented more in the square on the left in Fig. 7.1. Hence, the NTM seems like an ideal resource with which to quantify the landscape structure across England and use this to predict areas most likely to harbour ancient and veteran trees.

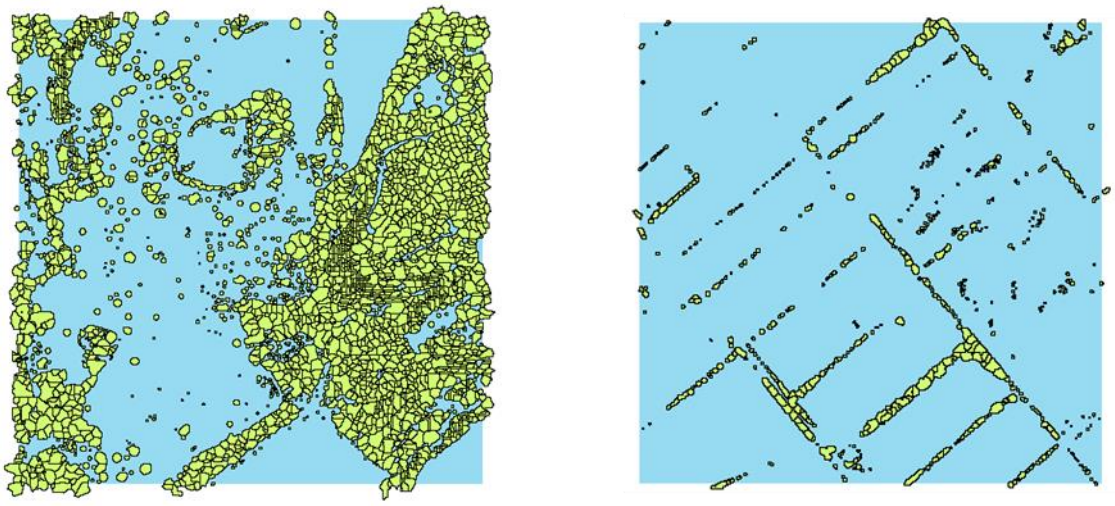


Fig. 7.1 *Two 1-km^2 grid squares from Suffolk, England with the overlaid National Tree Map canopies. Each polygon represents an individual canopy. Two very different landscapes are shown with different structures: on the left, the landscape has more canopy overall, including both areas with aggregated and dispersed trees, and less obvious influence of human activities; on the right the landscape appears to be more heavily influenced by human design, with linear rows of trees, as is typical of an agricultural or urban landscape.*

In this Chapter I aim to quantify landscape structure across the whole of England using the NTM as the selected habitat type, calculating a variety of landscape metrics at a 1-km^2 scale. I then use the resulting metrics as predictors of ancient and veteran tree distributions and abundance in Species Distribution Modelling (SDM). Many other studies have shown that landscape metrics improve SDM performance (Hopkins, 2009; Foltête et al., 2012; Hasui et al., 2017; Ortner and Wallentin, 2020) because they add fine-scale information which can indirectly predict ecological processes determining the species distribution. A key benefit of using the NTM is that the compilation of the canopy data is unaffected by

sampling bias, and the canopies are recorded with a higher level of geographical accuracy than the ATI records. I compare the distribution models fitted using the landscape metric predictors to models fitted only using environmental predictors (see Chapter 6), as well as models incorporating both types of predictors. As with Chapter 6, models were validated using independently collected field-survey data.

7.3 Methods

7.3.1 Study species and landscape metrics

The modelling processes in this chapter are based on the same 1-km grid and ATI ancient and veteran tree records across England first introduced in Chapter 4 (see Methods, Chapter 4) and also used in Chapter 6. Both the occurrence locations and abundance of ancient and veteran tree records per 1-km grid cell were used. All canopy polygons that intersected the 1-km study area grid across England (see Chapter 4) were obtained from the NTM (Bluesky, 2015, c/o the Woodland Trust). For each individual 1-km grid square, all canopy polygons that intersected that square were selected and converted to a single 1-km by 1-km raster layer with pixels at a 1-m resolution. Each canopy was considered to be an individual unique patch, so all pixels within each canopy polygon were allocated the same raster value ID number, unique for each canopy within the grid square. A total of 16 landscape metrics were calculated for each grid square based on the canopy raster layer using the R package 'landscapemetrics' (Hesselbarth et al., 2019) (Table 7.1). Seven metrics described characteristics of each individual canopy patch; the mean value for each of these was calculated across all canopies per grid square. Five metrics described the configuration or characteristics of all canopy patches per grid square, and four metrics related to the minimum or maximum characteristic of all canopies per grid square. Due to the extremely high number of canopies across England, High Processing Computing (HPC) was used to speed up computation, whereby all the metrics for 400 grid squares could be computed simultaneously through the use of a 400-core processor.

Table 7.1 *The original 16 landscape metrics calculated for each grid square in England based on the National Tree Map. Metrics highlighted in green indicate the final seven metrics remaining following collinearity reduction analysis using Variance Inflation Factor and Pearson's correlation coefficient.*

Metric	Abbr.	Description	Possible range	
			Min	Max
Mean canopy area	area_mn	Mean area (m^2) of all canopies per grid square.	0 (no canopies)	1-km^2
Number of canopies	no_canopies	Number of individual tree canopies per grid square.	0 (no canopies)	Unlimited
Total edge	totaledge	Sum of the lengths (m) of all tree canopy perimeters.	0 (no canopies)	Unlimited
Mean radius of gyration	gyrate_mn	Mean distance (m) between each pixel within a canopy and the canopy centroid averaged across all tree canopies per grid square.	0 (canopy is a single pixel)	Canopy covers entire grid cell
Mean related circumscribing circle	circle_mn	Mean related circumscribing circle of all canopies within a grid cell. The metric is a measure of patch elongation and is a ratio of canopy area to the area of the smallest circle that can possible surround and completely contain each canopy	0 (circular patches)	1 (elongated, linear patches)
Mean contiguity Index	contig_mn	Mean contiguity value of cells within a canopy, averaged across all canopies per grid square. Contiguity is a measure of spatial connectedness of all pixels within a canopy.	0 (canopy is single pixel)	1 (total patch contiguity)
Mean shape index	shape_mn	Mean shape index (patch perimeter/ minimum perimeter possible i.e. a square patch) of all canopies within a grid square. A measure of canopy irregularity.	1 (patch is square)	Unlimited
Mean perimeter area ratio	para_mn	Mean ratio of patch perimeter (m) to area (m^2) of all canopies within a grid square. Measure of canopy complexity.	0 (no canopies)	Unlimited
Mean fractal dimension index	frac_mn	Mean canopy complexity of all canopies per grid square. Calculated as natural log of canopy perimeter (m)/ natural log of canopy area (m^2).	1 (simple perimeter e.g. square)	2 (complex perimeter)
Landscape division index	division	Calculated as the probability that two randomly chosen pixels within a grid square are not in the same canopy.	0 (single canopy)	1 (each pixel is separate canopy)
Canopy cohesion index	cohesion	Measure of physical connectedness of the canopies within a grid square.	Not yet clarified as part of the R package.	
Splitting index	split	Measure of the number of canopies with a constant size when the landscape is divided into n patches (n = splitting index).	1 (single canopy)	No. pixels in landscape squared
Max. canopy area	max_area	Maximum area (m^2) of the largest canopy per grid square	0 (no canopies)	1-km^2
Mean canopy perimeter	mean_perim	Mean perimeter (m) of all canopies per grid square	0 (no canopies)	Unlimited
Max. canopy perimeter	max_perim	Largest perimeter (m) of a canopy per grid square	0 (no canopies)	Unlimited
Min. canopy perimeter	min_perim	Smallest perimeter (m) of a canopy per grid square	0 (no canopies)	Unlimited

7.3.2 Metric analysis and reduction

Many of the metrics were highly collinear (Pearson's correlation coefficients > 0.9), so two alternative methods of metric selection were compared. The first used both the Pearson correlation coefficients of the raw landscape metrics (with a threshold of 0.9) and the Variance Inflation Factor (VIF) (using a threshold of 10). VIF was calculated independently for each variable and provides a measure of collinearity between that predictor and all others in a set by estimating potential increases in the variance of the regression coefficients. A VIF value of 1 indicates no correlation with other predictors, and increases as collinearity becomes more severe, with values above 10 thought to be highly collinear (Franke, 2010). This process resulted in the retention of seven of the original 16 metrics (Table 7.1).

The second method involved carrying out Principal Component Analysis (PCA) on the 16 metrics. PCA is a method of reducing the dimensionality of a large dataset by combining variables into a set of new uncorrelated Principal Components (PCs) that capture as much of the information contained in the initial variables as possible. PCA was carried out on all 16 scaled landscape metrics using the 'prcomp' function in the 'stats' package in R (R Core Team, 2018). The first two PCs explained a combined total of 70% of the variance in the original dataset, and were retained. Eigenvalues of PC 3 and 4 were 1.17 and 1.03 respectively, falling to the right of the elbow on an elbow plot (Fig. 7.2), and all subsequent PC eigenvalues (5 onwards) fell below 1.

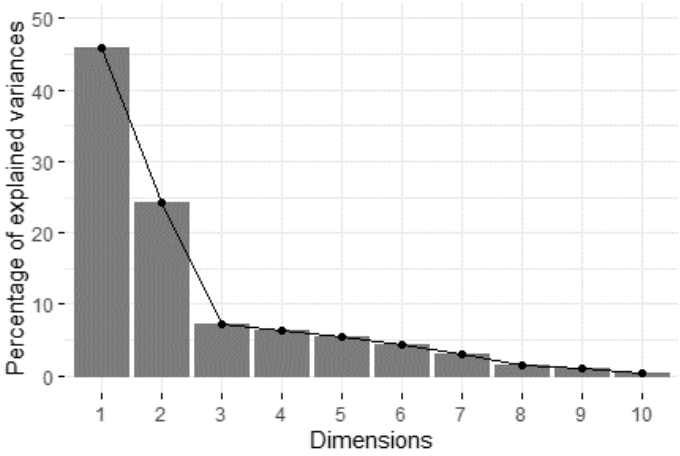


Fig. 7.2 *Percentage of explained variance of each principal component dimension shown from the Principal Component Analysis (PCA) of all 16 landscape metrics.*

7.3.3 MaxEnt Species Distribution Modelling

Raster layer predictors across the whole of England at a 1-km resolution were created from the values of a) each of the seven landscape metrics selected after collinearity analysis, and b) the two retained PCs. Maximum Entropy (MaxEnt) models were then fitted to the ancient and veteran tree occurrence records in relation to the two sets of predictors ('metrics' or 'PCs') at a 1-km resolution using 'ENMeval' package in R (Muscarella et al., 2014). As with the models in Chapter 6, models were fitted using 10,000 pseudo-absences background points randomly sampled across the study area, with all other MaxEnt parameters remaining at their default values (Phillips & Dudík, 2008). Model tuning was carried out based on the methods described in Appendix A6.1, with the best fitting models selected based on the corrected Akaike information criterion (AICc) (see Appendix A6.1). As in Chapter 6, model predictions were created and internally evaluated with 10-fold cross-validation (CV), using AICc and 'Area Under the Curve' (AUC) for the training and test data.

The landscape-metric models were additionally compared to the original distribution models of ancient and veteran tree occurrence fitted in Chapter 6 (the original model with no bias correction) using all of the environmental predictors (subsequently named the 'environment' model in this chapter) (see Table 6.1) to assess whether the use of landscape metrics in SDM provides a more accurate prediction of ancient and veteran tree presence across England. As a final step, two additional MaxEnt models were fitted following the same method as described, using the environmental predictors (Table 6.1) in combination with either the seven landscape metrics or the two PCs as predictors. Again models were evaluated internally using both AICc and training and test AUC.

7.3.4 Field surveys and model verification

The estimates of ancient and veteran tree presence-absence, raw abundance and tree density for each of the 52 1-km grid squares obtained from the independent field surveys as described in Chapter 6 were used to verify the model predictions (see Methods, Chapter 6 for more information). Presence-absence field verification estimates were analysed using AUC in relation to each of the model predictions of habitat suitability, and the raw abundance and tree density field estimates were analysed using Pearson

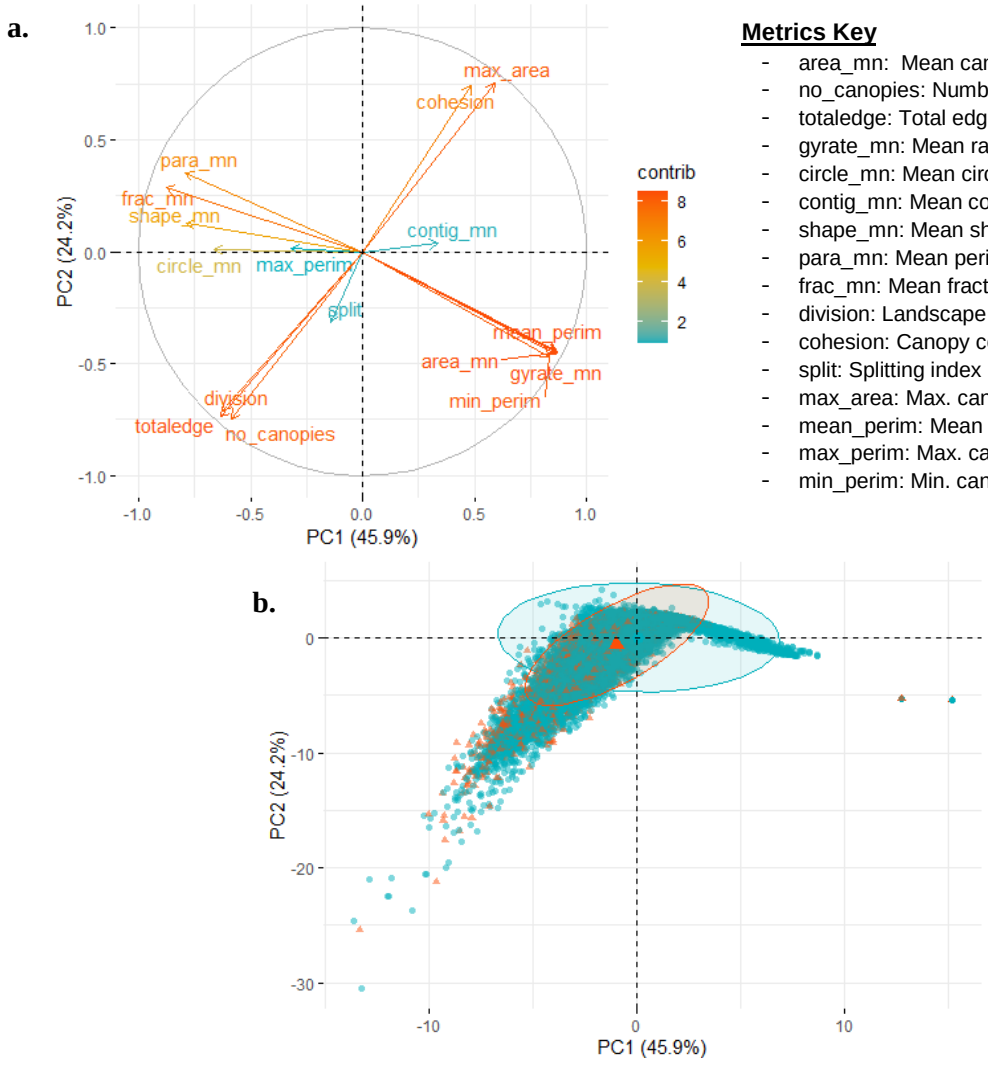
and Spearman correlation coefficients. As in Chapter 6, linear regression models of each distribution model prediction against field verification abundance estimates (either raw abundance or tree density) for the 52 surveyed squares were used to calibrate the models and produce predictions of the total number of ancient and veteran trees across England.

7.4 Results

7.4.1 Landscape metric analysis

Plots of each individual data point (grid square) grouped by whether each grid square contains ancient or veteran tree records or not (presence-absence) on the axes of PC1 and PC2 reveal distinct differences in the landscape metrics between the two groups (Fig. 7.3). Squares with records show strong associations with the total number of canopies, division index and total canopy edge (Fig. 7.3). This suggests that ancient and veteran tree presence is more likely in squares with a higher number of canopies that are more scattered (highly divided) and will therefore have a greater length of canopy perimeter within the square.

Plots of the abundance of ancient and veteran trees per 1-km grid cell against each of the seven retained original landscape metrics suggest there are peaks in abundance when mean contiguity index is closest to 1, number of canopies is less than 5000, mean related circumscribing circle is around 0.45-5, splitting index is closest to 1, mean shape index is around 1.2 and maximum perimeter of a canopy is around 20-m (Fig. 7.4). However, in contrast with the landscape-metrics PCA analysis, this suggests that abundance is highest when the canopies are highly connected, as well as when the canopies themselves are of an intermediate shape between circular and elongated, somewhat irregular and are relatively larger than other canopies (> 20 m).



Metrics Key

- area_mn: Mean canopy area
- no_canopies: Number of canopies
- totaledge: Total edge
- gyrate_mn: Mean radius of gyration
- circle_mn: Mean circumscribing circle
- contig_mn: Mean contiguity Index
- shape_mn: Mean shape index
- para_mn: Mean perimeter area ratio
- frac_mn: Mean fractal dimension index
- division: Landscape division index
- cohesion: Canopy cohesion index
- split: Splitting index
- max_area: Max. canopy area
- mean_perim: Mean canopy perimeter
- max_perim: Max. canopy perimeter
- min_perim: Min. canopy perimeter

Fig. 7.3a *Direction and contribution of each landscape metric to Principal Components (PCs) 1 and 2 (colours are shown to aid visualisation of metric contributions: red = high contribution, blue = low contribution) from the Principal Component Analysis (PCA) of all 16 original landscape metrics.* ***7.3b*** *Plot of each individual grid square coloured by ancient and veteran tree presence (present = orange, absent = blue) on the axes of PC1 and PC2 from the PCA of all 16 original landscape metrics. Two ellipses are shown in relation to grid squares with either presence or absence of ancient and veteran trees, and represent a multivariate normal distribution with a concentration level of 0.95. Each ellipse is centred on the means of the two types of grid square (presence or absence) and oriented in the direction of the first eigenvector of the covariance matrix.*

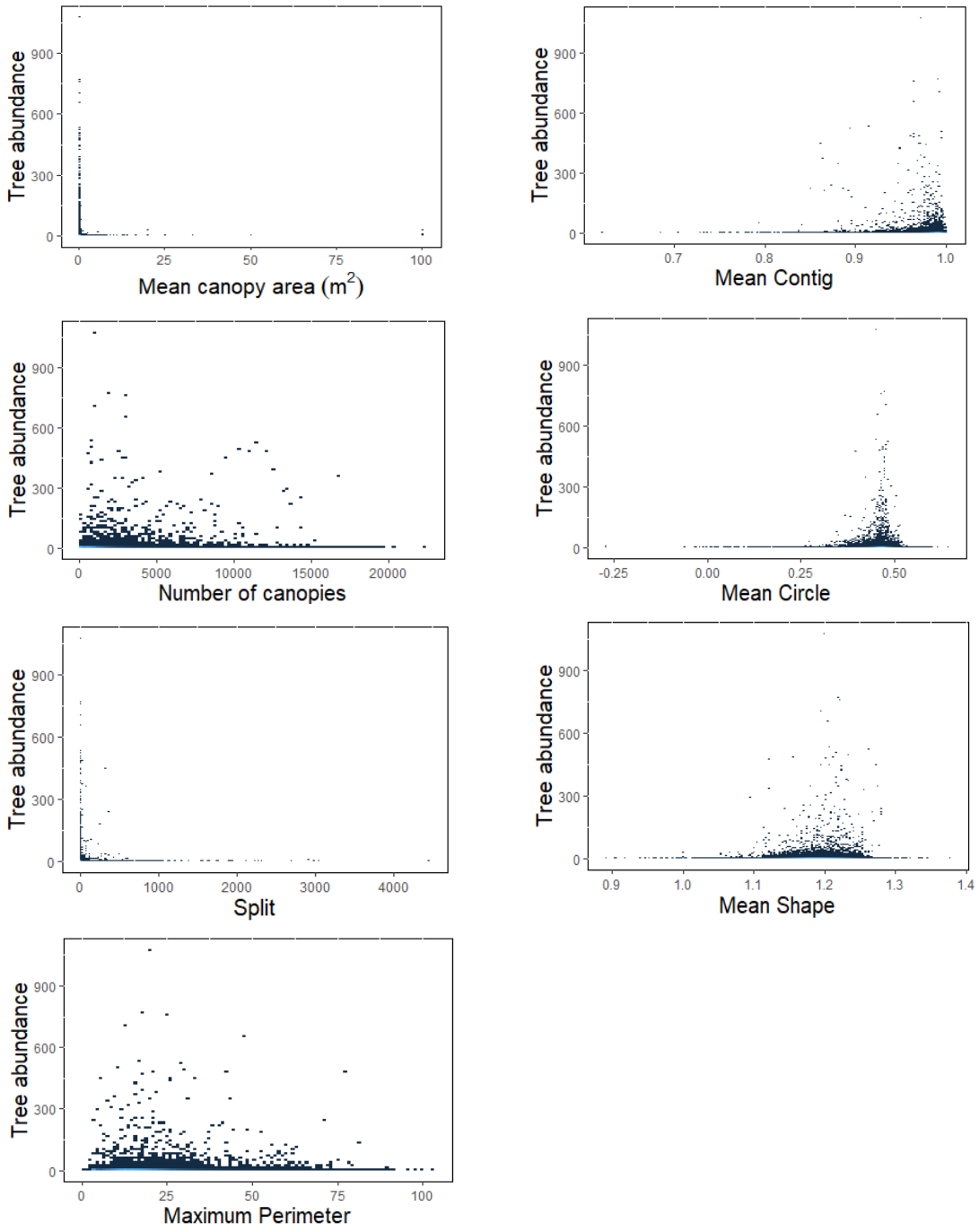


Fig. 7.4 *Scatterplots of ancient and veteran tree abundance in relation to the seven final landscape metrics.*

7.4.2 Internal validation of the distribution models

The use of all seven landscape metrics in the distribution models resulted in slightly higher performing models that produced better predictions based on both AIC and AUC respectively than when the landscape metric PCs were used, although the difference was relatively small (Fig. 7.5). Nevertheless, the best performing models in all cases involved the use of the environmental predictors, with the Environment model actually performing better than when used in combination with the landscape metrics (Fig. 7.5), although again differences between models were relatively small.

Distribution maps from each of the models show a wider range in habitat suitability for ancient and veteran trees across England when using the environment predictors. These models (Fig. 7.6 a, d and e) show high suitability in small sites, which are generally identified as parks, wood-pastures or forests (see Chapter 6 for more detail), scattered across the country. Both landscape metric distribution maps show less variability in habitat suitability, with the most suitable areas identified to the South West of London, centred on the South Downs National Park, Sherwood Forest in the East Midlands, the North York Moors National Park and Kielder Forest Park in the North of England (Fig. 7.6 b and c).

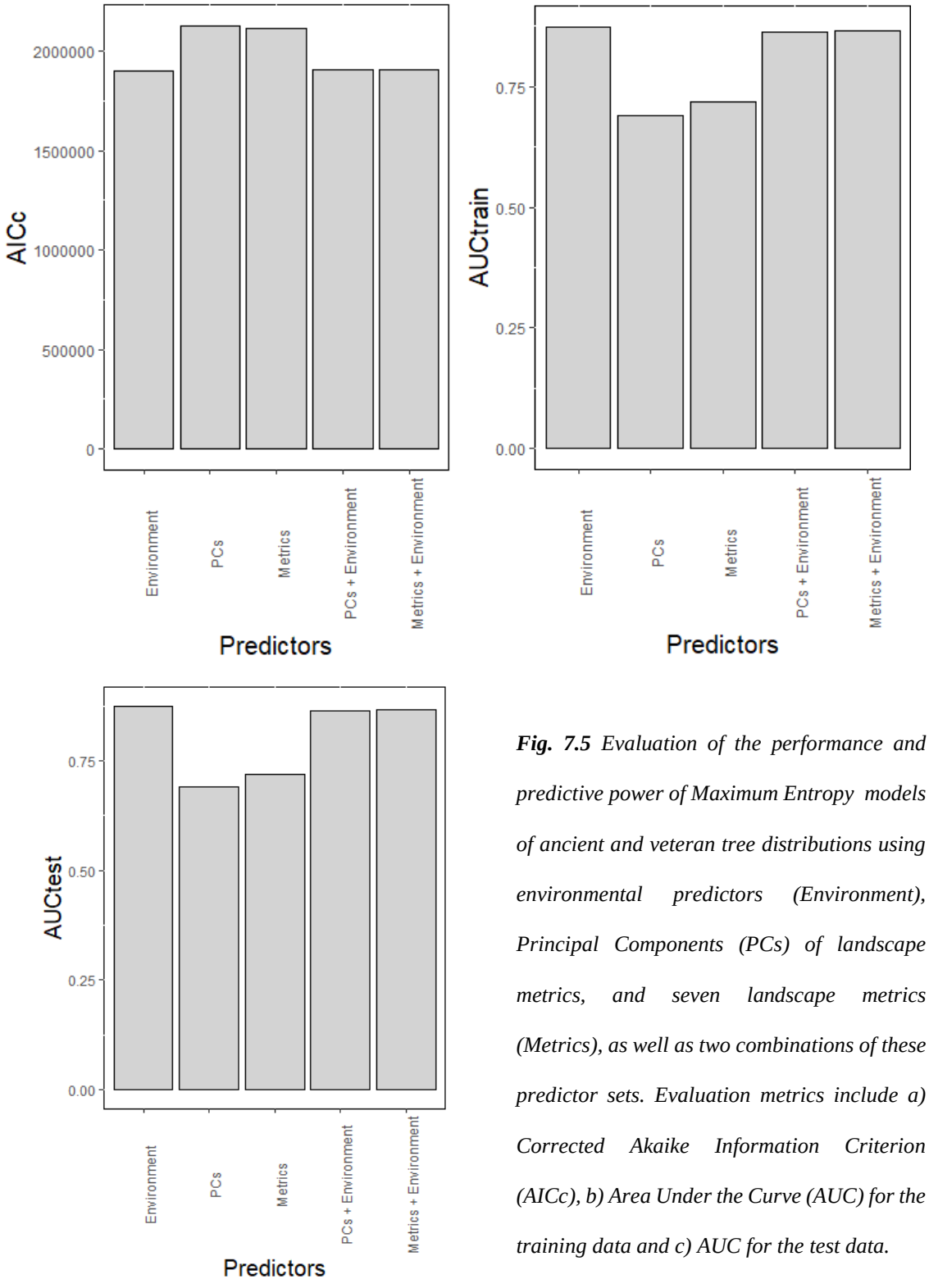


Fig. 7.5 *Evaluation of the performance and predictive power of Maximum Entropy models of ancient and veteran tree distributions using environmental predictors (Environment), Principal Components (PCs) of landscape metrics, and seven landscape metrics (Metrics), as well as two combinations of these predictor sets. Evaluation metrics include a) Corrected Akaike Information Criterion (AICc), b) Area Under the Curve (AUC) for the training data and c) AUC for the test data.*

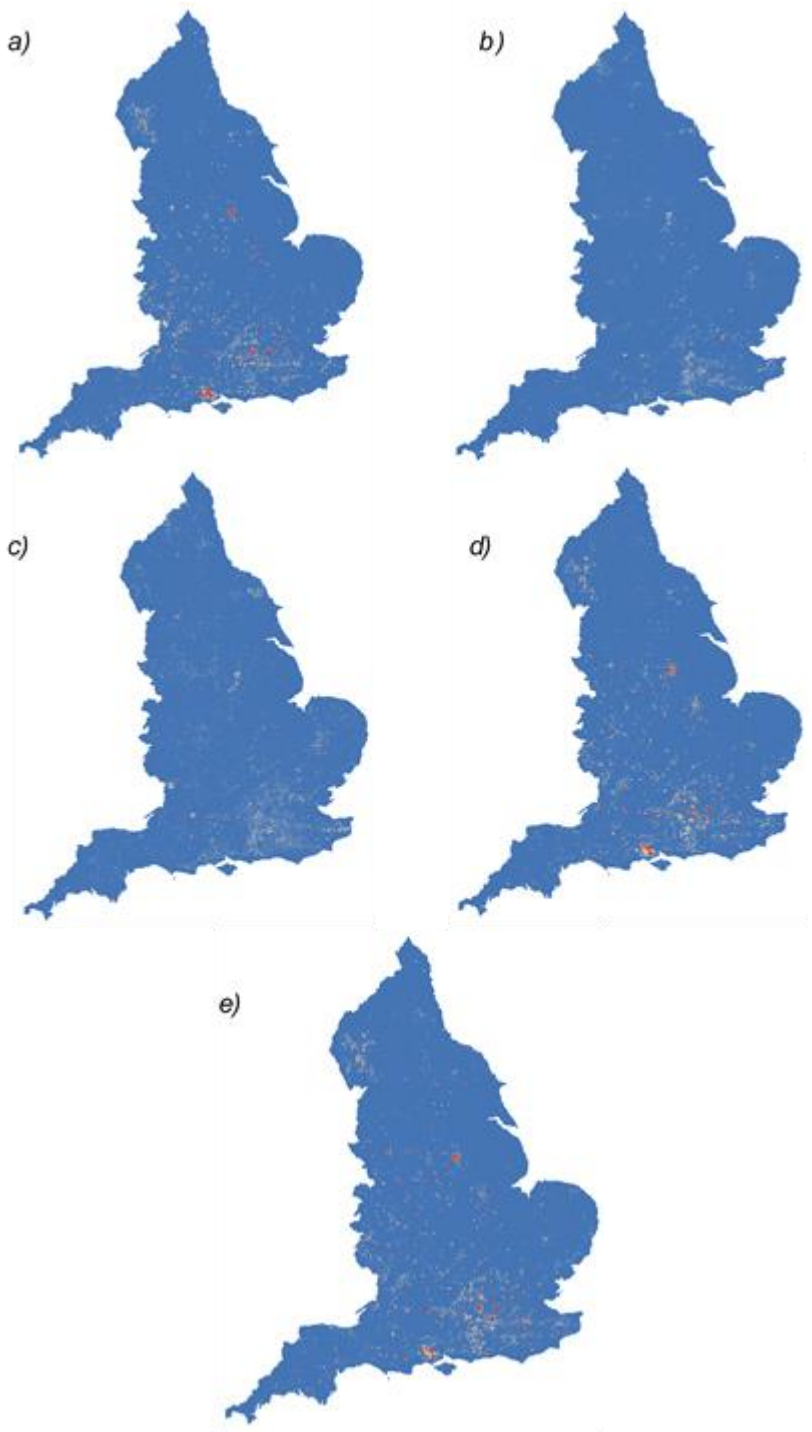


Fig. 7.6 *Distribution maps of ancient and veteran trees across England produced from Maximum Entropy models using five combinations of predictors: a) environmental predictors, b) seven original landscape metrics, c) Principal Components (PCs) of the landscape metrics, d) environmental + landscape metrics and e) environmental + landscape-metric PCs. Colour scales range from low habitat suitability of 0 (blue) to the highest habitat suitability of 0.164 (red).*

7.4.3 Field validation of the distribution models

Field validation supported the internal model evaluation: the model using only environmental predictors had the highest predictive power when compared to the field data (Table 7.2) and outperformed all other models when predicting presence-absence, abundance and density of ancient and veteran trees. The only exception was when assessing tree density predictions (accounting for variation in survey effort per grid square) using Pearson raw correlations: the model using the seven landscape metrics as predictors outperformed the model using environmental predictors. The models combining the two sets of predictors still performed worse than either individually (Table 7.2). Estimates of total ancient and veteran tree abundance across England using the field verification were relatively similar to those predicted by the models in Chapter 6, but they varied more, ranging from 911,842 to 1,456,555 based on raw abundance, and 1,867,480 to 2,489,774 based on tree density (Table 7.2).

Table 7.2 *Independent field evaluation of model predictions. Model predictions were evaluated against a) field verification estimates of the presence-absence (P-A) of ancient and veteran trees per square using Area Under the Curve (AUC), b) field estimates of raw tree abundance (total number of trees recorded per square) using Pearson (r) and Spearman (r_s) correlation coefficient tests and c) field estimates of tree density (number of trees in relation to estimated percentage survey effort for each square) also using Pearson and Spearman correlation coefficient tests. Values in bold represent those that are significant. Where indicated, significance levels are: $p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: **. For each model the total predicted abundance of ancient and veteran trees (T) across England was calculated from a linear regression model between the model predictions and field verification data (both raw tree abundance and tree density) for the 52 surveyed squares.*

	P-A	Tree abundance					Tree density				
Model	**AUC**	**r**		**r_s**		**T**	**r**		**r_s**		**T**
Environment	0.613	**0.589**	***	**0.339**	*	911,842	**0.490**	***	**0.410**	**	1,867,480
Metrics	0.541	**0.516**	***	0.241		1,456,555	**0.569**	***	**0.329**	*	2,489,774
PCs	0.581	-0.014		0.136		1,149,491	0.021		0.081		2,110,954
Metrics + Environment	0.611	**0.289**	*	**0.325**	*	1,036,337	0.241		**0.41**	**	1,987,616
PCs + Environment	0.557	-0.059		0.069		1,139,453	-0.095		-0.01		2,123,918

7.5 Discussion

Landscape metrics have been successfully used as predictors in many studies that map species distributions, and have improved model performance through their ability to add information about fine-scale processes affecting species distributions, for example habitat fragmentation, dispersal limitations and anthropogenic influences on the landscape, compared to typical broad-scale environmental SDM predictors (Westphal et al., 2003; Hopkins, 2009; Foltête et al., 2012; Chefaoui, 2014). However, their ability to boost distribution model performance is known to vary across both taxonomic groups (Hasui et al., 2017) and spatial scales (Wu et al., 2002; Chefaoui, 2014). The research in this chapter did not find that the addition of landscape metrics to SDM improved model performance or predictions of ancient and veteran tree distributions across England.

There are several factors which could explain the failure of landscape metrics to improve models of ancient and veteran trees. One is that using landscape metrics in SDM has been suggested to be most effective when modelling at a fine spatial scale, especially for smaller, sedentary organisms (Westphal et al., 2003; Schindler et al., 2013). In this study, fine-scale (1 m) information about the landscape reflected in detailed maps of tree canopies was converted (by averaging) to landscape metrics for use in distribution models at a larger scale (1 km), probably leading to the loss of information about the local habitat suitability, and underlying ecological processes affecting each individual tree might have been lost (Ortner and Wallentin, 2020). This has been found to be the case in other SDM studies (Hasui et al., 2017), and an approach which uses predictors at multiple scales has been suggested to be more appropriate (Foltête et al., 2012).

The majority of studies using landscape structural components in SDM often have more than one type of habitat from which to calculate the metrics, which may therefore represent a more complex picture of the whole landscape (Schindler et al., 2013; Morelli et al., 2018). As my landscape metrics were derived from only one type of habitat (tree canopy), it might be the case that they were unable to differentiate between habitat types which are actually ecologically very distinct. For example landscapes with narrow, linear patches of canopy might represent agricultural fields, but they equally

might represent urban or suburban residential gardens. Although trees are ubiquitous throughout our landscape, using another type of identifying habitat or land-use type within the model may have increased the accuracy of my classification of landscape structure, and hence improved model performance. I did use both agricultural and land class predictors in the combination models, as well as types of vegetation cover (forest, ancient woodland and orchard), but because these alternate landscape descriptors are at a 1-km resolution, they may be too coarse to describe accurately the landscape structure in a way that is ecologically important for trees.

The selection of specific landscape metrics used to quantify the landscape in relation to species distributions has been shown to be an important factor in the accuracy of distribution models (Li & Wu, 2004; Schindler et al., 2013). The choice of metrics from the hundreds available should ideally be made using prior knowledge of species ecology and landscape biodiversity (Schindler et al., 2013), and the optimal choice varies widely depending on the landscape or taxon (Fahrig, 2003; Walz, 2011). In particular, the metrics themselves are biologically meaningless unless selected appropriately at the correct scale for the target species. In this study, the initial selection of the 16 metrics was done using biological theory and consideration of the target organism, but the reduction analysis was carried out using purely statistical methods. My two alternate selection processes (PCA and collinearity reduction) produced distribution models that differed in both model fit and predictive performance. Although PCA of landscape metrics has been shown to be an effective selection method (Schindler et al., 2015), in my study, this method consistently produced the poorest models compared to using a subset of the raw metrics, or using the alternative environmental predictors. Models using a selection of seven raw metrics performed slightly better, and although it has been recommended that between eight and 15 metrics is the ideal number (Cushman et al., 2008), this has been criticised as potentially retaining correlated predictors (Schindler et al., 2015). Therefore, it seems unlikely that it is the number of metrics, rather than the biological information contained, that reduced model performance. As with many aspects of distribution modelling, detailed consideration of ecological theory is likely to result in the most accurate predictions, and future research should consider alternative subsets of metrics if computationally feasible (Schindler et al., 2015).

Although landscape metrics had a limited impact on the performance of the model predictions, the individual raw landscape metrics provided some interesting insights into aspects of the surrounding canopy that might increase the likelihood of ancient or veteran tree presence. The PCA highlighted more suitable squares as having many, scattered canopies, as might be found in wood-pasture habitat, which fits with our current knowledge of ancient and veteran tree distributions (Rackham, 1994; Nolan et al., 2020). However, analysis of the raw metrics is more indicative of a cohesive canopy structure for high ancient and veteran tree abundance, for example that of rural woodland, with canopies that are larger and more irregular. This highlights the potential importance of woodland and other more canopied areas for ancient and veteran trees, which can be overlooked in favour of other habitats such as wood-pasture, with which ancient trees are more commonly associated (Farjon, 2017; Nolan et al., 2020). Although there is no wildwood remaining in the UK from the ice age (Rackham, 1994), ancient woodland (woodland that has existed since at least the 16th century: Peterken, 1977) covers around 2.6% of land in England and Wales (Spencer and Kirby, 1992). Although the term 'ancient' refers only to the length of existence of the woodland, ancient trees can be found in ancient woodland (Rackham, 1980), but not usually on land that has been converted from ancient woodland to forest or plantation (Lonsdale, 2013). Nevertheless, it may be likely that sampling bias in relation to woodlands e.g. through either accessibility issues or a lack of survey interest, prevents many ancient and veteran trees from being found and recorded in these areas.

Predictions of the total number of ancient and veteran trees across England were similar to those obtained in Chapter 6, although the variation across models was greater. The distribution model fitted using the raw landscape metrics produced the highest estimate, with almost 2.5 million trees, whereas the best performing model (the environment model), estimated only ~1.9 million. This difference between predictions is quite significant, representing almost three times the current total number of records in the ATI (approximately 200,000). As discussed in Chapter 6, estimates in this range are probably not exaggerated, and even larger figures have been suggested by other studies (Fay, 2004). Based on model performance, the totals from the landscape-metric models are potentially overestimates.

Nevertheless, all of these models highlight the need to increase the rate of tree recording rapidly to ensure more trees are recorded and can be protected for the future.

7.6 Conclusion

Large-scale analysis of the landscape structure across England based on the NTM canopies revealed interesting insights into specific types of landscape and canopy structures that increase the likelihood of the presence of ancient and veteran trees. Nevertheless, the addition of landscape metrics to SDM did not increase the performance of predictive power of the models, probably a function of modelling distributions at a coarser scale than the metrics describe. Predictions of the total number of ancient and veteran trees across England from models using landscape metrics were potentially inflated, but still highlighted the need for rapid, targeted tree recording to find and protect the large number of undiscovered ancient and veteran trees in our landscape.

Chapter 8: Discussion and Conclusion.

8.1 Discussion

Ancient trees and other trees with veteran characteristics are keystone organisms and provide valuable ecological, historical and recreational functions around the world (Butler et al., 2002; ATF, 2008a; Lonsdale, 2013). Despite their importance, there are severe gaps in the knowledge about them and their distribution, and protection and conservation measures are lacking: as a result, their global decline is apparent (Read, 2000; ATF, 2005; Lindenmayer et al., 2012; Le Roux et al., 2014). This thesis uses one of the most important resources available currently in regards to ancient and veteran tree conservation, the UK Ancient Tree Inventory (ATI), containing over 200,000 tree records collected over the past 15 years (Nolan et al., 2020), to present the first national overview and analysis of ancient, veteran and notable trees in the UK. Although the ATI holds the largest collection of ancient and other noteworthy tree records to date, it has received little attention in scientific research either directly to address issues related to the trees and their distribution, or in regards to the numerous organisms and ecological processes supported by the trees. This thesis presents novel research using the ATI in quantitative scientific studies to discover important information about the true distribution of ancient and veteran trees and their key environmental determinants.

As with all trees, the true distribution of ancient and veteran trees is likely to be the result of both environmental factors and human influence (Rackham, 1980; Barnes et al., 2017; Farjon, 2017). Across multiple chapters of this thesis, a key predictor for ancient and veteran trees was the presence or coverage of wood-pasture habitat: with around 50% of all ancient tree records in the ATI falling within wood-pasture habitat, it is clear that there is an association between the two. Although this strong association is likely to be partially an artefact of sampling bias (recorders know wood-pasture contain ancient trees and will preferentially survey here), even when sampling bias is successfully accounted for in Chapter 7, wood-pasture is still a key predictor of tree presence. A second important predictor was coverage or distance to National Trust owned land, which as with wood-pasture is likely to be a

partial artefact of being favourable to survey (likely to have trees and also recreationally pleasing). In fact, many National Trust sites are also likely to contain some wood-pasture habitat (Harvey, 1987; Nolan et al., 2020). Therefore, even though these areas are probably predictors of sampling bias, extensive surveys of wood-pasture habitat or National Trust land should continue to be expanded due to the high number of ancient trees predicted to still be found in these areas.

The inclusion of historical predictors digitised directly from literature sources in distribution models is a relatively uncommon practice, not least because most organisms do not live to be hundreds of years old (Farjon, 2017). Nevertheless, factors such as distance to a moated site, Tudor deer park or type of historic countryside were important determinants of ancient and tree distributions in Chapter 3, 6 and 7. Inclusion of comparable predictors about the cultural landscape were also important in similar studies conducted by Hartel et al. (2013; 2018) and Moga et al. (2016), and show that consideration of the historical as well as the current landscape is necessary when studying organisms such as ancient and veteran trees. In fact, ancient and veteran trees are an unusual group of organisms: they consist of multiple species but are of a particular age group and life stage, they are very well recorded at the level of the individual, and they have a particularly unusual direct relationship with human society. Therefore, modelling this particular type of organism may require consideration of alternative predictors, and is likely that some of the methods and findings in relation to ancient and veteran tree research will not be readily applicable to other taxa and vice versa. For example, this may have been a reason why the characterisation of landscape structure in Chapter 7, which has been shown to be a successful SDM method with other taxa, was less successful at modelling ancient and veteran tree distributions than the model fitted using alternative historical predictors.

Although the usefulness of citizen science recording is undeniable (Dickinson et al., 2010a), as with many large species databases the collection of the ATI records is highly likely to be biased, a conclusive finding from Chapter 4. This chapter provided the first qualitative insight into potential causes of bias in the ATI, examining a wide range of environmental and anthropogenic factors. As suspected, the current ATI distribution is to some extent a reflection of the location of recorders, with the top recorder

locations significantly mirroring hotspots of ancient and veteran trees. Alongside recorder location and types of land use, ATI recording is also likely biased by distance to towns, cities, roads and watercourses, as well as altitude, echoing the findings of many previous studies on other species databases (Reddy and Dávalos, 2003; Kadmon et al., 2003; Kramer-Schadt et al., 2013; Mair and Ruete, 2016). Another likely predictor of bias in the ATI is the presence (or absence) of public rights of way across the landscape: this was considered a predictor in several models, but because coverage across the UK was patchy, it was not possible to incorporate it in any analysis. Nevertheless, accessibility to sites is likely to present a critical obstacle in the completion of the ATI: as shown in the field surveys, recorders had major difficulties in some of the 1-km grid squares, with coverage falling below 20% of the total area. Future targeted surveying would greatly benefit from obtaining permission from landowners to access and survey sites predicted to have high numbers of trees by the distribution models.

Species Distribution Modelling (SDM) is a key tool in conservation and protecting biodiversity, and new improvements are constantly being made to the various models and methodologies (Araújo and Guisan, 2006; Yu et al., 2020; Zurell et al., 2020; Carlson, 2020). Nevertheless, neglecting to account for sampling bias in SDM will result in predictions of distributions or key environmental determinants that reflect sampling processes as well as the true underlying ecology (Phillips et al., 2009; Syfert et al., 2013). In Chapter 5 of this thesis, I present a novel approach to SDM based on biased species data, using Zero-Inflated (ZI) models. I show in Chapter 5 and 6 that not only are these models able to produce more accurate distribution maps with reduced impacts of sampling bias, but ZI models can also provide inferences about unknown predictors of sampling bias in the raw data, something which few current SDM methods are able to do. In Chapter 6, when applied to a real-life case study using the ATI, ZI models perform well compared to other bias correction methods, and are especially good at producing predictions that scale well to the true abundance of ancient and veteran trees. Analysis of the ZI model zero component coefficients supported my findings from Chapter 4, also suggesting that bias in the ATI is likely caused by accessibility issues (for example coverage of roads, watercourses etc.) and the selective choosing of ‘interesting’ survey sites thought more likely to contain trees, such as National

Trust land or wood-pasture sites. The ZI 'zero' prediction map produced in Chapter 6 is invaluable in providing insight into areas of over- or undersampling across England, and highlights new areas, such as large parts of Cornwall, Devon and Norfolk, within which future surveying can be directed to assist in making the ATI more comprehensive. Therefore, I believe that ZI models can provide great benefits for conservation going forward, not only of ancient and veteran trees, but also many other at-risk taxa: they could represent a robust, valuable tool within the field of SDM.

One of the key goals of SDM is to produce the most accurate predictions of the true ecological niche and geographical distribution of a species (Dormann et al., 2007; Phillips et al., 2009; Elith et al., 2011). Correcting for sampling bias should form a crucial part of the process and there are a variety of tested methods for this (Phillips et al. 2009; Kramer-Schadt et al., 2013; Fourcade et al., 2014, Boria et al., 2014). In addition to the novel ZI models, spatial filtering of occurrence records consistently produced some of the best prediction maps in Chapter 6, especially when based on systematic sampling or removing records within a cluster. This method is likely successful due to the initial large number of records in the ATI, so that filtering occurrences does not significantly reduce the sample size and therefore model performance. The distribution maps of ancient and veteran trees produced using either this method or the ZI models highlighted many areas which they suggest are more suitable for tree presence compared to those suggested by an uncorrected model, or by other bias-correction methods. Particular areas with increased suitability, and therefore likely targets for immediate future surveys, include parts of East Anglia, Northumberland, Greater London, Herefordshire and the Lake District. Many large wood-pastures, some of which contain no records and are likely unsampled, are also highlighted as important potential hot-spots of trees. These prediction maps present the first quantitative and validated overview of the true unbiased distribution of ancient and veteran trees across England, and I believe the benefits to the discovery and future conservation of these organisms of having such maps will be great.

In this thesis I introduce two types of model validation in addition to internal model validation strategies: the historical mapping desk verification in Chapter 3, and the collection of independent field

data for Chapters 6 and 7. Model validation is an important part of SDM, especially in order to apply any results to practical conservation projects (Greaves et al., 2006; Costa et al., 2010; Fabri-Ruiz et al., 2019). Common internal model validation metrics such as Area Under the Curve (AUC) are the only validation method used in many SDM studies, yet they have received extensive criticism (Lobo et al., 2008; Peterson et al., 2008). My historical mapping validation strategy produced strong results in Chapter 3 that correlated highly with model predictions, and showed that this strategy has great potential for use in studies of tree distributions. Nevertheless, the possible high margin of error in the maps and selective recording (for example, trees in woodlands were typically designated as a single patch in historic maps), limits the use of this strategy, and although an improvement over internal model validation, is still likely to be less favourable compared to the collection of independent field data.

In Chapter 6, I show that using independent field data for model validation leads to significantly different inferences compared to internal model validation, highlighting the important of independently validating distribution models that will be used directly in conservation practice. Although not always feasible due to cost or time pressures, I show that relatively little field data (only 52 1-km squares, of which 13 were collected myself) were needed to validate the models accurately and produce robust conclusions about various distribution models. This was possible through the use of the large nationwide network of ATI recorders; similar networks are likely also to exist for many other large citizen-science projects, and could therefore be a very useful and underappreciated tool for scientific research and conservation. In agreement with other authors (Devictor et al., 2010; Tweddle et al., 2012; Newman et al., 2012), I believe there is a large scope not only for using data collected by citizen-science projects in scientific research, but also for involving the various networks of recorders in carrying out more strategic, targeted surveys for the purpose of model validation and selecting the most appropriate final species distribution prediction maps.

The final key output from this thesis is the novel creation of the first estimates of the total ancient and veteran tree numbers, both in wood-pasture habitat and across England, providing useful insights into the overall progress of the ATI project since its initiation and into how many trees are still unrecorded.

Interestingly, results from Chapter 3, 6 and 7 all reach a similar conclusion, predicting roughly that the ATI is currently at around a 10% completion level. The results suggest that the current recorded number of 10,000 ancient trees across all wood-pastures reflects an estimated 100,000 trees, and the ~200,000 ancient and veteran trees in the ATI represents an estimated 2 million such trees across the whole of England. Therefore, although the ATI is at the forefront of ancient and veteran tree recording worldwide, my results suggest that there is still a way to go to ensure the majority of trees are recorded and protected. Nevertheless, there are many positive signs that suggest ancient and veteran tree recording is increasing in popularity, and the redesign of the ATI website and a streamlined recording process in 2018/ 2019 will likely assist with this. Citizen science in general is also gaining in popularity as public awareness of conservation issues is expanding (Dickinson et al., 2010b; Newman et al., 2012), and in Chapter 2 I show that the number of records added annually to the ATI is also increasing. Therefore, by using the distribution maps produced from my research alongside the increased popularity of citizen-science recording and the ease of recording tree online directly in the field, the undiscovered ancient and veteran trees can be found more quickly and added to the ATI. Given the increasing interest and potential for targeted surveying, I think that it is very likely that in 15 more years the ATI will have reached a much higher completion level than just 20%.

The overall distribution and environmental niche of a tree is likely to be highly dependent on its taxonomic identity (Barnes et al., 2017), and I show in Chapter 2 that the genus or species is highly influential on the category of tree (ancient, veteran or notable), the threats each tree faces and where it is found. In Chapter 3 I model ancient tree abundance in wood-pasture for two genera separately, *Quercus* and *Fraxinus*, and highlight significant differences in their environmental determinants. However, due to the low frequency of other genera, statistical power was consistently too low to incorporate species differences into other models, especially when using the independently collected field data (where not all tree taxa were identified) to validate the models as in Chapters 6 and 7. Future research concentrated on interspecific differences between trees is likely to produce interesting results, and could be used to target surveying specific tree genera more at risk across our landscape than others.

Another future research opportunity identified in relation to this thesis would be the expansion of the models into Wales, Scotland and Ireland. In Chapter 2 I show that there are significant differences in tree characteristics between countries, and it is possible that alternative landscape and environmental predictors would be needed to capture effectively the ecological processes determining ancient and veteran tree distributions in these countries. As many of the predictors I used in my models in various chapters did not extend outside of England, especially the historical predictors (e.g. moated sites, historic forests etc.), alternatives would need to be found to capture this information - this was outside the scope and timeline of this project. Finally, the National Tree Map (NTM) is an amazing, albeit computationally demanding, resource that could be used to address many interesting ecological questions about ancient and veteran tree distributions. Currently the grid references of each ATI record are coarser than the resolution of the NTM, so each record cannot currently be matched to a single canopy. Future work could address this issue, allowing more fine-scale research about the localised environment of each ancient or veteran tree to be examined. The NTM, of which ancient and other noteworthy trees are a subset, could therefore be a great tool in investigating ecological processes about individual trees or local populations, as well as overall distributions.

8.2 Conclusion

This thesis presents the first overview and quantitative analysis of the true UK ancient, veteran and notable tree distribution using the globally renowned Ancient Tree Inventory (ATI). By using a variety of distribution modelling methods across varying scales (ranging from habitat level to large, national analyses), the true ancient and veteran tree distribution, their key environmental determinants and estimates of the total number of trees nationwide were successfully produced. These estimates are the first of their kind, and as I initially suspected, they suggest that the ATI is far from complete, with currently around only 10% of trees recorded. Identification of, and correction for sampling bias in the ATI using both novel and traditional bias correction methods was a key step in producing more accurate distribution maps, allowing new potential hot-spots of ancient and other noteworthy trees to be identified. In addition, validating model predictions using new, independently collected field data

provides an extra level of security in our ability to rely on these distribution maps for practical conservation applications, something which relatively few studies are able to claim. In conclusion, the research in this thesis provides the first crucial overview of ancient and other noteworthy trees across the UK, and has great implications for the conservation and protection of these valuable biological entities, helping to ensure their persistence and survival into the future.

References

Aberg, F., 1978. Medieval Moated Sites. Council for British Archaeology.

Akaike, H., 1973. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika 60, 255–265. https://doi.org/10.1093/biomet/60.2.255

Alexander, K., 1999. The invertebrates of Britain's wood pastures. British Wildlife. 11(2), 108-117.

Ancient Tree Forum c/o The Woodland Trust, 2005. Ancient Tree Guide No. 1: Trees and Farming. https://www.woodlandtrust.org.uk/publications/2005/01/ancient-tree-guide-1/. Accessed 17/10/17.

Ancient Tree Forum c/o The Woodland Trust, 2008a. Ancient Tree Guide No. 4: What are ancient, veteran and other trees of special interest? https://www.woodlandtrust.org.uk /publications/2008/11/what-are-ancient-veteran-and-trees-of-special-interest/. Accessed 17/10/17.

Ancient Tree Forum c/o The Woodland Trust, 2008b. Ancient Tree Guide No. 5: Trees and Climate Change. https://www.woodlandtrust.org.uk/publications/2008/12/ancient-tree-guide-5/. Accessed 17/10/17.

Ancient Tree Forum c/o The Woodland Trust, 2009. Ancient Tree Guide No. 7: Ancient trees for the future. https://www.woodlandtrust.org.uk/publications/2009/12/ancient-trees-for-the-future/. Accessed 17/10/17.

Ancient Tree Forum c/o The Woodland Trust, 2011. Ancient Tree Guide No. 3: Trees and Development. https://www.woodlandtrust.org.uk/publications/2011/12/ancient-tree-guide-3/. Accessed 17/10/17.

Araújo, M.B., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Araújo, M.B., Whittaker, R.J., Ladle, R.J., Erhard, M., 2005. Reducing uncertainty in projections of extinction risk from climate change. Glob. Ecol. Biogeogr. 14, 529–538. https://doi.org/10.1111/j.1466-822X.2005.00182.x

Aubad, J., Aragón, P., Rodríguez, M.Á., 2010. Human access and landscape structure effects on Andean forest bird richness. Acta Oecologica 36, 396–402. https://doi.org/10.1016/j.actao.2010.03.009

Austin, M.P., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157, 101–118. https://doi.org/10.1016/S0304-3800(02)00205-3

Baalman, P., Kirby, K.J., 1995. Trial measures of habitat (particularly woodland) fragmentation. English Nature Research Report 134, Peterborough, English Nature.

Bahn, V., McGill, B.J., 2013. Testing the predictive performance of distribution models. Oikos 122, 321–331. https://doi.org/10.1111/j.1600-0706.2012.00299.x

Baker, W.L., 1992. Effects of Settlement and Fire Suppression on Landscape Structure. Ecology 73, 1879–1887. https://doi.org/10.2307/1940039

Ballesteros, J. a., Stoffel, M., Bodoque, J. m., Bollschweiler, M., Hitz, O., Díez-Herrero, A., 2010. Changes in Wood Anatomy in Tree Rings of Pinus pinaster Ait. Following Wounding by Flash Floods. Tree-Ring Res. 66, 93–103. https://doi.org/10.3959/2009-4.1

Banks-Leite, C., Ewers, R.M., Kapos, V., Martensen, A.C., Metzger, J.P., 2011. Comparing species and measures of landscape structure as indicators of conservation importance. J. Appl. Ecol. 48, 706–714. https://doi.org/10.1111/j.1365-2664.2011.01966.x

Barbet-Massin, M., Jiguet, F., Albert, C.H., Thuiller, W., 2012. Selecting pseudo-absences for species distribution models: how, where and how many? Methods Ecol. Evol. 3, 327–338. https://doi.org/10.1111/j.2041-210X.2011.00172.x

Barnes, G., Pillatt, T., Williamson, T., 2017. Trees in England: Management and Disease since 1600. Univ of Hertfordshire Press.

Barry, S.C., Welsh, A.H., 2002. Generalized additive modelling and zero inflated count data. Ecol. Model. 157, 179–188. https://doi.org/10.1016/S0304-3800(02)00194-1

Baskent, E.Z., Jordan, G.A., 2011. Characterizing spatial structure of forest landscapes. Can. J. For. Res. https://doi.org/10.1139/x95-198

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. "Fitting Linear Mixed-Effects Models Using lme4." Journal of Statistical Software, 67, 1, 1–48.

Beaumont, L.J., Pitman, A.J., Poulsen, M., Hughes, L., 2007. Where will species go? Incorporating new advances in climate modelling into projections of species distributions. Glob. Change Biol. 13, 1368–1385. https://doi.org/10.1111/j.1365-2486.2007.01357.x

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Becker, N., Freeman, S., 2009. The economic value of old growth trees in Israel. For. Policy Econ. 11, 608–615. https://doi.org/10.1016/j.forpol.2009.08.004

Bevan-Jones, R., 2016. The Ancient Yew: A History of Taxus baccata. Windgather Press.

Biodiversity Reporting and Information Group (BRIG), 2011. UK Biodiversity Action Plan – Priority Habitat Descriptions. JNCC, Peterborough.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherspoon, S., Krkosek, M., Stuart-Smith, J.F., Pecl, G.T., Barrett, N., Frusher, S., 2014. Statistical solutions for error and bias in global citizen science datasets. Biol. Conserv. 173, 144–154. https://doi.org/10.1016/j.biocon.2013.07.037

Björkegren, A., Grimmond, C., 2018. Net carbon dioxide emissions from central London. Urban Clim., ICUC9: The 9th International Conference on Urban Climate 23, 131–158. https://doi.org/10.1016/j.uclim.2016.10.002

Bluesky International Limited, 2015. National Tree Map™ 2015. http://www.bluesky-world.com/#!national-tree-map/c1pqz. Accessed: 27/02/2019.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K., Mace, G.M., 2010. Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data. PLoS Biol. 8. https://doi.org/10.1371/journal.pbio.1000385

Boddy, L., 2001. Fungal Community Ecology and Wood Decomposition Processes in Angiosperms: From Standing Tree to Complete Decay of Coarse Woody Debris. Ecol. Bull. 43–56.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Boria, R.A., Olson, L.E., Goodman, S.M., Anderson, R.P., 2014. Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. Ecol. Model. 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Bouyer, Y., Rigot, T., Panzacchi, M., Moorter, B.V., Poncin, P., Beudels-Jamar, R., Odden, J., Linnell, J.D.C., 2015. Using Zero-Inflated Models to Predict the Relative Distribution and Abundance of Roe Deer Over Very Large Spatial Scales. Ann. Zool. Fenn. 52, 66–76. https://doi.org/10.5735/086.052.0206

Boyd, C., Brooks, T.M., Butchart, S.H.M., Edgar, G.J., Fonseca, G.A.B.D., Hawkins, F., Hoffmann, M., Sechrest, W., Stuart, S.N., Dijk, P.P.V., 2008. Spatial scale and the conservation of threatened species. Conserv. Lett. 1, 37–43. https://doi.org/10.1111/j.1755-263X.2008.00002.x

Brasier, C.M., 1996. Phytophthora cinnamomi and oak decline in southern Europe. Environmental constraints including climate change. Ann. Sci. For. 53, 347–358. https://doi.org/10.1051/forest:19960217

Briffa, K.R., 2000. Annual climate variability in the Holocene: interpreting the message of ancient trees. Quat. Sci. Rev. 19, 87–105. https://doi.org/10.1016/S0277-3791(99)00056-6

Brotons, L., Thuiller, W., Araújo, M.B., Hirzel, A.H., 2004. Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27, 437–448. https://doi.org/10.1111/j.0906-7590.2004.03764.x

Brunsdon, C., Fotheringham, S., Charlton, M., 1998. Geographically Weighted Regression. J. R. Stat. Soc. Ser. Stat. 47, 431–443. https://doi.org/10.1111/1467-9884.00145

Burnham, K.P., Anderson, D.R., 2003. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer Science & Business Media.

Butler, J., 2014. Mapping ancient and other trees of special interest: UK citizens' contribution to world tree heritage [WWW Document]. Eur. Wood-Pastures Transit. https://doi.org/10.4324/9780203797082-23

Butler, J., Alex, K., Green, T., 2002. Decaying Wood: An Overview of its Status and Ecology in the United Kingdom and Continental Europe. USDA Forest Service Gen.

Bystriakova, N., Peregrym, M., Erkens, R., Bezsmertna, O., Schneider, H., 2012. Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. Syst. Biodivers. 10. https://doi.org/10.1080/14772000.2012.705357

Cameron, A.C., Trivedi, P.K., 2013. Regression Analysis of Count Data. Cambridge University Press.

Carlson, C.J., 2020. embarcadero: Species distribution modelling with Bayesian additive regression trees in r. Methods Ecol. Evol. 11, 850–858. https://doi.org/10.1111/2041-210X.13389

Casalegno, S., Anderson, K., Hancock, S., Gaston, K.J., 2017. Improving models of urban greenspace: from vegetation surface cover to volumetric survey, using waveform laser scanning. Methods Ecol. Evol. 8, 1443–1452. https://doi.org/10.1111/2041-210X.12794

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society, Series A 158, 3, 419-466.

Chefaoui, R.M., 2014. Landscape metrics as indicators of coastal morphology: A multi-scale approach. Ecol. Indic. 45, 139–147. https://doi.org/10.1016/j.ecolind.2014.04.004

Chen, I.-C., Hill, J.K., Ohlemueller, R., Roy, D.B., Thomas, C.D., 2011. Rapid Range Shifts of Species Associated with High Levels of Climate Warming. Science 333, 1024–1026. https://doi.org/10.1126/science.1206432

Cherubini, P., Fontana, G., Rigling, D., Dobbertin, M., Brang, P., Innes, J.L., 2002. Tree-life history prior to death: two fungal root pathogens affect tree-ring growth differently. J. Ecol. 90, 839–850. https://doi.org/10.1046/j.1365-2745.2002.00715.x

Clement, L., Catzeflis, F., Richard-Hansen, C., Barrioz, S., de Thoisy, B., 2014. Conservation interests of applying spatial distribution modelling to large vagile Neotropical mammals. Trop. Conserv. Sci. 7, 202–223.

Cloke, P., Jones, O., 2020. Tree Cultures: The Place of Trees and Trees in Their Place. Routledge.

Connors, J.P., Galletti, C.S., Chow, W.T.L., 2013. Landscape configuration and urban heat island effects: assessing the relationship between landscape characteristics and land surface temperature in Phoenix, Arizona. Landsc. Ecol. 28, 271–283. https://doi.org/10.1007/s10980-012-9833-1

Cooke, R.C.& R., A.D.M., 1984. Ecology of Saprotrophic Fungi, 1st Edition edition. ed. UK: Longman 1984.

Costa, G.C., Nogueira, C., Machado, R.B., Colli, G.R., 2010. Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. Biodivers. Conserv. 19, 883–899. https://doi.org/10.1007/s10531-009-9746-8

Couch, S.M., 2012. Conservation of Avenue Trees. Arboric. J.

Cowley, M.J.R., Thomas, C.D., Thomas, J.A., Warren, M.S., 1999. Flight areas of British butterflies: assessing species status and decline. Proc. R. Soc. Lond. B Biol. Sci. 266, 1587–1592. https://doi.org/10.1098/rspb.1999.0819

Cozzi, G., Müller, C.B., Krauss, J., 2008. How do local habitat management and landscape structure at different spatial scales affect fritillary butterfly distribution on fragmented wetlands? Landsc. Ecol. 23, 269–283. https://doi.org/10.1007/s10980-007-9178-3

Crall, A.W., Newman, G.J., Stohlgren, T.J., Holfelder, K.A., Graham, J., Waller, D.M., 2011. Assessing citizen science data quality: an invasive species case study. Conserv. Lett. 4, 433–442. https://doi.org/10.1111/j.1755-263X.2011.00196.x

Crane, M., Lindenmayer, D.B., Cunningham, R.B., Stein, J.A.R., 2017. The effect of wildfire on scattered trees, 'keystone structures', in agricultural landscapes. Austral Ecol. 42, 145–153. https://doi.org/10.1111/aec.12414

Cristofoli, S., Monty, A., Mahy, G., 2010. Historical landscape structure affects plant species richness in wet heathlands with complex landscape dynamics. Landsc. Urban Plan. 98, 92–98. https://doi.org/10.1016/j.landurbplan.2010.07.014

Cunningham, R.B., Lindenmayer, D.B., 2005. Modeling Count Data of Rare Species: Some Statistical Issues. Ecology 86, 1135–1142. https://doi.org/10.1890/04-0589

Cushman, S.A., McGarigal, K., Neel, M.C., 2008. Parsimony in landscape metrics: Strength, universality, and consistency. Ecol. Indic. 8, 691–703. https://doi.org/10.1016/j.ecolind.2007.12.002

Dénes, F.V., Silveira, L.F., Beissinger, S.R., 2015. Estimating abundance of unmarked animal populations: accounting for imperfect detection and other sources of zero inflation. Methods Ecol. Evol. 6, 543–556. https://doi.org/10.1111/2041-210X.12333

Dennis, R.L.H., Thomas, C.D., 2000. Bias in Butterfly Distribution Maps: The Influence of Hot Spots and Recorder's Home Range. J. Insect Conserv. 4, 73–77. https://doi.org/10.1023/A:1009690919835

Devictor, V., Whittaker, R.J., Beltrame, C., 2010. Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. Divers. Distrib. 16, 354–362.

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010a. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dickinson, J.L., Zuckerberg, B., Bonter, D.N., 2010b. Citizen Science as an Ecological Research Tool: Challenges and Benefits. Annu. Rev. Ecol. Evol. Syst. 41, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Dormann, C., M. McPherson, J., B. Araújo, M., Bivand, R., Bolliger, J., Carl, G., G. Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., R. Peres-Neto, P., Reineking, B., Schröder, B., M. Schurr, F., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30, 609–628.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., McClean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D., Lautenbach, S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Dorresteijn, I., Hartel, T., Hanspach, J., Wehrden, H. von, Fischer, J., 2013. The Conservation Value of Traditional Rural Landscapes: The Case of Woodpeckers in Transylvania, Romania. PLOS ONE 8, e65236. https://doi.org/10.1371/journal.pone.0065236

Dudík, M., E. Schapire, R., J. Phillips, S., 2005. Correcting sample selection bias in maximum entropy density estimation, in: Advances in Neural Information Processing Systems.

Dwyer, R.G., Carpenter-Bundhoo, L., Franklin, C.E., Campbell, H.A., 2016. Using citizen-collected wildlife sightings to predict traffic strike hot spots for threatened species: a case study on the southern cassowary. J. Appl. Ecol. 53, 973–982. https://doi.org/10.1111/1365-2664.12635

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A. et al., 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29, 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Burgman, M.A., Regan, H.M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecol. Model. 157, 313–329. https://doi.org/10.1016/S0304-3800(02)00202-8

Elith, J., Kearney, M., Phillips, S., 2010. The art of modelling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elith, J., Phillips, S.J., Hastie, T., Dudík, M., Chee, Y.E., Yates, C.J., 2011. A statistical explanation of MaxEnt for ecologists. Divers. Distrib. 17, 43–57. https://doi.org/10.1111/j.1472-4642.2010.00725.x

ESRI, 2018. ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Everett, S., Parakoottathil, D.J., 2018. Transformation, meaning-making and identity creation through folklore tourism: the case of the Robin Hood Festival. J. Herit. Tour. 13, 30–45. https://doi.org/10.1080/1743873X.2016.1251443

Fabri-Ruiz, S., Danis, B., David, B., Saucède, T., 2019. Can we generate robust species distribution models at the scale of the Southern Ocean? Divers. Distrib. 25, 21–37. https://doi.org/10.1111/ddi.12835

Fahrig, L., 2003. Effects of habitat fragmentation on biodiversity. Annu. Rev. Ecol. Evol. Syst. 34, 487–515. https://doi.org/10.1146/annurev.ecolsys.34.011802.132419

Farjon, A., 2017. Ancient Oaks in the English landscape. Kew Publishing.

Fay, N., 2002. Environmental Arboriculture, Tree Ecology and Veteran Tree Management. Arboric. J. 26, 213–238. https://doi.org/10.1080/03071375.2002.9747336

Fay, N., 2004. Survey methods & development of innovative arboricultural techniques, The Trees of History, Proceedings of the International Congress; University of Torino, Italy.

Fenton, J., 1984. The SWT Highland birchwood survey. Edinburgh: Scottish Ecological Consultants on behalf of Scottish Wildlife Trust.causing legal issues (ATF website, 2019).

Fielding, A.H., Bell, J.F., 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conserv. 24, 38–49.

Fischer, J., Stott, J., Zerger, A., Warren, G., Sherren, K., Forrester, R.I., 2009. Reversing a tree regeneration crisis in an endangered ecoregion. Proc. Natl. Acad. Sci. 106, 10386–10391. https://doi.org/10.1073/pnas.0900110106

Fischer, J., Zerger, A., Gibbons, P., Stott, J., Law, B.S., 2010. Tree decline and the future of Australian farmland biodiversity. Proc. Natl. Acad. Sci. 107, 19597–19602. https://doi.org/10.1073/pnas.1008476107

Fisher, R.A., 1941. The Negative Binomial Distribution. Ann. Eugen. 11, 182–187. https://doi.org/10.1111/j.1469-1809.1941.tb02284.x

Fitzpatrick, M.C., Gotelli, N.J., Ellison, A.M., 2013. MaxEnt versus MaxLike: empirical comparisons with ant species distributions. Ecosphere 4, art55. https://doi.org/10.1890/ES13-00066.1

Fitzpatrick, M.C., Preisser, E.L., Ellison, A.M., Elkinton, J.S., 2009. Observer bias and the detection of low-density populations. Ecol. Appl. 19, 1673–1679. https://doi.org/10.1890/09-0265.1

Foltête, J.-C., Clauzel, C., Vuidel, G., Tournant, P., 2012. Integrating graph-based connectivity metrics into species distribution models. Landsc. Ecol. 27, 557–569. https://doi.org/10.1007/s10980-012-9709-4

Forejt, M., Skalos, J., Pereponova, A., 2017. Changes and continuity of wood-pastures in the lowland landscape in Czechia. Applied Geography, 79:235–244.

Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27, 245–256. https://doi.org/10.1111/geb.12684

Fourcade, Y., Engler, J.O., Besnard, A.G., Rödder, D., Secondi, J., 2013. Confronting expert-based and modelled distributions for species with uncertain conservation status: A case study from the corncrake (Crex crex). Biol. Conserv. 167, 161–171. https://doi.org/10.1016/j.biocon.2013.08.009

Fourcade, Y., Engler, J.O., Rödder, D., Secondi, J., 2014. Mapping Species Distributions with MAXENT Using a Geographically Biased Sample of Presence Data: A Performance Assessment of Methods for Correcting Sampling Bias. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0097122

Fox, J., Weisberg, S., 2011. An R Companion to Applied Regression, Second edition. ed. SAGE Publications, Inc, Thousand Oaks, Calif.

Franke, G.R., 2010. Multicollinearity, in: Wiley International Encyclopedia of Marketing. American Cancer Society. https://doi.org/10.1002/9781444316568.wiem02066

Freeman, E.A., Moisen, G.G., 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol. Model. 217, 48–58. https://doi.org/10.1016/j.ecolmodel.2008.05.015

Freitag, S., Hobson, C., Biggs, H.C., Jaarsveld, A.S. van, 1998. Testing for potential survey bias: the effect of roads, urban areas and nature reserves on a southern African mammal data set. Anim. Conserv. 1, 119–127. https://doi.org/10.1111/j.1469-1795.1998.tb00019.x

Fulford, T., 1995. Wordsworth's 'Yew-Trees': Politics, Ecology, and Imagination. Romanticism 1, 272–288. https://doi.org/10.3366/rom.1995.1.2.272

Fuller, R.J., Warren, M.S., 1993. Coppiced Woodlands: Their Management for Wildlife. Joint Nature Conservation Committee.

Gardiner, M.M., Allee, L.L., Brown, P.M., Losey, J.E., Roy, H.E., Smyth, R.R., 2012. Lessons from lady beetles: accuracy of monitoring data from US and UK citizen-science programs. Front. Ecol. Environ. 10, 471–476. https://doi.org/10.1890/110185

Gardner, W., Mulvey, E.P., Shaw, E.C., 1995. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychol. Bull. 118, 392–404. https://doi.org/10.1037/0033-2909.118.3.392

Getz, W.M., Marshall, C.R., Carlson, C.J., Giuggioli, L., Ryan, S.J., Romañach, S.S., Boettiger, C., Chamberlain, S.D., Larsen, L., D'Odorico, P., O'Sullivan, D., 2018. Making ecological models adequate. Ecol. Lett. 21, 153–166. https://doi.org/10.1111/ele.12893

Gibbons, P., Lindenmayer, D.B., Fischer, J., Manning, A.D., Weinberg, A., Seddon, J., Ryan, P., Barrett, G., 2008. The Future of Scattered Trees in Agricultural Landscapes. Conserv. Biol. 22, 1309–1319. https://doi.org/10.1111/j.1523-1739.2008.00997.x

Gonzalez, S.C., Soto-Centeno, J.A., Reed, D.L., 2011. Population distribution models: species distributions are better modeled using biologically relevant data partitions. BMC Ecol. 11, 20. https://doi.org/10.1186/1472-6785-11-20

Gouraguine, A., Moranta, J., Ruiz-Frau, A., Hinz, H., Reñones, O., Ferse, S.C.A., Jompa, J., Smith, D.J., 2019. Citizen science in data and resource-limited areas: A tool to detect long-term ecosystem changes. PLOS ONE 14, e0210007. https://doi.org/10.1371/journal.pone.0210007

Greaves, G.J., Mathieu, R., Seddon, P.J., 2006. Predictive modelling and ground validation of the spatial distribution of the New Zealand long-tailed bat (Chalinolobus tuberculatus). Biol. Conserv. 132, 211–221. https://doi.org/10.1016/j.biocon.2006.04.016

Griffith, J.A., Martinko, E.A., Price, K.P., 2000. Landscape structure analysis of Kansas at three scales. Landsc. Urban Plan. 52, 45–61. https://doi.org/10.1016/S0169-2046(00)00112-2

Guisan, A., Zimmermann, N.E., Elith, J., Graham, C.H., Phillips, S., Peterson, A.T., 2007. What Matters for Predicting the Occurrences of Trees: Techniques, Data, or Species' Characteristics? Ecol. Monogr. 77, 615–630. https://doi.org/10.1890/06-1060.1

Gustafson, E.J., Parker, G.R., 1992. Relationships between landcover proportion and indices of landscape spatial pattern. Landsc. Ecol. 7, 101–110. https://doi.org/10.1007/BF02418941

Haines-Young, R., Chopping, M., 1996. Quantifying landscape structure: a review of landscape indices and their application to forested landscapes. Prog. Phys. Geogr. Earth Environ. 20, 418–445. https://doi.org/10.1177/030913339602000403

Hall, S.J.G., Bunce, R.G.H., 2011. Mature trees as keystone structures in Holarctic ecosystems – a quantitative species comparison in a northern English park. Plant Ecol. Divers. 4, 243–250. https://doi.org/10.1080/17550874.2011.586735

Harley, J.B., 1968. Error and Revision in Early Ordnance Survey Maps. Cartogr. J. 5, 115–124. https://doi.org/10.1179/caj.1968.5.2.115

Hartel, T., Dorresteijn, I., Klein, C., Máthé, O., Moga, C.I., Öllerer, K., Roellig, M., von Wehrden, H., Fischer, J., 2013. Wood-pastures in a traditional rural region of Eastern Europe: Characteristics, management and status. Biol. Conserv. 166, 267–275. https://doi.org/10.1016/j.biocon.2013.06.020

Hartel, T., Hanspach, J., Moga, C.I., Holban, L., Szapanyos, Á., Tamás, R., Hováth, C., Réti, K.-O., 2018. Abundance of large old trees in wood-pastures of Transylvania (Romania). Sci. Total Environ. 613–614, 263–270. https://doi.org/10.1016/j.scitotenv.2017.09.048

Hartel, T., Plieninger, T., 2014. European Wood-pastures in Transition: A Social-ecological Approach. Routledge.

Hartesveldt, R., Harvey, H., Shellhammer, H., Stecker, R., 1975. The Giant Sequoia of the Sierra Nevada. Rep. NPS 120, USDI National Park Service, Washington, DC.

Harvey, H.J., 1987. Changing attitudes to nature conservation: The National Trust. Biol. J. Linn. Soc. 32, 149–159. https://doi.org/10.1111/j.1095-8312.1987.tb00421.x

Hastie, T., Fithian, W., 2013. Inference from presence-only data; the ongoing controversy. Ecography 36, 864–867. https://doi.org/10.1111/j.1600-0587.2013.00321.x

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media.

Hasui, É., Silva, V.X., Cunha, R.G.T., Ramos, F.N., Ribeiro, M.C., Sacramento, M., Coelho, M.T.P., Pereira, D.G.S., Ribeiro, B.R., 2017. Additions of landscape metrics improve predictions of occurrence of species distribution models. J. For. Res. 28, 963–974. https://doi.org/10.1007/s11676-017-0388-5

Helliwell, D.R., 1989. Lime Trees in Britain. Arboric. J. 13, 119–123. https://doi.org/10.1080/03071375.1989.9756409

Hernandez, P.A., Graham, C.H., Master, L.L., Albert, D.L., 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. Ecography 29, 773–785. https://doi.org/10.1111/j.0906-7590.2006.04700.x

Hertzog, L.R., Besnard, A., Jay-Robert, P., 2014. Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. Divers. Distrib. 20, 1403–1413. https://doi.org/10.1111/ddi.12249

Hijmans, R. J., Phillips, S., Leathwick, J., Elith, J., 2011. Package ‘dismo’. Available online at: http://cran.r-project.org/web/packages/dismo/index.html.

Hijmans, R.J., 2012. Cross-validation of species distribution models: removing spatial sorting bias and calibration with a null model. Ecology 93, 679–688.

Hijmans, R.J., Graham, C.H., 2006. The ability of climate envelope models to predict the effect of climate change on species distributions. Glob. Change Biol. 12, 2272–2281. https://doi.org/10.1111/j.1365-2486.2006.01256.x

Hirzel, A.H., Helfer, V., Metral, F., 2001. Assessing habitat-suitability models with a virtual species. Ecol. Model. 145, 111–121. https://doi.org/10.1016/S0304-3800(01)00396-9

Hjältén, J., Johansson, T., Alinvi, O., Danell, K., Ball, J.P., Pettersson, R., Gibb, H., Hilszczański, J., 2007. The importance of substrate type, shading and scorching for the attractiveness of dead wood to saproxylic beetles. Basic Appl. Ecol. 8, 364–376. https://doi.org/10.1016/j.baae.2006.08.003

Hodge, S.J., Peterken, G.F., 1998. Deadwood in British forests: priorities and a strategy. For. Int. J. For. Res. 71, 99–112. https://doi.org/10.1093/forestry/71.2.99

Holdenrieder, O., Pautasso, M., Weisberg, P.J., Lonsdale, D., 2004. Tree diseases and landscape processes: the challenge of landscape pathology. Trends Ecol. Evol. 19, 446–452. https://doi.org/10.1016/j.tree.2004.06.003

Hopkins, R.L., 2009. Use of landscape pattern metrics and multiscale data in aquatic species distribution models: a case study of a freshwater mussel. Landsc. Ecol. 24, 943–955. https://doi.org/10.1007/s10980-009-9373-5

Howard, C., Stephens, P.A., Pearce-Higgins, J.W., Gregory, R.D., Willis, S.G., 2014. Improving species distribution models: the value of data on abundance. Methods Ecol. Evol. 5, 506–513. https://doi.org/10.1111/2041-210X.12184

Humphrey, J.W., 2005. Benefits to biodiversity from developing old-growth conditions in British upland spruce plantations: a review and recommendations. For. Int. J. For. Res. 78, 33–53. https://doi.org/10.1093/forestry/cpi004

Imdad Ullah, M., Muhammad, A., Altaf, S., 2016. mctest: An R Package for Detection of Collinearity among Regressors. R J. 8, 499–509. https://doi.org/10.32614/RJ-2016-062

Isaac, N.J.B., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods Ecol. Evol. 5, 1052–1060. https://doi.org/10.1111/2041-210X.12254

Jiménez-Valverde, A., Aragón, P., Lobo, J.M., 2021. Deconstructing the abundance–suitability relationship in species distribution modelling. Glob. Ecol. Biogeogr. 30, 327–338. https://doi.org/10.1111/geb.13204

Johnston, A., Fink, D., Hochachka, W.M., Kelling, S., 2018. Estimates of observer expertise improve species distributions from citizen science data. Methods Ecol. Evol. 9, 88–97. https://doi.org/10.1111/2041-210X.12838

Johnston, A., Fink, D., Reynolds, M.D., Hochachka, W.M., Sullivan, B.L., Bruns, N.E., Hallstein, E., Merrifield, M.S., Matsumoto, S., Kelling, S., 2015. Abundance models improve spatial and temporal prioritization of conservation resources. Ecol. Appl. 25, 1749–1756. https://doi.org/10.1890/14-1826.1

Jonsson, B.G., Kruys, N., Ranius, T., 2005. Ecology of species living on dead wood : lessons for dead wood management.

Jönsson, M.T., Fraver, S., Jonsson, B.G., 2009. Forest history and the development of old-growth characteristics in fragmented boreal forests. J. Veg. Sci. 20, 91–106. https://doi.org/10.1111/j.1654-1103.2009.05394.x

Kadmon, R., Farber, O., Danin, A., 2003. A Systematic Analysis of Factors Affecting the Performance of Climatic Envelope Models. Ecol. Appl. 13, 853–867.

Kadmon, R., Farber, O., Danin, A., 2004. Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models. Ecol. Appl. 14, 401–413.

Kelly, P.E., Cook, E.R., Larson, D.W., 1992. Constrained Growth, Cambial Mortality, and Dendrochronology of Ancient Thuja occidentalis on Cliffs of the Niagara Escarpment: An Eastern Version of Bristlecone Pine? Int. J. Plant Sci. 153, 117–127. https://doi.org/10.1086/297013

Kirby, K., 2015. What might a sustainable population of trees in wood-pasture sites look like? Hacquetia, 14, 43-52.

Kirby, K., Watkins, C., 2015. Europe's Changing Woods and Forests: From Wildwood to Managed Landscapes. CABI.

Kirby, K.J., Thomas, R.C., Key, R.S., McLEAN, I.F.G., Hodgetts, N., 1995. Pasture-woodland and its conservation in Britain. Biol. J. Linn. Soc. 56, 135–153. https://doi.org/10.1111/j.1095-8312.1995.tb01129.x

Kleiber, C., Zeileis, A., 2016. Visualizing Count Data Regressions Using Rootograms. Am. Stat. 70, 296–303. https://doi.org/10.1080/00031305.2016.1173590

Komori, O., Eguchi, S., Saigusa, Y., Kusumoto, B., Kubota, Y., 2020. Sampling bias correction in species distribution models by quasi-linear Poisson point process. Ecol. Inform. 55, 101015. https://doi.org/10.1016/j.ecoinf.2019.101015

Kosmala, M., Wiggins, A., Swanson, A., Simmons, B., 2016. Assessing data quality in citizen science. Front. Ecol. Environ. 14, 551–560. https://doi.org/10.1002/fee.1436

Kramer-Schadt, S., Niedballa, J., Pilgrim, J.D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A.K., Augeri, D.M., Cheyne, S.M., Hearn, A.J., Ross, J., Macdonald, D.W., Mathai, J., Eaton, J., Marshall, A.J., Semiadi, G., Rustam, R., Bernard, H., Alfred, R., Samejima, H., Duckworth, J.W., Breitenmoser-Wuersten, C., Belant, J.L., Hofer, H., Wilting, A., 2013. The importance of correcting for sampling bias in MaxEnt species distribution models. Divers. Distrib. 19, 1366–1379. https://doi.org/10.1111/ddi.12096

Kuemmerlen, M., Schmalz, B., Guse, B., Cai, Q., Fohrer, N., Jähnig, S., 2014. Integrating catchment properties in small scale species distribution models of stream macroinvertebrates. Ecol. Model. 277, 77–86. https://doi.org/10.1016/j.ecolmodel.2014.01.020

Kupfer, J.A., 2012. Landscape ecology and biogeography: Rethinking landscape metrics in a post-FRAGSTATS landscape. Prog. Phys. Geogr. Earth Environ. 36, 400–420. https://doi.org/10.1177/0309133312439594

Lachat, T., Bouget, C., Bütler, R., Müller, J., 2013. Deadwood: quantitative and qualitative requirements for the conservation of saproxylic biodiversity, in: Integrative Approaches as an Opportunity for the Conservation of Forest Biodiversity. pp. 92–102.

Lambert, D., 1992. Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. Technometrics 34, 1–14. https://doi.org/10.1080/00401706.1992.10485228

Lamine, S., Petropoulos, G.P., Singh, S.K., Szabó, S., Bachari, N.E.I., Srivastava, P.K., Suman, S., 2018. Quantifying land use/land cover spatio-temporal landscape pattern dynamics from Hyperion using

SVMs classifier and FRAGSTATS®. Geocarto Int. 33, 862–878. https://doi.org/10.1080/10106049.2017.1307460

Lanner, R.M., 2002. Why do trees live so long? Ageing Res. Rev. 1, 653–671. https://doi.org/10.1016/S1568-1637(02)00025-9

Laurance, W.F., Delamônica, P., Laurance, S.G., Vasconcelos, H.L., Lovejoy, T.E., 2000. Conservation: Rainforest fragmentation kills big trees. Nature 404, 35009032. https://doi.org/10.1038/35009032

Law, B., Caccamo, G., Roe, P., Truskinger, A., Brassil, T., Gonsalves, L., McConville, A., Stanton, M., 2017. Development and field validation of a regional, management-scale habitat model: A koala Phascolarctos cinereus case study. Ecol. Evol. 7, 7475–7489. https://doi.org/10.1002/ece3.3300

Lewington, A., 2012. Ancient Trees: Trees that live for a thousand years. Pavilion Books.

Li, H., Wu, J., 2004. Use and misuse of landscape indices. Landsc. Ecol. 19, 389–399. https://doi.org/10.1023/B:LAND.0000030441.15628.d6

Liang, W., Papeş, M., Tran, L., Grant, J., Washington-Allen, R., Stewart, S., Wiggins, G., 2018. The effect of pseudo-absence selection method on transferability of species distribution models in the context of non-adaptive niche shift. Ecol. Model. 388, 1–9. https://doi.org/10.1016/j.ecolmodel.2018.09.018

Lindén, A., Mäntyniemi, S., 2011. Using the negative binomial distribution to model overdispersion in ecological count data. Ecology 92, 1414–1421. https://doi.org/10.1890/10-1831.1

Lindenmayer, D. B., Laurance, W. F., Franklin, J. F., et al., 2014. New policies for old trees: Averting a global crisis in a keystone ecological structure. Conservation Letters, 7, 61–69.

Lindenmayer, D.B., Laurance, W.F., Franklin, J.F., 2012. Global Decline in Large Old Trees. Science 338, 1305–1306. https://doi.org/10.1126/science.1231070

Linder, P., Östlund, L., 1998. Structural changes in three mid-boreal Swedish forest landscapes, 1885–1996. Biol. Conserv. 85, 9–19. https://doi.org/10.1016/S0006-3207(97)00168-7

Linkie, M., Chapron, G., Martyr, D.J., Holden, J., Leader-Williams, N., 2009. Assessing the viability of tiger subpopulations in a fragmented landscape. J. Appl. Ecol. 576–586. https://doi.org/10.1111/j.1365-2664.2006.01153.x@10.1111/(ISSN)1365-2664.ECOLASIA

Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28, 385–393. https://doi.org/10.1111/j.0906-7590.2005.03957.x

Lobo, J.M., Jiménez-Valverde, A., Real, R., 2008. AUC: a misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lõhmus, A., Kinks, R., Soon, M., 2010. The Importance of Dead-Wood Supply for Woodpeckers in Estonia. Balt. For. 16, 11.

Loiselle, B.A., Jørgensen, P.M., Consiglio, T., Jiménez, I., Blake, J.G., Lohmann, L.G., Montiel, O.M., 2008. Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? J. Biogeogr. 35, 105–116. https://doi.org/10.1111/j.1365-2699.2007.01779.x

Lonsdale, D., 2013. Ancient and other veteran trees: further guidance on management. Ancient Tree Forum, London, 1–212. http://ancienttreeforum.co.uk/wp-content/uploads/2015/02/ATF_book.pdf

Luna, S., Gold, M., Albert, A., Ceccaroni, L., Claramunt, B., Danylo, O., Haklay, M., Kottmann, R., Kyba, C., Piera, J., Radicchi, A., Schade, S., Sturm, U., 2018. Developing mobile applications for environmental and biodiversity citizen science: considerations and recommendations. Joly, A., Vrochidis, S., Karatzas, K., Karppinen, A., Bonnet, P., (eds) Multimedia Tools and Applications for Environmental & Biodiversity Informatics. (pp. 9-30) Springer, Cham.

Luoto, M., Virkkala, R., Heikkinen, R.K., 2007. The role of land cover in bioclimatic models depends on spatial resolution. Glob. Ecol. Biogeogr. 16, 34–42. https://doi.org/10.1111/j.1466-8238.2006.00262.x

Lustig, A., Stouffer, D.B., Doscher, C., Worner, S.P., 2017. Landscape metrics as a framework to measure the effect of landscape structure on the spread of invasive insect species. Landsc. Ecol. 32, 2311–2325. https://doi.org/10.1007/s10980-017-0570-3

Lyashevska, O., Brus, D.J., van der Meer, J., 2016. Mapping species abundance by a spatial zero-inflated Poisson model: a case study in the Wadden Sea, the Netherlands. Ecol. Evol. 6, 532–543. https://doi.org/10.1002/ece3.1880

MacKenzie, D.I., 2005. What are the issues with Presence-Absence data for wildlife managers? J. Wildl. Manag. 69, 849–860. https://doi.org/10.2193/0022-541X(2005)069[0849:WATIWP]2.0.CO;2

Mair, L., Ruete, A., 2016. Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa. PLOS ONE 11, e0147796. https://doi.org/10.1371/journal.pone.0147796

Major, R., 1967. The Ginkgo, the most ancient living tree–the resistance of Ginkgo biloba L. to pests accounts in part for the longevity of this species. Science 157, 1270–1273.

Manning, A.D., Fischer, J., Lindenmayer, D.B., 2006. Scattered trees are keystone structures – Implications for conservation. Biol. Conserv. 132, 311–321. https://doi.org/10.1016/j.biocon.2006.04.023

Martin, T.G., Wintle, B.A., Rhodes, J.R., Kuhnert, P.M., Field, S.A., Low-Choy, S.J., Tyre, A.J., Possingham, H.P., 2005. Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecol. Lett. 8, 1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

Mateo, R.G., Felicisimo, A.M., Munoz, J., 2011. Species distributions models: A synthetic revision. Rev. Chil. Hist. Nat. 84, 217–240.

McGarigal, K., 1995. FRAGSTATS: Spatial Pattern Analysis Program for Quantifying Landscape Structure. U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station.

McInnes, R.N., Hemming, D., Burgess, P., Lyndsay, D., Osborne, N.J., Skjøth, C.A., Thomas, S., Vardoulakis, S., 2017. Mapping allergenic pollen vegetation in UK to study environmental exposure and human health. Sci. Total Environ. 599–600, 483–499. https://doi.org/10.1016/j.scitotenv.2017.04.136

Meynard, C.N., Kaplan, D.M., 2013. Using virtual species to study species distributions and model performance. J. Biogeogr. 40, 1–8. https://doi.org/10.1111/jbi.12006

Meynard, C.N., Leroy, B., Kaplan, D.M., 2019. Testing methods in species distribution modelling using virtual species: what have we learnt and what are we missing? Ecography 42, 2021–2036. https://doi.org/10.1111/ecog.04385

Minami, M., Lennert-Cody, C.E., Gao, W., Román-Verdesoto, M., 2007. Modeling shark bycatch: The zero-inflated negative binomial regression model with smoothing. Fish. Res. 84, 210–221. https://doi.org/10.1016/j.fishres.2006.10.019

Mitchell, R.J., Beaton, J.K., Bellamy, P.E., Broome, A., Chetcuti, J., Eaton, S., Ellis, C.J., Gimona, A., Harmer, R., Hester, A.J., et al., 2014. Ash dieback in the UK: A review of the ecological and conservation implications and potential management options. Biol. Conserv. 175, 95–109. https://doi.org/10.1016/j.biocon.2014.04.019

Moga, C.I., Samoilă, C., Öllerer, K., Băncilă, R.I., Réti, K.-O., Craioveanu, C., Poszet, S., Rákosy, L., Hartel, T., 2016. Environmental determinants of the old oaks in wood-pastures from a changing traditional social–ecological system of Romania. Ambio 45, 480–489. https://doi.org/10.1007/s13280-015-0758-1

Moir, A., 2013. The exceptional yew trees of England, Scotland and Wales. Q. J. For.

Morelli, F., Benedetti, Y., Šímová, P., 2018. Landscape metrics as indicators of avian diversity and community measures. Ecol. Indic. 90, 132–141. https://doi.org/10.1016/j.ecolind.2018.03.011

Morin, R.S., Liebhold, A.M., Tobin, P.C., Gottschalk, K.W., Luzader, E., 2007. Spread of beech bark disease in the eastern United States and its relationship to regional forest composition. Can. J. For. Res. 37, 726–736. https://doi.org/10.1139/X06-281

Mota-Vargas, C., Rojas-Soto, O.R., 2016. Taxonomy and ecological niche modeling: Implications for the conservation of wood partridges (genus Dendrortyx). J. Nat. Conserv. 29, 1–13. https://doi.org/10.1016/j.jnc.2015.10.003

Mountford, E.P., Peterken, G.F., 2003. Long-term change and implications for the management of wood-pastures: experience over 40 years from Denny Wood, New Forest. For. Int. J. For. Res. 76, 19–43. https://doi.org/10.1093/forestry/76.1.19

Muscarella, R., Galante, P.J., Soley-Guardia, M., Boria, R.A., Kass, J.M., Uriarte, M., Anderson, R.P., 2014. ENMeval: An R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. Methods Ecol. Evol. 5, 1198–1205. https://doi.org/10.1111/2041-210X.12261

Naimi, B., Araújo, M.B., 2016. sdm: a reproducible and extensible R platform for species distribution modelling. Ecography 39, 368–375.

National Planning Policy Framework, 2019. Ministry of Housing, Communities and Local Government https://www.gov.uk/government/publications/national-planning-policy-framework--2. Accessed August 2019.

Neilson, N., 1940. The Forests: 1327 – 1336 in The English Government at Work (Willard and Morris, 1940) The Medieval Academy of America, Massachusetts.

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., Crowston, K., 2012. The future of citizen science: emerging technologies and shifting paradigms. Front. Ecol. Environ. 10, 298–304. https://doi.org/10.1890/110294

Nolan, V., Reader, T., Gilbert, F., Atkinson, N., 2020. The Ancient Tree Inventory: a summary of the results of a 15 year citizen science project recording ancient, veteran and notable trees across the UK. Biodivers. Conserv. 29, 3103–3129. https://doi.org/10.1007/s10531-020-02033-2

O'Neill, R.V., Krummel, J.R., Gardner, R.H., Sugihara, G., Jackson, B., DeAngelis, D.L., Milne, B.T., Turner, M.G., Zygmunt, B., Christensen, S.W., Dale, V.H., Graham, R.L., 1988. Indices of landscape pattern. Landsc. Ecol. 1, 153–162. https://doi.org/10.1007/BF00162741

Ortner, O., Wallentin, G., 2020. Integration of landscape metric surfaces derived from vector data improves species distribution models. Ecol. Model. 431, 109160. https://doi.org/10.1016/j.ecolmodel.2020.109160

Owen, K., Alderman, D., 2008. Ancient Tree Hunt: The minimum girth of ancient trees – a guide for verifiers. Ancient Tree Hunt (Ancient Tree Forum, Woodland Trust, Tree Register of the British Isles).

Paradis, E., Schliep, K., 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pardon, P., Reubens, B., Reheul, D., Mertens, J., De Frenne, P., Coussement, T., Janssens, P., Verheyen, K., 2017. Trees increase soil organic carbon and nutrient availability in temperate agroforestry systems. Agric. Ecosyst. Environ. 247, 98–111. https://doi.org/10.1016/j.agee.2017.06.018

Parnell, J. a. N., Simpson, D.A., Moat, J., Kirkup, D.W., Chantaranothai, P., Boyce, P.C., Bygrave, P., Dransfield, S., Jebb, M.H.P., Macklin, J., Meade, C., Middleton, D.J., Muasya, A.M., Prajaksood, A., Pendry, C.A., Pooma, R., Suddee, S., Wilkin, P., 2003. Plant collecting spread and densities: their potential impact on biogeographical studies in Thailand. J. Biogeogr. 30, 193–209. https://doi.org/10.1046/j.1365-2699.2003.00828.x

Pautasso, M., Chiarucci, A., 2008. A Test of the Scale-dependence of the Species Abundance–People Correlation for Veteran Trees in Italy. Ann. Bot. 101, 709–715. https://doi.org/10.1093/aob/mcn010

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133, 225–245. https://doi.org/10.1016/S0304-3800(00)00322-7

Pearce, J.L., Boyce, M.S., 2006. Modelling distribution and abundance with presence-only data. J. Appl. Ecol. 43, 405–412. https://doi.org/10.1111/j.1365-2664.2005.01112.x

Peterken, G., 1977. Habitat conservation priorities in British and European woodlands. Biological Conservation, 11, 223-236.

Peterson, A.T., Papeş, M., Soberón, J., 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecol. Model. 213, 63–72. https://doi.org/10.1016/j.ecolmodel.2007.11.008

Petit, S., Watkins, C., 2003. Pollarding Trees: Changing Attitudes to a Traditional Land Management Practice in Britain 1600–1900. Rural Hist. 14, 157–176. https://doi.org/10.1017/S0956793303001018

Phillips, S.J., 2008. Transferability, sample selection bias and background data in presence-only modelling: a response to Peterson et al. (2007). Ecography 31, 272–278. https://doi.org/10.1111/j.0906-7590.2008.5378.x

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 190, 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S.J., Dudík, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175. https://doi.org/10.1111/j.0906-7590.2008.5203.x

Phillips, S.J., Dudik, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J., Ferrier, S., 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecol. Appl. 19, 181–197. https://doi.org/10.1890/07-2153.1

Pigott, D., 1992. The clones of common lime (Tilia×vulgaris Hayne) planted in England during the seventeenth and eighteenth centuries. New Phytol. 121, 487–493. https://doi.org/10.1111/j.1469-8137.1992.tb02949.x

Pino, J., Rodà, F., Ribas, J., Pons, X., 2000. Landscape structure and bird species richness: implications for conservation in rural areas between natural parks. Landsc. Urban Plan. 49, 35–48. https://doi.org/10.1016/S0169-2046(00)00053-0

Plieninger, T., Hartel, T., Martín-López, B., Beaufoy, G., Bergmeier, E., Kirby, K., Montero, M.J., Moreno, G., Oteros-Rozas, E., Van Uytvanck, J., 2015. Wood-pastures of Europe: Geographic coverage, social–ecological values, conservation management, and policy implications. Biol. Conserv. 190, 70–79. https://doi.org/10.1016/j.biocon.2015.05.014

Pocock, M.J.O., Evans, D.M., 2014. The Success of the Horse-Chestnut Leaf-Miner, Cameraria ohridella, in the UK Revealed with Hypothesis-Led Citizen Science. PLoS ONE 9. https://doi.org/10.1371/journal.pone.0086226

Polyakov, M., Majumdar, I., Teeter, L., 2008. Spatial and temporal analysis of the anthropogenic effects on local diversity of forest trees. For. Ecol. Manag. 255, 1379–1387. https://doi.org/10.1016/j.foreco.2007.10.052

Ponder, W.F., Carter, G.A., Flemons, P., Chapman, R.R., 2001. Evaluation of Museum Collection Data for Use in Biodiversity Assessment. Conserv. Biol. 15, 648–657. https://doi.org/10.1046/j.1523-1739.2001.015003648.x

Potts, J.M., Elith, J., 2006. Comparing species abundance models. Ecol. Model., Predicting Species Distributions 199, 153–163. https://doi.org/10.1016/j.ecolmodel.2006.05.025

Powers, R.P., Jetz, W., 2019. Global habitat loss and extinction risk of terrestrial vertebrates under future land-use-change scenarios. Nat. Clim. Change 9, 323–329. https://doi.org/10.1038/s41558-019-0406-z

Quelch, P., 2002. An illustrated guide to ancient wood pasture in Scotland. Glasgow. http://frontpage.woodland-trust.org.uk/ancient-treeforum/atfresources/images/guide28_54pp.pdf

Quelch, P., 2013. Upland Wood Pastures, in: Rotherham, I.D. (Ed.), Cultural Severance and the Environment: The Ending of Traditional and Customary Practice on Commons and Landscapes Managed in Common, Environmental History. Springer Netherlands, Dordrecht, pp. 419–430. https://doi.org/10.1007/978-94-007-6159-9_30

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org.

Rackham, O., 1967. The History & Effects of Coppicing as a woodland Practice, in The Biotic Effects of Public Pressures on the Environment, Nature Conservancy, London.

Rackham, O., 1976. Trees and woodland in the British landscape. London: J. M. Dent.

Rackham, O., 1980. Ancient woodland its history, vegetation and uses in England. Arnold, London.

Rackham, O., 1986. History of the countryside. Dent, London.

Rackham, O., 1994. The Illustrated History of the Countryside. Orion Publishing Group, London.

Radosavljevic, A., Anderson, R.P., 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. J. Biogeogr. 41, 629–643. https://doi.org/10.1111/jbi.12227

Ranius, T., 2002. Influence of stand size and quality of tree hollows on saproxylic beetles in Sweden. Biol. Conserv. 103, 85–91. https://doi.org/10.1016/S0006-3207(01)00124-0

Ranius, T., 2006. Measuring the dispersal of saproxylic insects: a key characteristic for their conservation. Popul. Ecol. 48, 177–188. https://doi.org/10.1007/s10144-006-0262-3

Ranius, T., Johansson, P., Berg, N., Niklasson, M., 2008. The influence of tree age and microhabitat quality on the occurrence of crustose lichens associated with old oaks. J. Veg. Sci. 19, 653–662. https://doi.org/10.3170/2008-8-18433

Rasey, A., 2004. Priority woodland in the landscape for two bat BAP species: the importance of ancient trees and woodlands. Iale (uk), Int Assoc Landscapeecol, Lymm.

Read, H., 2000. Veteran Trees: A guide to good management. English Nature, Peterborough: www.naturalengland.gov.uk

Read, H.J., Wheater, C.P., Forbes, V., Young, J., 2010. The current status of ancient pollard beech trees at Burnham Beeches and evaluation of recent restoration techniques. Q. J. For. 104, 109–120.

Reddy, S., Dávalos, L.M., 2003. Geographical sampling bias and its implications for conservation priorities in Africa. J. Biogeogr. 30, 1719–1727. https://doi.org/10.1046/j.1365-2699.2003.00946.x

Renner, I.W., Elith, J., Baddeley, A., Fithian, W., Hastie, T., Phillips, S.J., Popovic, G., Warton, D.I., 2015. Point process models for presence-only analysis. Methods Ecol. Evol. 6, 366–379. https://doi.org/10.1111/2041-210X.12352

Ridout, M., Hinde, J., Demétrio, C.G.B., 2001. A Score Test for Testing a Zero-Inflated Poisson Regression Model Against Zero-Inflated Negative Binomial Alternatives. Biometrics 57, 219–223. https://doi.org/10.1111/j.0006-341X.2001.00219.x

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929. https://doi.org/10.1111/ecog.02881

Rocchini, D., Hortal, J., Lengyel, S., Lobo, J.M., Jiménez-Valverde, A., Ricotta, C., Bacaro, G., Chiarucci, A., 2011. Accounting for uncertainty when mapping species distributions: The need for maps of ignorance. Prog. Phys. Geogr. 35, 211–226. https://doi.org/10.1177/0309133311399491

Rodríguez, J.P., Brotons, L., Bustamante, J., Seoane, J., 2007. The application of predictive modelling of species distribution to biodiversity conservation. Divers. Distrib. 13, 243–251. https://doi.org/10.1111/j.1472-4642.2007.00356.x

Rohner, B., Bugmann, H., Bigler, C., 2013. Estimating the age–diameter relationship of oak species in Switzerland using nonlinear mixed-effects models. Eur. J. For. Res. 132, 751–764. https://doi.org/10.1007/s10342-013-0710-5

Roper, C., 2003. Historical Mapping is Still Under-valued and Under-used. Cartogr. J. 40, 131–134. https://doi.org/10.1179/000870403235001502

Rosenthal, G., Schrautzer, J., Eichberg, C., 2012. Low-intensity grazing with domestic herbivores: A tool for maintaining and restoring plant diversity in temperate Europe. TUEXENIA 32, 167–205.

Rossi, J.-P., Garcia, J., Roques, A., Rousselet, J., 2016. Trees outside forests in agricultural landscapes: spatial distribution and impact on habitat connectivity for forest organisms. Landsc. Ecol. 31, 243–254. https://doi.org/10.1007/s10980-015-0239-8

Roux, D.S.L., Ikin, K., Lindenmayer, D.B., Manning, A.D., Gibbons, P., 2014. The Future of Large Old Trees in Urban Landscapes. PLOS ONE 9, e99403. https://doi.org/10.1371/journal.pone.0099403

Rubino, D.L., McCarthy, B.C., 2003. Composition and ecology of macrofungal and myxomycete communities on oak woody debris in a mixed-oak forest of Ohio. Can. J. For. Res. 33, 2151–2163. https://doi.org/10.1139/x03-137

Ruczynski, I., Bogdanowicz, W., 2008. Summer roost selection by tree-dwelling bats Nyctalus noctula and N-leisleri: A multiscale analysis. J. Mammal. 89, 942–951. https://doi.org/10.1644/07-MAMM-A-134.1

Rust, S., Roloff, A., 2002. Reduced photosynthesis in old oak (Quercus robur): the impact of crown and hydraulic architecture. Tree Physiol. 22, 597–601.

Saltre, F., Duputie, A., Gaucherel, C., Chuine, I., 2015. How climate, migration ability and habitat fragmentation affect the projected future distribution of European beech. Glob. Change Biol. 21, 897–910. https://doi.org/10.1111/gcb.12771

Sarkar, C., Webster, C., Gallacher, J., 2018. Residential greenness and prevalence of major depressive disorders: a cross-sectional, observational, associational study of 94 879 adult UK Biobank participants. Lancet Planet. Health 2, e162–e173. https://doi.org/10.1016/S2542-5196(18)30051-2

Schindler, S., von Wehrden, H., Poirazidis, K., Hochachka, W.M., Wrbka, T., Kati, V., 2015. Performance of methods to select landscape metrics for modelling species richness. Ecol. Model., Use of ecological indicators in models 295, 107–112. https://doi.org/10.1016/j.ecolmodel.2014.05.012

Schindler, S., von Wehrden, H., Poirazidis, K., Wrbka, T., Kati, V., 2013. Multiscale performance of landscape metrics as indicators of species richness of plants, insects and vertebrates. Ecol. Indic., Linking landscape structure and biodiversity 31, 41–48. https://doi.org/10.1016/j.ecolind.2012.04.012

Schmeller, D.S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F., Lengyel, S., Nowicki, P., Déri, E., Budrys, E., et al., 2009. Advantages of volunteer-based biodiversity monitoring in Europe. Conserv. Biol. J. Soc. Conserv. Biol. 23, 307–316. https://doi.org/10.1111/j.1523-1739.2008.01125.x

Schulman, L., Toivonen, T., Ruokolainen, K., 2007. Analysing botanical collecting effort in Amazonia and correcting for it in species range estimation. J. Biogeogr. 34, 1388–1399. https://doi.org/10.1111/j.1365-2699.2007.01716.x

Sebek, P., Altman, J., Platek, M., Cizek, L., 2013. Is Active Management the Key to the Conservation of Saproxylic Biodiversity? Pollarding Promotes the Formation of Tree Hollows. PLOS ONE 8, e60456. https://doi.org/10.1371/journal.pone.0060456

Seibold, S., Bässler, C., Brandl, R., Gossner, M.M., Thorn, S., Ulyshen, M.D., Müller, J., 2015. Experimental studies of dead-wood biodiversity — A review identifying global gaps in knowledge. Biol. Conserv. 191, 139–149. https://doi.org/10.1016/j.biocon.2015.06.006

Seibold, S., Hagge, J., Müller, J., Gruppe, A., Brandl, R., Bässler, C., Thorn, S., 2018. Experiments with dead wood reveal the importance of dead branches in the canopy for saproxylic beetle conservation. For. Ecol. Manag. 409, 564–570. https://doi.org/10.1016/j.foreco.2017.11.052

Seibold, S., Thorn, S., 2018. The Importance of Dead-Wood Amount for Saproxylic Insects and How It Interacts with Dead-Wood Diversity and Other Habitat Factors, in: Ulyshen, M.D. (Ed.), Saproxylic Insects: Diversity, Ecology and Conservation, Zoological Monographs. Springer International Publishing, Cham, pp. 607–637. https://doi.org/10.1007/978-3-319-75937-1_18

Siitonen, J., 2001. Forest Management, Coarse Woody Debris and Saproxylic Organisms: Fennoscandian Boreal Forests as an Example. Ecol. Bull. 11–41.

Sileshi, G., Hailu, G., Nyadzi, G.I., 2009. Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. Ecol. Model. 220, 1764–1775. https://doi.org/10.1016/j.ecolmodel.2009.03.024

Sist, P., Mazzei, L., Blanc, L., Rutishauser, E., 2014. Large trees as key elements of carbon storage and dynamics after selective logging in the Eastern Amazon. For. Ecol. Manag. 318, 103–109.

Smith, A.N.H., Anderson, M.J., Millar, R.B., 2012. Incorporating the intraspecific occupancy–abundance relationship into zero-inflated models. Ecology 93, 2526–2532. https://doi.org/10.1890/12-0460.1

Snäll, T., Kindvall, O., Nilsson, J., Pärt, T., 2011. Evaluating citizen-based presence data for bird monitoring. Biol. Conserv. 144, 804–810. https://doi.org/10.1016/j.biocon.2010.11.010

Sólymos, P., Lele, S., Bayne, E., 2012. Conditional likelihood approach for analyzing single visit abundance survey data in the presence of zero inflation and detection error. Environmetrics 23, 197–205. https://doi.org/10.1002/env.1149

Song, Z., Seitz, S., Li, J., Goebes, P., Schmidt, K., Kühn, P., Shi, X., Scholten, T., 2019. Tree diversity reduced soil erosion by affecting tree canopy and biological soil crust development in a subtropical forest experiment. For. Ecol. Manag. 444, 69–77. https://doi.org/10.1016/j.foreco.2019.04.015

Speed, J., 1989. The Counties of Britain: A Tudor Atlas (Nicolson and Hawkyard, 1989). Pavilon, London, United Kingdom.

Speight, M., 1989. Saproxylic invertebrates and their conservation. Nature and environment. Strasbourg: Council of Europe.

Spencer, J.W., Kirby, K.J., 1992. An inventory of ancient woodland for England and Wales. Biol. Conserv. 62, 77–93. https://doi.org/10.1016/0006-3207(92)90929-H

Stagoll, K., Lindenmayer, D.B., Knight, E., Fischer, J., Manning, A.D., 2012. Large trees are keystone structures in urban parks. Conserv. Lett. 5, 115–122. https://doi.org/10.1111/j.1755-263X.2011.00216.x

Stahle, D., 1996. Tree rings ancient forest relics. Arnoldia 56, 2–10.

Stevenson, R. L., 1875-1876. Forest Note in Collected Essays. eBooks@Adelaide, Australia.

Stockwell L, D., Peterson, 1999. The GARP modelling system: problems and solutions to automated spatial prediction. Int. J. Geogr. Inf. Sci. 13, 143–158. https://doi.org/10.1080/136588199241391

Sullivan, B.L., Wood, C.L., Iliff, M.J., Bonney, R.E., Fink, D., Kelling, S., 2009. eBird: A citizen-based bird observation network in the biological sciences. Biol. Conserv. 142, 2282–2292.

Sutherland, J., 2012. Error analysis of Ordnance Survey map tidelines, UK. Proc. ICE - Marit. Eng. 165, 189–197. https://doi.org/10.1680/maen.2011.10

Svensson, M., Johansson, V., Dahlberg, A., Frisch, A., Thor, G., Ranius, T., 2016. The relative importance of stand and dead wood types for wood-dependent lichens in managed boreal forests. Fungal Ecol. 20, 166–174. https://doi.org/10.1016/j.funeco.2015.12.010

Syfert, M.M., Smith, M.J., Coomes, D.A., 2013. The Effects of Sampling Bias and Model Complexity on the Predictive Performance of MaxEnt Species Distribution Models. PLOS ONE 8, e55158. https://doi.org/10.1371/journal.pone.0055158

Teacher, A., Griffiths, D., Hodgson, D., Inger, R., 2013. Smartphones in ecology and evolution: a guide for the app-rehensive. 3, 5268–5278. Ecol. And Evol. https://doi.org/10.1002/ece3.888.

Thompson, C.G., Kim, R.S., Aloe, A.M., Becker, B.J., 2017. Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. Basic Appl. Soc. Psychol. 39, 81–90. https://doi.org/10.1080/01973533.2016.1277529

Threlfall, C.G., Law, B., Banks, P.B., 2012. Influence of Landscape Structure and Human Modifications on Insect Biomass and Bat Foraging Activity in an Urban Landscape. PLOS ONE 7, e38800.

Thuiller, W., Brotons, L., Araújo, M.B., Lavorel, S., 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27, 165–172. https://doi.org/10.1111/j.0906-7590.2004.03673.x

Thuiller, W., Lavorel, S., Sykes, M.T., Araújo, M.B., 2006. Using niche-based modelling to assess the impact of climate change on tree functional diversity in Europe. Divers. Distrib. 12, 49–60. https://doi.org/10.1111/j.1366-9516.2006.00216.x

Tiago, P., Ceia-Hasse, A., Marques, T.A., Capinha, C., Pereira, H.M., 2017a. Spatial distribution of citizen science casuistic observations for different taxonomic groups. Sci. Rep. 7, 1–9. https://doi.org/10.1038/s41598-017-13130-8

Tiago, P., Pereira, H.M., Capinha, C., 2017b. Using citizen science data to estimate climatic niches and species distributions. Basic Appl. Ecol. 20, 75–85. https://doi.org/10.1016/j.baae.2017.04.001

Troll, C., 1968. Landschaftsökologie. In Pflanzensoziologie und Landschaftsökologie, pp. 1–21. Edited by R. Tüxen. Dr W. Junk Publishers, The Hague.

Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J., Martin, T.G., 2013. Realising the full potential of citizen science monitoring programs. Biol. Conserv. 165, 128–138.

Turner, M.G., 1989. Landscape Ecology: The Effect of Pattern on Process. Annu. Rev. Ecol. Syst. 20, 171–197. https://doi.org/10.1146/annurev.es.20.110189.001131

Turner-Skoff, J.B., Cavender, N., 2019. The benefits of trees for livable and sustainable communities. PLANTS PEOPLE PLANET 1, 323–335. https://doi.org/10.1002/ppp3.39

Tweddle, J.C., Robinson, L.D., Pocock, M.J.O., Roy, H.E., 2012. Guide to citizen science: developing, implementing and evaluating citizen science to study biodiversity and the environment in the UK. NERC/Centre for Ecology & Hydrology, Wallingford.

Václavík, T., Meentemeyer, R.K., 2012. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Divers. Distrib. 18, 73–83. https://doi.org/10.1111/j.1472-4642.2011.00854.x

Vailshery, L.S., Jaganmohan, M., Nagendra, H., 2013. Effect of street trees on microclimate and air pollution in a tropical city. Urban For. Urban Green. 12, 408–415. https://doi.org/10.1016/j.ufug.2013.03.002

Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37, 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Veloz, S.D., 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J. Biogeogr. 36, 2290–2299. https://doi.org/10.1111/j.1365-2699.2009.02174.x

Venables, W., Ripley, B., 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Visser, F., 2014. Rapid mapping of urban development from historic Ordnance Survey maps: An application for pluvial flood risk in Worcester. J. Maps 10, 276–288. https://doi.org/10.1080/17445647.2014.893847

Vollering, J., Halvorsen, R., Mazzoni, S., 2019. The MIAmaxent R package: Variable transformation and model selection for species distribution models. Ecol. Evol. 9, 12051–12068. https://doi.org/10.1002/ece3.5654

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57, 307 – 333.

Walz, U., 2011. Landscape Structure, Landscape Metrics and Biodiversity. Living Rev Landsc. Res 5.

Welsh, A.H., Cunningham, R.B., Donnelly, C.F., Lindenmayer, D.B., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. 88, 297–308.

Wenger, S.J., Freeman, M.C., 2008. Estimating Species Occurrence, Abundance, and Detection Probability Using Zero-Inflated Distributions. Ecology 89, 2953–2959. https://doi.org/10.1890/07-1127.1

Wenger, S.J., Olden, J.D., 2012. Assessing transferability of ecological models: an underappreciated aspect of statistical validation. Methods Ecol. Evol. 3, 260–267.

Westphal, M.I., Field, S.A., Tyre, A.J., Paton, D., Possingham, H.P., 2003. Effects of landscape pattern on bird species distribution in the Mt. Lofty Ranges, South Australia. Landsc. Ecol. 18, 413–426. https://doi.org/10.1023/A:1026115807529

White, J., 1998. Estimating the Age of Large and Veteran Trees in Britain. Forestry Practice. Information Note FCIN 12.

Williams, M.R., Yates, C.J., Stock, W.D., Barrett, G.W., Finn, H.C., 2016. Citizen science monitoring reveals a significant, ongoing decline of the Endangered Carnaby's black-cockatoo Calyptorhynchus latirostris. Oryx 50, 626–635. https://doi.org/10.1017/S0030605315000320

Williamson, T., Barnes, G., Pillatt, T., 2017. Trees in England:Management and disease since 1600. University of Hertfordshire Press.

Wilson, K.A., McBride, M.F., Bode, M., Possingham, H.P., 2006. Prioritizing global conservation efforts. Nature 440, 337–340. https://doi.org/10.1038/nature04366

Wisz, M.S., Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecol. 9, 8. https://doi.org/10.1186/1472-6785-9-8

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H., Guisan, A., NCEAS Predicting Species Distributions Working Group, 2008. Effects of sample size on the performance of species distribution models. Divers. Distrib. 14, 763–773. https://doi.org/10.1111/j.1472-4642.2008.00482.x

Woodland Trust, 2001. Position statement on Ancient Trees. Woodland Trust, Grantham, UK. https://www.woodlandtrust.org.uk/publications/2013/09/ancient-trees-parkland-and-wood-pasture/

Woodland Trust, 2017. Trees outside woods: in contributing to the ecological connectivity and functioning of landscapes. Grantham.

Woodland Trust, undated. Reference to Birnam Oak. https://www.woodlandtrust.org.uk/visiting-woods/woods/murthly-and-strathbraan-birnam-walk/

Wu, J., Shen, W., Sun, W., Tueller, P.T., 2002. Empirical patterns of the effects of changing scale on landscape metrics. Landsc. Ecol. 17, 761–782. https://doi.org/10.1023/A:1022995922992

Yackulic, C.B., Chandler, R., Zipkin, E.F., Royle, J.A., Nichols, J.D., Campbell Grant, E.H., Veran, S., 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods Ecol. Evol. 4, 236–243. https://doi.org/10.1111/2041-210x.12004

Yu, H., Cooper, A.R., Infante, D.M., 2020. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. Ecol. Model. 432, 109202. https://doi.org/10.1016/j.ecolmodel.2020.109202

Yuan, Y., Zeng, G., Liang, J., Li, X., Li, Z., Zhang, C., Huang, L., Lai, X., Lu, L., Wu, H., Yu, X., 2014. Effects of landscape structure, habitat and human disturbance on birds: A case study in East Dongting Lake wetland. Ecol. Eng. 67, 67–75. https://doi.org/10.1016/j.ecoleng.2014.03.012

Yunyun, H., Xingang, K., Junhui, Z., 2009. Variable Relationship between Tree Age and Diameter at Breast Height for Natural Forests in Changbai Mountains. J. Notheast For. Univ. 37, 38-42.

Zeileis, A., Hothorn, T., 2002. Diagnostic checking in regression relationships. R News 2, 3, 7-10. https://CRAN.R-project.org/doc/Rnews/

Zeileis, A., Kleiber, C., Jackman, S., 2008. Regression Models for Count Data in R. J. Stat. Softw. 27, 1–25. https://doi.org/10.18637/jss.v027.i08

Zhang, X., Cui, G., Liu, X., Zhang, Z., Xi, M., Li, J., Lu, J., 2017. The Characteristics of Ancient and Famous Trees in Qingdao City, Shandong Province, China and Possible Conservation Measures. Fresenius Environ. Bull. 26, 2016–2024.

Zuquim, G., Tuomisto, H., Jones, M.M., Prado, J., Figueiredo, F.O.G., Moulatlet, G.M., Costa, F.R.C., Quesada, C.A., Emilio, T., 2014. Predicting environmental gradients with fern species composition in Brazilian Amazonia. J. Veg. Sci. 25, 1195–1207. https://doi.org/10.1111/jvs.12174

Zurell, D., Zimmermann, N.E., Gross, H., Baltensweiler, A., Sattler, T., Wüest, R.O., 2020. Testing species assemblage predictions from stacked and joint species distribution models. J. Biogeogr. 47, 101–113. https://doi.org/10.1111/jbi.13608

Zuur, A., Ieno, E.N., Smith, G.M., 2007. Analyzing Ecological Data, Statistics for Biology and Health. Springer-Verlag, New York.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Zero-Truncated and Zero-Inflated Models for Count Data, in: Mixed Effects Models and Extensions in Ecology with R, Statistics for Biology and Health. Springer, New York, NY, pp. 261–293. https://doi.org/10.1007/978-0-387-87458-6_11

Appendices

A2: Appendix 2 – Additional tables and figures from Chapter 2

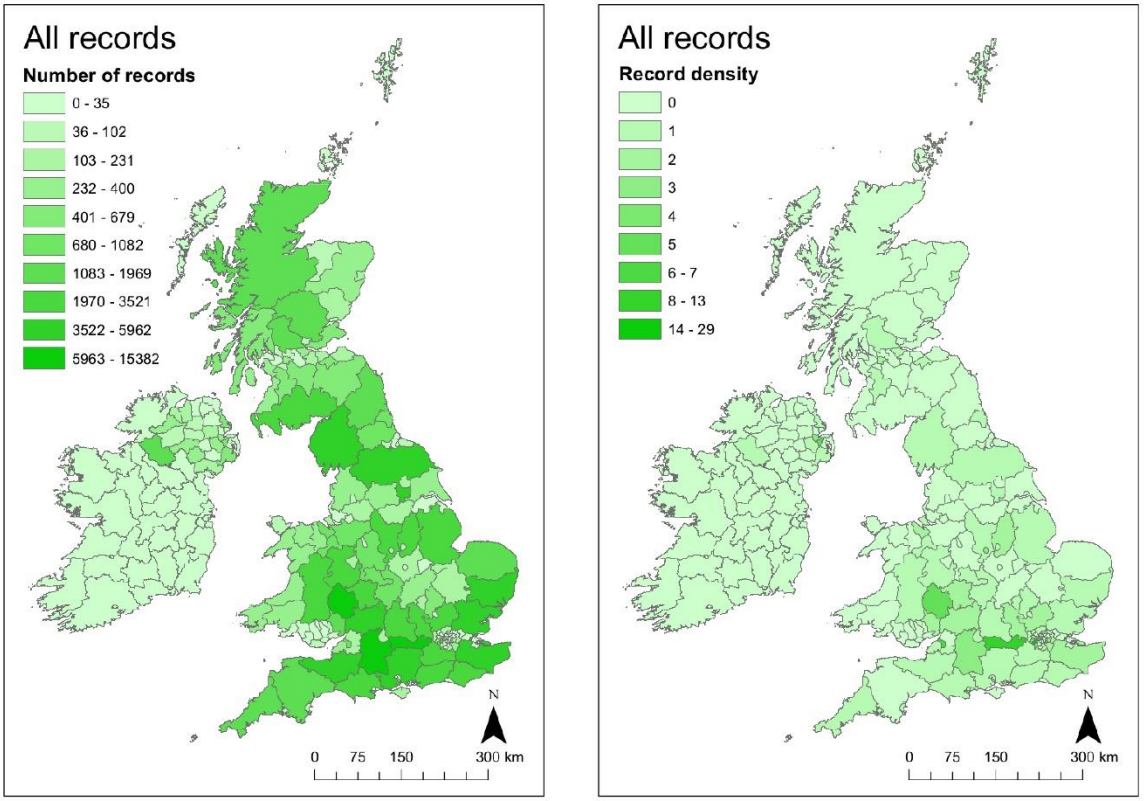


Fig. A2.1 *The number of records and the record density (number of records per km^2) in the ATI, shown for each county or unitary authority throughout England and Wales, council area in Scotland and Northern Ireland and administrative counties in the Republic of Ireland. Records that do not fall within any county boundary i.e. with incorrect grid references and records in Jersey, Guernsey or the Isle of Man are excluded.*

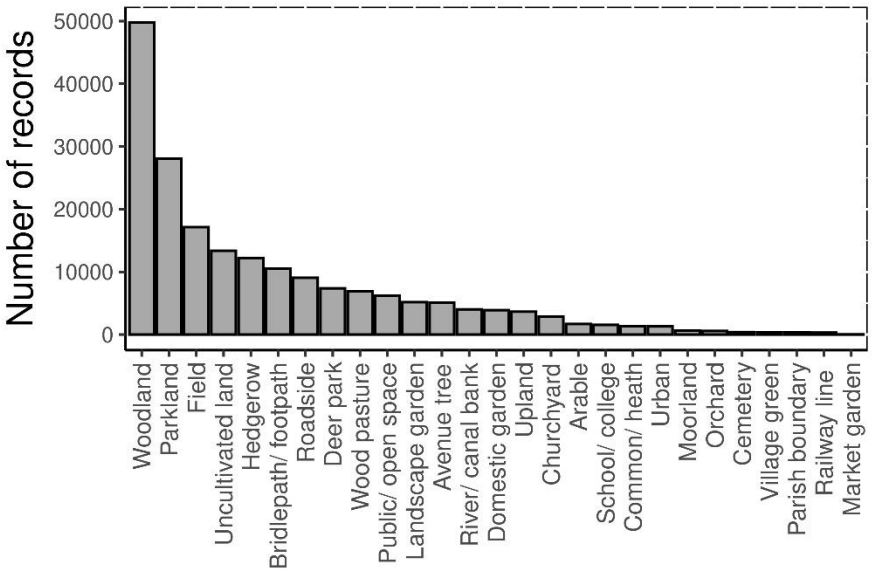


Fig. A2.2 *The frequency of records in the ATI with a particular habitat characteristic noted alongside that record*

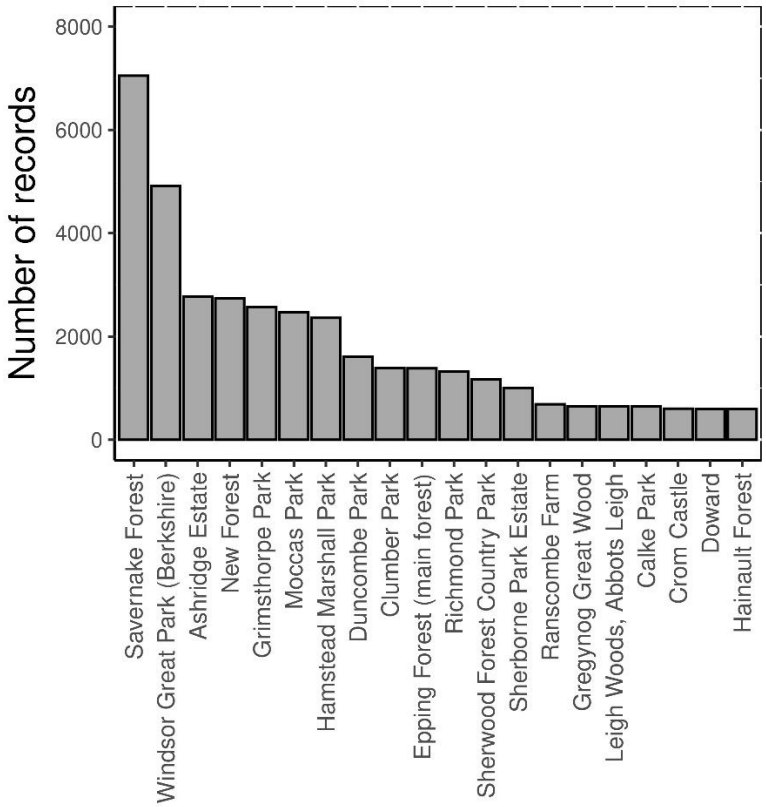


Fig. A2.3 *The 20 individual sites containing the highest number of records in the ATI. Forests, parks and large estates form the majority of sites with large numbers of records, as well as farms, castles and larger areas of public land e.g. the Dowards in Herefordshire (shown here as Doward).*

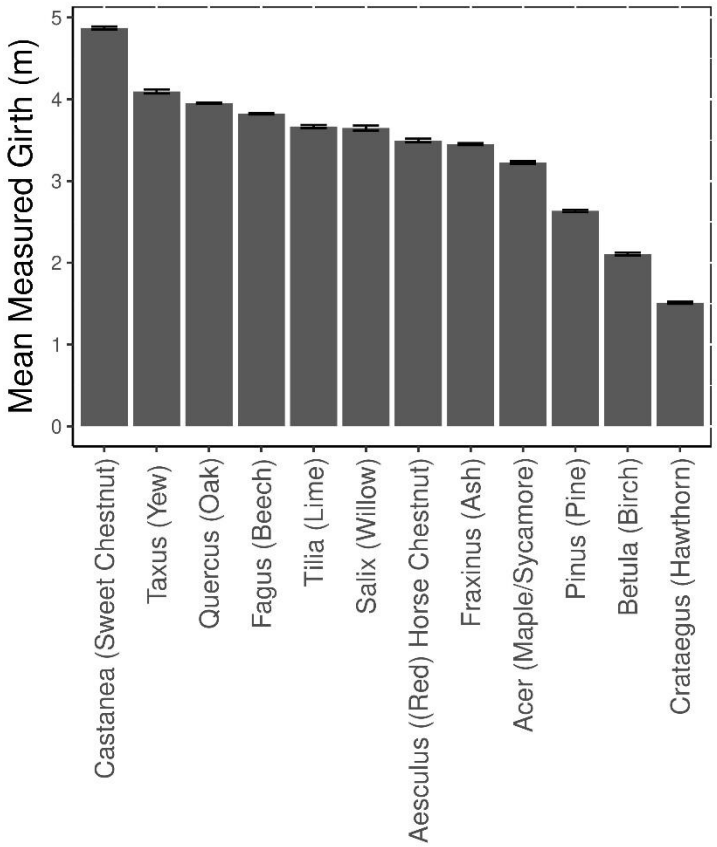


Fig. A2.4 *Mean measured girth (m) of the 12 most frequent genera of tree in the ATI.*

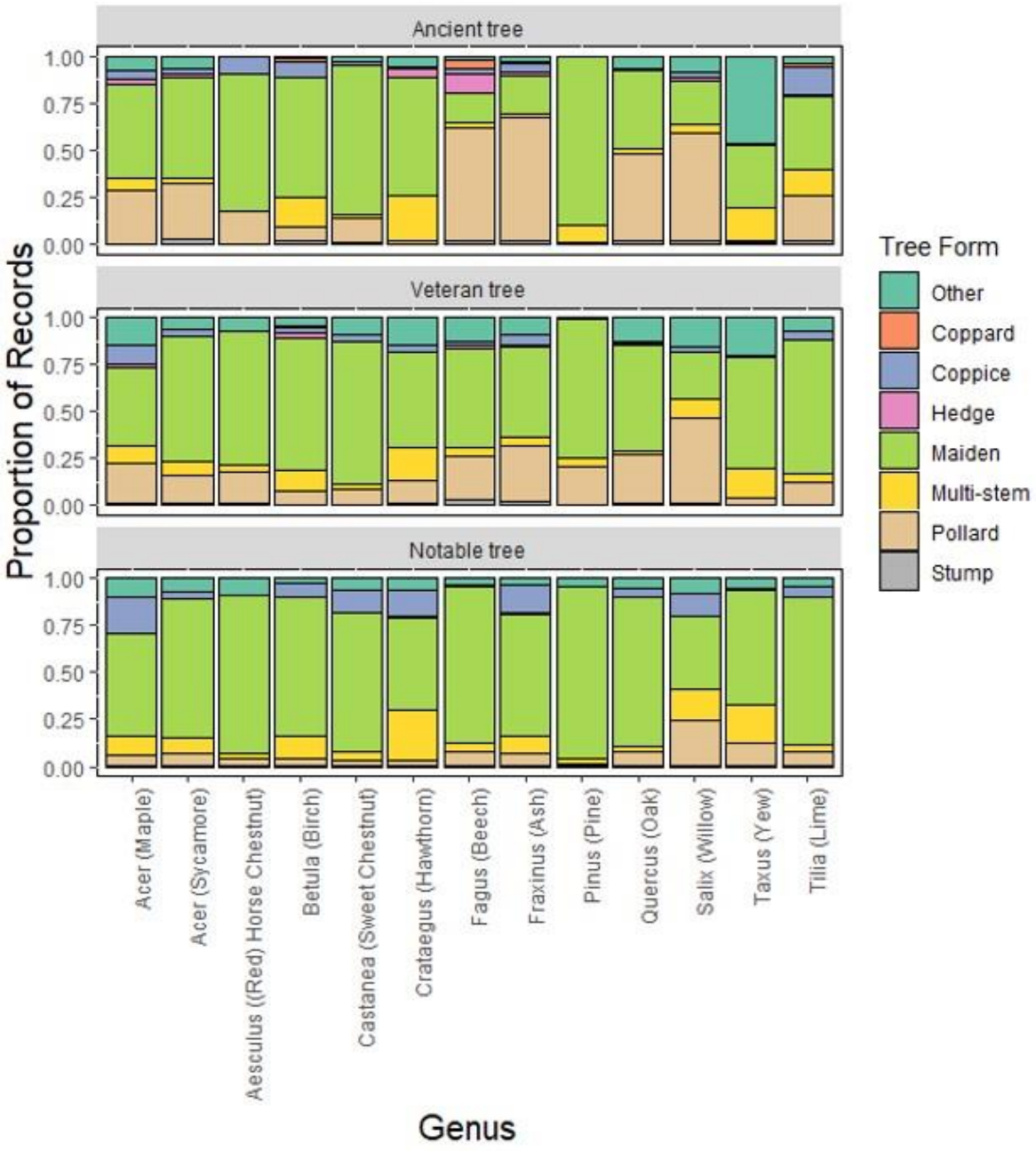


Fig. A2.5 *The relative proportion of ATI records for the 12 most common genera across three categories (ancient, veteran and notable), and eight tree forms.*

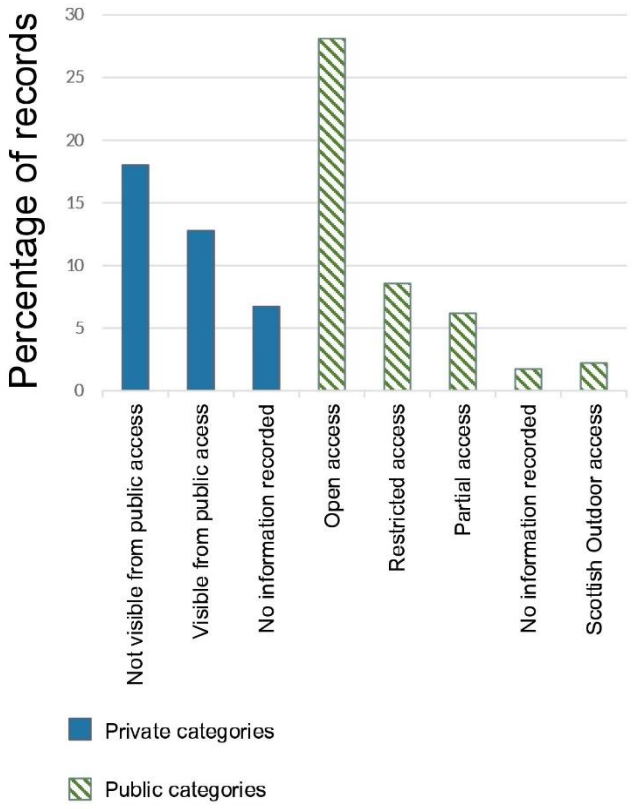


Fig. A2.6 *Percentage of records within each public or private accessibility categories in the ATI.*

Table A2.1*. Name and number of records of all fungi species recorded in the ATI.*

Common name	Latin name	Number of records
Southern bracket	*Ganoderma australe*	787
Beafsteak fungus	*Fistulina hepatica*	425
Oak bracket	*Inonotus dryadeus*	368
Chicken of the woods	*Laetiporus sulphureus*	202
Shaggy bracket	*Inonotus hispidus*	130
Dryad's sadle	*Polyporus squamosus*	57
Giant polypore	*Meripilus giganteus*	54
Birch polypore fungus	*Piptoporus betulinus*	51
Blushing bracket	*Daedaleopsis confragosa*	15
Dyer's mazegill	*Phaeolous schweinitzii*	10

Table A2.2 *A comprehensive list of all 211 taxa recorded in the ATI, along with the level of identification and number of records of each taxa.*

Taxa	No. of Records	Family	Genus	Species	Sub-species/ var.
Alder	2315	*Betulaceae*	*Alnus*	*glutinosa*	
Alder Buckthorn	3	*Rhamnaceae*	*Frangula*	*alnus*	
Apple	61	*Rosaceae*	*Malus*		
Ash	9,840	*Oleaceae*	*Fraxinus*		
Aspen	93	*Salicaceae*	*Populus*	*tremula*	
Atlas Cedar (Blue)	153	*Pinaceae*	*Decrus*	*atlantica*	
Austrian/ Corsican/ Black Pine	223	*Pinaceae*	*Pinus*	*nigra*	
Bay Willow	10	*Salicaceae*	*Salix*	*pentandra*	
Beech	14,296	*Fagaceae*	*Fagus*		
Bhutan Pine	10	*Pinaceae*	*Pinus*	*wallichiana*	
Birch	1,288	*Betulaceae*	*Betula*		
Bird Cherry	45	*Rosaceae*	*Prunus*	*padus*	
Black Mulberry	150	*Moraceae*	*Morus*	*nigra*	
Black Walnut	35	*Juglandaceae*	*Juglans*	*nigra*	
Blackthorn	26	*Rosaceae*	*Prunus*	*spinosa*	
Blue Gum	5	*Myrtaceae*	*Eucalyptus*		
Box	52	*Buxaceae*	*Buxus*	*sempervirens*	
Broad Leaved Lime var. Rubra	11	*Malvaceae*	*Tilia*	*platyphyllos*	*var. rubra*
Cappadocian Maple	16	*Sapindaceae*	*Acer*	*cappadocicum*	
Caucasian Elm	12	*Ulmaceae*	*Zelkova*	*carpinifolia*	
Caucasian/ Nordmann Fir	2	*Pinaceae*	*Abies*	*nordmanniana*	
Caucasian Wingnut	13	*Juglandaceae*	*Pterocarya*	*fraxinifolia*	
Cedar	218	*Pinaceae*	*Cedrus*		
Cedar of Lebanon	537	*Pinaceae*	*Cedrus*	*libani*	
Cherry	337	*Rosaceae*	*Prunus*		
Cherry Plum	14	*Rosaceae*	*Prunus*	*cerasifera*	
Chestnut leaved Oak	6	*Fagaceae*	*Quercus*	*castaneifolia*	
Chinese Juniper	2	*Cupressaceae*	*Juniperus*	*chinensis*	
Cider Gum	2	*Myrtaceae*	*Eucalyptus*	*gunnii*	
Coast Redwood	142	*Cupressaceae*	*Sequoia*	*sempervirens*	
Colorado White Fir	1	*Pinaceae*	*Abies*	*concolor*	
Common Ash	1,828	*Oleaceae*	*Fraxinus*	*excelsior*	
Common Beech	6,209	*Fagaceae*	*Fagus*	*sylvatica*	

Common Hawthorn	338	*Rosaceae*	*Crataegus*	*monogyna*	
Common Hornbeam	261	*Betulaceae*	*Carpinus*	*betulus*	
Common Juniper	374	*Cupressaceae*	*Juniperus*	*comnunis*	
Common Laburnum	7	*Fabaceae*	*Laburnum*	*anagryoides*	
Common Lime	3,374	*Malvaceae*	*Tilia*	*Tilia* × *europaea*	
Common Pear	120	*Rosaceae*	*Pyrus*	*communis*	
Common Walnut	136	*Juglandaceae*	*Juglans*	*regia*	
Common Whitebeam	101	*Rosaceae*	*Sorbus*	*aria*	
Common/ English Yew	3,028	*Taxaceae*	*Taxus*	*baccata*	
Copper/ Purple Beech	444	*Fagaceae*	*Fagus*	*sylvatica*	
Coral Bark Willow	2	*Salicaceae*	*Salix*	*alba*	*var. 'Vitellina'*
Cork Oak	26	*Fagaceae*	*Quercus*	*suber*	
Cornish Elm	3	*Ulmaceae*	*Ulmus*	*minor*	*var. 'Stricta'*
Cotoneaster	2	*Rosaceae*	*Cotoneaster*		
Crab Apple	708	*Rosaceae*	*Malus*	*slyvestris*	
Crack Wilow	574	*Salicaceae*	*Salix*	*fragilis*	
Cricket Bat Willow	16	*Salicaceae*	*Salix*	*alba*	*var. 'Caerulea'*
Crimean Pine	18	*Pinaceae*	*Pinus*	*nigra*	*subsp. Pallasiana*
Cypress	56	*Cupressaceae*			
Cypress Oak	2	*Fagaceae*	*Quercus*	*robus*	*var. 'Fastigiata'*
Dawn Redwood	33	*Cupressaceae*	*Metasequoia*		
Deodar cedar	88	*Pinaceae*	*Cedrus*	*deodara*	
Dogwood	3	*Cornaceae*	*Cornus*		
Douglas Fir	339	*Pinaceae*	*Pseudotsuga*	*menziesii*	
Dove Tree/ Handkerchief Tree	7	*Nyssaceae*	*Davidia*	*involucrata*	
Doward Whitebeam	2	*Rosaceae*	*Sorbus*	*eminentiformis*	
Downy Birch	387	*Betulaceae*	*Betula*	*pubescens*	
Dutch Elm	3	*Ulmaceae*	*Ulmus*		
Eastern Hemlock	7	*Pinaceae*	*Tsuga*	*canadensis*	
Elder	131	*Adoxaceae*	*Sambucus*	*nigra*	
Elm	533	*Ulmaceae*	*Ulmus*		
English Elm	97	*Ulmaceae*	*Ulmus*	*minor*	*var. 'Atinia'*
Eucalyptus	8	*Myrtaceae*	*Eucalyptus*		
European Larch	115	*Pinaceae*	*Larch*	*decidua*	
European Silver Fir	55	*Pinaceae*	*Abies*	*alba*	
European White Elm	12	*Ulmaceae*	*Ulmus*	*laevis*	
Evans Whitebeam	1	*Rosaceae*	*Sorbus*	*evansii*	
Exeter Elm	1	*Ulmaceae*	*Ulmus*	*Exoniensis*	

False Acacia	113	*Fabaceae*	*Robinia*	*pseudoacacia*	
Fern-leaved Beech	50	*Fagaceae*	*Fagus*	*sylvatica*	*var. 'Asplenifolia'*
Field Maple	2,195	*Sapindaceae*	*Acer*	*campestre*	
Fig	10	*Moraceae*	*Ficus*		
Fir	102	*Pinaceae*	*Abies*		
Giant Sequoia	1,332	*Cupressaceae*	*Sequoiadendron*	*giganteum*	
Ginkgo/ Maidenhair tree	62	*Ginkgoaceae*	*Ginkgo*	*biloba*	
Goat Willow/ Sallow	470	*Salicaceae*	*Salix*	*caprea*	
Golden Ash	3	*Oleaceae*	*Fraxinus*	*excelsior*	*var. 'Jaspidea'*
Golden Weeping Willow	26	*Salicaceae*	*Salix*	*Salix x Sepulcralis*	*var. 'Chrysocoma'*
Grand Fir	39	*Pinaceae*	*Abies*	*grandis*	
Grecian Fir	9	*Pinaceae*	*Abies*	*cephalonica*	
Grey Poplar	83	*Salicaceae*	*Populus*	*P. × canescens*	
Grey Willow	36	*Salicaceae*	*Salix*	*cinerea*	
Guelder Rose	4	*Adoxaceae*	*Virburnum*	*opulus*	
Hawthorn	2,756	*Rosaceae*	*Crataegus*		
Hazel	652	*Betulaceae*	*Corylus*		
Hemlock	8	*Apiaceae*	*Conium*	*maculatum*	
Herefordshire Whitebeam	1	*Rosaceae*	*Sorbus*	*herefordensis*	
Holly	1,467	*Aquifoliaceae*	*Ilex*		
Holm Oak	387	*Fagaceae*	*Quercus*	*ilex*	
Hornbeam	1,854	*Betulaceae*	*Carpinus*		
Horse Chestnut	2,952	*Sapindaceae*	*Aesculus*	*hippocastanum*	
Hungarian Oak	24	*Fagaceae*	*Quercus*	*frainetto*	
Huntingdon Elm	19	*Ulmaceae*	*Ulmus*	*Ulmus × hollandica*	*var. Major'*
Hybrid Black Poplar	91	*Salicaceae*	*Populus*	*Populus x canadensis*	
Hybrid Black Poplar Regenerata	29	*Salicaceae*	*Populus*	*Populus × canadensis*	*var. 'Regenerata'*
Hybrid Black Poplar Robusta	22	*Salicaceae*	*Populus*	*Populus × canadensis*	*var. 'Robusta'*
Hybrid Black Poplar Serotina	77	*Salicaceae*	*Populus*	*Populus × canadensis*	*var. 'Serotina'*
Hybrid Sessile and English Oak	407	*Fagaceae*	*Quercus*	*Q. × rosacea*	
Incense Cedar	18	*Cupressaceae*	*Calocedrus*	*decurrens*	

Indian Bean	41	*Bignoniaceae*	*Catalpa*	*bignonioides*	
Irish Yew	54	*Taxaceae*	*Taxus*	*baccata*	*var. 'Fastigiata'*
Italian Cypress	6	*Cupressaceae*	*Cupressus*	*sempervirens*	
Ivy	31	*Araliaceae*	*Hedera*		
Japanese Larch	2	*Pinaceae*	*Larix*	*kaempferi*	
Japanese Red Cedar	45	*Cupressaceae*	*Cryptomeria*	*Japonica*	*var. 'Elegans'*
Jeffrey Pine	3	*Pinaceae*	*Pinus*	*jeffreyi*	
Judas Tree	14	*Fabaceae*	*Cercis*	*siliquastrum*	
Juniper	58	*Cupressaceae*	*Juniperus*		
Laburnum	21	*Fabaceae*	*Laburnum*		
Larch	77	*Pinaceae*	*Larix*		
Large Leaved Lime	324	*Malvaceae*	*Tilia*	*platyphyllos*	
Lawson Cypress	98	*Cupressaceae*	*Chamaecyparis*	*lawsoniana*	
Lawson Cypress Erecta	2	*Cupressaceae*	*Chamaecyparis*	*lawsoniana*	*var. 'Erecta Viridis'*
Leylandii Leighton Green	1	*Cupressaceae*	*Cupressus*	*C. × leylandii*	*var. 'Leighton Green'*
Lime	2,276	*Malvaceae*	*Tilia*		
Liquidambar/ Sweetgum	15	*Altingiaceae*	*Liquidambar*		
Lombardy Poplar	60	*Salicaceae*	*Populus*	*nigra*	*var. 'Italica'*
London Plane	778	*Platanaceae*	*Platanus*	*P. × acerifolia*	
Low's Fir	2	*Pinaceae*	*Abies*	*concolor*	*subsp. 'lowiana'*
Lucombe Oak	144	*Fagaceae*	*Quercus*	*Q. x hispanica*	*var. 'Lucombeana'*
Manna Ash	10	*Oleaceae*	*Fraxinus*	*ornus*	
Maple	160	*Sapindaceae*	*Acer*		
Maritime Pine	30	*Pinaceae*	*Pinus*	*pinaster*	
Midland Hawthorn	15	*Rosaceae*	*Crataegus*	*laevigata*	
Mirbeck's Oak	6	*Fagaceae*	*Quercus*	*canariensis*	
Monkey Puzzle	165	*Araucariaceae*	*Araucaria*	*araucana*	
Monterey Cypress	86	*Cupressaceae*	*Cupressus*	*macrocarpa*	
Monterey Pine	149	*Pinaceae*	*Pinus*	*radiata*	
Morinda Spruce	5	*Pinaceae*	*Picea*	*smithiana*	
Mulberry	38	*Moraceae*	*Morus*		
Nikko Fir	1	*Pinaceae*	*Abies*	*homolepis*	
Noble Fir	48	*Pinaceae*	*Abies*	*procera*	
Nootka Cypress	6	*Cupressaceae*	*Cupressus*	*nootkatensis*	
Norway Maple	193	*Sapindaceae*	*Acer*	*platanoides*	
Norway Spruce	52	*Pinaceae*	*Picea*	*abies*	
Oak	40,336	*Fagaceae*	*Quercus*		

Orchard Apple	299	*Rosaceae*	*Malus*		
Oriental Plane	86	*Platanaceae*	*Platanus*	*orientalis*	
Oriental Spruce	9	*Pinaceae*	*Picea*	*orientalis*	
Osier	3	*Salicaceae*	*Salix*	*viminalis*	
Pear	174	*Rosaceae*	*Pyrus*		
Pedunculate Oak	29,204	*Fagaceae*	*Quercus*	*robur*	
Pin Oak	18	*Fagaceae*	*Quercus*	*palustris*	
Pine	155	*Pinaceae*	*Pinus*		
Plane	40	*Platanaceae*	*Platanus*		
Plantier's Poplar	9	*Salicaceae*	*Populus*	*nigra*	*var. 'Plantierensis'*
Plum	111	*Rosaceae*	*Prunus*		
Pomegranate	1	*Lythraceae*	*Punica*		
Pondersa Pine	2	*Pinaceae*	*Pinus*	*ponderosa*	
Poplar	311	*Salicaceae*	*Populus*		
Purging Buckthorn	10	*Rhamnaceae*	*Rhamnus*	*cathartica*	
Purple Sycamore	4	*Sapindaceae*	*Acer*	*pseudoplatanus*	*var. 'Purpureum'*
Red Horse Chestnut	24	*Sapindaceae*	*Aesculus*	*Aesculus* × *carnea*	
Red Oak	144	*Fagaceae*	*Quercus*	*rubra*	
Redwood	51	*Cupressaceae*			
Roble/ Southern Beech	4	*Nothofagaceae*	*Nothofagus*	*obliqua*	
Rowan/ Mountain Ash	887	*Rosaceae*	*Sorbus*		
Sapporo Autumn Gold Elm	7	*Ulmaceae*	*Ulmus*	*Davidiana var. japonica* × *pumila*	*var. 'Sapporo Autumn Gold'*
Sawara Cypress	5	*Cupressaceae*	*Chamaecyparis*	*pisifera*	
Scotch Laburnum	12	*Fabaceae*	*Laburnum*	*alpinum*	
Scots Pine	3,508	*Pinaceae*	*Pinus*	*sylvestris*	
Service Tree	15	*Rosaceae*	*Sorbus*	*torminalis*	
Service Tree of Fontainebleau	13	*Rosaceae*	*Sorbus*	*latifolia*	
Sessile Oak	3,909	*Fagaceae*	*Quercus*	*petraea*	
Silver Birch	957	*Betulaceae*	*Betula*	*pendula*	
Silver Lime	10	*Malvaceae*	*Tilia*	*tomentosa*	
Silver Maple	42	*Sapindaceae*	*Acer*	*saccharinum*	
Silver Pendant Lime	9	*Malvaceae*	*Tilia*	*tomentosa*	*var. 'Petiolaris'*
Single Leaved Ash	3	*Oleaceae*	*Fraxinus*	*anomala*	
Sitka Spruce	60	*Pinaceae*	*Picea*	*sitchensis*	
Small Leaved Lime	860	*Malvaceae*	*Tilia*	*cordata*	

Smooth Japanese Maple	2	*Sapindaceae*	*Acer*	*palmatum*	
Smooth-leaved Elm	154	*Ulmaceae*	*Ulmus*	*minor*	*minor*
Southern Beech	11	*Nothofagaceae*	*Nothofagus*		
Spindle	13	*Celastraceae*	*Euonymus*		
Spruce	22	*Pinaceae*	*Picea*		
Stone Pine	8	*Pinaceae*	*Pinus*	*pinea*	
Strawberry Tree	12	*Ericaceae*	*Arbutus*	*unedo*	
Swamp Cypress	35	*Cupressaceae*	*Taxodium*	*distichum*	
Sweet Chestnut	7,098	*Fagaceae*	*Castanea*	*sativa*	
Sycamore	4,149	*Sapindaceae*	*Acer*	*pseudoplatanus*	
Symonds Yat Whitebeam	1	*Rosaceae*	*Sorbus*	*saxicola*	
Tree of Heaven	9	*Simaroubaceae*	*Ailanthus*	*altissima*	
Tulip Tree	140	*Magnoliaceae*	*Liriodendron*		
Turkey Oak	562	*Fagaceae*	*Quercus*	*cerris*	
Variegated Sycamore	12	*Sapindaceae*	*Acer*	*pseudoplatanus*	*var. 'Simon-Louis Freres'*
Walnut	157	*Juglandaceae*	*Juglans*		
Wayfaring Tree	2	*Adoxaceae*	*Viburnum*	*lantana*	
Weeping Ash	3	*Oleaceae*	*Fraxinus*	*excelsior*	*var. 'Pendula'*
Weeping Beech	28	*Fagaceae*	*Fagus*	*sylvatica*	*var. 'Pendula'*
Western Hemlock	47	*Pinaceae*	*Tsuga*	*heterophylla*	
Western Red Cedar	129	*Cupressaceae*	*Thuja*	*plicata*	
Weymouth Pine	9	*Pinaceae*	*Pinus*	*strobus*	
Wheatley Elm	5	*Ulmaceae*	*Ulmus*	*minor*	*var. 'Sarniensis'*
White Poplar	32	*Salicaceae*	*Populus*	*alba*	
White Willow	352	*Salicaceae*	*Salix*	*alba*	
Whitebeam	167	*Rosaceae*	*Sorbus*	*aria*	
Wild Black Poplar	1,144	*Salicaceae*	*Populus*	*nigra*	
Wild Cherry	551	*Rosaceae*	*Prunus*	*avium*	
Wild Cherry Double Gean	17	*Rosaceae*	*Prunus*	*avium*	*var. 'Plena'*
Wild Pear	42	*Rosaceae*	*Pyrus*	*pyraster*	
Wild Service Tree	163	*Rosaceae*	*Sorbus*	*torminalis*	
Willow	911	*Salicaceae*	*Salix*		
Wingnut	4	*Juglandaceae*	*Pterocarya*		
Wych Elm	462	*Ulmaceae*	*Ulmus*	*glabra*	
Wych Elm var. Pendula	4	*Ulmaceae*	*Ulmus*	*glabra*	*var. 'Pendula'*
Yew	1505	*Taxaceae*	*Taxus*		
Zelkova	4	*Ulmaceae*	*Zelkova*		

Table A2.3*. Guide to the broad categories of land class and general included habitats. Adapted from 'Land Cover Map 2015 – Dataset documentation V.1.0 (CEH, 2017)'.*

Broad Land class	Habitats included
Broadleaf woodland	Broadleaved, mixed and yew woodland
Coniferous woodland	Coniferous woodland
Arable	Arable and horticulture
Improved grassland	Improved grassland
Semi-natural grassland	Neutral, calcareous and acid grassland. Fen, marsh and swamp.
Mountain, heath, bog	Dwarf shrub heath (heather and heather grassland), bog and inland rock
Saltwater	Saltwater
Freshwater	Freshwater
Coastal	Supra-littoral and littoral rock, supra-littoral and littoral sediment, saltmarsh
Built-up areas/ gardens	Urban and suburban

Table A2.4. *Guide to the historic types of countryside, as defined in Rackham, 1976 - Trees and Woodland in the British Landscape. Each countryside type is deemed to be mutually exclusive, although a distinction between Highland and Highland – Cornwall has been made, although the landscape is deemed to be similar.*

Countryside Type	Broad Description
Ancient	Lowland countryside. Hedged and walled landscape that can be traced back often to even the Bronze age. Fields are irregular and of varied origin with varied and thick hedgerows, hamlets, medieval farms, pollards and many ancient trees.
Planned	Lowland countryside. Regular fields, straight roads and small woods, derived following the Enclosure Acts in the 18th and 19th centuries. Features exposed buildings, thin hawthorn hedgerows, few roads and big villages. Medieval woods and ancient trees remain in places where they were failed to be destroyed after the enclosures.
Highland (including highland in Cornwall)	Coverage of moors, dales and mountains. Ancient woods are generally composed of Oak (*Quercus* spp.) and management declined earlier in these areas than in the lowland ancient or planned countryside.

Table A2.5. *Guide to the WRB Reference Soil Groups (RSG). Adapted from the 'World reference base for soil resources 2014: International soil classification system for naming soils and creating legends for soil maps (2015)'.*

Characteristic	**RSG**
Soils with thick organic layers	Histosols
Soils with strong human influence (e.g. intensive agriculture, containing artefacts)	Anthrosols, Technosols
Soils with limitations to Root Growth (e.g. permafrost, high concentration of soluble salts or Na, thin soil, alternate wet-dry conditions)	Cryosols, Leptosols, Solonetz, Vertisols, Solonchaks
Soils distinguished by Fe/Al Chemistry (e.g. stagnating water, presence of oxides or humus, accumulation of Fe etc.)	Gleysols, Andosols, Podzols, Plinthosols, Nitisols, Ferralsols, Planosols, Stagnosols
Accumulation of organic matter in mineral topsoil (e.g. dark topsoil, secondary carbonates etc.)	Chernozems, Kastanozems, Phaeozems, Umbrisols
Accumulation of moderately soluble salts or non-saline substances. (e.g. accumulation of secondary silica or carbonates)	Durisols, Gypsisols, Calcisols
Soils with clay-enriched subsoil	Retisols, Acrisols, Lixisols, Alisols, Luvisols
Soils with little or no profile differentiation (e.g. moderately developed, sandy, marine or sediments)	Cambisols, Arenosols, Fluvisols, Regosols

***Table A2.6**. Guide to the agricultural classes. Adapted from 'Agricultural Land Classification of England and Wales - Revised guidelines and criteria for grading the quality of agricultural land' (Ministry of Agriculture, Fisheries and Food, 1988).*

Agricultural Class	Broad Description
Grade 1	Excellent quality agricultural land
Grade 2	Very good quality agricultural land
Grade 3a	Good quality agricultural land
Grade 3b	Moderate quality agricultural land
Grade 4	Poor quality agricultural land
Grade 5	Very poor quality agricultural land
Urban	Housing, industry, commercial, education, transport etc.
Non-agricultural	Golf courses, parkland, sports fields, allotments etc.
Woodland	Commercial and non-commercial woodland
Agricultural buildings	Permanent agricultural buildings, glasshouses etc.
Land not surveyed	Agricultural land that has not been surveyed
Open water	Lakes, ponds and rivers

A3: Appendix 3 – Additional tables and figures from Chapter 3

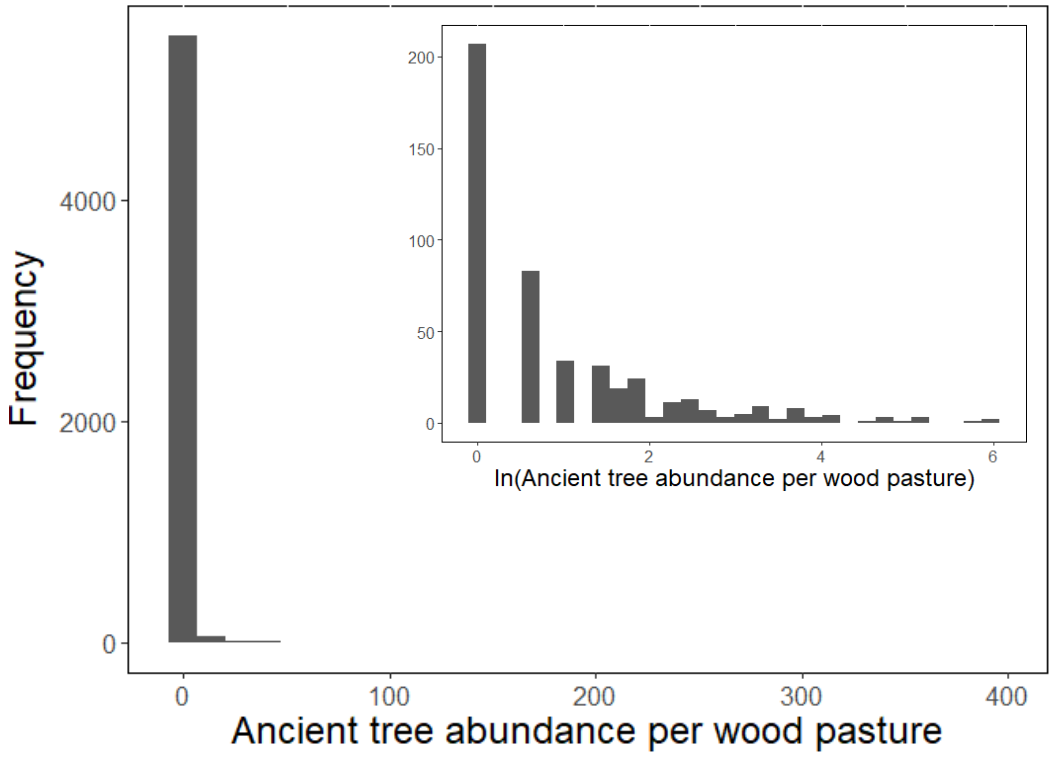


Fig. A3.1 *Histograms of ancient tree abundance (number of ancient trees per wood pasture) in England. Main: All wood pastures including wood pastures with and without ancient tree records. Ancient tree abundance ranges from 0 to 392, with 91.4% of wood pastures containing no ancient tree records. Inset: Only wood pastures that contain one or more ancient tree records are shown on a natural log (ln) scale.*

Table A3.1 *Model coefficients (and standard errors = SE), associated Z values and p values (p < 0.05:*, p < 0.01:**, p < 0.001:***) for the 'count' and 'zero' components of the ZIP and ZINB models of ancient tree abundance in wood-pastures across England.*

a. Reference category is 'Countryside type - Ancient countryside'

b. Reference category is 'Soil Type – Clay'

c. Reference category is 'Land Classification - Arable

	ZIP				***ZINB***			
	Count		***Zero***		***Count***		***Zero***	
	Estimate (SE)	***Z***	***Estimate (SE)***	***Z***	***Estimate (SE)***	***Z***	***Estimate (SE)***	***Z***
Wood-pasture area (km^2)	0.163 (0.006)	26.24***	-1.068 (0.063)	-16.90***	0.705 (0.087)	8.124***	-2.875 (0.299)	-7.211***
Distance from nearest town (km)	0.210 (0.026)	8.201***	0.116 (0.079)	1.462	0.203 (0.1220	1.667	0.279 (0.195)	1.429
Distance from nearest city (km)	0.200 (0.023)	8.598***	-0.053 (0.070)	-0.760	0.216 (0.102)	2.114*	-0.070 (0.180)	-0.390
Distance from a royal forest (km)	-0.093 (0.024)	-3.880***	-0.114 (0.070)	-1.628	0.284 (0.121)	2.353*	0.183 (0.181)	1.011
Distance from a moated site (km)	-0.186 (0.039)	-4.836***	-0.101 (0.087)	-1.160	0.053 (0.170)	0.309	0.072 (0.279)	0.259
Distance from a Tudor deer park (km)	-0.214 (0.026)	-8.118***	0.105 (0.070)	1.498	-0.353 (0.090)	-3.924***	0.009 (0.182)	0.051
Distance from a medieval deer park (km)	-0.268 (0.029)	-9.330***	0.032 (0.074)	0.420	-0.028 (0.097)	-0.290	0.212 (0.175)	1.214
Distance from a commons (km)	-0.006 (0.019)	-0.297	0.084 (0.062)	1.350	-0.210 (0.088)	-2.393*	-0.106 (0.176)	-0.599
Cover of ancient woodland (%)	-0.021 (0.014)	-1.525	-0.256 (0.058)	-4.393***	0.050 (0.119)	0.423	-0.335 (0.183)	-1.834
Cover of traditional orchard (%)	-0.122 (0.061)	-2.008*	-0.191 (0.129)	-1.481	-0.093 (0.082)	-1.133	-1.062 (0.768)	-1.383
Cover of forest or woodland (%)	0.367 (0.031)	11.71***	0.540 (0.092)	5.886***	0.189 (0.187)	1.010	0.778 (0.230)	3.390***
Countryside type - Highland[a]	0.595 (0.152)	3.910***	0.294 (0.380)	0.774	0.105 (0.778)	0.135	-0.110 (1.540)	-0.072
Countryside type - Highland Cornwall[a]	0.351 (0.162)	2.172*	0.336 (0.380)	0.884	0.613 (0.782)	0.784	0.971 (1.612)	0.603
Countryside type - Planned[a]	0.622 (0.159)	3.917***	0.617 (0.389)	1.586	0.325 (0.790)	0.411	0.967 (1.614)	0.599
Soil Type – Fe/Al[b]	-0.695 (0.056)	-12.07***	0.570 (0.183)	3.110**	-0.545 (0.267)	-2.042*	1.127 (0.471)	2.391*
Soil Type – No Profile[b]	-0.042 (0.040)	-1.040	-0.094 (0.139)	-0.677	-0.204 (0.207)	-0.989	-0.103 (0.354)	-0.291
Soil Type – Limited Root Growth[b]	-0.683 (0.107)	-6.355***	-0.313 (0.282)	-1.112	0.157 (0.505)	0.311	0.254 (0.871)	0.292
Soil Type – Other[b]	0.655 (0.055)	11.975***	0.075 (0.267)	0.282	1.076 (0.413)	2.605**	1.293 (0.599)	2.157*
Distance from nearest major road (km)	-0.260 (0.041)	-6.358***	-0.031 (0.086)	-0.356	-0.241 (0.128)	-1.883	-0.193 (0.205)	-0.940
Land Classification – Broadleaved[c]	0.537 (0.065)	8.208***	-0.810 (0.234)	-3.465***	0.538 (0.377)	1.429	-1.096 (0.648)	-1.691
Land Classification – Other[c]	-2.250 (0.174)	-12.97***	-1.982 (0.457)	-4.337***	-0.545 (0.442)	-1.233	-3.083 (1.061)	-2.905**
Land Classification – Grassland[c]	0.147 (0.053)	2.755**	-0.589 (0.140)	-4.212***	0.120 (0.217)	0.552	-0.694 (0.349)	-1.985*
Land Classification – Urban[c]	-1.826 (0.139)	-13.14***	-0.459 (0.335)	-1.372	-1.289 (0.442)	-2.916**	-1.136 (0.734)	-1.549
Agricultural Classification – Agricultural	-0.921 (0.044)	-20.79***	-0.431 (0.212)	-2.032*	-0.722 (0.287)	-2.513*	-1.224 (0.512)	-2.389*
Altitude (m)	-0.220 (0.024)	-9.168***	0.056 (0.081)	0.683	-0.175 (0.120)	-1.466	-0.026 (0.209)	-0.123
Percent cover of buildings (%)	-1.321 (0.084)	-15.68***	-0.562 (0.148)	-3.792***	-0.141 (0.184)	-0.763	-0.023 (0.233)	-0.099
Minor road length per km^2	0.566 (0.059)	9.677***	0.447 (0.146)	3.062**	-0.836 (0.283)	-2.953**	0.060 (0.423)	0.143
National Trust owned land - TRUE	0.298 (0.038)	7.860***	-0.974 (0.192)	-5.065***	0.695 (0.288)	2.416*	-0.871 (0.691)	-1.260
Distance from nearest watercourse (km)	-0.074 (0.030)	-2.518*	0.027 (0.066)	0.411	-0.105 (0.144)	-0.729	-0.266 (0.285)	-0.934
Theta					-1.808 (0.092)	-19.67***		
Log Likelihood	-5327.016				-2282.361			
Number of parameters	60				61			
AIC	10774.03				4686.721			

Table A3.2. *Model coefficients (and standard errors = SE), associated Z values and p values (p < 0.05:*, p < 0.01:**, p < 0.001:***) for the 'count' and 'zero' components of the ZINB models of ancient tree abundance of the two most common genera of ancient tree (Quercus (Oak) and Fraxinus (Ash)) in wood-pastures across England.*

a. Reference category is 'Countryside type - Ancient countryside'

b. Reference category is 'Soil Type – Clay'

c. Reference category is 'Land Classification – Arable

	Zero				***Count***			
	Quercus		***Fraxinus***		***Quercus***		***Fraxinus***	
	Estimate (SE)	***Z***	***Estimate (SE)***	***Z***	***Estimate (SE)***	***Z***	***Estimate (SE)***	***Z***
Wood-pasture area (km^2)	-3.790 (0.900)	-4.210 ***	0.938 (0.531)	1.766	0.759 (0.111)	6.830 ***	1.523 (0.235)	6.515 ***
Distance from nearest town (km)	-0.111 (0.273)	-0.406	1.756 (1.350)	1.301	0.123 (0.143)	0.857	0.459 (0.376)	1.221
Distance from nearest city (km)	-0.158 (0.311)	-0.507	1.744 (1.104)	1.579	0.200 (0.142)	1.320	0.991 (0.265)	3.739 ***
Distance from a royal forest (km)	0.447 (0.309)	1.447	2.315 (1.953)	1.185	0.117 (0.162)	0.718	0.840 (0.487)	1.726
Distance from a moated site (km)	0.416 (0.383)	1.085	-0.795 (0.883)	-0.901	0.019 (0.175)	0.110	-0.015 (0.347)	-0.042
Distance from a Tudor deer park (km)	0.019 (0.261)	0.072	-0.736 (0.656)	-1.121	-0.189 (0.130)	-1.453	-0.339 (0.312)	-1.086
Distance from a medieval deer park (km)	0.098 (0.267)	0.366	-2.210 (1.513)	-1.461	0.028 (0.121)	0.235	-1.282 (0.405)	-3.164 **
Distance from a commons (km)	0.196 (0.294)	0.669	-0.185 (0.520)	-0.356	-0.073 (0.130)	-0.568	-0.253 (0.267)	-0.946
Cover of ancient woodland (%)	-0.634 (0.294)	-2.158 *	0.031 (0.328)	0.095	0.004 (0.120)	0.031	0.094 (0.191)	0.490
Cover of traditional orchard (%)	-0.934 (0.774)	-1.207	2.357 (1.222)	1.929	-0.161 (0.169)	-0.951	2.189 (0.692)	3.164 **
Cover of forest or woodland (%)	0.942 (0.357)	2.634 **	-0.134 (0.936)	-0.143	0.152 (0.203)	0.750	-0.803 (0.385)	-2.087 *
Countryside type - Highland[a]	2.177 (1.813)	1.201	-1.694 (3.865)	-0.438	0.602 (0.850)	0.709	0.450 (1.595)	0.282
Countryside type - Highland Cornwall[a]	1.764 (1.868)	0.945	2.018 (3.047)	0.662	0.489 (0.765)	0.639	2.075 (1.589)	1.306
Countryside type - Planned[a]	3.190 (1.947)	1.639	7.329 (3.283)	2.232 *	0.469 (0.810)	0.579	5.203 (2.103)	2.474 *
Soil Type – Fe/Al[b]	1.957 (0.844)	2.319 *	-0.625 (1.453)	-0.430	-0.627 (0.321)	-1.955	-0.882 (0.709)	-1.244
Soil Type – No Profile[b]	0.770 (0.533)	1.443	0.309 (0.961)	0.321	0.165 (0.258)	0.640	0.808 (0.565)	1.429
Soil Type – Limited Root Growth[b]	-3.011 (2.377)	-1.266	-7.461 (4.003)	-1.864	-1.705 (0.491)	-3.476 ***	-2.743 (1.168)	-2.348 *
Soil Type – Other[b]	2.652 (0.889)	2.981 **	-7.537 (3.743)	-2.013 *	1.298 (0.496)	2.819 **	-0.954 (1.209)	-0.789
Distance from nearest major road (km)	0.394 (0.441)	0.892	0.572 (0.478)	1.196	-0.009 (0.237)	-0.039	0.280 (0.252)	1.109
Land Classification – Broadleaved[c]	-1.206 (0.886)	-1.360	-1.467 (2.211)	-0.664	0.875 (0.430)	2.032 *	1.096 (0.911)	1.203
Land Classification – Other[c]	-4.364 (1.938)	-2.252 *	0.735 (2.351)	0.313	0.082 (0.523)	0.157	0.835 (1.511)	0.553
Land Classification – Grassland[c]	0.206 (0.533)	0.388	1.033 (1.012)	1.021	0.443 (0.274)	1.614	1.548 (0.563)	2.750 **
Land Classification – Urban[c]	-1.497 (1.123)	-1.333	0.772 (3.758)	0.206	-1.180 (0.531)	-2.224 *	-1.228 (1.645)	-0.747
Agricultural Classification – Agricultural	-2.118 (0.924)	-2.292 *	-5.156 (2.881)	-1.790	-0.580 (0.334)	-1.739	-2.393 (0.910)	-2.629 **
Altitude (m)	0.033 (0.314)	0.104	1.085 (0.519)	2.090 *	-0.554 (0.159)	-3.491 ***	1.191 (0.223)	5.251 ***
Percent cover of buildings (%)	-0.063 (0.372)	-0.169	4.810 (2.842)	1.692	-0.288 (0.219)	-1.315	1.208 (1.679)	0.719
Minor road length (km)	-0.555 (0.624)	-0.890	3.270 (2.830)	1.155	-1.120 (0.325)	-3.443 ***	-0.058 (1.409)	-0.041
National Trust owned land - TRUE	-0.581 (0.872)	-0.665	5.716 (2.504)	2.282 *	0.945 (0.356)	2.656 **	4.237 (0.946)	5.005 ***
Distance from nearest watercourse (km)	-0.919 (0.366)	-2.513 *	-2.352 (1.245)	-1.889	-0.308 (0.121)	-2.552 *	-0.913 (0.390)	-2.341 *

A5.1: Appendix 5.1 - Additional tables and figures from Chapter 5

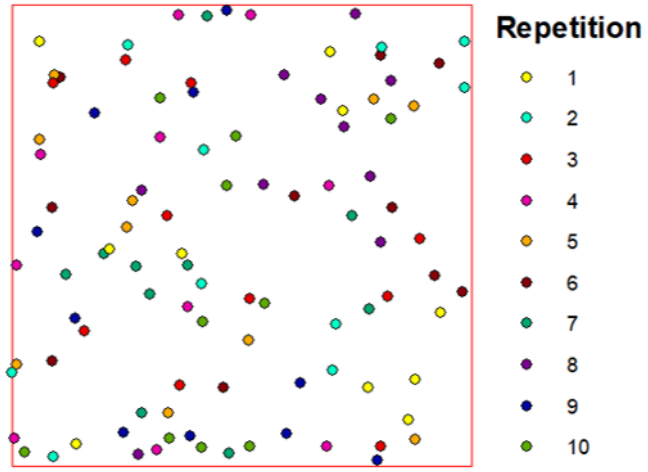


Fig. A5.1.1 *Spatial positions of the 10 randomly placed hypothetical 'town centres' across the simulation study area for each of the 10 simulation repetitions.*

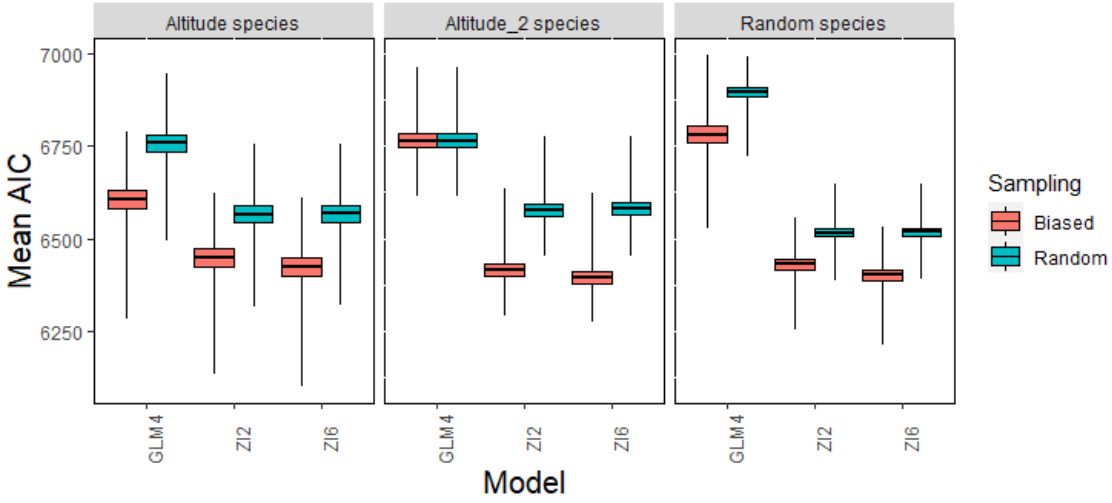


Fig. A5.1.2 *AIC evaluation of model performance for a non-zero-inflated generalised linear model (GLM4) including both the bias and biological predictor, and two zero-inflated models which either exclude (ZI2) or include (ZI6) the bias predictor in the zero component. Mean AIC values (± SE and data range) are shown across the 10 repetitions of randomly placed 'town centres' for three hypothetical simulated species: one species simulated randomly with no biological preferences (random species) and two with biological preferences of high altitude (altitude species and altitude_randomised (here termed altitude_2) species (no spatial autocorrelation)). Two different sampling strategies (random and biased) are considered.*

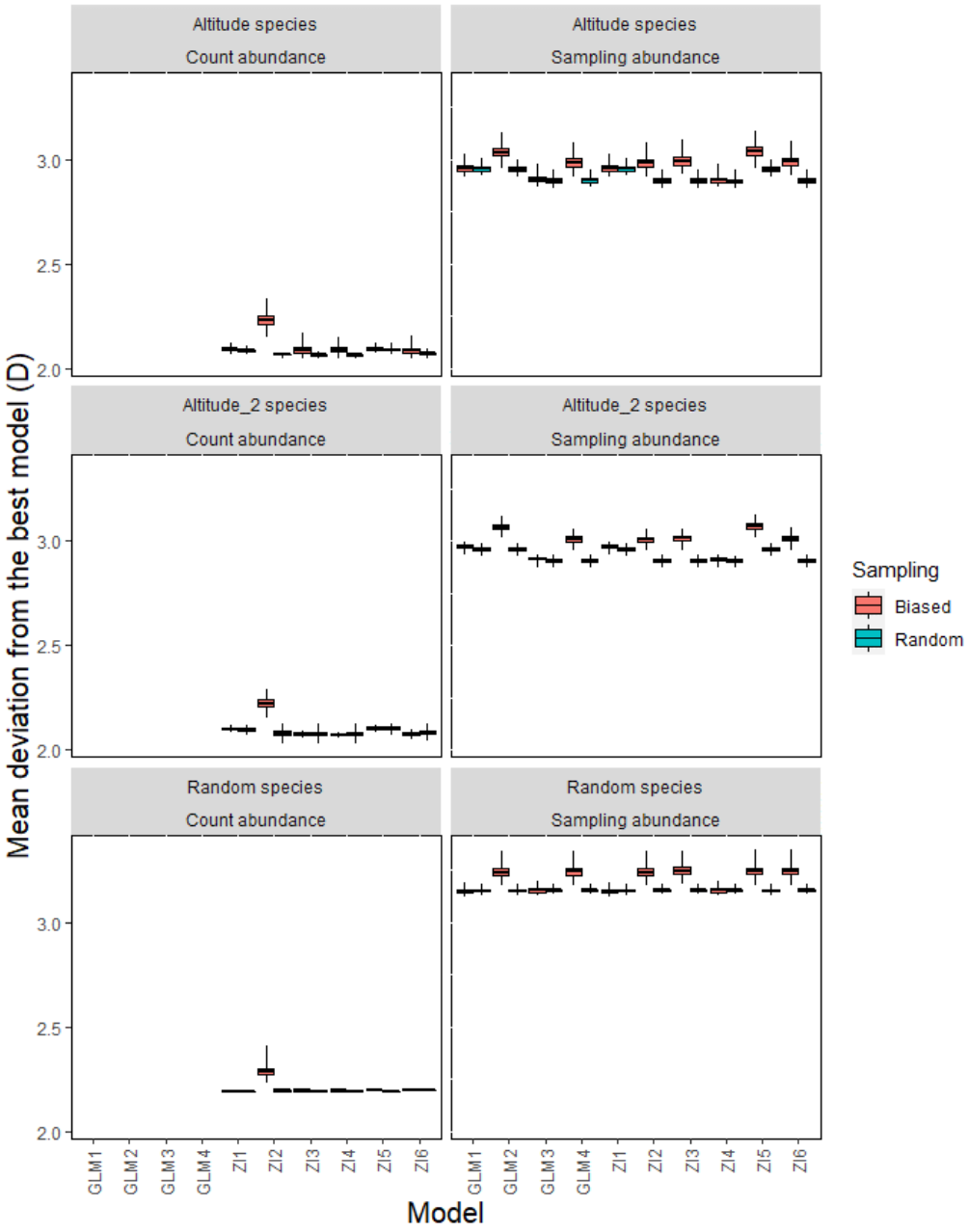


Fig. A5.1.3 *Evaluation of model predictions of abundance (based on D = 'deviation from the best model') for three hypothetical organisms (one with randomly simulated occurrences = random species, and two with occurrences simulated based on biological preferences = altitude species or altitude_randomised species (here termed altitude_2 species)). Mean D (± SE and data range) is shown for each sampling strategy (random or biased) across 10 different sets of hypothetical 'town centres' for each model. There are four non-zero-inflated generalised linear models, and six zero-inflated (ZI) models. For explanations of the structure of each model, see Tab. 3. Two types of prediction were evaluated: the count abundance predictions from the count component of the ZI models and the sampling abundance predictions from the whole of the ZI models or from the GLMs. Note the different scales on the vertical axes for the two types of predictions.*

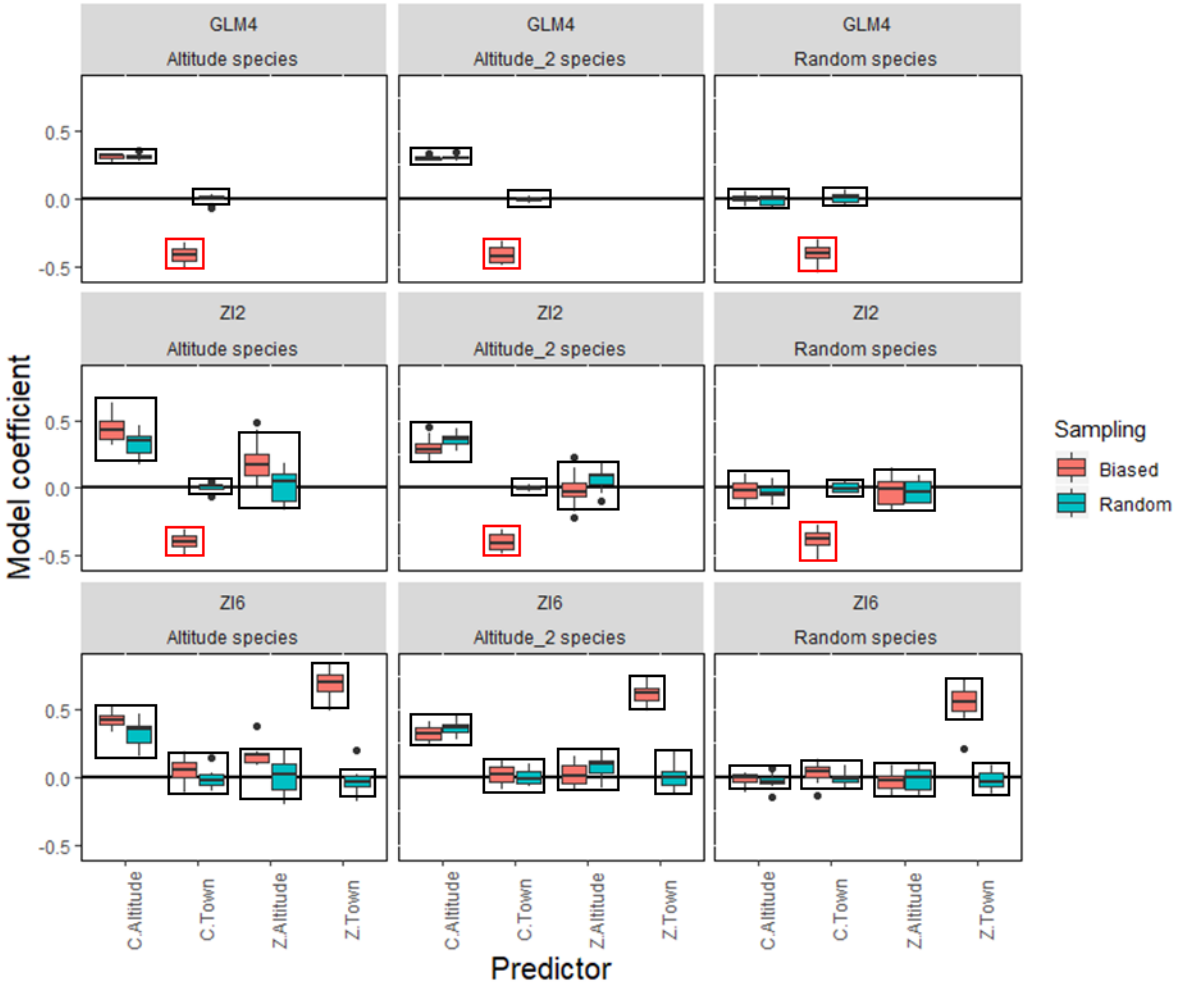


Fig. A5.1.4 *Model coefficients estimating the effects of a biological predictor (altitude or altitude_randomised (here termed altitude_2)) and a sampling bias predictor (distance to nearest hypothetical town) on the abundance of a hypothetical organism from a non-zero-inflated generalised linear model containing both the bias and biological predictor (GLM4), and two zero-inflated models which either exclude (ZI2) or include (ZI6) the bias predictor in the zero component. Zero-inflated (ZI) models include components which model both the count (C) of organisms per grid cell, and excess zeros (Z) caused by zero-inflation. For explanations of the structure of each model, see Tab. 3. Median model coefficients and range are shown for models fitted with data simulated using two different sampling strategies: random sampling and biased sampling. Results highlighted in red boxes indicate where the model is including the bias variable as a predictor of abundance where it should not. Black boxes are results that are correctly predicted.*

Altitude species

Predictor	Model	Sampling: Random						Sampling: Biased					
		True abundance		Sampling abundance/ occurrence		Zero predictions		True abundance		Sampling abundance/ occurrence		Zero predictions	
		r	se	r	se	r	se	r	se	r	se	r	se
Altitude	GLM3			0.999	0.000					0.999	0.000		
	GLM4			0.992	0.003					0.603	0.052		
	ZI2	0.979	0.007	0.993	0.003	0.183	0.254	0.738	0.029	0.623	0.050	0.822	0.082
	ZI6	0.944	0.013	0.993	0.003	0.071	0.198	0.960	0.012	0.607	0.053	0.187	0.072
Distance from town	GLM3			-0.037	0.071					-0.037	0.070		
	GLM4			-0.046	0.082					-0.781	0.019		
	ZI2	-0.054	0.081	-0.042	0.080	0.016	0.060	-0.648	0.039	-0.765	0.020	0.002	0.054
	ZI6	0.069	0.091	-0.044	0.081	0.262	0.152	0.069	0.086	-0.769	0.021	0.958	0.009

Altitude_2 species

Predictor	Model	Sampling: Random						Sampling: Biased					
		True abundance		Sampling abundance/ occurrence		Zero predictions		True abundance		Sampling abundance/ occurrence		Zero predictions	
		r	se	r	se	r	se	r	se	r	se	r	se
Altitude	GLM3			0.999	0.000					0.999	0.000		
	GLM4			0.985	0.007					0.576	0.020		
	ZI2	0.980	0.005	0.986	0.007	0.497	0.194	0.565	0.042	0.593	0.021	-0.289	0.238
	ZI6	0.944	0.011	0.986	0.007	0.414	0.159	0.957	0.013	0.584	0.022	0.045	0.046
Distance from town	GLM3			-0.002	0.003					-0.003	0.003		
	GLM4			-0.057	0.043					-0.787	0.017		
	ZI2	-0.044	0.035	-0.053	0.043	0.002	0.004	-0.781	0.032	-0.773	0.018	0.002	0.004
	ZI6	-0.025	0.085	-0.052	0.042	0.012	0.188	0.056	0.072	-0.774	0.020	0.980	0.006

Fig. A5.1.5 *Spearman's Rank correlation coefficients (r_s) between the model predictors (altitude/ altitude_randomised (here termed altitude_2) and distance from nearest town) and model predictions under two sampling strategies (random and biased). The top panel represents results for altitude species, whereas the bottom panel represents results for altitude_randomised species. These predictions are either abundance predictions from the whole model (shown for the generalised linear models (GLMs), sampling abundance predictions from the zero-inflated (ZI) models, count abundance predictions of true abundance (shown for the ZI models) and predictions of the probability an observation is an excess zero (shown for the ZI models). GLM3 and the zero component of ZI2 do not include the bias predictor, whereas GLM4 and the zero component of ZI6 do contain the bias predictor. Values represent the mean coefficients (including standard error (se)) across the 10 simulated sets of 'town centres' Coefficients are colour-coded based on strength: the darker the colour, the stronger the correlation. Red values represent positive correlations, whereas blue represent negative correlations.*

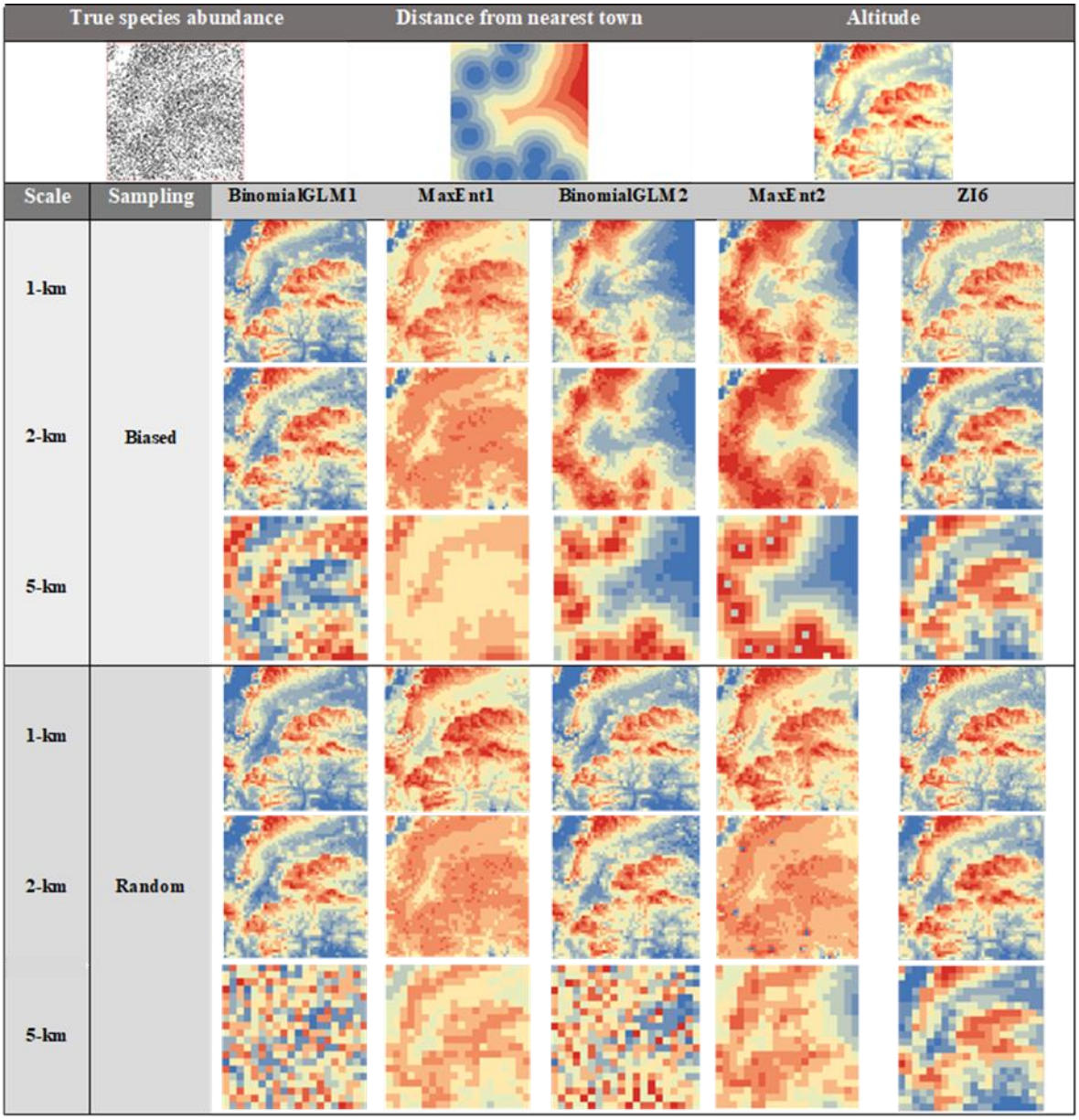


Fig. A5.1.6 *Example distribution maps for a hypothetical species whose occurrence is positively influenced by altitude (altitude species) from two binomial generalised linear models (GLMs) and two Maximum Entropy (MaxEnt) models. Maps are compared to maps of predicted abundance produced using the count abundance predictions (from the count component only) from a zero-inflated (ZI) model (ZI6) which inclues the bias in predictor in both components of the model. BinomialGLM1 and MaxEnt1 include only the biological predictor of altitude, whereas BinomialGLM2 and MaxEnt2 also include the bias predictor of distance from the nearest town. Unlike the zero-inflated (ZI) model, only one prediction can be obtained from the whole model and therefore will contain influences of sampling bias if present. Models were built with either data collected by random or biased sampling. Individual cells are colour coded based on abundance for the ZI abundance predictions or on probability of presence for the binomial GLM and MaxEnt predictions (high = red, low = blue).*

Table A5.1.1 *Number of grid cells (at 1km^2 resolution) across the study area above each altitude threshold (m) used for Simulation 2.*

Threshold (m)	Number of cells above threshold
0	10,000
50	9,068
100	5,329
125	3,364
150	1,993
175	1,036
200	396

A5.2: Appendix 5.2 - Simulation methods and results using average temperature as an alternative biological predictor.

Methods

Simulation 1 (Accuracy of species distribution maps from ZI models) was repeated using an alternative biological predictor to altitude - average temperature in ^{o}C across the study area between 1970-2000 obtained from WorldClim (WorldClim, accessed 10/05/18) at a 30-second resolution, and then converted to a 1km^2 resolution. Following the protocol of Simulation 1, a species with 5000 occurrence points was simulated across the study area based on the temperature layer converted to a probability layer using a logarithmic scale; the species was simulated to prefer higher average temperatures (Fig. A5.2.1). The same bias predictor of distance to nearest town centre was used, and the simulation was again repeated 10 times for each set of town centres. All of the model structures in Table 5.3 were used. Model predictive power was assessed using 'deviation from the best model' ('*D*').

Results

Results from the alternative run of Simulation 1 with the species preferring high temperatures echo those using altitude, in that the count abundance predictions provide the most accurate estimates (according to the metric '*D*') of true species abundance (Fig. A5.2.2). Again, the GLMs and the ZI sampling abundance predictions perform poorly in comparison and are unable to capture the effect of sampling bias or model successfully the excess zeros. Of the ZI models, all with the exception of ZI2 (where the bias predictor is omitted from the zero component but included in the count component), are able to provide good estimates of true species abundance. The zero component of the ZI models is again able effectively to identify and model the sampling bias (Fig. A5.2.1). Although the correlations between distance from nearest town centre and average temperature are higher than for altitude (which is reflected in the zero component of the ZI6 models which include the bias predictor in this component), the ZI models are still able to produce accurate abundance maps using the count abundance predictions.

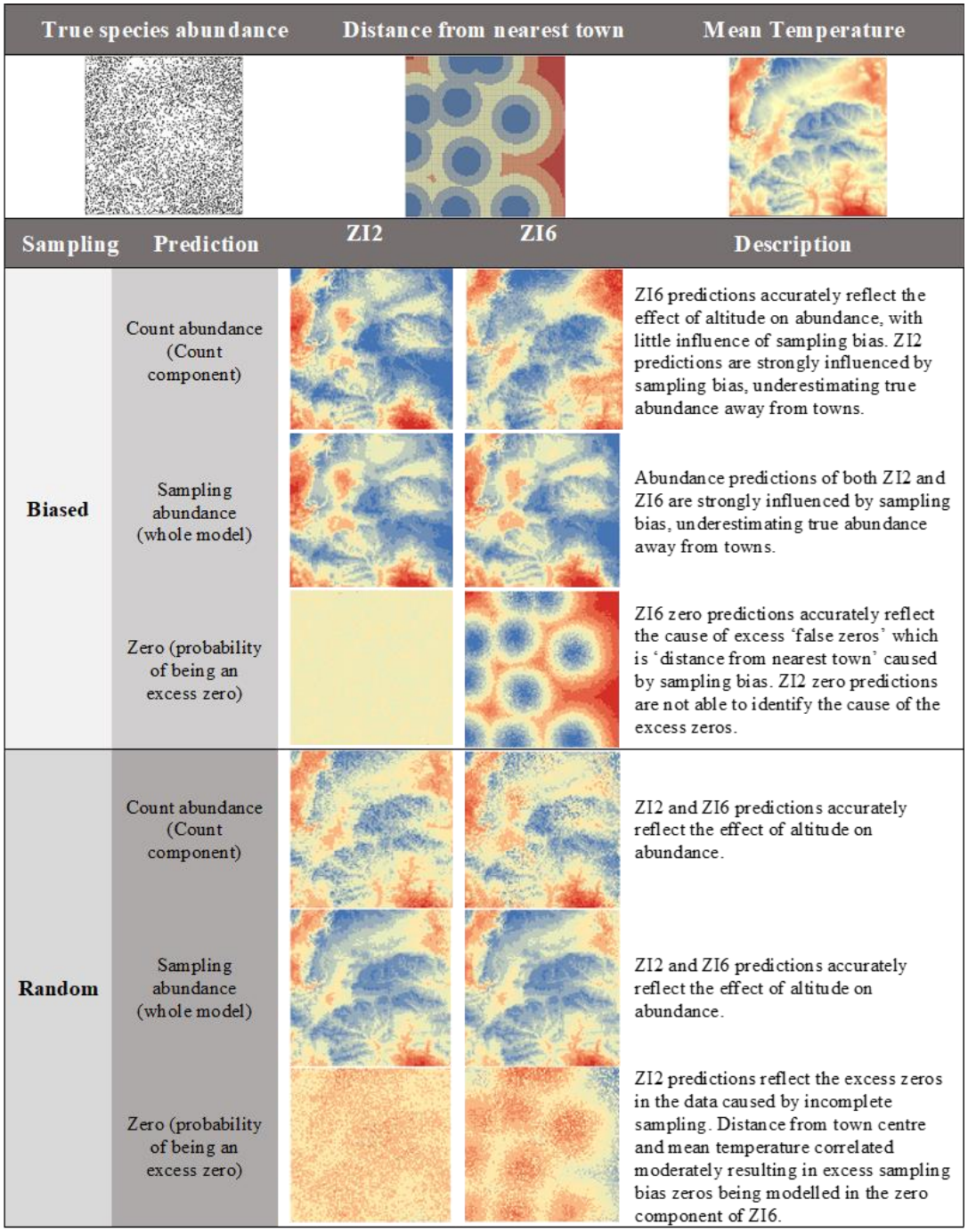


Fig. A5.2.1 *Example maps showing predicted abundance (count abundance and sampling abundance– see main text) and excess zeros (zero) for a hypothetical species whose occurrence is positively influenced by mean annual temperature, from two zero-inflated models (ZI2 and ZI6). Both models include a biological predictor (mean temperature) of both abundance and excess zeros, and a bias predictor (distance from the nearest town) as a predictor of abundance. ZI6 also includes distance from the nearest town as a predictor of excess zeros. Models were built with either data collected by randomly sampling grid cells (random) or with sampling bias (biased). Individual cells are colour coded based on abundance for the abundance predictions or on probability of being an excess zero for the zero predictions (high = red, low = blue).*

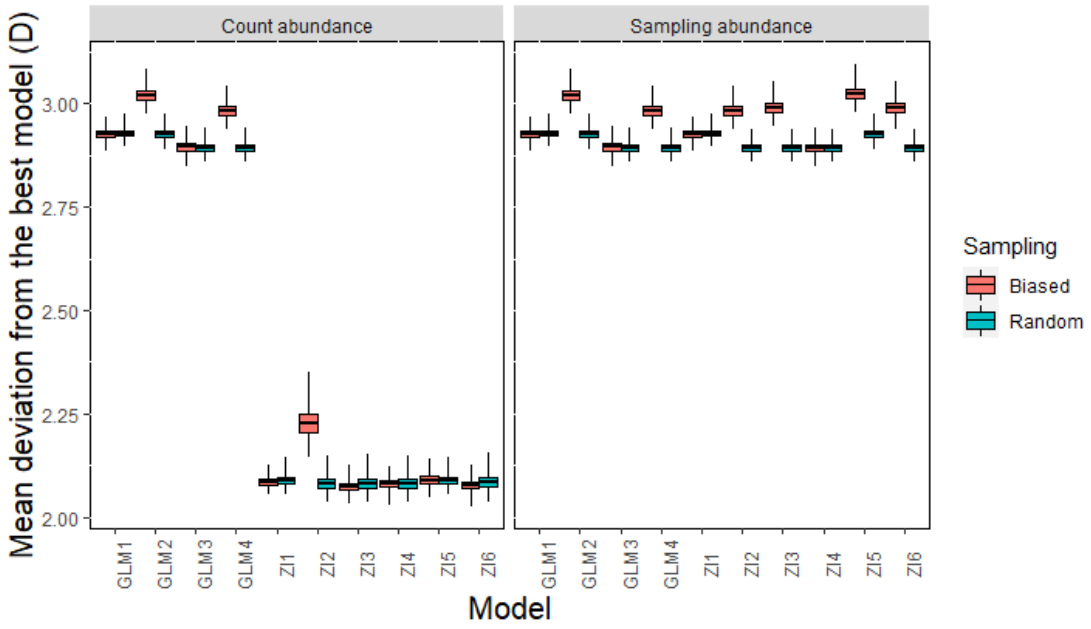


Fig. A5.2.2 *Evaluation of model predictions of abundance (based on D = 'deviation from the best model') for a hypothetical organism with a biological preference for warm temperatures. Mean D (± SE and data range) is shown for each sampling strategy (random or biased) across 10 different sets of hypothetical 'town centres' for each model. There are four non-zero-inflated generalised linear models, and six zero-inflated (ZI) models. For explanations of the structure of each model, see Tab. 3. Two types of prediction were evaluated: the count abundance predictions from the count component of the ZI models and the sampling abundance predictions from the whole of the ZI models or from the GLM*

A5.3: Appendix 5.3 - Derivation of D: 'Deviation from the best model'

Model predictions from Simulation 1 and Simulation 2 were evaluated using a novel metric derived from first principles that I named 'deviation from the best model' (*D*). This metric compares the probability of obtaining the true (raw) abundance (i.e. before sampling occurs) in each cell based on the model prediction, with the probability of obtaining a prediction equal to the true (raw) abundance i.e. predictions produced by a perfect model (Eq. 5.1). For each grid cell *i*, the probability of obtaining the true (raw) abundance (pA_i) was estimated from a Poisson probability distribution with a mean equal to the predicted mean abundance ($\bar{A}i$) for that cell. The summed natural logs of these probabilities across the study area represents the overall probability of obtaining the true (raw) abundances under the model predictions. This is then expressed as a ratio against the summed natural log probabilities for each cell ($q\bar{A}_i$) that would be obtained for a perfect model where the true (raw) abundance is equal to the predicted mean abundance.

Eq. 5.1
$$D = \frac{\sum(\ln(pA_i|\bar{A}_i))}{\sum(\ln(q\bar{A}_i|\bar{A}_i))}$$

The rationale behind creating a new evaluation metric is that the generation of occurrence points was based on a Poisson process and was deliberately zero-inflated, so there are an extremely large proportion of 0's and 1's, and the highest 'abundance count' is only six. Therefore, a metric based on the probability of obtaining the raw data from the model, rather than a direct assessment of the actual values, would provide a more appropriate measure of model performance and fit that was not as weighted by the large proportion of zeroes in the data. Using traditional binary presence-absence classification metrics would results in penalties against large predictions of abundance in comparison to the true raw abundance: the observed values of 0 and 1 which are most common in this dataset would score more highly using binary classification methods. The magnitude of the difference between prediction and true (raw) abundance observations will scale with the mean, leading to greater "inaccuracy" in cells with large numbers of individuals, regardless of the model used. Therefore, a metric that considers the actual abundance value rather than just presence-absence is likely to provide a more accurate assessment of model predictive power.

A6.1: Appendix 6.1 - Species distribution model parameter tuning and evaluation of alternative methods of splitting of training and test data.

Introduction

When fitting any species distribution model (SDM) the choice of parameters can highly influence the models accuracy and performance (Fourcade et al., 2018). Maximum Entropy modelling has a variety of variable parameters that can be used to tune the model and produce a model of best fit (Phillips et al., 2009). Two of these parameters, feature class (FC) and regularisation measure (RM) are among the most useful and allow optimisation between overfitting and goodness of fit (Muscarella et al., 2014; Fourcade et al., 2018). In addition, the method of splitting training and test data for model evaluation has been shown to have strong influences on the models performance (Wenger & Olden, 2012; Bahn & McGill, 2013). One of the most common methods involves selecting a random proportion of occurrence (and background) records, usually between 20-50% as a 'pseudo-independent' test data set (Fielding & Bell, 1997). Alternative methods involve splitting the data into k number of groups (k-fold cross validation) and using each group to subsequently act as the test data, and the other groups as the training set. However, these methods of random splitting have been critised for underestimating model error and being affected by spatial autocorrelation problems (Burnham & Anderson, 2003; Araújo et al., 2005). Instead, non-random, geographical splitting of the data may be more appropriate, and can test the extrapolation ability of the model (Radosavljevic & Anderson, 2014; Roberts et al., 2017). Initial analysis was carried out to evaluate the best method to split the test and training data, as well as the best tuning parameter combination of FCs and RMs for the baseline SDM of ancient and veteran trees across England with no bias correction method.

Methods and analysis

MaxEnt models of ancient and veteran tree distributions across England were tuned and fitted in R (R Core Team, 2018) using the 'ENMeval' package in R (Muscarella et al., 2014). Initial model tuning using combinations of FCs 'Linear (L)', 'Linear and Quadratic (LQ)', 'Linear, Quadratic and Product (LQP)' or 'Linear, Quadratic, Product, Threshold and Hinge (LQPTH)' and RMs of 0.5, 1, 2, 3, 4, and

5 was undertaken for the model with no bias correction method applied. Model predictive power was evaluated using three methods of splitting the data into training and test data. The first method involved geographic splitting of the data into four spatial blocks, from which one was randomly assigned as test data and the others as training data ('Block'). The second method was similar but split the data into a spatial checkerboard design ('Check'), dividing the area into bins at the resolution of the raster predictors. The final method used 10 fold cross validation ('Kfold'). The splitting was carried out 10 times, with a separate model run for each, resulting in a total of 720 models (four feature classes (FC) x six regularisation methods (RM) x 3 splitting methods x 10 repetitions). Model performance was evaluated using corrected Akaike information criterion (AICc) and 'Area Under the Curve' (AUC). Generalised Linear Mixed Models (GLMMs) were used to analyse significant differences between model performance (AICc and AUC) in relation to FC, RM and splitting methods. GLMMs were fitted in R using package 'lme4' (Bates et al., 2015) separately for training and test data specifying a Gaussian distribution, and included splitting method, FC and RM as fixed factors, and repetition run as a random factor. Backward selection based on AIC was used to find the most parsimonious model with the most influential predictors.

Results

Model performance and predictive power differed significantly across splitting method, FC and RM (Table A6.1.1). When considering each parameter separately, the most effective tuning parameters based on both mean AICc and AUC (train and test) were the 'Kfold' splitting method, FC 'LQ' and RM 5 (Fig. A6.1.1 & A6.1.2). However, when considering interactions between parameters, an increase in RM only had a significantly positive influence on model performance (AICc) across FCs 'LQP' or 'LQPTH', and had little effect on model with 'L' or 'LQ' FCs (Fig. A6.1.1). Therefore, the choice of RM when using either 'L' or 'LQ' FCs appears to be of little consequence, and the default version of 1 may be the best choice. Additionally, there was a significant interaction between splitting method and FC (Table A6.1), with significantly poorer model performances with FC 'LQP' or 'LQPTH', particularly for the 'Block' splitting method (Fig. A6.1.1). Therefore, based on AICc, the selection of

the best tuning parameters should be based on a 'Kfold' splitting method and FC 'LQ', with any RM. When considering AUC, all parameters and interactions had a significant influence on the model predictive power (Table A6.1.1). Again, the worst performing models used the 'Block' splitting method, 'LQP' and 'LQPTH' FCs and lower RMs), particularly when assessing the test data (Fig. A6.1.2).

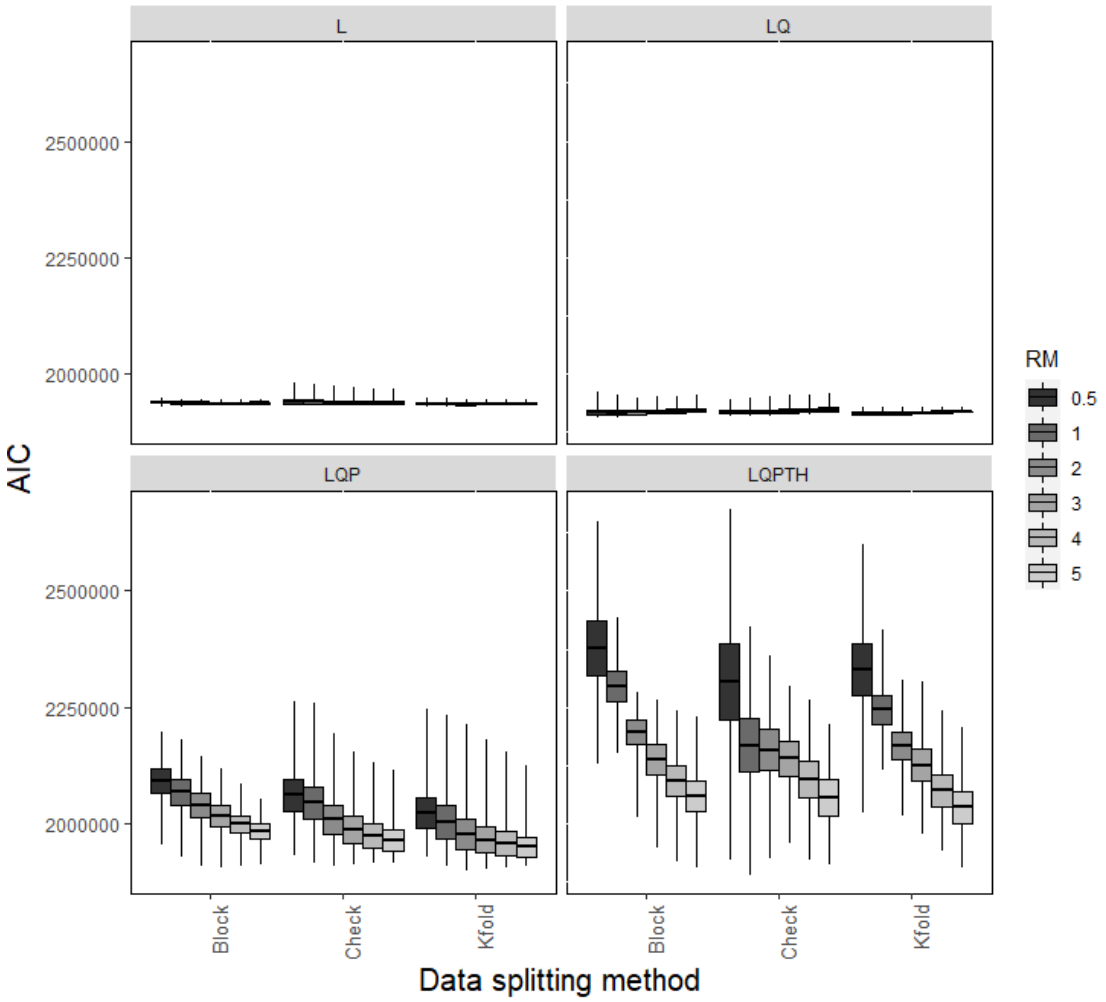


Fig. A6.1.1. *Corrected Akaike Information Criterion (AICc) for each of the model tuning combinations of splitting method into training and test data ('Block, 'Check' or 'Kfold'), feature class (FC) ('Linear (L)', 'Linear and Quadratic (LQ)', 'Linear, Quadratic and Product (LQP)' or 'Linear, Quadratic, Product, Threshold and Hinge (LQPTH)') and regularisation measure (RM) (0.5, 1, 2, 3, 4, 5). Mean values (±SE) are shown across the 10 repetitions of model fitting.*

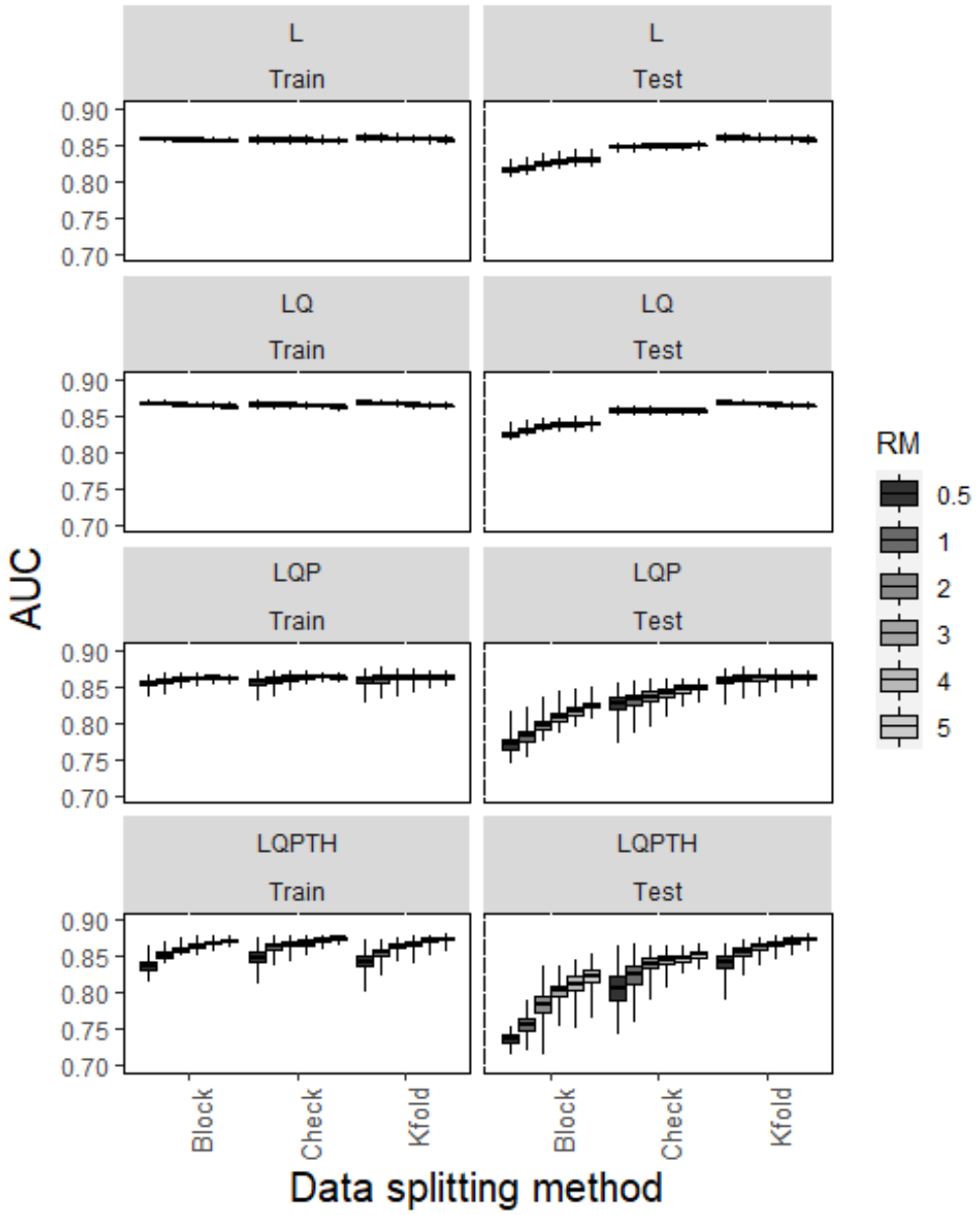


Fig. A6.1.2 *Area Under the Curve (AUC) for each of the model tuning combinations of splitting method into training and test data ('Block, 'Check' or 'Kfold'), feature class (FC) ('Linear (L)', 'Linear and Quadratic (LQ)', 'Linear, Quadratic and Product (LQP)' or 'Linear, Quadratic, Product, Threshold and Hinge (LQPTH)') and regularisation measure (RM) (0.5, 1, 2, 3, 4, 5). Mean values (±SE) are shown across the 10 repetitions of model fitting for both the training and test data set.*

Table A6.1.1 *Significance of tuning parameters in relation to model fitting (corrected Akaike Information Criterion: AICc) and model training and testing predictive power (Area Under the Curve: AUC). Parameters tested include the method of splitting data into training and testing sets ('Block', 'Check' or 'Kfold' methods), regularisation measures (RM) (0.5, 1, 2, 3, 4 and 5) and feature classes (FC) ('Linear (L)', 'Linear and Quadratic (LQ)', 'Linear, Quadratic and Product (LQP)' or 'Linear, Quadratic, Product, Threshold and Hinge (LQPTH)'). Results shown are based on a Type III analysis of variance (ANOVA) F test (with degrees of freedom) and significance levels of the p value (* < 0.05, ** < 0.01, *** < 0.001).*

	AICc	**AUCtrain**	**AUCtest**
RM	28.26 (5,635)***	20.18 (1,687)***	184.1 (1,687)***
FC	447.9 (3,635)***	19.77 (3,687)***	129.3 (3,687)***
Splitting method	5.915 (2,635)**	7.423 (2,687)***	358.5 (2,687)***
RM:FC	14.08 (15,635)***	21.26 (3,687)***	61.16 (3,687)***
RM: Splitting method	0.336 (10,635)	3.114 (2,687)*	48.33 (2,687)***
FC: Splitting method	2.399 (6,635)*	5.996 (6,687)***	28.97 (6,687)***
RM: FC: Splitting method	0.367 (30,635)	3.045 (6,687)**	7.450 (6,687)***

Conclusion

The choice of tuning parameters is an important step in model fitting, as well as the division of the training and test data for model evaluation. Choice of parameters is highly model specific and should be carried out before fitting and interpreting any SDM. In all cases, 'Kfold' data splitting was the most effective way to divide training and test data, regardless of any other parameter. Therefore, in all subsequent models of bias correction I have chosen to use this method. For the baseline model of ancient and veteran tree distributions with no bias correction, the combination of parameters which produce the model with both the highest performance and fit, as well as predictive power were using FC 'LQ' and RM 5, hence these parameters are the chosen ones for this model. For all other sampling bias corrected distribution models, models were fitted using all combinations of FC and RM as the best combination is likely to be highly variable across models. The best model for each bias correction method was then chosen based on AICc.

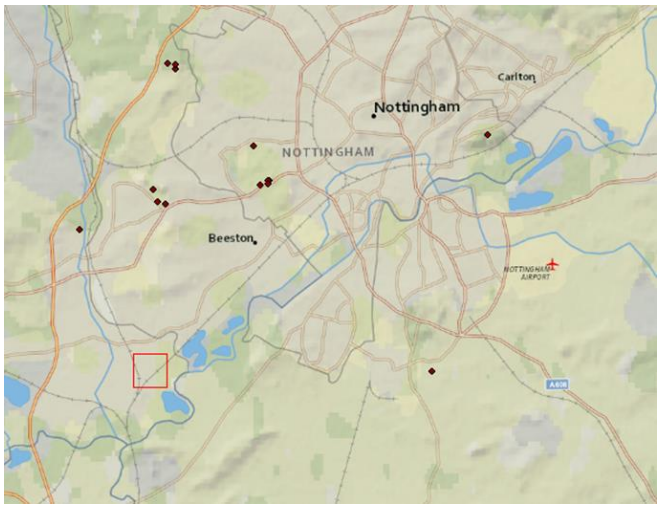
A6.2: Appendix 6.2 – Example survey form and instructions for field survey volunteers

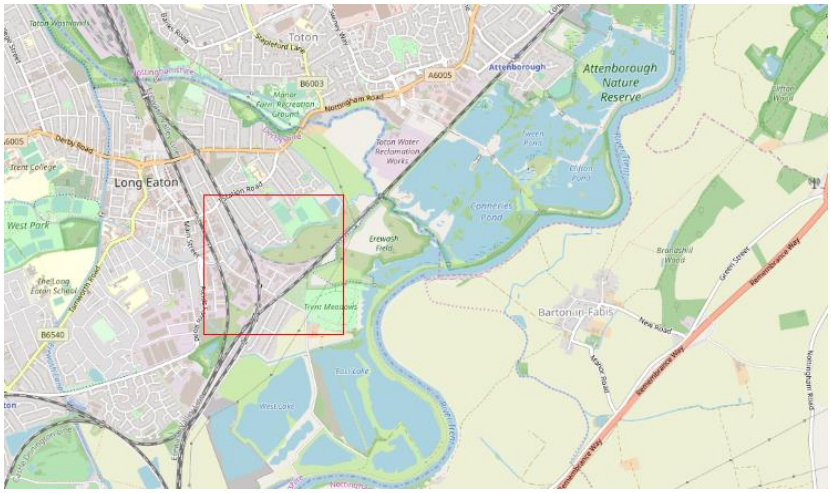
Please note this square was not used in the field model validation work, it was purely selected for trial purposes due to its location and to use as the example to inform the volunteers of the methods and types of areas they were required to survey. Therefore, not all parts of the square (priority areas or roads) were outlined here, as they would have been for a true verification square: only a sample of areas and roads were digitised for demonstration purposes.

A worked example: Sections completed by a volunteer are in RED

1.1 - Your Grid Square and maps.

Grid square ID: 37
Central Grid Reference (i.e. for the centre of the grid square): AB 12345 67890
Please see maps of this grid square below





Key

Orange dots = potential parking spaces/ areas

Green lines = paths or roads to survey

Blue lines = can take a quick look from afar if possible/ time permitting

Shaded areas = examples of potential priority areas

Please use these maps to mark on the areas (left) or roads (right) you have covered.

1.2 - Survey summary

Please complete this table and clearly and neatly as possible.

Area	Area Priority Rating	Time spent on area (min)	Estimated % cover of area	Number of ancient trees	Number of veteran trees
1	Survey whole area where possible. Priority area.	*30*	*80*	*0*	*1*
2	Survey whole area where possible. Priority area.	*45*	*95*	*1*	*2*
3	Survey whole area if accessible and possible	*No access*	*0*	*0*	*0*
4	Survey if time permits	*20*	*30*	*0*	*0*
5	Survey if time permits	*15*	*45*	*0*	*0*
6	Survey if time permits	*No time*	*0*	*0*	*0*
7	Survey if time permits	*No time*	*0*	*0*	*0*
8	Check accessibility and survey if time permits	*No access*	*0*	*0*	*0*
9	Check accessibility and survey if time permits	*No access*	*0*	*0*	*0*
…	…	…	…	…	…
TOTAL				*1*	*3*

1.3 – Tree Recording Form

Please record any ancient and veteran trees that you find using the table below.

Please make sure that you specify which "area number" each tree is in. This should correspond to the "area numbers" in the map in section 1.1.

Record as much information as you can.
However, if all you can record is the area number and veteran status (ancient / veteran) then this is fine for the purposes of this project.

(example text is shown in red)

Please complete this table and clearly and neatly as possible.

Area	Tree No.	Ancient/ Veteran	Species	Grid Reference	Photo (Y/N)	Comments
1	*1*	*Veteran*	*Oak (pedunculate)*	*AB XXXXX XXXXX*	*Y*	
2	*2*	*Ancient*	*Oak (pedunculate)*	*AB XXXXX XXXXX*	*Y*	
2	*3*	*Veteran*	*Beech*	*AB XXXXX XXXXX*	*Y*	
2	*4*	*Veteran*	*Willow*	*AB XXXXX XXXXX*	*N*	*Not close enough to take photo*

Please use this space for any other comments on the survey or trees…

No further comments

1.4 - Map of Approximate Tree Locations

During the survey, please plot the locations of any ancient and veteran trees that you find during the survey.

Please also make sure that there is a 10 figure grid reference for each tree that you record.

Please make sure that you label each tree e.g. 1, 2, 3, 4 etc. This number should correspond to the tree number on your tree recording form above.

A6.3: Appendix 6.3 – Additional figures from Chapter 6

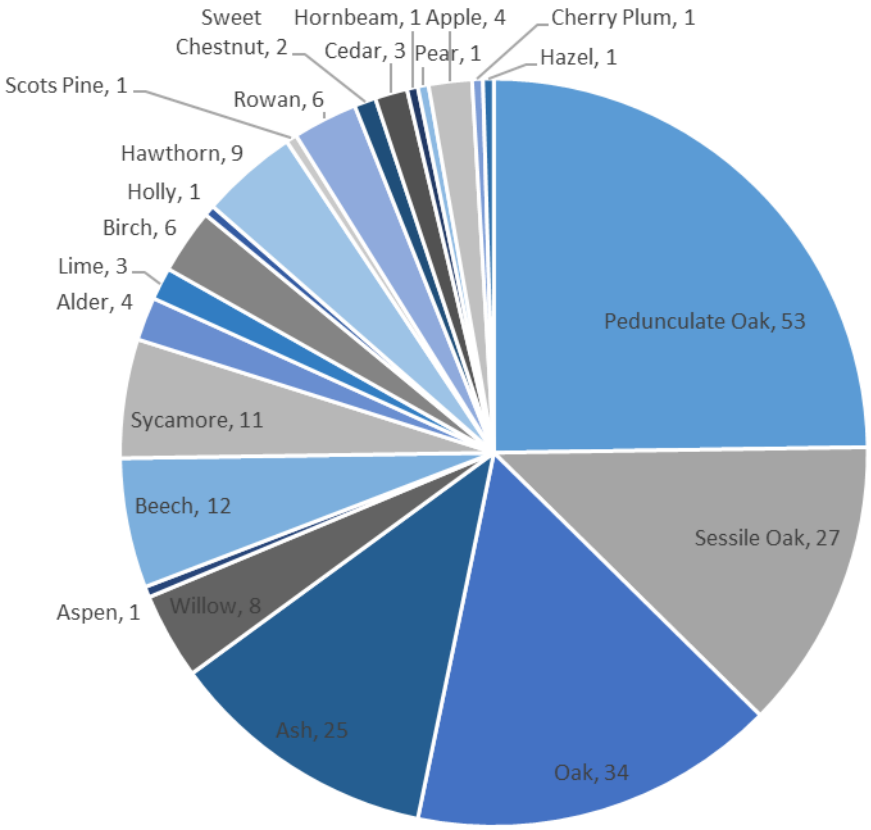


Fig. A6.3.1 *Number of each species/ genera (common names shown) of tree that was able to be identified out of the 52 surveyed grid squares from the field verification work.*

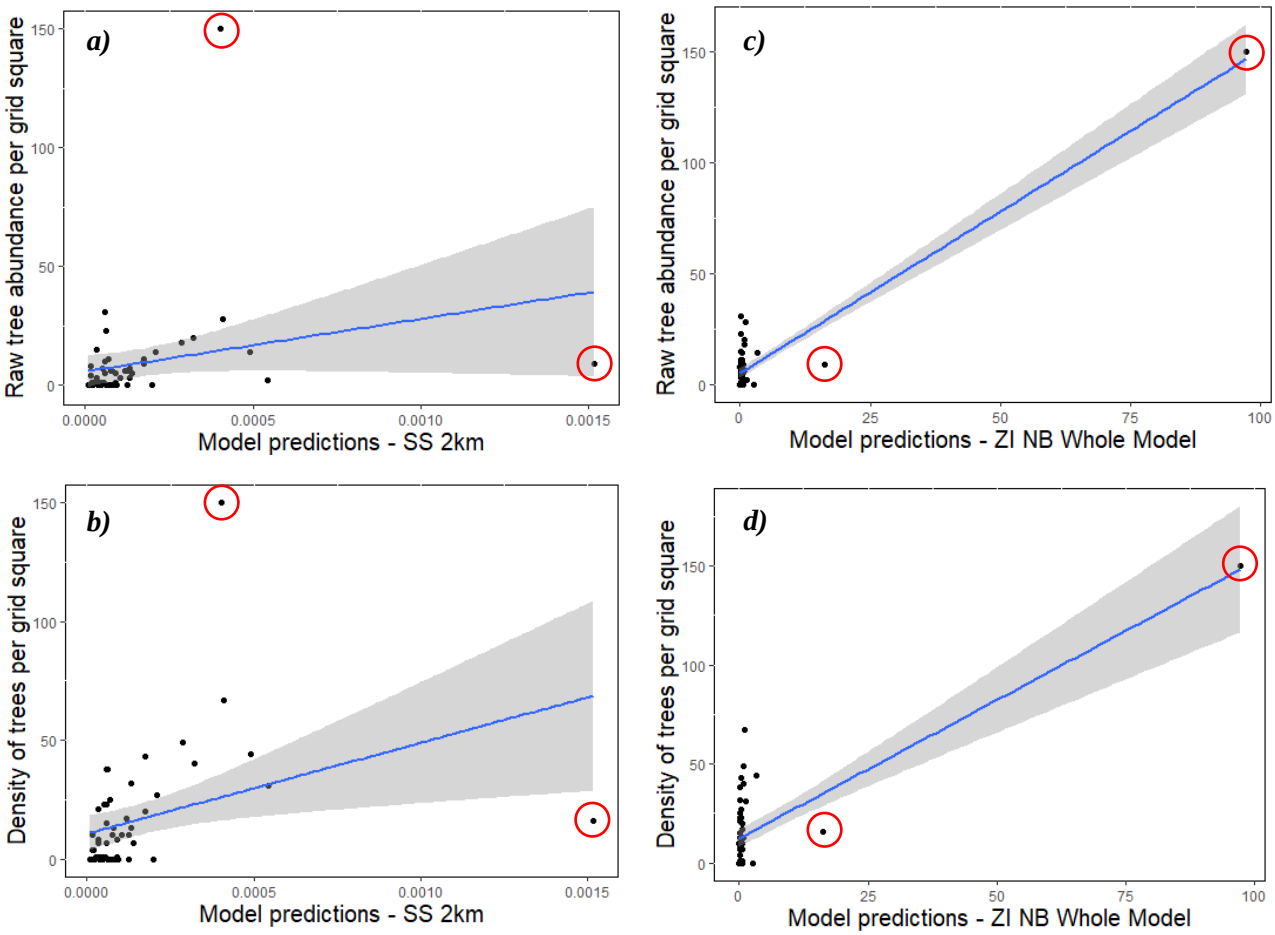


Fig. A6.3.2 *Scatterplots of model predictions (±SE) from the ancient and veteran tree distribution model fitted using systematic sampling (SS) at a 2-km resolution (i.e. the best overall performing Maximum Entropy (MaxEnt) model) or the Zero-Inflated (ZI) negative binomial (NB) model in relation to estimates from the 52 surveyed grid squares of a & c) the raw field verification abundance estimates and b & d) estimates of density of trees (including estimates of survey effort). Two grid squares (circled in red) are deemed to be outliers.*

Fig. A6.3.3 *Predicted maps of ancient and veteran tree distributions (or abundance from the ZI models) from each model with and without a bias correction method. Predicted areas of high suitability are represented in red, whereas predicted areas of low suitability are represented in blue. Map predictions from each model are not shown to the same colour scale.*

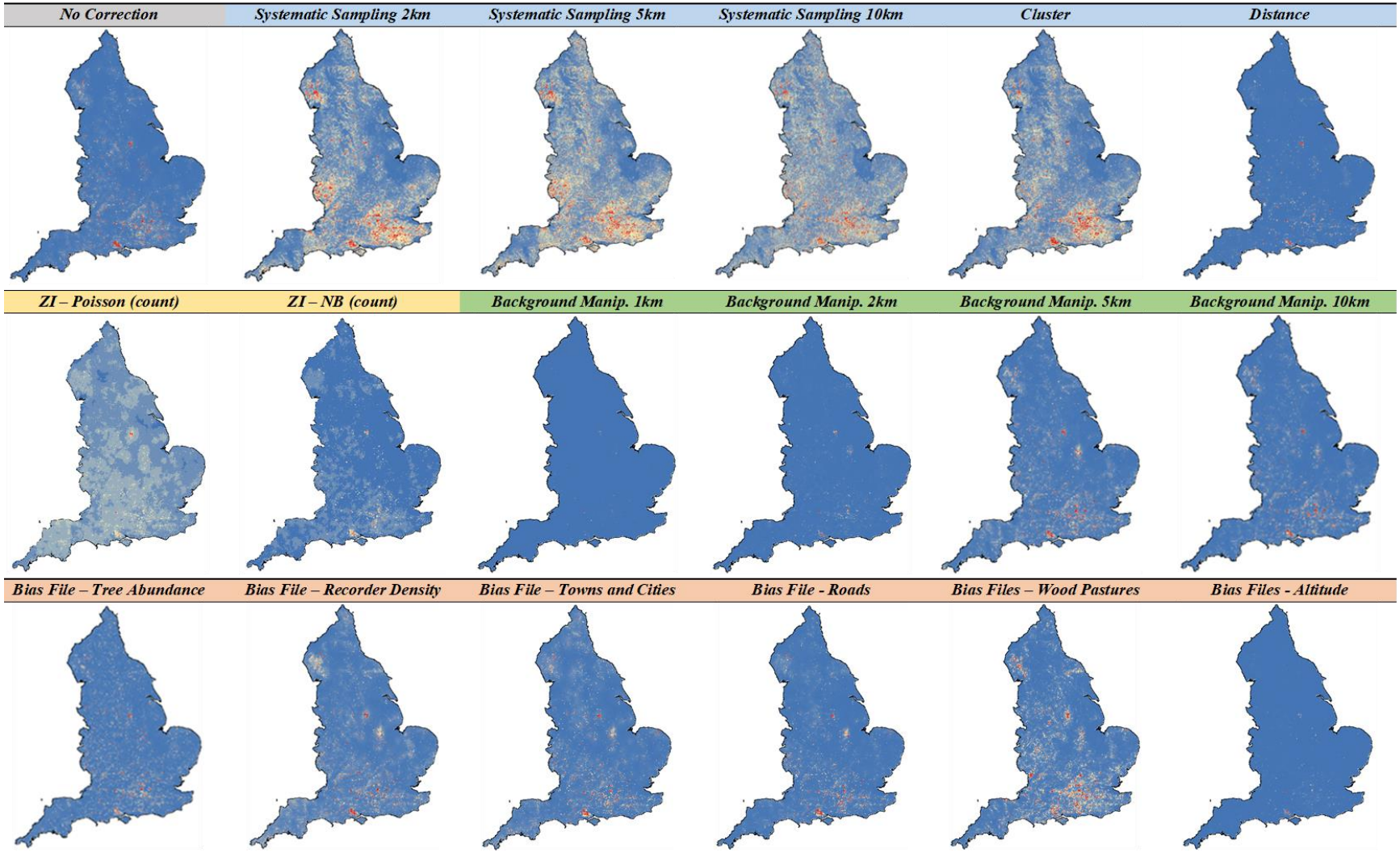


Fig. A6.3.4 *Calculated differences between predicted maps of ancient and veteran tree distributions with no correction and each predicted distribution map from a model using a bias correction method. Abundance predictions from the ZI models were first scaled between 0 and 1 before calculating their difference from the probability predictions from the model with no bias correction, Blue squares represent areas that are predicted to be less suitable following the application of bias correction. Difference maps from each model are not shown to the same colour scale.*