Forest Research UNIVERSITY OF Southampton The Scottish Forestry Trust

University of Southampton

Faculty of Social Sciences
School of Mathematical Sciences

Investigating the effect of survey designs on urban forest population estimates: A simulation based approach using Bayesian spatial models

by

Philip Wells
MSc

A thesis for the degree of
Doctor of Mathematical Sciences

July 2024

University of Southampton

Abstract

Faculty of Social Sciences
School of Mathematical Sciences

Doctor of Mathematical Sciences

Investigating the effect of survey designs on urban forest population estimates: A simulation based approach using Bayesian spatial models

by Philip Wells

Field-based survey methods are a commonly used ecological approach for developing a greater understanding of the benefits of urban forests. Such surveys often aim to collect information on trees contained in several pre-specified plot locations. However, uncertainty exists on the total minimum number of survey plots required to effectively quantify urban forest benefits. By optimising the number of survey plots we ensure that surveys of urban forests in UK towns and cities can be carried out as quickly and cheaply as possible. In this thesis we propose a simulation-based approach for exploring the optimal number of survey plots required for urban forest surveys. Our approach uses state of the art Bayesian spatial modelling to account for the spatial nature of the survey data and characteristics of the city such as many features of the prevailing landscape and their spatial properties. We illustrate our models using bespoke code written in the STAN software language, which allows for modelling of spatially dependent data. Simulations from our models are then used to explore a variety of different survey plot designs, by considering the efficacy of the survey plot design in estimating total tree populations. We illustrate our methods using both survey data and tree locations derived from areal photography. Using the proposed simulation methodology, we obtain robust results and compare those with similar results reported by other authors using non-model based methods. Assessment of population errors from the simulations, highlighted the need for more survey plots in areas with higher variation in the rate of trees. Relative population errors simulated under a range of different conditions have been produced, with conclusions on the minimum number of survey plots required dependent on which simulation conditions are deemed most appropriate. Generally, the accuracy in tree population estimates increased at a lower rate after 200 survey plots, suggesting further plots may not provide much additional information, however this is subject to personal interpretation of an acceptable level of estimation accuracy. Stratified survey designs are found to have little impact on the accuracy of urban forest population estimates in our research, however are likely to result in more representative surveys.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission

Signed:.. Date:..................

Acknowledgements

Thanks to my supervisors, Sujit Sahu, Malcolm Hudson and Kieron Doick for agreeing to take me as their PhD student and all their help and support over the last few years.

Thanks to everyone I encountered in Forest Research and the Urban Forest Research Group, the Scottish Forestry Trust and the University of Southampton. The help, friendship and information provided was invaluable in completing this Thesis.

I'd like to acknowledge Cambridge City Council, Bluesky World and everyone contributing to i-Tree Eco surveys, without which this research would not be possible.

Finally, thanks to all my friends and family for their invaluable love and support since I started in 2019.

Chapter 1

Ecological background

1.1 Defining urban forests

The term urban forest can be broadly defined as 'all the trees in the urban realm – in public and private spaces, along linear routes and waterways and in amenity areas. It contributes to green infrastructure and the wider urban ecosystem' (Doick et al., 2016). A more specific definition follows this same broad outline but is dependent on our definitions of the terms 'urban areas' and 'trees'. 'Urban areas' is a term largely used to refer to built up areas containing high population densities. In the context of this thesis 'urban areas' will refer to towns and cities within the UK, with a boundary specified according to some local authority district dataset provided by the Office for National Statistics (ONS, 2020). The term 'tree' is broadly defined by the Oxford English Dictionary as:

> '*A perennial plant having a self-supporting woody main stem or trunk (which usually develops woody branches at some distance from the ground), and growing to a considerable height and size. (Usually distinguished from a bush or shrub by size and manner of growth)*' (OED, 2023)

As noted in the definition, a more specific criteria is required to differentiate trees from bushes and shrubs. This criteria differs depending on the context of a dataset. Therefore, more specific definitions of the term 'tree' are given in Sections 1.3.1 and 1.3.2 in the context of the data being discussed.

One of the central reasons for assessing existing urban forests is to consider their impact on 'Ecosystem services'. A definition of the terms ecosystem and ecosystem services is laid out in the Millennium Ecosystem Assessment (MEA), which states that:

> ‘*An ecosystem is a dynamic complex of plant, animal, and microorganism communities and the nonliving environment interacting as a functional unit . . . Ecosystem services are the benefits people obtain from ecosystems.*’ (Millennium Ecosystem Assessment, 2005)

The MEA divides the Ecosystem services offered into four broad groups. These include:

- Provisioning services, such as food and fuel
- Regulating services, such as climate and flood regulation
- Cultural services, such as recreational and aesthetic benefits
- Supporting services, such as soil formation and photosynthesis.

The research in this thesis will consider ecosystems of urban forests. We therefore begin by providing a brief overview of a few ecosystem services provided by urban forests.

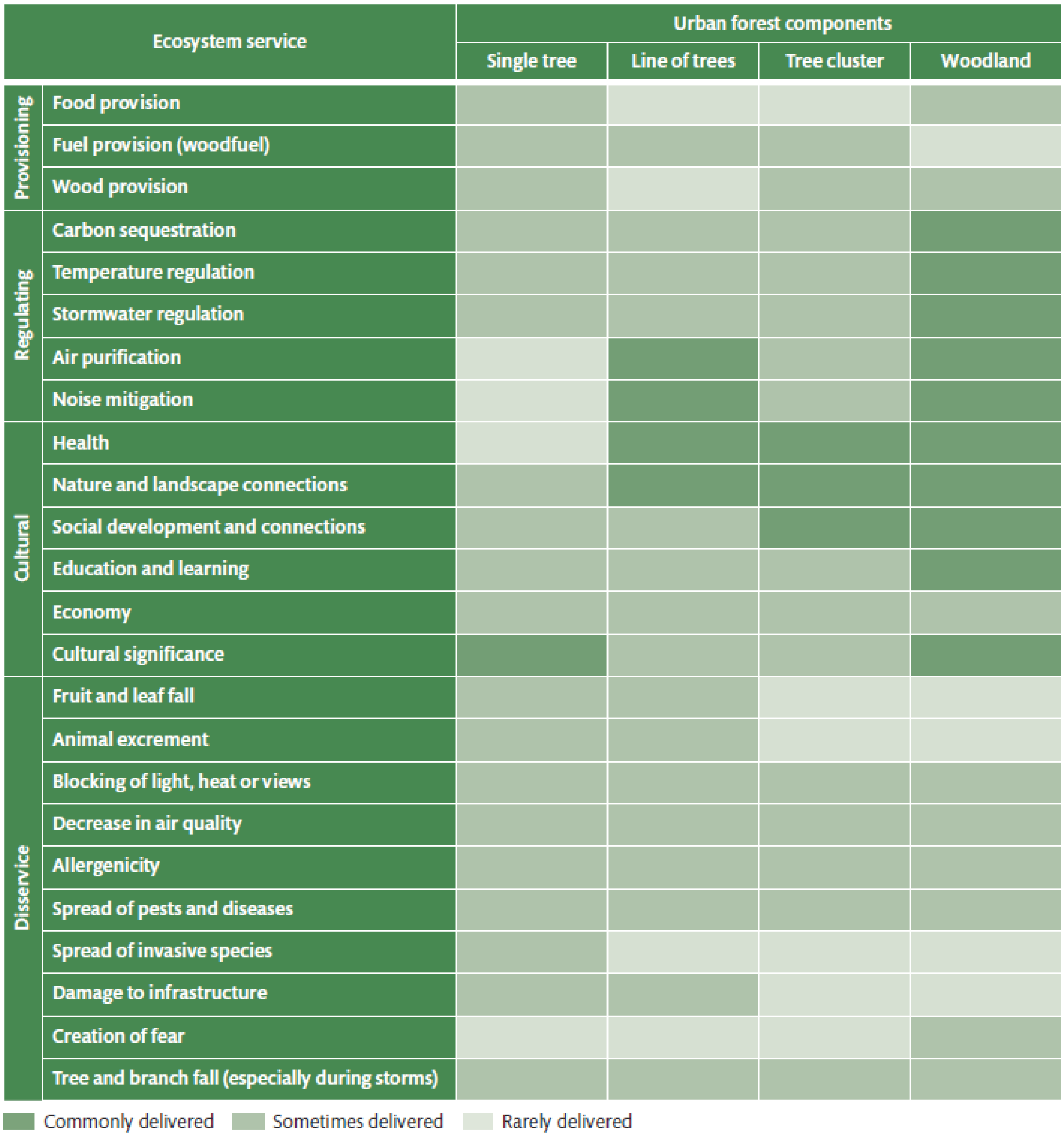
1.2 Urban forests and ecosystem services

Urban forests provide a number of ecosystem services to the local area. These services are commonly dependent on a number of different urban forest characteristics, such as the canopy cover, height and structure of the trees within an urban forest. Table 1.1 provides a number of different ecosystem services that can be offered by urban forests and illustrates whether the ecosystem service can be provided by single trees, lines of trees, tree clusters and woodlands (Davies et al., 2017a).

In recent years, there has been a particular emphasis on ensuring people live within close proximity to green spaces. As part of this, the UK government has plans to ensure that the entire population lives within 15 minutes of a green space or water (Briggs, 2023). This idea is developed further through the proposed 3-30-300 rule, which suggests that people should be within 300m from the nearest park or greenspace, each neighbourhood should have 30% canopy cover and 3 trees should be visible from each home (Konijnendijk, 2022). The 3-30-300 rule is motivated by the ecosystem services provided and also suggests there is an interest in the location of trees within urban forests.

There is evidence to suggest that increases in the number of people living in urban areas could have a major impact on the structure of urban forests and in turn the ecosystem services provided. According to 2018 estimates, the proportion of the UK’s population living in urban areas, such as towns and cities, could rise from 83% in 2018 to 90% by 2050 (UN, 2018). If not considered carefully, this population growth could have

Table 1.1: Matrix of the relationship between ecosystem services and urban forest components. Table adapted from (Davies et al., 2017a)



Ecosystem service		Urban forest components			
		Single tree	Line of trees	Tree cluster	Woodland
Provisioning	Food provision	▒	░	░	▒
	Fuel provision (woodfuel)	▒	▒	▒	░
	Wood provision	▒	░	▒	▒
Regulating	Carbon sequestration	▒	▒	▒	▓
	Temperature regulation	▒	▒	▒	▓
	Stormwater regulation	▒	▒	▒	▓
	Air purification	░	▓	▒	▓
	Noise mitigation	░	▓	▒	▓
Cultural	Health	░	▓	▓	▓
	Nature and landscape connections	▒	▓	▓	▓
	Social development and connections	▒	▒	▓	▓
	Education and learning	▒	▒	▒	▓
	Economy	▒	▒	▒	▒
	Cultural significance	▓	▒	▒	▓
Disservice	Fruit and leaf fall	▒	▒	░	░
	Animal excrement	▒	▒	░	░
	Blocking of light, heat or views	▒	▒	▒	▒
	Decrease in air quality	▒	▒	▒	▒
	Allergenicity	▒	▒	▒	▒
	Spread of pests and diseases	▒	▒	▒	▒
	Spread of invasive species	▒	░	░	░
	Damage to infrastructure	▒	▒	░	░
	Creation of fear	░	░	░	▒
	Tree and branch fall (especially during storms)	▒	▒	▒	▒

▓ Commonly delivered ▒ Sometimes delivered ░ Rarely delivered

a significant impact on the ecosystems in urban areas. The following provides a few examples of how population growth in urban areas could impact ecosystems.

- A larger urban population may lead to gardens being converted into car parking surfaces, due to lack of space and increases in car ownership. As car parking surfaces are generally less permeable than gardens, it would be expected that the frequency and severity of flooding will increase (Warhurst et al., 2014).

- Characteristics of urban areas commonly result in a phenomenon known as urban heat islands, whereby urban areas have a higher temperature than surrounding rural areas. The urban heat island effect has been shown to be directly correlated with an urban area's population size and density. Rising populations in urban areas, along with the impact of climate change, is therefore likely to exacerbate the urban heat island effect. The presence of vegetation, such as trees, is an effective way of reducing the impact of urban heat islands and regulating the temperature in urban areas (Wang et al., 2021).

- Urban areas can contain a number of chemical pollutants and particulates, which have a negative effect on the health of an areas inhabitants and the surrounding environment. Vegetation, including urban forests, has been shown to be an effective way of removing pollutants and improving the quality of air (Nowak et al., 2006). As there is evidence to suggest that a correlation exists between an urban area's population and the amount of air pollution present (Borck and Schrauth, 2021), urban forests could play a vital part in reducing the amount of pollution and ensuring suitable levels of air quality in urban areas.

It is therefore of particular interest to monitor and understand the current ecosystem services provided, so as to allow policy makers and community groups to develop proactive strategies for maintaining the benefits provided by ecosystem services as urban areas develop into the future.

Despite the benefits, green infrastructure in urban areas is often considered as a 'development luxury or afterthought' and generally receives less funding in the UK when compared to infrastructure for housing and transport (Mell et al., 2013). Tree officers face a number of challenges in providing regulating ecosystem services from urban forests, highlighted through interviews carried out with staff members responsible for tree management decisions in urban areas across Britain (Davies et al., 2017b). Amongst the interviewed participants 'widespread dissatisfaction' was noted for the reactive approaches to urban forest management, with participants instead preferring more proactive management approaches which enhance the provision of ecosystem services. The introduction of proactive urban forest management is in part dependent on political support and funding. Many of the interviewed tree officers expressed that a key approach for gaining political support is through comprehensive summaries of local ecosystem services delivered by trees and the resulting economic benefits. There is therefore interest in approaches that use the structure and properties of existing trees in urban areas to establish and quantify estimates of the ecosystem services delivered and the resulting economic impact.

1.3 Quantifying urban forest benefits

To quantify some of the ecosystem services offered by urban forests, it is important to first consider the structure and characteristics of an urban forest. In this thesis we consider remote sensing and the i-Tree Eco program as two separate approaches for capturing information on existing urban forests.

1.3.1 Remote sensing

The term remote sensing refers to the process in which the earth's surface is imaged, often with the use of observation satellites. Using remote sensing it is then possible to build up some image of how an urban forest's structure is expected to look. Remote sensing data can come from a variety of different locations including the European Space Agency's Sentinel program and the Environment Agency's LiDAR data (Fassnacht et al., 2023). Remote sensing and satellite approaches are particularly adept at summarising canopy cover, defined as the area of ground covered by a tree canopy. An example of this is the i-Tree canopy study, which asked participants to select whether random points contain tree canopy cover or not based on webmap images. The responses at different points were then used to provide an overall estimate of the canopy cover for the total study area (Sales et al., 2023). Areal images are however less successful at capturing ground level information, such as the species of tree and shrub data, which can be vital when trying to summarise the ecosystem services offered by urban forests.

The remote sensing data used to explore urban forest structures throughout this thesis will be the National Tree Map (BlueSky, 2020a) and ProximiTREE (BlueSky, 2020b) datasets. Using aerial photography, accurate terrain and surface data, the National Tree Map/ProximiTREEE data is designed to capture locations, heights and canopy/crown extents of all trees within a set area. Trees are only included in the National Tree map data if they are over 3m in height, with a canopy width larger than 10 metres and only included in the ProximiTREE data if they are over 1m in height. The ProximiTREE data has been funded by the Interreg 2 Seas Programme 2014-2020 co-funded by the European Regional Development Fund under subsidy contract No. 2S05 -048 and supplied by Cambridge City Council as part of its Cambridge Canopy Project commitments. Petersfield NTM data was supplied by Bluesky International Limited.

1.3.2 i-Tree Eco

An alternative approach to remote sensing for exploring urban forests is the use of survey plots. Unlike remote sensing, survey plots generally only allow for a small proportion of the total area under investigation to be observed. Inference surrounding the entire urban area is then made based on the observed, sampled locations. Whilst surveying approaches don't give explicit information on the urban forest across the entirety of an urban area, surveying does allow for ground level information to be collected that would not necessarily be available using remote sensing techniques. It is for this reason that remote sensing approaches are designed to 'augment rather than replace existing protocols, such as iTree Eco' (Baines et al., 2020).

The surveying approach followed and explored in this thesis is the one laid out in the i-Tree Eco program. In short, i-Tree Eco is a survey based approach for assessing the structure and function of trees within urban areas. As the data is collected at ground level, trees and bushes are differentiated through the Diameter at Breast Height (DBH) as opposed to either canopy cover, or tree height. The DBH is a measurement of a tree diameter provided at approximately breast height, or more specifically a height of 4.5 feet (1.37 meters) from the ground. Within i-Tree Eco data, trees are defined as woody material with a DBH larger than or equal to 7cm. A more detailed overview of the i-Tree Eco program is given on the i-Tree Eco website, which states:

> *'i-Tree Eco version 6 is a flexible software application designed to use data collected in the field from single trees, complete inventories, or randomly located plots throughout a study area along with local hourly air pollution and meteorological data to quantify forest structure, environmental effects, and value to communities' (i-Tree Eco, 2021)*

The underlying model assessing the structure and function of trees in i-Tree Eco is referred to as the Urban FORest Effects (UFORE) model (Nowak et al., 2008a). By collecting data relating to the land use, tree cover, meteorology and pollution concentration, the UFORE model is able to provide estimates for a variety of information, both relating to the forest structure (e.g. number of trees, species composition, tree health) and several regulating services functions, such as air pollution removal, carbon storage and sequestration. Furthermore, i-Tree Eco is then able to calculate and present economic summaries associated with some of the ecosystem services, summarising an urban forest as an asset with an appreciable return (Mutch et al., 2017).

A variety of information is collected from the i-Tree Eco survey plots and the observed trees for the UFORE model to work effectively. We summarise information collected for i-Tree Eco surveys using two tables observed in the UFORE model literature (Nowak et al., 2008a). Table 1.2 details each of the collected variables for the survey plots, whilst Table 1.3 details each of the collected variables for the observed trees.

1.4 Machine learning

An alternative approach for summarising urban forest characteristics across an urban area is presented in a paper employing machine learning (Baines et al., 2020). Machine learning can be broadly described as an approach which uses computational algorithms to automatically examine datasets and then make resulting inferences. (Baines et al.,

Table 1.2: General plot information collected for the Urban Forest Effects(UFORE) Model. Table adapted from the UFORE model literature (Nowak et al., 2008a)

Variable	Description
Plot ID^z	Unique identifier
Plot address^y	
Date and crew	
Photo number	Used to help identify plot
Measurement units^z	Units for all measurement in the plot; metric (m/cm) or English (ft/in)
Reference objects^y	At least two objects that will assist in locating plot center for future plot remeasurements
Distance to reference object^y	Distance from plot center to each reference object (ft or m)
Direction to object^y	Direction from plot center to each reference object (degrees)
Tree measurement point (TMP)^y	If plot center falls on a building or other surface (such as a highway) where plot center cannot be accessed, the plot is not moved; all distances and directions to trees are measured and recorded from a recorded fixed point (e.g., building corner) referred to as the TMP
Percent measured^z	Proportion of the plot that is actually measured as portions of plot may be denied access
Land use^z	As determined by crew in the field from a standard list of land uses
Percent in^z	Proportion of the plot in each land use to nearest 1%
Tree cover^z	Percent of plot area covered by tree canopies estimated to nearest 5%
Shrub cover^z	Percent of plot area covered by shrub canopies estimated to nearest 5%
Plantable space	Percent of plot area that is plantable for trees (i.e., plantable soils space not filled with tree canopies) and tree planting would not be restricted as a result of land use (footpath, baseball field, and so on); to nearest 5%

^{z}Required for Urban Forest Effects(UFORE) analysis.
^{y}Required for permanent reference of plot

Table 1.3: Tree variables collected for Urban Forest Effects(UFORE) analysis. Table adapted from the UFORE model literature (Nowak et al., 2008a)

Variable	Description
Tree ID	Unique tree number
Distance (ft/m) and direction (degrees) from plot center or TMP[y]	Used to identify and locate trees for future measurements; TMP is tree measurement point
Species code[z]	Species code from standard list currently containing over 10,000 tree and shrub species
Number of dbhs recorded[z]	For multistemmed trees
DBH[z]	Diameter at breast height (in/cm) for all recorded stems
DBH measurement height	Recorded if DBH is not measured at 1.37 m (4.5 ft)
Total height[z]	Height to top of tree (ft/m)
Height to crown base[z]	Height to base of live crown (ft/m)
Crown width[z]	Recorded by two measurements: N-S (north–south) and E-W (east–west) widths (ft/m)
Percent canopy missing[z]	The percent of the crown volume that is not occupied by leaves; two perpendicular measures of missing leaf mass are made and the average result is recorded; recorded to nearest 5%
Dieback[z]	Percent crown dieback to nearest 5%
Percent impervious beneath canopy	Percent of land area beneath entire tree canopy's drip line that is impervious
Percent shrub cover beneath canopy	Percent of land area beneath canopy drip line that is occupied by shrubs
Crown light exposure[z]	Number of sides of the tree receiving sunlight from above; used to estimate competition and growth rates
Distance (ft/m) and direction (degrees) to space-conditioned residential buildings[z]	Measured for trees at least 6.1 m (20 ft) tall and within 18.3 m (60 ft) of structures three stories or less in height
Street tree	Y/N; used to estimate proportion of population that is street trees
Tree status	Indicates if tree is new or removed from last measurement period

[z] Required for Urban Forest Effects(UFORE) analysis.

[y] Required for permanent reference of plot

2020) demonstrates how machine learning techniques can be used to estimate three urban forest characteristics which include, canopy cover, canopy height and tree density across urban areas. i-Tree Eco datasets were considered, however models were primarily developed using remote sensing, LiDAR data. LiDAR data is an areal imaging technique used here to build three dimensional images of the canopy structure. It is of note that the LiDAR and i-Tree Eco datasets were only ever considered separately when modelling. The (Baines et al., 2020) paper notes that the modelling approach used can result in overestimates of the total number of observed trees. While steps are taken to try and reduce the issue of overestimation, the results and discussion do not provide any information regarding the efficacy of the steps taken. The machine learning approaches used in the paper are also unable to consider spatial effects. As spatial effects have been shown to influence results throughout ecology (Di Zio et al., 2004; Sahu et al., 2007; Du et al., 2017), we would preferably like to develop an approach that accounts for the presence of any spatial effects. As the presence of spatial effects can be inconsistent for different areas (Zhou et al., 2017) it is important to assess the requirement of whether a spatial component is needed. Details of how spatial analysis can be considered within our analysis are provided in Chapter 2.

1.5 Survey plot design considerations in i-Tree Eco

The accuracy of any i-Tree Eco results will be dependent on a suitable area being observed by the survey plots. If the surveyed area is a poor representation of the entire urban forest, then any results will be unrepresentative of the area under investigation. The surveyed area may be unrepresentative if either the area characteristics are markedly different between the full and the surveyed areas, or the surveyed area is not sufficiently large enough to draw any reasonable conclusions. To ensure the surveyed area is sufficiently large, we could include larger numbers of survey plots in the survey plot design, however this also makes the entire survey more expensive to conduct.

Limited resources, including tight funding, within urban forestry means that it is particularly advantageous for i-Tree Eco surveys to be carried out as cheaply as possible, whilst ensuring accurate results. The problem of limited resources within urban forestry was raised multiple times in interviews with people involved in i-Tree Eco studies, with one participant stating:

> *'I guess the only barriers to implementing any recommendations would obviously be financial, those would be the key barriers' (Hall et al., 2018)*

By ensuring the costs of i-Tree Eco studies are kept to a minimum, this could allow for more money to be spent elsewhere such as on the analysis and dissemination of the results.

When discussing the financial cost of i-Tree Eco surveys, it is important to clarify how much the surveys generally cost to conduct. The topic of i-Tree Eco costs is addressed in a paper on an i-Tree Eco survey which took place in Petersfield (Moffat and Doick, 2019). The Petersfield i-Tree Eco survey had a much lower cost in comparison to other surveys due to a use of 'citizen science' volunteers for data collection. However, the paper notes that using a professional arboricultural firm consisting of two-person teams surveying 200 i-Tree Eco plots over a period of approximately five weeks is expected to cost approximately £20,000 (2017 prices). This cost assumed cost-saving approaches, such as minimising the travelling time for each day and areal pre-assessments so that more surveyors can be applied to densely covered plots and vice-versa. Furthermore, it is suggested by Forest research that an i-Tree Eco study consisting of 200 plots, delivered by a lead co-ordinator and sub-contracted field-surveyors is expected to cost approximately £35,000 (2017 prices). The survey design, specifically the number of survey plots required, is therefore key in calculating how much time is required to complete the i-Tree Eco survey. By optimising the number of survey plots required, we will minimise the time, and therefore the cost, required for completion of an i-Tree Eco survey, while still ensuring accuracy in the results.

1.6 Existing studies on survey design effects in the ecology literature

Optimising survey designs so as to provide results as efficiently as possible is a subject that has previously been addressed in the ecology literature. In this section we summarise some of the findings and general approaches from the ecology literature, to consider how our research will be conducted and how our research will fit into the wider literature.

1.6.1 Survey context

Before exploring the details of surveying approaches in the wider ecological literature, it is first worth highlighting the context dependent nature of efficient sampling procedures. This idea is explored in a paper on the context of monitoring biodiversity (Yoccoz et al., 2001), which suggests that the following three questions should be considered at the sampling design stage :

- 'Why monitor?'
- 'What should be monitored?'
- 'How should monitoring be carried out?'

A broad approach does little to answer these specific questions and therefore further developments of survey designs are often required for specific contexts.

In their exploration of surveying approaches across ecology, (Kenkel et al., 1989) noted the importance of context when considering sampling procedures, emphasising how sampling considerations should naturally follow from consideration of the investigation objectives. When reviewing sampling designs in the wider ecological literature it is therefore important for us to consider how the techniques used to explore sampling procedures are applicable in the context of our research and sampling objectives.

(Kenkel et al., 1989) further suggested that sampling approaches in ecology arise from three dichotomies which include, parameter estimation versus pattern detection, univariate outcomes versus multivariate outcomes and discrete versus continuous sampling universes. The meanings of these terms are summarised as follows:

- Parameter estimation: The survey is intended to determine estimates for some kind of parameter of interest. Examples of parameters could include species diversity or the effect of air quality. Commonly parameter estimation should also consider the amount of variation associated with a parameter estimate.
- Pattern detection: Considers underlying patterns observed in the data. This can often take the form of some spatial analysis in an ecological context.
- Univariate: Refers to instances where we have one outcome of interest.
- Multivariate: Refers to instances where we have multiple outcomes which are assessed simultaneously.
- Discrete sampling universe: Sampling units are natural, distinct and recognisable. Examples include individual plants or geographic units such as islands.
- Continuous sampling universe: Sampling units are defined as part of the sampling design and are not natural recognisable units. In a continuous sampling universe, consideration needs to be given to the location, number and sizes of sampling units.

For the i-Tree Eco data, we note that a multivariate, parameter estimation approach, with continuous sampling is used. Whilst pattern detection analysis could be conducted from i-Tree Eco data, this is not part of the UFORE model used to summarise the data. The data has multivariate outcomes as the UFORE model considers multiple different variables to produce a variety of different summary information. i-Tree Eco surveys have a continuous sampling universe as the survey design is decided at random in advance of conducting the survey, without the sampling units occurring naturally.

While (Kenkel et al., 1989) provides a useful overview of different sampling considerations, we believe that some aspects of the paper are either oversimplified or outdated.

For example, it could be argued that pattern analysis should be considered as part of parameter estimation, as observed patterns may in turn affect parameter estimates. It could therefore be appropriate to view pattern analysis and parameter estimates simultaneously, however it is still important to consider that both are accounted for at the design stage of the sampling. It is also of note that the ecology literature has expanded considerably since the paper's publication, resulting in the literature addressing more specific and relevant sampling areas as opposed to more generalised sampling recommendations. It is however still key to ensure that the context of the sampling literature is carefully considered.

Sample size, plot size, plot shape and plot locations were each highlighted as sampling features that should be considered at the design stage (Kenkel et al., 1989). We therefore consider how each of these sampling features are dealt with in the ecology literature.

1.6.2 Survey sample size

We begin our review of the environmental survey literature by examining different approaches to optimising the number of survey plots, otherwise known as the sample size. (Hoffmann et al., 2019) demonstrates an approach for identifying the optimal size and number of survey plots required for quantifying and understanding biodiversity within some alpine grassland areas. This aim is summarised as selecting the size and number of survey plots so as to provide 'maximal information via minimal effort'. Despite looking more specifically at the topic of biodiversity and studying alpine grassland areas as opposed to urban forests, the overall aim of the paper is very similar to one of the key aims being investigated in our research. However, rather than being based on data observed within a select number of small locations, findings are instead based on nine 20m × 20m squares in which the required information has been fully observed. Each of these squares is then divided up into 100 smaller grids of size 2m × 2m, referred to in the paper as subplots. These subplots provide a useful framework, from which multiple different design strategies can be assessed. For example to investigate a sampling strategy of 12 subplots, 12 subplots could be selected from within the entire grid and compared to the findings observed throughout the entire 20m × 20m grid. Information collected for the paper included the diversity metric, Shannon's information entropy, and species richness. For each of the assessed sample sizes, medians and 95% intervals were also presented, so as to give some estimate of the associated level of certainty.

While the number of survey plots is one of the key focuses of the paper, the impact of the size and location of the survey plots was also considered. The size was directly accounted for by considering progressively larger grid sizes in addition to different sample sizes. The findings between grid sizes were then compared using a very similar approach to the comparison of sample sizes.

The impact of subplot locations on any findings is acknowledged, however is not explicitly addressed in comparison to survey plot numbers and sizes. (Steinbauer et al., 2012) is cited as highlighting how distance decay between species communities has been observed in the literature, suggesting that more similar communities of species are observed closer together in comparison to distances further away. Areas with a large number of nearby communities can then be referred to as clustered locations, in which we could expect to find more similar species. To address the effect of clustering, each selection procedure was carried out 10,000 times with locations being randomly selected. Summaries were then presented across the entirety of the sampling locations, so that the variation resulting from the subplot locations was captured in the summary. While this presents a useful overview, it may have been of interest to also consider the effectiveness of unclustered subplot locations, an idea explored in more depth in Section 1.6.3.

In general, the approach outlined by (Hoffmann et al., 2019) provides a useful framework for assessing the impact of different sampling designs. However, the approach is difficult to apply for i-Tree Eco datasets where the urban forest has not been fully observed. We instead consider an alternative approach in which sampling designs can be assessed without the requirement of a fully observed set of data.

As an alternative to using fully observed data, (Schweiger et al., 2016) provides a simulation based approach for assessing different sampling designs. These sampling designs are not placed within a specific ecological context, with the paper instead providing a more general summary of sampling designs. This broad approach extends to the response variable, which is defined as a biotic response and used to represent responses ranging from species richness to biomass. These responses are simulated from a range of modelling distributions, with random error components added to provide 'noise' that would commonly be observed in ecological data. The term simulation is used here to refer to data that has been artificially created in such a way that it is a reasonable approximation of the reality. The three error terms considered include

- Gradient error: Some kind of expected variation that can be fully explained. For example some kind of area characteristic.
- Systematic error: a constant but unknown error in the data. Could for example be a result of spatial clustering.
- Random error: Error that cannot be explained. This is used to represent the presence of random variation in the response.

Like the (Hoffmann et al., 2019) paper, data is represented using grids with different survey plot combinations repeatedly assessed for a range of different sample sizes. The suitability of the survey designs for pattern detection are then explored by fitting linear models to the data found in the survey plots. The parameters included in the linear models were decided through the calculation of the AICc, a prediction error estimator designed to compare models by measuring the trade-off between model fit and model complexity. We note that replicates of survey plots, in which data is repeatedly measured at the survey plot locations, are considered in the paper but deemed largely inappropriate in the context of urban forests.

The derivation of some of the error terms provided, appear to be unreliable. For example it is suggested that the random noise is not expected to account for more than 25% of the total variation, based on eddy flux measurements which contain very high levels of uncertainty. This appears to be quite a crude method of establishing the proportion of random noise accounted for in the simulations, which may not be appropriate for all ecological designs. Additionally, the paper appears to suggest that if the random error has a normal distribution with standard deviation of 0.25, then 25% of the total variation is accounted for by the error term. It is unclear how this conclusion is reached, due to the normal distribution being centered around an undefined predictor level and interpretations of the associated standard deviation being unclear without further information.

We propose that for i-Tree Eco data a simulation based approach be used to investigate different sampling designs. This will consist of simulating tree locations over an entire area, based on observed data in select locations. Using this approach provides the density of trees in locations across the area of interest, without requiring a fully observed dataset of tree locations. The simulated data can then be used as a framework for assessing different sampling designs. Simulation based approaches can also be considered for the National Tree Map/ProximiTREE data as whilst the National Tree Map/ProximiTREE data is effectively observed in full, simulation approaches allow us to consider alternative potential urban forest structures which follow from the data, but have not been directly observed. We note here that the i-Tree Eco data and National Tree Map/ProximiTREE data will be considered separately due to disagreements between the datasets and differences between the dataset's definitions of the term tree.

1.6.3 Survey plot locations

The topic of survey plot locations is explored for a number of different scenarios, throughout the ecology literature. A primary reason for exploring survey locations is that findings in nearby survey plots are often more likely to share similarities than survey plots which are further away, an occurrence referred to as spatial clustering. Spacial clustering is therefore often accounted for at the design stage, by using careful selection of survey locations. However, there is variation in the approaches and conclusions reached about how survey plot locations should be decided within the ecology literature.

(Bacaro et al., 2015) compared the effect of three sampling 'shapes' in estimating plant diversity, using a complete, existing dataset in the Siena Province, Italy. As before, the observed area was split up into subplots which were then used to represent different sampling designs. As the data is fully observed, the subplot survey findings can be compared to the 'real' values and used to make inference about different survey designs. In this case 604 10m × 10m plots were each divided up into 16 2.5m × 2.5m subplots. Sampling designs were then assessed by selecting four subplots in each plot so as to be arranged as either squares, rectangles, or placed at random locations. For squares, four subplots are selected so that all of the squares are touching and with each subplot representing a corner of a square. For rectangles, the subplots are instead laid out in one connected straight line resembling a rectangle. We note that the interpretation of the word 'shape' in the context of this paper, refers to the shape of the sampling designs as opposed to the shapes of the subplots. The findings of the paper suggested that the shape of the survey design should be dependent on the survey objective, although a random placement approach generally resulted in much higher species richness values. The use of square survey designs was said to be more suitable if the objective was to explore species composition amongst more homogeneous vegetation, whereas rectangular survey designs were suited for recording more species than squares whilst still ensuring the results are pooled over one large area. These results can be attributed to the distance decay in similarities of species composition (Steinbauer et al., 2012), whereby species composition is likely to be more similar in nearby areas, such as observed in the square survey designs, compared to further away. It is therefore more appropriate to consider randomly located survey plots for surveying urban forest data, so as to capture as much information over the entire area as possible.

(Güler et al., 2016) conducted a similar approach for investigating the impact of plot locations on species richness counts. Once again, square and rectangular survey plot designs were considered and compared to randomly located survey plot designs, referred to as discontiguous sampling designs. However, it was concluded that the species richness findings for contiguous and discontiguous survey designs were incomparable due to the presence of clustering in the underlying data. (Güler et al., 2016) instead proposed that species richness calculations from discontiguous survey designs, be referred to using the term 'cumulative richness'. For i-Tree Eco data we are intending to summarise over a wider area and therefore are more interested in species richness across the entire area as opposed to select locations.

The impact of survey plot locations were also considered in the context of sampling bias. (Leitão et al., 2011) considers surveys in which survey plot locations are biased so as to explore primarily special protection areas and areas near to roads. The context of the paper explores habitat models of species distributions for a large Steppe bird dataset in Southern Portugal. These models are designed so as to provide information on the spatial pattern of species and biodiversity. Once again observations are removed from some underlying baseline dataset to form sampling designs which are then compared to the original findings at baseline. Due to the nature of the data (birds are not stationary) the underlying baseline dataset is instead based on an intensive random sample as opposed to a fully observed dataset. Analysis was conducted by comparing fitted models for the sampling designs to fitted models at baseline, with the results suggesting that model performance was dependent on the locations selected under the sampling design. More specifically (Leitão et al., 2011) suggests that 'it is logical that the greater the geographical bias in a dataset, the greater the resulting environmental bias will be'. These findings are echoed in the plant community literature (Chiarucci, 2007), where randomisation techniques were found to be more reliable than preferential sampling techniques. As i-Tree Eco datasets are generally subjected to randomisation procedures in the sampling design, we would not expect environmental bias to be present. However, (Leitão et al., 2011) notes that environmental bias can occur under even carefully designed studies. It is therefore suggested that surveys should be assessed to ensure suitability for extrapolating results. As a result, different underlying characteristics of the datasets will be explored prior to any analysis in this report. For sampling designs, we should ensure that randomisation designs such as stratification are considered and that the survey design covers a sufficient area.

1.6.4 Other survey design considerations

The ecology literature also considers other aspects of survey design which will not be addressed in our analysis for practical reasons. For example, the size of survey plots in sampling designs has been highlighted in the literature as an important consideration when exploring ecological data (Levin, 1992; Chave, 2013). The interest in survey plot sizes is due in part to the existence of ecological communities, described as 'the living organisms present within a space-time unit of any magnitude' (Palmer and White, 1994). It could be expected for ecological communities to share similar characteristics, however these communities are difficult to scale due to a lack of a clear classification. It is therefore often impossible to select the size and location of a survey plot so as to ensure an entire community has been observed. Within the literature, similar approaches were taken to assess the impact of survey plot sizes, as to assessing appropriate survey plot locations and sample sizes. Approaches involved comparison of sampling designs to both simulated (Steinbauer et al., 2012) and observed (Dengler et al., 2009; Hoffmann et al., 2019) baseline datasets and generally concluded that the size of survey plots could influence outcomes. (Dengler et al., 2009) further suggested that in the context of observing species richness, it is beneficial for uniform plot sizes to be applied. The literature did not appear to address some of the practical concerns surrounding the use of large survey plots in an urban forest context. For example, larger survey plots in residential areas are likely to require more permissions in comparison to smaller survey plots covering less land. This could result in parts of larger survey plots being inaccessible and increase the time taken for data collection. Additionally, survey plots in woodland areas can contain dense tree and shrub populations, resulting in the data taking much longer than usual to collect. We should therefore provide some consideration of the time data collection takes when considering survey plot sizes and numbers, to ensure estimates are optimised according to available resources. It is due to practical concerns, that the size of survey plots used will not be explored in further detail within this thesis.

The shape of survey plots in ecological sampling designs, has also received some attention in the literature. For example (Paul et al., 2019) compared the use of square and circular survey plots for collecting forestry information. The use of square plots were shown to be positively biased, whereas circular plots were less biased. This bias could be attributed to data collection being much easier to carry out practically with circular plots than square plots. To explain further, circular plot areas can be easily defined as within a certain distance of some centre point, whereas the exact boundary of a square plot may often be defined by the surveyor, which could lead to the introduction of biases. In some instances the use of grids may be more appropriate, particularly if it is advantageous for the survey plots to be able to tessellate (Keeley and Fotheringham, 2005; Bacaro et al., 2015). For the outlined practical purposes, only the use of circular survey plots will be considered within this thesis.

1.6.5 Survey designs in i-Tree Eco

There are a few papers in the Ecology literature which directly address the appropriateness of survey designs in i-Tree Eco. Perhaps the most central paper on the subject, (Nowak et al., 2008b), examines the effect of plot and sample sizes on the timing and precision of urban forest assessments and reaches the conclusion that the use of 200 circular one-tenth acre (0.04 ha) survey plots provides a reasonable population estimate. This conclusion is reached in the paper using survey data collected for the UFORE model in 14 different cities, with plot sizes ranging from 110 to 220 survey plots. The errors were standardised using a population size of 200 plots and relative standard errors, proportional to the population estimates, were calculated. From these findings the paper suggests that if a 12% relative standard error is deemed appropriate, then 200 one-tenth acre plots will produce a reasonable population estimate. However, this conclusion is based on the average relative standard error, whereas the relative standard error across the 14 cities can be seen to vary between 8.1 and 19.2, suggesting that 200 plots is far from guaranteed to result in a relative standard error of 12%. Furthermore, this variation in the relative standard error highlights how the number of plots required is often dependent on characteristics of the area under observation and that a simple ‘one size fits all’ approach to assessing the number of survey plots needed is inappropriate. It should also be noted that the paper’s results are based on urban areas within the US and may not be generalisable within the UK. For example grid based road systems have generally been adopted much more widely in the US than the UK which could have an impact on the underlying ecosystem, including the urban forest structure.

(Nowak et al., 2008b) examines how the relative standard error changes as the number of plots increases. The relative standard error results appear to be based on manipulating the standard errors observed within the 14 cities, as opposed to using surveys with the appropriate number of survey plots. For instance, the relative standard error is presented up to 500 survey plots, however none of the surveys used in the analysis contains more than 220 survey plots. Ideally results would be based on the standard errors observed for the requisite number of plots as opposed to extrapolating the results of existing studies. Despite issues surrounding the analysis of the population estimates, the paper offers useful information on time considerations when carrying out the surveys. Amongst these considerations are the set up time and number of permissions required from landowners for the plots, the time it takes to travel between plots and the relationship between the assessment time and the plot size. These considerations should be accounted for when developing a plot design structure, so as to ensure that the design is economically feasible. For example, a hypothetical design that prioritises plot placement in areas which are expected to have a high density of trees may provide a better population estimate using less plots, however the increased time it would take to assess densely covered plots (Nowak et al., 2008b) could mean that the approach provides little or no economic benefit. This concern can be addressed by ensuring results hold for a range of different survey plot designs.

In an investigation into the effects of sampling on quantifying urban forest structures, (Jin and Yang, 2020) found reason to reject the suggestion that 200 survey plots should ensure a population estimate with a 12% relative standard error. The paper investigated the benefits and drawbacks of three different sampling designs. These designs include simple random sampling, whereby locations were decided at random, stratified sampling, whereby locations were located randomly according to some set strata or sub populations, and systematic sampling, whereby locations were randomly set but with a fixed periodic interval. Analysis for this paper was carried out by attempting to extract the spatial locations of trees within urban areas and then using the results to build an underlying urban forestry structure in Philadelphia and Beijing. Surveys were then simulated according to the different sampling strategies and repeated for each sample size between 200 and 500 survey plots. Unlike (Nowak et al., 2008b), this approach has the advantage of allowing the results of different survey plot designs to be repeatedly tested for the data and does not rely on extrapolating the standard error of existing studies. Once the study designs were simulated, the error in the total population estimates was calculated based on comparisons to the population estimate from the extracted tree locations. The results of the analysis concluded that smaller sample sizes, particularly surveys with plot sizes less than 200, resulted in population estimates with low levels of accuracy. Additionally, stratified random sampling designs were generally found to produce more reliable estimates when compared to simple random sampling and systematic sampling designs. As discussed, there are drawbacks associated with only drawing tree locations from areal images, with (Jin and Yang, 2020) noting how the real distribution patterns of the tree species are likely to be much more complex than those constructed for the analysis. Generalisability of the results to urban areas in the UK was also quite low, with assessment only carried out on two cities located outside the UK. It is therefore of interest to establish an approach that considers the appropriateness of using 200 survey plots and whether or not a different survey design approach would be more efficient.

1.7 Thesis outline

The aim of this thesis is to consider the use of Bayesian spatial modelling techniques, to assess the number of survey plots required for accurate estimates of urban forest populations using i-Tree Eco data. For each survey plot design we consider not only whether the design is generally appropriate, but if the results are stable under a wider variety of conditions. The outlined approach is considered for i-Tree Eco, ProximiTREE and National Tree Map (NTM) data only, however could be easily adapted for alternative datasets so as to assess a range of ecological surveying processes.

Chapter 2 introduces and summarises the existing statistical techniques that have been employed to obtain our results. We summarise the Bayesian modelling approach in general, before detailing some of the spatial models considered for analysis. For each spatial modelling approach we consider the suitability of the model, including statistical benefits and drawbacks of the approach. Practical application of the discussed spatial models is considered for the programming language R, before concluding the chapter with an introduction to the basic methodology used for simulating data.

In Chapter 3, application of the discussed spatial modelling techniques is illustrated in full for some completely observed data, which uses areal imaging as a basis, in the UK city of Cambridge. We begin the chapter with a full summary of completely observed data, the ProximiTREE data set, including specific reference to findings in the Cambridge area. Spatial exploratory analysis is conducted on the ProximiTREE data, to provide some initial evidence in favour of including a spatial component in our models. All environmental covariates explored for our fitted models are summarised in full, with example tables and illustrations given for the area of Cambridge. The fitted model is then detailed in full, along with assessment of convergence diagnostics, the model fit, the estimated populations and the estimated cell values. A brief summary of findings for a similar, completely observed, set of data in Petersfield is also provided.

The application of spatial modelling techniques to partially observed i-Tree Eco survey data is explored in Chapter 4. A detailed summary of the i-Tree Eco data is provided, with a particular focus on i-Tree Eco data collected for the UK city of Southampton.

The Southampton i-Tree Eco data contained approximately 400 survey plots, double the previously recommended sample size for i-Tree Eco surveys, providing a much larger observed area for our models. Our novel spatial modelling approach for partially observed survey data, is detailed in full and illustrated for the Southampton i-Tree Eco data. As in Chapter 3, our fitted model is assessed in full, before considering the results of applying the model to i-Tree Eco data for the Cambridge and Petersfield areas.

Using the models fitted in Chapters 3 and 4, we explore the efficacy of different survey designs in estimating total tree populations from simulations drawn from the fitted models. The processes used to simulate survey plot locations, tree densities and estimates of population errors associated with survey plot structures are detailed and considered in full. The results of the simulation approach are considered under a number of different conditions including, the data driving model simulations, the area being assessed, the amount of variation in the tree density values, the simulations drawn from the model, the use of stratification in the survey plot designs and the approach used to calculating population error.

We conclude the thesis with a discussion of both the simulation and model results. Further work that could be conducted from the topics explored in our thesis is also considered. Appendices containing additional code, tables and figures for Chapters 3, 4 and 5 are provided. The appendices provide information used to obtain our results, which are not included in the main body of the thesis for the sake of brevity.

Chapter 2

Statistical Theory

Throughout this thesis we make use of existing statistical methodologies to construct an approach for assessing survey plot designs. In this chapter, we summarise the existing statistical theories which form a backbone to the analysis we have conducted. The chapter begins with a summary of what the term Bayesian modelling means, before detailing different sampling techniques for implementing Bayesian models. We then proceed to discuss a variety of spatial analysis techniques, with a particular focus on approaches for incorporating a spatial component into Bayesian modelling methodologies. Practical approaches for how the described Bayesian spatial models can be fitted in the programming language R is provided. The chapter concludes with a brief summary of the approach taken to simulating data in the thesis. We note that this chapter is intended to provide background to the ideas developed in later chapters of the thesis and that further specifics on applying the statistical techniques is given in later chapters, within the context of the data.

2.1 Bayesian statistics

Throughout this thesis we have generally applied a Bayesian approach for conducting statistical analysis, where parameters of interest are assumed to have distributions of possible values. The Bayesian approach contrasts the frequentist approach to statistical analysis, with the frequentist approach instead assuming that parameters are fixed but unknown values. As a result, the Bayesian approach allows parameter uncertainty to be accounted for explicitly as part of the analysis. The consideration of parameters as distributions rather than fixed points means Bayesian statistics is an ideal framework for producing simulated values, an idea explored further in Section 2.8.

A further benefit of using a Bayesian approach for conducting statistical analysis, is that Bayesian statistics allows for the inclusion of prior information. Prior information refers to additional information and beliefs outside of the observed data which can be incorporated into analysis using Bayesian statistics. The basic framework, allowing for both observed and prior information to be considered, is given by Bayes theorem and is introduced as follows.

2.1.1 Bayes' theorem

We introduce Bayes' theorem by first considering how the theorem can be used to calculate the probability of an event occurrence. Taking A as an event with some associated probability of occurrence and $B_1, B_2, \ldots, B_n$ as a set of mutually exclusive and exhaustive events, then Bayes' theorem can be written for any i, where $i = 1, \ldots, n$, as,

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)} \tag{2.1}$$

Under the definition of Bayes' theorem above, additional information can be incorporated into the probability calculation through the term $P(B_i)$. The term $P(B_i)$ is referred to as our prior information which we note is derived independently of event A.

For the purposes of Bayesian modelling, the Bayes theorem given in Equation 2.1 can be generalised for random variables. In the generalisation we replace B_i with a set of model parameters, θ, and replace A with some observed data y. Our prior information takes the form of $p(\theta)$, the likelihood derived from our observed data is represented by $f(y|\theta)$ and our posterior distribution is represented as $\pi(\theta|y)$. More formally, Bayes theorem can be written for random variables as,

$$\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(y|\theta)\pi(\theta)d\theta} \tag{2.2}$$

where $-\infty < \theta < \infty$.

We note that the denominator, often known as the normalising constant, of Equation 2.2 does not contain θ as the term is integrated out of the expression. As the normalising constant can be tricky to calculate, it is commonly removed when writing the generalisation of Bayes theorem for random variables. By rewriting Equation 2.2 to remove the normalising constant, Bayes theorem is given as,

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta) \tag{2.3}$$

Or in words, the posterior is proportional to the likelihood multiplied by the prior.

2.1.2 Markov chain Monte Carlo (MCMC)

A common way of obtaining posterior inference for Bayesian models is through the use of the Markov chain Monte Carlo (MCMC) methodology. As the name suggests, the MCMC methodology uses a combination of both Markov chains and Monte Carlo integration to derive posterior inference for Bayesian models.

The term Markov chain is used to describe a process in which a sequence of numbers are generated from a transition distribution, with each sequence entry based only on the result of the previous sequence entry, following some initial value in the sequence. Formally we can write a Markov chain for a parameter, θ as,

$$\theta^{(i+1)} \sim p(\theta|\theta^{(i)}) \tag{2.4}$$

where i is a whole number representing the sequence position and $p()$ represents a transition kernel associated with the chain.

The term Monte Carlo integration essentially refers to a process whereby integrals of distributions can be estimated by drawing a large number of samples from the distribution. By applying Monte Carlo techniques to Markov chains estimating some parameter of interest, we have the basis for the MCMC methodology. In essence, we say that MCMC methods provide an approach for drawing large numbers of random samples from some parameters of interest, $\boldsymbol{\theta}$, in order to provide accurate estimates of a posterior distribution. MCMC methods are generally viewed as a computationally intensive approach for gaining estimated posterior samples, due to the high level of sampling required by Monte Carlo integration techniques. Additionally, the number of samples required for MCMC methods should account for the fact that MCMC techniques often take time to converge to the posterior estimate. As a result, the first set of iterations are commonly referred to as 'burn in' and are not included in final MCMC estimates.

2.1.3 Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is an approach which incorporates MCMC methodology for providing samples to estimate some posterior distribution. A general version of the algorithm in a Bayesian modelling context is given as follows:

STEP 1: Set the initial value $\boldsymbol{\theta^0} = (\theta_1^0, \theta_2^0, \ldots, \theta_n^0)$

STEP 2: Generate ϕ, the candidate point, from a proposal distribution, $q(\phi|\boldsymbol{\theta})$ for the first parameter, θ_1, at iteration $t+1$.

STEP 3: Calculate the acceptance probability for the first parameter using,

$$\alpha(\theta_1, \phi) = \min\left\{1, \frac{\pi(\phi|\theta_1)q(\theta_1|\phi)}{\pi(\theta_1|y)q(\phi|\theta_1)}\right\}$$

STEP 4: With probability $\alpha(\theta_1, \phi)$ we accept the proposed value and set $\theta_1^{(t+1)} = \phi$, else $\theta_1^{(t+1)} = \theta_1^{(t)}$

STEP 5: Repeat steps 2-4 for all other parameters

STEP 6: Repeat steps 2-5 for total number of iterations

We note that while the above approach updates each of the parameters separately, it is also possible to simultaneously update a set of parameters with each step if required.

2.1.4 Gibbs sampler

The Gibbs sampler is a widely used MCMC sampling approach that represents a special case of the Metropolis-Hastings algorithm where the proposal distribution is set as equal to the posterior distribution. As a result, the acceptance probability calculated for the Metropolis-Hastings algorithm is equal to one for the Gibbs sampler and all samples are accepted. The Gibbs sampler algorithm can be used to generate samples from some posterior distribution, $\pi(\theta)$ as follows,

STEP 1: Set the initial value $\boldsymbol{\theta^0} = (\theta_1^0, \theta_2^0, \ldots, \theta_n^0)$

STEP 2: Update parameters for step $t+1$ as follows

$$\begin{array}{ccc}
\theta_1^{(t+1)} & \text{is sampled from} & \pi(\theta_1|\theta_2^{(t)}, \ldots, \theta_n^{(t)}) \\
\theta_2^{(t+1)} & \text{is sampled from} & \pi(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)} \ldots, \theta_n^{(t)}) \\
\vdots & \vdots & \vdots \\
\theta_n^{(t+1)} & \text{is sampled from} & \pi(\theta_n|\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{n-1}^{(t+1)})
\end{array}$$

STEP 3: Repeat STEP 2 until t is equal to the total number of iterations.

2.1.5 Hamiltonian Monte Carlo

Another sampling approach which avoids some of the slow exploration associated with random walk techniques is the Hamiltonian Monte Carlo(HMC) sampler (Neal, 2012).

The HMC sampler uses the concept of Hamiltonian dynamics, a concept originating in physics, to prioritise areas with higher posterior probabilities, whilst a Metropolis Hastings step in the sampler ensures that areas with lower posterior probabilities are not ignored completely. As a result, the HMC sampler is often considered a more efficient sampling approach when compared to alternative approaches such as the Gibbs sampler.

2.1.6 Assessing convergence and comparing model performance

After using some MCMC sampling approach to obtain estimates of model parameters, it is important to consider the efficacy of both the model and the sampling process. A number of techniques exist which allow the user to consider whether posterior samples have converged effectively. We briefly detail some of these techniques as follows,

> *Multiple chains* - Currently the MCMC process has been defined using only one chain generated from a set of initial values. By running multiple chains from a variety of different initial values, we can assess whether the sampler has converged through comparison between chains. Chains can be run simultaneously using different cores, to ensure that multiple chains can be obtained without a significant time penalty. It is generally recommended that four chains be used to ensure convergence without large time and computational penalties (Stan Development Team, 2018).
>
> *Trace plots* - A simple visual approach for considering whether a sampler has converged. Trace plots involve tracing the line between parameter values sampled at each iteration. Traceplots should look like ‘hairy caterpillars’ centered around the average parameter estimate, with alternatives suggesting potential issues with the sampler. By plotting chains simultaneously with different colours, the trace plots should also confirm whether any chains have mixed well.
>
> $\hat{R}$ - The $\hat{R}$ diagnostic is an approach for comparing the between and within chain convergence of parameter estimates. Between chain convergence is assessed, as chains may converge to different distributions and within chain convergence is assessed as chains may cover the same area without converging to a distribution. $\hat{R}$ values close to one suggest the chains have mixed well, whereas values larger than one suggest the chains have not mixed well (Vehtari et al., 2021).
>
> *Geweke diagnostic* - The Geweke diagnostic is a convergence diagnostic comparing the mean at the start of a chain, typically the first 10%, to the mean at the end of a chain, typically the last 50%. Mean comparison is conducted through the calculation of a test statistic for equal means. Convergence is suggested if the means at the beginning and end of a chain are approximately equal (Geweke, 1991).

Thinning - Thinning is an approach whereby samples are systematically removed according to some pattern, e.g. every tenth sample. Thinning allows for posterior samples to take up less memory, whilst still allowing the sampler to run for the intended number of total samples.

Credible intervals - Credible intervals refer to probability intervals associated with some parameter. For example, we would say with probability, p, that some parameter lies between two values. These two values define the upper and lower end of our credible intervals. Credible intervals should be interpreted in the context of the fitted model, however the inclusion of zero in a parameter's credible interval can be indicative of the parameter failing to have a significant effect. Credible intervals are often provided at the 95% probability level, however the probability level should be specified.

Effective sample size - Autocorrelation within chains can be responsible for increasing uncertainty in parameter estimates. For each parameter estimate, the effective sample size provides the number of independent samples with the same power as the autocorrelated samples. In Stan, the effective sample size calculations are similar to the $\hat{R}$ calculations and incorporate both between chain and within chain calculations (Stan Development Team, 2018).

Deviance Information Criterion (DIC) - The Deviance Information Criterion (DIC) is a model comparison statistic, commonly used for the comparison of Bayesian hierarchical models. The DIC assesses model efficacy while penalising model complexity to ensure effective models that afford overfitting. Models with lower DIC values are generally considered as superior to models with higher DIC values.

Watanabe-Akaik Information Criterion (WAIC) - The Watanabe-Akaik Information Criterion (WAIC) is an alternative to the DIC, which is considered 'a more fully Bayesian approach'. The WAIC is considered 'more Bayesian' as the WAIC derives values by averaging over the posterior as opposed to using point estimates. The derivation of the WAIC can be more difficult to calculate than that of the DIC, however is often considered more appropriate, particularly for hierarchical models (Gelman et al., 2013).

Leave-One-Out-Cross-Validation (LOOCV) - Leave-One-Out-Cross-Validation (LOOCV) is another model comparison statistic commonly used for comparing Bayesian hierarchical models. As a basis for LOOCV, we consider removing one observation from a set of n observations. Our fitted model is then used to estimate the missing observation, based on the remaining $n-1$ observations. Under LOOCV this process is repeated for all n observations, with the accuracy of model estimates used to calculate the LOOCV estimate. Like the DIC and WAIC, the LOOCV is interpreted by comparing the values obtained under different models. Stable LOO calculations can be calculated quickly from existing simulation draws using a procedure known as Pareto-smoothed importance sampling (PSIS). PSIS (Vehtari et al., 2016). The PSIS procedure is an approach which uses importance sampling to attach weights which smooth over long tails that may not be capturing the target distribution (Vehtari et al., 2022). More detailed information on how the LOOCV and PSIS procedures are calculated can be found in the literature.

2.2 Spatial analysis introduction

The term spatial analysis refers to analysis that not only accounts for characteristics of an observed area, but also findings in nearby areas. The reason for conducting spatial analysis is perhaps best explained through Tobler's first law of Geography, which states 'everything is related to everything else, but near things are more related than distant things'. As spatial analysis often considers some response values in the context of other nearby response values, it is common for spatial analysis to consider autocorrelation. As autocorrelation is a common consideration in time series analysis, time series and spatial analysis techniques frequently overlap in their methodologies.

The spatial analysis techniques that can be applied are generally dependent on the structure of a dataset. The following section explores some forms the data can take and gives a basic overview of some spatial analysis techniques that can be applied.

2.3 Types of spatial data

Spatial data can typically be classified as one of the following three types, point-reference, areal or point pattern data (Banerjee et al., 2014). The first of these types, Point-reference data, refers to data collected from observations at a number of randomly selected locations within some area of interest. Point-reference data can be written as a stochastic process with $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$, where $\mathcal{D}$ refers to a fixed geographical area, $\boldsymbol{s}$ a finite set of locations found within $\mathcal{D}$ (written as $\boldsymbol{s} \in \mathcal{D}$) and $Y(\boldsymbol{s})$ the number of observations at location s. The proximity of locations from each other provides a basis for defining a spatial component when conducting point reference analysis. Surveyed i-Tree Eco data can be placed in a format suitable for point-reference analysis by treating our locations, $\boldsymbol{s}$, as the centre points for each survey plot and our response, $Y(\boldsymbol{s})$, as the total number of trees observed within the boundaries of each survey plot. An illustration of how this point-reference setup looks is given in Figure 2.1 for some i-Tree Eco data collected in the UK city of Southampton. By applying point-reference approaches to surveyed i-Tree Eco data as described, we ensure that the locations of the survey plots are used but lose information on the precise locations of trees within each of the survey plots.

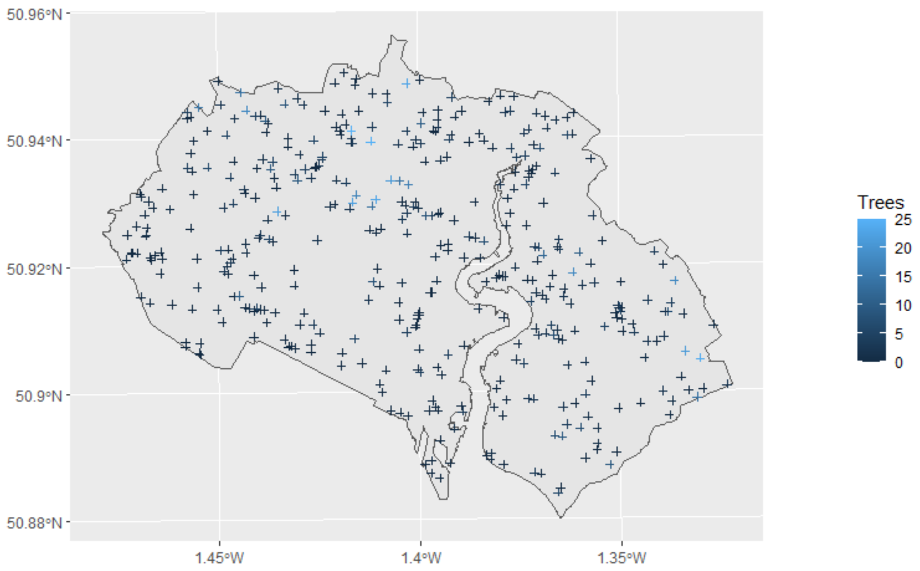


FIGURE 2.1: Example of a point-reference setup approach for some i-Tree Eco survey data located in Southampton

The term areal data refers to data relating to smaller areal units within a larger area of interest. More formally a two dimensional polygon, $\mathcal{D}$, is partitioned into a number of smaller areal units, $\boldsymbol{\mathcal{C}}$. We use $Y(\boldsymbol{\mathcal{C}})$ to refer to some response variable within each areal unit, with the total number of areal units written as n_C. By using the proximity of areal units, $\boldsymbol{\mathcal{C}}$, from each other as a basis, spatial analysis can be conducted on areal data. Frequently, areal data is only available in instances where information has been anonymised for privacy such as illness, household income and voting habits. The availability of areal data means that areal analysis techniques are commonly used in disease mapping (Lee, 2011; Obaromi, 2019), however other examples include analysis of voting data (Lauderdale and Clark, 2016) and analysis of motor vehicle crashes (Morris et al., 2019). In Figure 2.2 we illustrate how some areal data could be expected to look using manipulated i-Tree Eco data taken from the UK city of Southampton. To achieve the illustration, the area of Southampton has been divided up into polygons based on Medium Layer Super Output Areas and a total rate of trees per Ha has been calculated for each polygon based on the survey data found within the polygon boundaries.

Point pattern data has similarities to point reference data, but with points being indicative of an event occurrence as opposed to a fixed location. A key part in the analysis of point pattern data is often whether observations exhibit clustering, regularity or Complete Spatial Randomness (CSR) (Cressie, 1994). Clustering is identified if points are spatially grouped together, with regularity indicated by points being approximately equally spaced out across the area of interest. In the absence of both clustering and regularity, CSR would be concluded. An intensity parameter, λ, is used to represent the

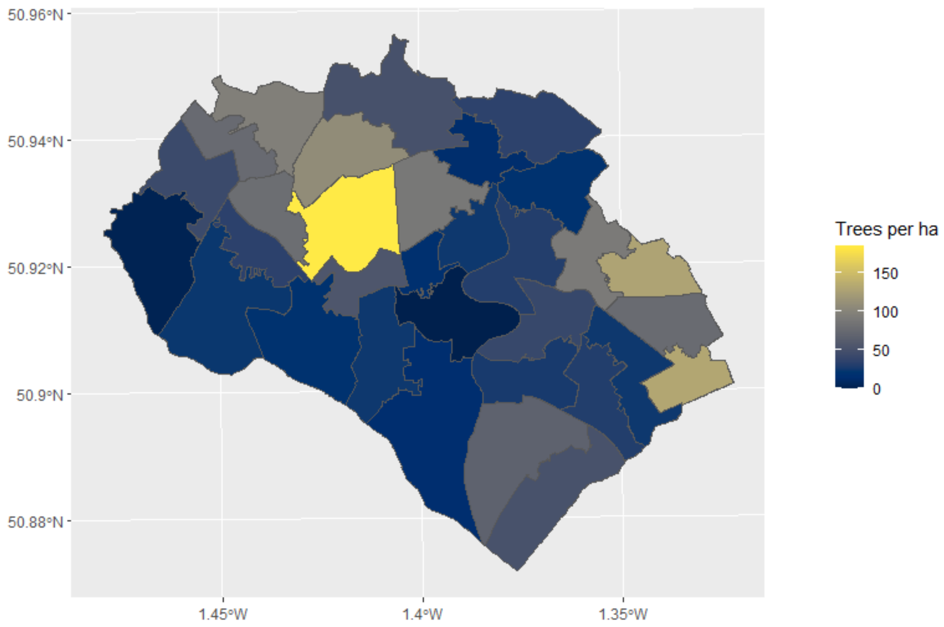


Figure 2.2: Example of an areal setup approach for some i-Tree Eco survey data located in Southampton

rate of occurrences within a nearby area and is expected to be homogeneous across the area of interest in the absence of any spatial effects. Examples of point pattern data include earthquake epicentres (Ouchi and Uekawa, 1986), wildlife locations (Khaemba, 2001) and tree locations within a forest (Law et al., 2009). The trees located within the survey plots would be considered point pattern data, an example of which is illustrated in Figure 2.3 for a survey plot with a particularly high level of trees observed.

For the analysis conducted in this thesis, we will largely focus on applying statistical methods used for modelling areal data. Justification for this decision is given within the context of the data being analysed in chapters 3 and 4, along with consideration to the benefits and drawbacks of the methodology being used.

2.4 Neighbourhood definitions

The basis commonly used to account for the spatial structure when exploring and modelling areal data is referred to as the neighbourhood, or proximity, matrix. This is essentially a matrix where each entry is used to provide information on the spatial relationship observed between each possible pair of areal units. More formally the matrix consists of weights, W, which represent the spatial association between the different areal units $1, 2, \ldots, G$. Weights, w_{jk}, are assigned within the matrix, to represent the spatial association between the areal units j and k (Banerjee et al., 2014). Commonly these weights are defined as binary values based on whether the two areal units share a

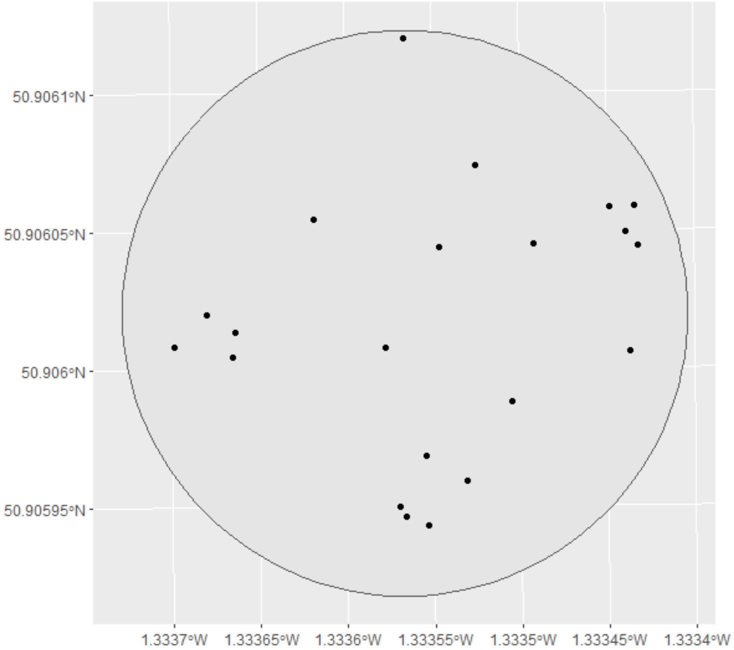
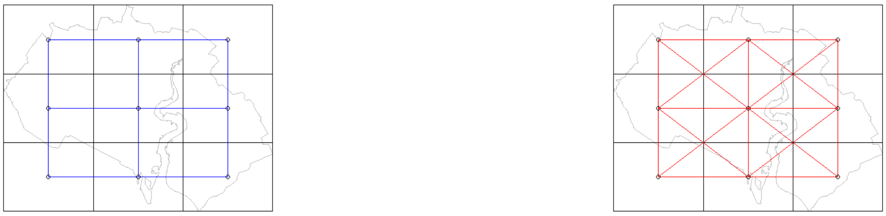


FIGURE 2.3: Plot of trees for a single plot in the i-Tree Eco survey data, to illustrate point pattern data. Note that a plot with a particularly high tree coverage has been selected for illustrative purposes.

border, however the weighting function could instead be designed so as to incorporate other spatial information, such as the distances between areal units. If required, additional proximity matrices can be defined for different orders, whereby the order dictates the proximity of the areal units. For instance we may have a first order proximity matrix representing the direct neighbours of an areal unit, a second order proximity matrix representing both the first order areal units and neighbours of the first order areal units and so on.

When working with areal data, where the proximity matrix is defined based on touching areal units, it is useful to specify whether 'queen' or 'rook' based neighbours are being used. In the R package SPDEP, 'queen' based neighbours refer to any touching areal units, whereas 'rook' based neighbours use the stricter criteria that both areal units must share an edge (Bivand and Wong, 2018). An example of how this will look using both queen and rook based neighbours can be seen for a simple 3×3 grid in Figure 2.4. 'Queen' based neighbours have been applied for the proximity matrices used throughout this thesis, so as to account for the entirety of any surrounding cells.

As an alternative to proximity matrices, (Morris et al., 2019) recommends the use of graph edgesets for defining the neighbourhood relationships when conducting spatial modelling in the programming platform, Stan. Graph edgesets consist of two vectors in which each vector row defines neighbouring cells for all of the neighborhood relationships present within a graph. An example illustration of how an edgeset can be defined from a simple graph and proximity matrix is given in Figure 2.5.



(A) Neighbourhood structure using ‘Rook’ weights

(B) Neighbourhood structure using ‘Queen’ rates

FIGURE 2.4: Example neighbourhood structures for a 3×3 grid

Graph edgesets are recommended as an alternative to proximity matrices as storing the neighbourhood definitions as an edgeset requires less memory than specifying a full proximity matrix in cases where the proximity matrix is sparse. Additionally, the use of proximitry matrices when modelling is often much slower than the use of edgesets, as manipulating large matrices is much more computationally expensive compared to a pair of vectors.

In this thesis we will exclusively consider first order proximity matrices with binary weights, as this structure allows the neighbour relationships to be easily defined as a graph edgeset. Note that despite the use of the phrase edgeset here, ‘Queen’ based neighbour definitions are still being applied.

2.5 Exploratory spatial techniques

Before fitting a statistical model, we can conduct a number of spatial exploratory analysis techniques to provide preliminary evidence on whether a spatial term is required in our model. We will be considering spatial exploratory techniques associated with areal and point pattern data. For areal data, exploratory spatial techniques can be used to establish preliminary evidence on whether some observations, $\boldsymbol{Y}$, are more similar in nearby areas compared to observations further away. This preliminary evidence can then be used to provide an initial justification for whether a spatial component should be considered in any statistical modelling approaches. For point pattern data, spatial exploratory techniques can be used to asses the extent that fully observed tree locations exhibit signs of clustering.

2.5.1 Areal exploratory techniques

Moran’s I and Geary’s C are two techniques that can be conducted as spatial exploratory analysis for areal data. These techniques provide exploratory analysis on whether a

Here we illustrate how a spatial structure can be defined for use in spatial modelling. We begin with a numbered diagram of a 3×2 grid:

1	2	3
4	5	6

From the above spatial structure we define the following proximity matrix. This proximity matrix uses first order queen based neighbours with binary weights:

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Alternatively, our binary weights matrix can be written as a pair of vectors, known as an edge set:

$$\left\{\begin{matrix} 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 & 5 & 5 & 5 & 5 & 5 & 6 & 6 & 6 \\ 2 & 4 & 5 & 1 & 3 & 4 & 5 & 6 & 2 & 5 & 6 & 1 & 2 & 5 & 1 & 2 & 3 & 4 & 6 & 2 & 3 & 5 \end{matrix}\right\}^T$$

Figure 2.5: Example illustration of how a proximity matrix and an edge set can be created from a basic spatial structure

spatial association exists between areal units. Evidence of a spatial association can then provide justification on whether a spatial component should be considered in any statistical models.

Moran's I (Moran, 1950), acts as an adaptation of Pearson's correlation coefficient and summarises the level of spatial autocorrelation present. The level of spatial autocorrelation present is calculated for Moran's I by comparing an observed area unit to its neighbouring areas using the weights, w_{jk}, as defined in section 2.4. More specifically the formula for Moran's I can be written as:

$$I = \frac{G}{S_0} \frac{\sum_{j=1}^{G} \sum_{k=1}^{G} w_{jk}(y_j - \bar{y})(y_k - \bar{y})}{\sum_{j=1}^{G}(y_j - \bar{y})^2} \tag{2.5}$$

Where $S_0 = \sum_{j=1}^{n} \sum_{k=1}^{n} w_{jk}$ is the sum of all the neighbours, G is the total number of areal units and y_j is the response value associated with areal unit j. Moran's I usually ranges between -1 and $+1$, with values close to -1 indicative of regularity and values close to $+1$ indicative of clustering. Under the null hypothesis of no autocorrelation, the expected value of Moran's I is not always equal to zero. It is therefore common to assess whether spatial autocorrelation is present by conducting a statistical test comparing the observed I, $\hat{I}$, to the value of I in the instance of no autocorrelation, I_0 (Paradis and Schliep, 2019). Alternatively, the observed Moran's I could be assessed through comparison to other Moran's I values drawn using a Monte Carlo approach. Under a Monte Carlo approach, observed responses are randomly distributed to existing areal units a number of times and used as a basis for calculating a number of alternative Moran's I values. The original Moran's I can be compared to the Moran's I values generated by the Monte Carlo approach, to assess for the presence of spatial autocorrelation (Gimond, 2023).

For further investigation into whether a spatial component is present, Geary's C can be calculated (Geary, 1954). Like Moran's I, Geary's C acts as a measure of spatial autocorrelation present in the data. More formally we write Geary's C as:

$$C = \frac{(G-1) \sum_{j=1}^{G} \sum_{k=1}^{n} w_{jk}(y_j - y_k)^2}{2S_0 \sum_{j=1}^{n}(y_j - \bar{y})^2} \tag{2.6}$$

where once again $S_0 = \sum_{j=1}^{n} \sum_{k=1}^{n} w_{jk}$ is the sum of all the neighbours, G is the total number of areal units and y_j is the response value associated with areal unit j. Under the absence of spatial correlation it's expected that Geary's C would be approximately equal to one and therefore spatial correlation is identified when Geary's C tends to less than one. A Monte Carlo approach can also be conducted for Geary's C using the same basic outline conducted for Moran's I (Bivand and Wong, 2018).

2.5.2 Point pattern exploratory techniques

To assess whether some point pattern data exhibits CSR, we can apply techniques known as the G function and Ripley's K to the data. We note that in the context of the G function and Ripley's K, the term neighbours refers to nearby locations as opposed to the definition given in section 2.4 for areal data.

To assess for the presence of CSR amongst the point reference locations, the G function uses the distance, r, between the nearest neighbours for each observation in a point pattern dataset and compares this to what would be expected under CSR. Informally the function, $G(r)$, can be thought of as a 'nearest neighbour distribution' where $G(r) = Pr(nearest\, event \ \leq \ r)$ and which represents the cumulative distribution function of the distance between a typical random point and its nearest neighbour (Banerjee et al., 2014). For observations close to the border of an area of interest, $\mathcal{D}$, we are likely to observe edge effects, whereby potential surrounding observations that lie outside the border are not accounted for. An edge correction can be built into the G function estimate using the following adaptation for each location, s_i,

$$\hat{G}(r) \ = \ \frac{\sum_i I(r_i \leq r \leq b_i)}{\sum_i I(r < b_i)} \tag{2.7}$$

where r_i refers to the neighbour distance for each i, b_i refers to the distance from s_i to the edge of the area of interest, $\mathcal{D}$, and s_i refers to the i^{th} location amongst the set of locations, $\boldsymbol{s}$.

Whether the locations exhibit CSR can be established by comparison of the estimated G function to the G function under CSR defined as $G(r) \ = \ 1 - exp(-\lambda \pi r^2)$, for a constant intensity, λ.

Another distance based exploratory method applicable to point reference data is Ripley's K (Ripley, 1977). As opposed to calculating the distance between a point and its nearest neighbour, Ripley's K instead counts the number of observations within distance, d of a point. We again calculate an estimate of Ripley's K based on the observed data and compare this to Ripley's K under CSR. More formally, (Kiskowski et al., 2009) specifies Ripley's K for distance, r, as,

$$K(r) = \frac{1}{h} \sum_{i=1}^{h} H_{s_i}(r)/\lambda \tag{2.8}$$

where s_i is the ith plot location, λ is the intensity and the function $H(r)$ is the expected number of points within distance, r, and the sum is taken over h points.

For more convenient use, (Besag, 1977) accounted for the fact that the area of the circle under CSR is πr^2 and proposed,

$$L(r) = \sqrt{\frac{K(r)}{\pi}} \tag{2.9}$$

A common way to interpret $L(r)$ is to plot d against $L(d)$, so as to illustrate how the function behaves as the distance increases. Under CSR we would have $L(d) = d$, suggesting that much larger and smaller values than expected would indicate the existence of clustering and regularity respectively. To account for edge effects, Ripley's K can be adjusted so as to be calculated relative to the circle area within the study area.

2.6 Bayesian spatial models

There are a wide variety of approaches for fitting Bayesian spatial models to datasets with a range of different structures. In the following sections we consider areal modelling techniques which employ a Conditional Autoregressive (CAR) spatial component.

2.6.1 Conditional Autoregressive models

A popular way of defining a spatial component for areal data is through the use of Conditional Autoregresive (CAR) models. Sometimes known as Besag models, CAR models smooth over neighbouring areal units in an attempt to remove the noise attributed to spatial variation. An approach for defining the CAR model is laid out by (Banerjee et al., 2014) and considers the CAR model for each ϕ_j as follows,

$$\phi_j | \phi_k,\, k \neq j \sim N\left(\sum_k b_{jk}\phi_k, \sigma^2_{CAR_j}\right),\ j = 1, \ldots, G \tag{2.10}$$

with b_{jk} representing values of a spatial weights matrix, $\boldsymbol{B}$, and $\sigma^2_{CAR_j}$ representing the unknown variance of the CAR model for each areal unit, j. From Equation 2.10, we say that each ϕ_j is Normally distributed around the sum of the weighted value of its neighbours with some unknown variance, $\sigma^2_{CAR_j}$.

Using Brook's lemma (Besag, 1974), it can then be show that when $(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\sigma^2_{CAR}}$ is positive definite, then $\boldsymbol{\phi} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma_{CAR}})$, where $\boldsymbol{\Sigma_{CAR}}$ is a covariance matrix equal to $(\boldsymbol{I} - \boldsymbol{B})^{-1}\boldsymbol{\sigma^2_{CAR}}$. For $\boldsymbol{\Sigma_{CAR}}$ to be considered a valid covariance matrix, the following conditions must hold (Ver Hoef et al., 2017):

- $\boldsymbol{I} - \boldsymbol{B}$ has positive eigenvalues,

- $\boldsymbol{\Sigma_{CAR}}$ is diagonal with positive diagonal elements,

- $b_{i,i} = 0 \; \forall i$ and
- $b_{i,j}/\sigma^2_{CAR_{i,i}} = b_{i,j}/\sigma^2_{CAR_{j,j}} \; \forall i, j.$

By defining $\boldsymbol{Q}$ as a precision matrix equal to the inverse of our covariance matrix $\boldsymbol{\Sigma_{CAR}}$, we can then write $\boldsymbol{\phi} \sim N(\boldsymbol{0}, \boldsymbol{Q^{-1}})$. The precision matrix $\boldsymbol{Q}$ is constructed using two matrices, $\boldsymbol{D}$ and $\boldsymbol{A}$ (Morris et al., 2019). The matrix $\boldsymbol{D}$ represents a diagonal matrix in which the off-diagonal entries are equal to zero and the diagonal entries, d_{jj}, represent the number of neighbours observed at the j^{th} areal unit. Meanwhile, $\boldsymbol{A}$ represents an adjacency matrix, with entries equal to one if areal units are neighbours and zero otherwise. Following the conditions laid out for a valid covariance matrix above, it should be ensured that the precision matrix, $\boldsymbol{Q}$ is a positive definite matrix (Ver Hoef et al., 2017). The proximity matrix can then be written as:

$$\boldsymbol{Q} = \boldsymbol{D}\left(\boldsymbol{I} - \alpha \boldsymbol{A}\right) \tag{2.11}$$

where $\boldsymbol{I}$ is the identity matrix and α is a parameter used to control the amount of spatial dependence, with $\alpha = 0$ representing spatial independence (Banerjee et al., 2014).

By setting $\alpha = 1$, we have a special case of the CAR model known as the Intrinsic Conditional Autoregressive (ICAR) model, where the proximity matrix is now defined as $\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{A}$. As the new definition of $\boldsymbol{Q}$ allows for some zero eigenvalues, we say that $\boldsymbol{Q}$ is improper (Lavine and Hodges, 2012). As $\boldsymbol{Q}$ is improper, we are unable to use the ICAR model as a model for data and instead used the model as a prior distribution.

(Morris et al., 2019) highlights the advantage of using the ICAR model over the CAR model in terms of computation time. It is noted that the log probability density of ϕ under the CAR model can be written as:

$$\frac{n}{2} log(det(\boldsymbol{Q})) - \frac{1}{2}\phi^T \boldsymbol{Q} \phi \tag{2.12}$$

where n is the number of components in each graph.

Computing the determinant for $\boldsymbol{Q}$ requires G^3 operations, where G is the total number of areal units. An MCMC sampler will need to recalculate the probability density of ϕ for every new proposal and therefore using the CAR model can be computationally expensive. Meanwhile, under the ICAR model, the term $\frac{n}{2} log(det(\boldsymbol{Q}))$ is constant, resulting in only G^2 operations being required. In summary, while the ICAR model can only be used as a prior distribution, it is much more computationally efficient when using MCMC methods.

Under the ICAR model the conditional specification of ϕ_j can be written as:

$$\phi_j \sim N\left(\frac{\sum_{j \sim k} \phi_j}{d_j}, \frac{\tau_{\phi_j}^2}{d_j}\right) \tag{2.13}$$

Additionally, it can be particularly useful for model fitting in programs such as Stan to use the joint specification of the ICAR model (Morris et al., 2019). The joint specification for ϕ is given as:

$$p(\phi) \propto exp\left(-\frac{1}{2}\sum_{j\sim k}(\phi_j - \phi_k)^2\right) \tag{2.14}$$

where as the term $(\phi_j - \phi_k)^2$ is dependent on the difference between the values of neighbouring cells, it follows that minimising this term will result in spatial smoothing. Furthermore, by centering the model using the constraint $\sum_G \phi_j = 0$, we ensure that the log probability density will be defined, as the domain of integration is restricted to only the set of parameters summing to one (Morris et al., 2019).

A limitation of the Besag model is that it exclusively accounts for spatial variation and not other explanatory parameters. We therefore consider models that contain a spatial CAR component alongside an independent random error term and model covariates.

2.6.2 Besag, York and Molie model

One adaptation of the Besag model that allows for the inclusion of a a spatial CAR component, independent random error component and model covariates, is the Besag, York and Molié (BYM) model (Besag et al., 1991). The BYM model is a lognormal Poisson model, often used for modelling count data, $\boldsymbol{Y}_j$ for each area j. Often a Poisson distribution with mean $\boldsymbol{E}_j\boldsymbol{\eta}_j$ is used, where $\boldsymbol{E}_j$ is an expected count and $\boldsymbol{\eta}$ is some relative risk. This setup allows for easy interpretation of the relative risk, quantifying in each area whether the average risk is higher ($\boldsymbol{\eta}_j > 1$) or lower ($\boldsymbol{\eta}_j < 1$) than the average risk in the standard population (Moraga, 2019). The ease of interpreting the relative risk means the BYM model is popular in disease mapping contexts. It is important to highlight the use of the log link in the BYM model, which is particularly adept at modelling the occurrence of rare events.

In full, the BYM model can be formally written as

$$Y_j|\eta_j \sim Poisson(E_j\eta_j), \quad j = 1, \ldots, n_c \tag{2.15}$$

$$\log(\eta_j) = \mu + \beta_1 X_{j1} + ... + \beta_p X_{jp} + \sigma_j \tag{2.16}$$

$$\sigma_j = u_j + v_j \tag{2.17}$$

where μ represents the intercept and overall risk level, $\boldsymbol{X}_j = (X_{j,1}, X_{j,2}, ..., X_{j,p})$ is a vector of p covariates relating to area, j and $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^{'}$ is a vector of associated coefficients, u_j represents a spatial component assigned with an ICAR distribution, v_j is some uncorrelated non-spatial random effects term and σ_j is some total error term comprising the sum of the spatial and non-spatial error terms.

Following our definition in section 2.6.1 we assign a prior distribution of $\boldsymbol{u} \sim N\left(0, \tau_u^{-1}\boldsymbol{Q}^{-}\right)$ where $\boldsymbol{Q}^{-}$ represents the inverse of the precision matrix $\boldsymbol{Q}$ and τ_u represents a precision parameter associated with the spatial random effect terms, $\boldsymbol{u}$. A similar structure is employed for the independent random error term which is assigned a prior distribution of $\boldsymbol{v} \sim N\left(0, \tau_v^{-1}\boldsymbol{I}\right)$, where $\boldsymbol{I}$ is the identity matrix and τ_v is some precision parameter associated with the independent random effect terms, $\boldsymbol{v}$.

(Riebler et al., 2016) highlights some of the issues surrounding the interpretation of parameters in the BYM model. As either $\boldsymbol{u}$ or $\boldsymbol{v}$ can be responsible for most of the variation, it is difficult to know how hyperpriors, i.e. priors placed on parameters of prior distributions, should be set for the precision parameters τ_u and τ_v. Furthermore, only the sum of the random effects, $\boldsymbol{\sigma}$ is identifiable under the BYM model, as the spatial, $\boldsymbol{u}$, and non-spatial, $\boldsymbol{v}$, error terms cannot be viewed independently of each other (MacNab, 2011). We therefore consider alternatives to the BYM model, which maintain the inclusion of a CAR prior whilst ensuring parameters can be clearly defined and interpreted.

2.6.3 Leroux model

An alternative approach to the BYM model when considering the implementation of a CAR component is given by the Leroux model (Riebler et al., 2016). Under the Leroux model, only the sum of the spatial and independent error terms, $\boldsymbol{\sigma}$, is considered. Spatial and non-spatial error is accounted for with the introduction of a mixing parameter, ρ, which ranges between 0 and 1. When $\rho = 0$ we say that the model accounts for non-spatial random effects only, whereas for $\rho = 1$ the model reduces down to the Besag model. By considering the proportion of spatial to non-spatial error through a mixing parameter, the Leroux model avoids the issue of separate identifiable error terms found with the BYM model. Under the Leroux model, we define our σ term to be normally distributed around a mean of of 0 with a covariance matrix of:

$$Var(\boldsymbol{\sigma}|\tau_\sigma, \rho) = \tau_\sigma^{-1}\left((1-\rho)\boldsymbol{I} + \rho\boldsymbol{Q}\right)^{-1} \tag{2.18}$$

where τ_σ represents a precision parameter associated with our overall random effects term, $\boldsymbol{\sigma}$, $\boldsymbol{I}$ represents the identity matrix and $\boldsymbol{Q}$ represents a precision matrix as defined in Section 2.6.1.

From Equation 2.18, the conditional expectation of σ_j can be written as a weighted average of our independent random error model and the Besag model, considering for all other random effects. Furthermore, the conditional variance of σ is the weighted average of $\frac{1}{\tau_\sigma}$ and $\frac{1}{\tau_\sigma \cdot n}$ where n is the number of neighbours at cell j (Riebler et al., 2016).

The primary advantage of the Leroux model compared to the BYM model is in the interpretation of the random effects $\boldsymbol{u}$ and $\boldsymbol{v}$. Using the BYM model only the sum of the random effects, $\boldsymbol{u}+\boldsymbol{v}$ is identifiable, meaning that we are unable to see the spatially structured component independently of the random error component (MacNab, 2011). In contrast the σ and ρ parameters can be considered independently of each other, lending the parameters to a much easier interpretation.

2.6.4 Scaling the spatial component

A potential problem with both the BYM and the Leroux models is that the spatial component is not scaled. By scaling Intrinsic Gaussian Markov Random Fields (IGMRFs), such as the CAR model, we allow for a much better understanding of our precision parameters, τ_u, τ_v and τ_σ. Without scaling, the precision parameters commonly become confounded, complicating any assignation of hyperpriors. If the hyperprior selected for the precision parameters is too large, then the spatial variation may become 'blurred', whereas if the selected hyperprior is too small, the spatial variation may be overfitted as a result of large local variations. Furthermore, by scaling the precision parameters across different applications, interpretation of the the precision parameters becomes identical across applications, as opposed to being dependent on an underlying neighbourhood matrix (Sørbye and Rue, 2014; Riebler et al., 2016).

The benefits of scaling mean that it is desirable to scale the Leroux model, however this is impossible as scaling would be dependent on the value of ρ (Riebler et al., 2016). We therefore consider an alternative model to the Leroux, which retains the easily identifiable ρ and $\boldsymbol{\sigma}$ parameters, whilst allowing for the model to be scaled.

2.6.5 BYM2 model

By combining the ideas laid out in Section 2.6 so far, we can define a model commonly referred to in the literature as the BYM2 model. The BYM2 model reparameterises the BYM model so as to include both a mixing parameter, ρ, and overall error term, $\boldsymbol{\sigma}$, as defined for the Leroux model, while maintaining the error terms $\boldsymbol{u}$ and $\boldsymbol{v}$ defined for the BYM model (Dean et al., 2001). The BYM2 model uses a similar structure to the BYM model, only differing in the definition of $\boldsymbol{\sigma}$, which is given for the BYM2 model as,

$$\boldsymbol{\sigma} = \frac{1}{\sqrt{\tau_\sigma}}\left(\sqrt{\rho}\,\boldsymbol{u}_\star + \sqrt{1-\rho}\,\boldsymbol{v}\right) \tag{2.19}$$

with covariance matrix defined as,

$$Var(\boldsymbol{\sigma}|\tau_\sigma,\rho) = \tau_\sigma^{-1}\left((1-\rho)\boldsymbol{I} + \rho\boldsymbol{Q}_*^{-}\right) \tag{2.20}$$

where $\boldsymbol{u}_{\star}$ is a scaled version of the spatial random effects term and $\boldsymbol{Q}_{*}^{-}$ is the scaled version of the inverted proximity matrix.

For $\boldsymbol{\sigma}$ to legitimately be a standard deviation for overall error, the variance of both $\boldsymbol{u}_{\star}$ and $\boldsymbol{v}$ should be approximately equal to 1. This is achieved through the appropriate selection of a scaling factor, s. As the scaling factor is only dependent on the underlying neighbourhood structure, as opposed to outcomes of the model fit, the scaling factor can be introduce into the model as data prior to running the model (Morris et al., 2019). By introducing the scaling factor as data, we avoid the need for the scaling factor to be repeatedly calculated at each iteration of an MCMC sampler.

2.7 Model fitting in R

The models introduced within this section can be modelled using a number of different approaches within the programming language R. We consider four different approaches, each of which uses Bayesian modelling techniques and includes a CAR component to deal with spatial variation. Comparisons between the different approaches is presented, with the findings lifted in part from a paper comparing different software implementations for spatial disease mapping (Vranckx et al., 2019).

2.7.1 OpenBUGS (Open Bayesian inference Using Gibbs Sampling)

BUGS, or Bayesian inference Using Gibbs Sampling, is a piece of software designed for the analysis of Bayesian models using Markov Chain Monte Carlo (MCMC) methods (Lunn et al., 2009). OpenBUGS is a more recent open source addition to the BUGS project which allows MCMC methods to be fitted using a number of different sampling approaches including Gibbs, Metropolis-Hastings or slice sampling (Lunn et al., 2000). Using the package `R2OpenBUGS` allows for OpenBUGS code to be called directly from, and subsequently analysed in, R (Sturtz et al., 2005).

CAR models can be fitted in OpenBUGS using the built in function `car.normal()` (Spiegelhalter et al., 2003). The `car.normal()` function takes four values as inputs and are defined as follows:

adj[]: The second column of a graph edgeset as described in Section 2.4. This is a vector listing the ID's of neighbouring cells in order.

weights[]: A vector containing the weight associated with each entry in the **adj**[] vector. These weights follow the definition given to the weights matrix defined in Section 2.4.

num[]: a vector containing the number of neighbours in each cell.

tau: A scalar argument representing the precision parameter of the CAR prior, defined in Section 2.6.1 for the CAR model as τ_u.

The ability to fit CAR models using the `car.normal()` function can be easily extended to allow the BYM and BYM2 models to be fitted using BUGS, however is more complex for models such as the Leroux. In comparison to other software implementations, OpenBUGS was found to produce similar results but at much slower rates (Vranckx et al., 2019).

2.7.2 Integrated Nested Laplace Approximations (INLA)

Integrated Nested Laplace Approximations (INLA) refers to a Bayesian model fitting methodology which acts as an alternative to MCMC based approaches. INLA instead uses Laplace approximations to calculate posterior estimates for approximate Bayesian inference of a class of models known as latent Gaussian models. Generally, the term latent Gaussian models refers to a subclass of models taking the basic form of a variable, y_j, which is assumed to follow some distribution family, a link function used to model some structured additive predictor required for the distribution family and a third stage allowing for prior and hyperprior assignment (Rue et al., 2009). We say that the BYM model can be referred to as a latent Gaussian model, as the response is assumed to follow a Poisson distribution, with a log link function used for modelling the Poisson rates η_j, with priors and hyperpriors assigned to each of the requisite parameters. Furthermore, the Besag, Leroux and BYM2 models can be classified as latent Gaussian models and therefore be fitted using INLA. The absence of any random sampling process results in the INLA methodology providing deterministic approximations of the parameter marginals.

The main benefit of INLA's deterministic approach is that the parameter estimates are produced much quicker than the MCMC approaches applied in our other software implementations. The deterministic nature of the INLA methodology also ensures that sample convergence and mixing do not need to be assessed as they would generally be for MCMC methods (Moraga, 2019). Despite being computationally less intensive, there is evidence to suggest that estimates produced using INLA, share similar levels of accuracy to estimates produced using MCMC methods (Smedt et al., 2015). Fitting and assessing models using the INLA methodology is easily accessible within R through the R-INLA package, which directly allows for the specification of Besag, BYM, Leroux and BYM2 models.

R-INLA has however been criticised due to difficulties in explicitly defining parts of models. For example it has been noted that hyperpriors cannot be defined and implemented easily, particularly when compared to other software implementations such as OpenBUGS (Carroll et al., 2015). Alternative approaches to the R-INLA package may therefore be more appropriate for defining and developing on existing modelling approaches.

2.7.3 CARBayes

CARBayes (Lee, 2013) is a spatial modelling package in R, which uses MCMC methods for fitting spatial models with CAR priors. Of the models presented in Section 2.6, the CARBayes package is capable of fitting the BYM and Leroux models through the functions `S.CARbym()` and `S.CARleroux()` respectively. Furthermore the Besag model can be fitted using the `S.CARleroux()` function and setting the value of ρ equal to one. The CARBayes package allows for a range of distribution families including Binomial, Poisson and Zero-Inflated Poisson (ZIP) distributions. In addition to the `S.CARbym()` and `S.CARleroux()` functions, each of the CARBayes models can also be fitted through the function `Bcartime()`, found within the R package bmstdr (Sahu, 2022). Using the `Bcartime()` function allows for easy spatial modelling and analysis of areal unit data through both CARBayes and R-Inla.

CARBayes results were found to be roughly comparable to those in other software packages, albeit with marginally wider credible intervals (Vranckx et al., 2019). Unlike R-Inla, the CARBayes modelling functions cannot contain areal units without any neighbours. It is of further note that missing response values are generally predicted by the model under CARBayes, with the exception of the ZIP distribution family which requires that all response values are observed. While CARBayes is a very useful tool for fitting existing spatial modelling structures, an alternative software implementation should be used when developing bespoke models that fall outside of the CARBayes framework.

2.7.4 Stan

The final software implementation that we consider is that of Stan (Carpenter et al., 2017). Stan is a C++ based package that uses MCMC algorithms to fit a variety of Bayesian models specified by the user. Stan primarily uses a No-U-Turn Sampler (NUTS) approach for obtaining samples from a model's posterior distribution. NUTS is a sampler intended as an extension to Hamilton Monte Carlo that avoids retracing previous steps so that the sampler should perform more efficiently (Hoffman and Gelman, 2014). Stan code is typically split into a number of different blocks, each of which specify different parts of the model. Some examples include:

data: A block used to specify all of the inputs required for the model.

transformed data: A block specifying any transformations that are required. For example, to use the log of any inputs in the model.

parameters: A block specifying any variables which are to be directly sampled in Stan.

transformed parameters: Additional variables calculated from the data,transformed data and parameters blocks. The variables specified in the transformed parameters blocks are included as output for each draw.

model: The block used to define the model fitted in Stan.

generated quantities: A block used for generating posterior inference directly in Stan. Alternatively, this posterior inference can often be calculated from the Stan output in R if required.

For each variable introduced in Stan it is important to declare the data type (e.g. integer, vector, matrices) along with any further required information (e.g. a minimum integer value, the total number of vector entries, the number of rows and columns included in a matrix). Stan can be run directly in R through the use of the R-Stan package.

While Stan does not contain any direct functions for fitting models with CAR priors, the BYM2 model can still be fitted in Stan (Morris et al., 2019). To save memory and use a computationally less expensive approach, it is recommended that neighbour relations are defined as edgesets as opposed to full adjacency matrices. The use of edgesets lends itself particularly well to sparse matrices such as those commonly used when defining the spatial weights for areal modelling. The CAR component can be defined from the edgeset in Stan using the joint specification definition of ϕ introduced in Section 2.6 (Equation 2.14). A light sum to zero constraint is imposed on ϕ by ensuring the mean of ϕ is Normally distributed around zero with little variation.

Compared to other software packages using CAR-based modelling approaches, Stan was found to produce reliable estimates quickly, albeit at a slower rate than INLA (Vranckx et al., 2019). The ability to specify each stage of the modelling process means that Stan is particularly adept at offering a flexible approach for fitting accurate models with CAR priors. We therefore suggest that the Stan software is ideal for fitting bespoke models which incorporate CAR priors.

2.8 Simulation methodology

Simulations can provide an ideal framework for observing the performance of a process under a range of different conditions. By using Bayesian models as a basis, we ensure that simulations are informed by the data and prior information included in the model. Using the MCMC based sampling techniques discussed earlier in the chapter provides a range of sample values for each parameter and our response value. By adjusting observed sample parameters, we can also consider simulations produced under systematically different conditions to the observed data. Further detail on how simulations have been applied for this thesis are given in Chapter 5 in the context of our data and resulting models.

Chapter 3

Spatial model application to completely observed areal imaging data

Using the techniques introduced in Chapter 2, we propose fitting a spatial model to tree location data collected across an entire area of interest. As i-Tree Eco data is only collected at surveyed locations, we instead use tree locations taken from the ProximiTREE and National Tree Map (NTM) datasets for the areas of Cambridge and Petersfield respectively. By using fully observed data, we expect less variation in our model parameters due to the availability of more information and the absence of extrapolation for unobserved areas. A drawback to the ProximiTREE and NTM data is that trees are defined based on some minimum height, whereas trees are defined for i-Tree Eco based on the diameter at breast height. The difference in tree definitions could result in differing conclusions based on the fitted model and any resulting simulations. We therefore present an approach for modelling fully observed ProximiTREE and NTM data in this chapter, before proceeding to present a modelling approach for i-Tree Eco data in Chapter 4.

Before detailing the spatial model approach taken for completely observed data, we first introduce the data being modelled, followed by the data used as covariates in both modelling chapters.

3.1 Initial data set introduction: Cambridge ProximiTREE

As discussed in Section 1.3.1, ProximiTREE (BlueSky, 2020b) data can be used to derive tree locations using a number of techniques, such as areal photography. Under the ProximiTREE data, trees are defined based on a height over 1m, a much looser definition compared to the National Tree Map (NTM) definition (3m or higher) and the i-Tree Eco definition (Diameter at breast height of 7cm or larger). Due to the tree definition used for the ProximiTREE data, we would expect that the rate of trees observed by the ProximiTREE dataset would generally be higher than the rate of trees observed in i-Tree Eco and the NTM datasets for comparable areas.

In this chapter, we explore ProximiTREE data taken from the Cambridge area. Cambridge was selected for analysis due in part to the availability of both i-Tree Eco and ProximiTREE datasets. Furthermore, as Cambridge is a mid-sized city, with a total area of approximately 4,000 Hectares, this allows for easier comparison between the findings in the Cambridge and Southampton areas later in the Thesis.

In total, the Cambridge ProximiTREE data contains 335,972 trees, with 226,354(67%) trees of height larger than 3m. On average, 82.6 trees were observed per Hectare across the Cambridge ProximiTREE data, much higher than the estimated 52.2 trees per Hectare from the Cambridge i-Tree Eco data. Spatial exploratory analysis using Moran's I and Geary's C suggested highly significant evidence of clustering, using both a statistical test and Markov Chain Monte Carlo methods.

Tree densities were summarised from the ProximiTREE data across the Cambridge area by overlaying hexagonal cells of size 0.5Ha as described in Chapter 2. Discussion and justification for the selected cell shape and size, is provided in Section 4.8. The number of trees observed within the cells, ranged between 0 and 441 trees, with Figure 3.1 illustrating the number of trees observed in all cells across Cambridge. Visual inspection of the number of trees in each cell suggests a possible spatial effect, with nearby cells often appearing to contain similar numbers of trees.

Around the border of Cambridge, cells are not entirely located within the area of interest. To ensure our model accounts for the proportion of observed area within the cells, we introduce the expected number of trees in each cell as an offset. The expected number of trees is calculated from the observed area for each cell using internal standardisation, a process detailed in Section 4.2. The inclusion of an offset avoids the model penalising results from cells around the border, which could otherwise be attributed to the area observed in the cell rather than the area characteristics.

Consideration of the G function and Ripley's K, strongly suggested the presence of clustering amongst the tree rates in the cells (Figures A.1 and A.2 in Appendix A). We therefore suggest that the inclusion of a spatial component should be considered when modelling the Cambridge ProximiTREE data.

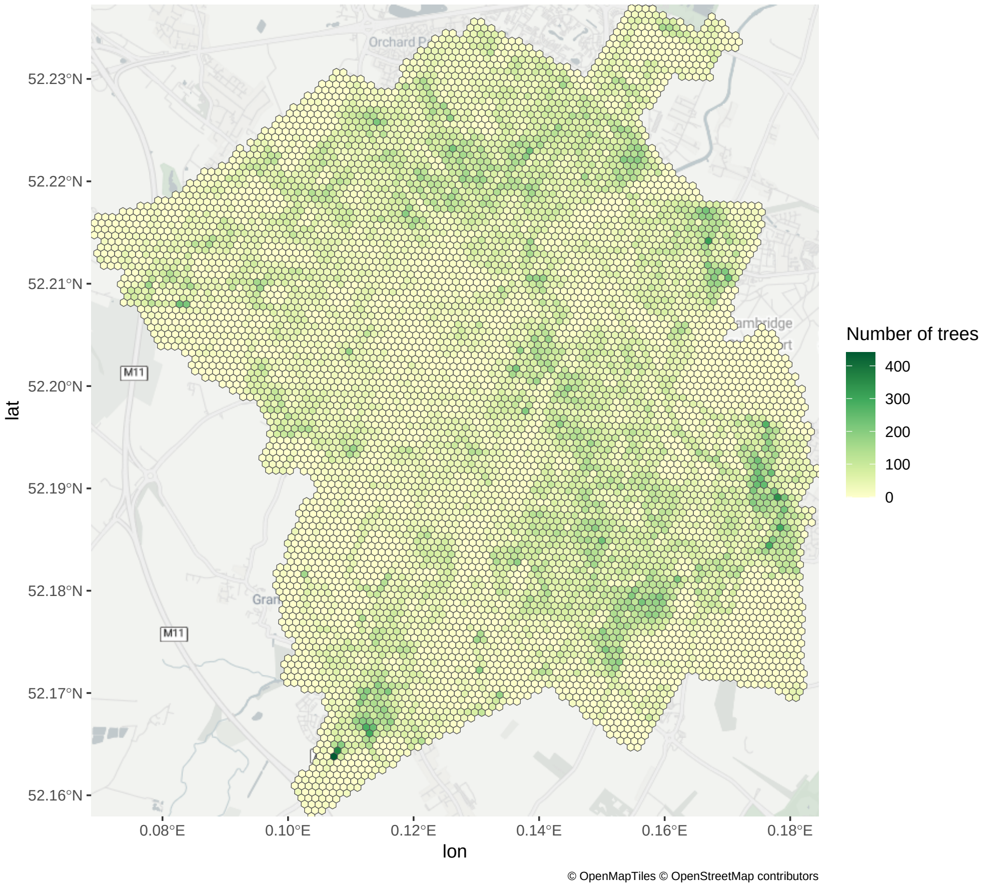


FIGURE 3.1: Plot of the number of ProximiTREE trees located within overlaid hexagonal cells of size 0.5Ha in Cambridge

3.2 Environmental covariates

In the following sections we give an overview of the covariates considered in our models. The following section discusses the covariates associated with the Cambridge area only, however tables and figures for Petersfield are provided in Appendix A. Environmental covariates were often considered based on their inclusion in the stratification process for the Southampton i-Tree Eco survey plots. Further details of the stratification process are provided in Section 4.1.

It is of note that the Normalized difference vegetation index (NDVI) was also explored as a possible covariate in the model, but is not described in detail here. The NDVI is a remote sensing based vegetation index which ranges between -1 and $+1$, with values approaching $+1$ generally indicative of dense healthy vegetation and lower values indicative of less or no vegetation (GISGeography, 2024). The NDVI data (NASA, 2016) was explored to assess whether the amount of vegetation in the cells, best explained the density of trees. Calculating the correlation coefficient between the tree density and NDVI values for each cell gave a value of 0.0538, indicating little association between the NDVI and the tree densities. Similar findings were found in other areas and an

NDVI covariate was explored but not included in the final model for any model selection processes conducted.

3.2.1 Land use

The density of trees within an area is expected to be strongly linked with the make up of the environment. For example, we would expect a higher number of trees to be observed in woodland areas compared to more industrial areas. For the purposes of modelling in this thesis, area environments are placed into broad Land use categories adapted from a 2015 land cover map from the UK Centre for Ecology and Hydrology (Rowland et al., 2017). Under the 2015 land cover map, multiple interlinking polygons representing different land use categories are used to build up a summary of the underlying environment within an area of interest. For ease of use in our analysis, the land use categories have been condensed using the broad habitat categories offered by the 2015 Land cover map (UKCEH, 2017). Additional Land use categories that could not be easily grouped together, were classified using an 'other' category. Details of the categories used to construct the final categories can be observed in Table 3.1.

TABLE 3.1: Summary of 2015 Land Cover Map categories used to build the final Land Use categories

Final categories	**Original Categories**
Grassland	Rough low-productivity grassland Fen marsh and swamp Neutral grassland Improved grassland
Suburban	Suburban
Urban/Urban industrial	Bare Urban industrial Urban
Woodland	Broad leaved, mixed and yew woodland Coniferous woodland
Other	Littoral sediment Supra-littoral sediment Freshwater Salt water Inland rock Arable and horticulture

By illustrating the land use coverage in Cambridge (Figure 3.2) we can observe that a large proportion of the area is suburban with some urban areas clustered in the centre and grassland spread throughout the city. We note that there appears to be very few woodland areas recorded in Cambridge. These findings are supported by our summary of land use coverage presented in Table 3.2, which suggests suburban areas encompass 42.3% of Cambridge compared to woodland area encompassing just 2.7% of Cambridge. The rate of trees within each land use category is also presented in Table 3.2 and illustrates the differences between the number of trees observed in each land use category. Of particular note is that the suburban tree rate is particularly high, with more trees per Ha observed in suburban areas than woodland areas. In our research, the high rate of trees in suburban areas is unique to the Cambridge ProximiTREE data and does not appear to be reflected in the Petersfield National Tree Map data or any of the i-Tree Eco data. The discrepancy between tree rates under different land use categories resulted in a land use category being included when carrying out a model selection process for the Cambridge ProximiTREE data.

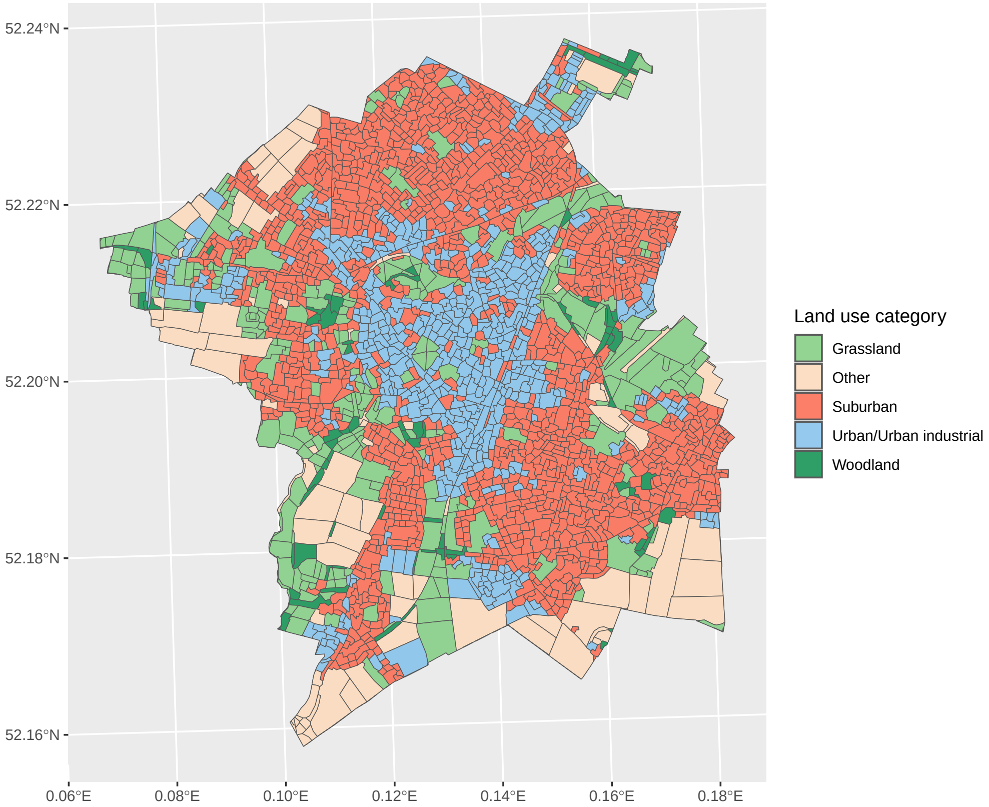


FIGURE 3.2: Plot of Land Use categories in Cambridge

3.2.2 OS MasterMap Topography

As an alternative to the land use categories presented in 3.1, we explore the use of the OS MasterMap Topography Layer from the Ordinance Survey (OS, 2019). The OS

Table 3.2: Summary of land use categories, adapted from the 2015 Land Cover Map. Numbers and rates of trees produced from ProximiTree data

	Land use category				
	Grassland	**Other**	**Suburban**	**Urban/ Urban industrial**	**Woodland**
Cambridge area coverage in Ha (%)	776.6 (19.1%)	682.3 (16.8%)	1,756.6 (43.2%)	743.4 (18.3%)	110.9 (2.7%)
Number of trees (%)	29,566 (8.8%)	19,030 (5.7%)	229,187 (68.2%)	48,191 (14.3%)	9,998 (3.0%)
Rate of trees per Ha	38.1	27.9	130.5	64.8	90.1

Mastermap Topography Layer data, henceforth referred to as the MasterMap data, is a frequently maintained framework which is used for the the referencing of geographic information across Great Britain (OS, 2017). The MasterMap data contains a range of environmental characteristics represented using points, lines and polygons as appropriate. The data product guide notes that features are represented using nine different themes, listed as:

- Administrative boundaries
- Buildings
- Heritage and antiquities
- Land
- Rail
- Roads, tracks, and paths
- Structures
- Terrain and height
- Water

For our modelling purposes, we have extracted data from the land theme which uses polygons as an indicator of areas classed as 'natural'. Our belief is that the natural polygons are more likely to contain highers densities of trees inside them. To assess this belief, we intend to explore whether natural polygons are required as part of the modelling. The high quantity of natural polygons means that this data must be reinterpreted for computationally feasible analysis. This has been achieved by calculating the proportion of each cell covered by the natural polygons and storing the result as a value ranging from zero to one. It is expected that the natural proportions will contain similar information to the land use data and a decision on which best represents the information will be decided as part of the model selection process.

Illustrations of the natural proportions for each cell suggest less natural area within the centre of Cambridge (Figure 3.3), similar to the high levels of urban land use observed in the centre of Cambridge. Generally the rate of trees was found to increase as the natural proportions got higher, with an exception amongst some of the highest proportions (Table 3.3). This exception may be explained away by the inclusion of other variables as part of the modelling process. Assessed natural cover categories were defined with zero representing cells with no natural coverage and categories one to four representing quartiles of the non-zero natural coverage values.

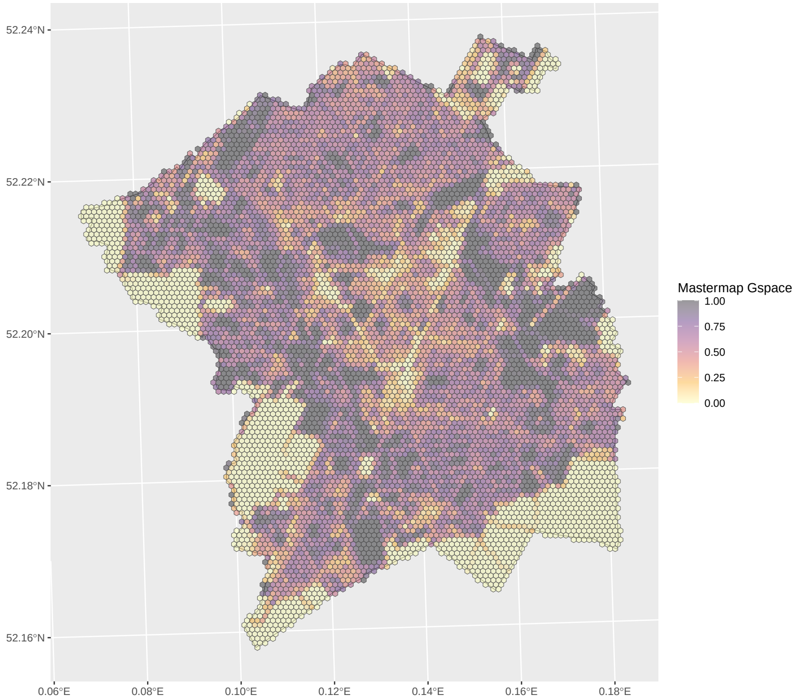


FIGURE 3.3: Plot of Greenspace coverage within overlaid hexagonal cells of size 0.5Ha in Cambridge. Greenspace coverage defined from OS MasterMap Topology data

TABLE 3.3: Summary of OS MasterMap Natural data for Cambridge. Numbers and rates of trees produced from ProximiTREE data

	Natural coverage category				
	0	**1**	**2**	**3**	**4**
Number of trees (%)	6,800 (2.0%)	53,491 (15.9%)	104,176 (31.0%)	117,529 (35.0%)	53,976 (16.1%)
Rate of trees per Ha	13.8	60.3	115.2	130.8	60.8

Using an identical approach to calculating the natural proportions, we also calculate the proportion of each cell covered in buildings from the Mastermap data. Using the proportion of buildings is again intended as an alternative interpretation of the land use data at a smaller level. As almost half of the cells have a building proportion value equal to zero, we have rewritten the building proportion variable into a categorical format, more conducive for analysis. The categorical format is created by storing the zero values in one category and defining four other categories based on the quartiles of the non-zero building proportion values.

In contrast to the natural proportion findings, the building proportions were generally found to be much higher in the urban areas in the centre of Cambridge (Figure 3.4). Furthermore, the rate of trees in cells with buildings was found to to be higher than cells without buildings (3.4). This again could be explained in a modelling context with consideration to other variables.

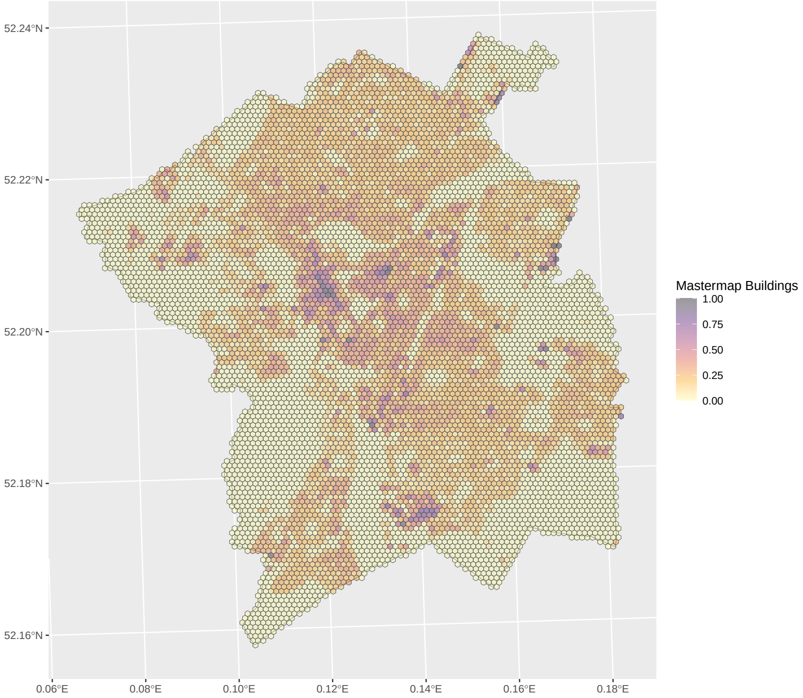


FIGURE 3.4: Plot of Building coverage within overlaid hexagonal cells of size 0.5Ha in Cambridge. Building coverage defined from OS MasterMap Topology data

TABLE 3.4: Summary of OS MasterMap Buildings data for Cambridge. Numbers and rates of trees produced from ProximiTREE data

	Buildings coverage category				
	0	**1**	**2**	**3**	**4**
Number of trees (%)	40,579 (12.1%)	61,961 (18.4%)	94,642 (28.2%)	86,662 (25.8%)	52,128 (15.5%)
Rate of trees per Ha	31.6	89.6	135.4	124.2	75.0

3.2.3 Air Quality Management Areas (AQMAs)

Air Quality Management Area (AQMA) data uses polygons to identify areas with poor levels of air quality (Defra, 2020). These areas are established through a review and assessment of the air quality by local authorities within the UK and indicate areas where the air quality is unlikely to meet the national air quality objectives. These objectives relate to the concentration of various pollutants, such as sulphur dioxide, and include dates for which the objective should be reached and then maintained. AQMAs are represented using polygons, which are generally indicative of areas with poorer air quality. We note that in some areas, such as Petersfield, no AQMAs are present, whereas in other areas, such as Birmingham, the entire city is classified as an AQMA (BCC, 2021).

AQMAs were selected as one of the strata when stratifying the survey plots due to the important role trees play in improving the air quality in urban areas. For example, the role that vegetation, particularly trees, could play in mitigating the effects of PM_{10} pollution has been highlighted in the literature (Tiwary et al., 2009). There is therefore interest in understanding the existing vegetation found in areas with particularly poor air quality (highlighted by the AQMAs).

In Cambridge, the AQMAs are largely confined to one large area of size 661Ha in the centre of the city (Figure 3.5). A slight difference in the rate of trees between AQMAs and non-AQMAs has been observed, with lower rates of trees observed in AQMAs. While the variation in tree rates may be explained away when accounting for other variables, there is some justification for investigating whether an AQMA variable is needed when modelling.

Table 3.5: Summary of Air Quality Management Areas (areas with poorer levels of air quality) in Cambridge. Numbers and rates of trees produced from ProximiTREE data

	Air Quality Management Area		
	Yes	**No**	**Total**
Cambridge coverage, ha (%)	662.6 (16.3%)	3,407.3 (83.7%)	4,069.9 (100%)
Number of trees (%)	43,642 (13.0%)	292,330 (87.0%)	335,972 (100%)
Rate of trees per ha	65.8	85.8	82.6

3.2.4 Indicies of Multiple Deprivation (IMD)

Indices of Multiple Deprivation(IMD)(MHCLG, 2019) are used to indicate the levels of deprivation at the Lower Layer Super Output Area (LSOA) level, a geospatial unit used throughout England and Wales for the reporting of statistics in small areas. The IMD consists of the following domains, each of which represents a different domain of deprivation: (Noble et al., 2019):

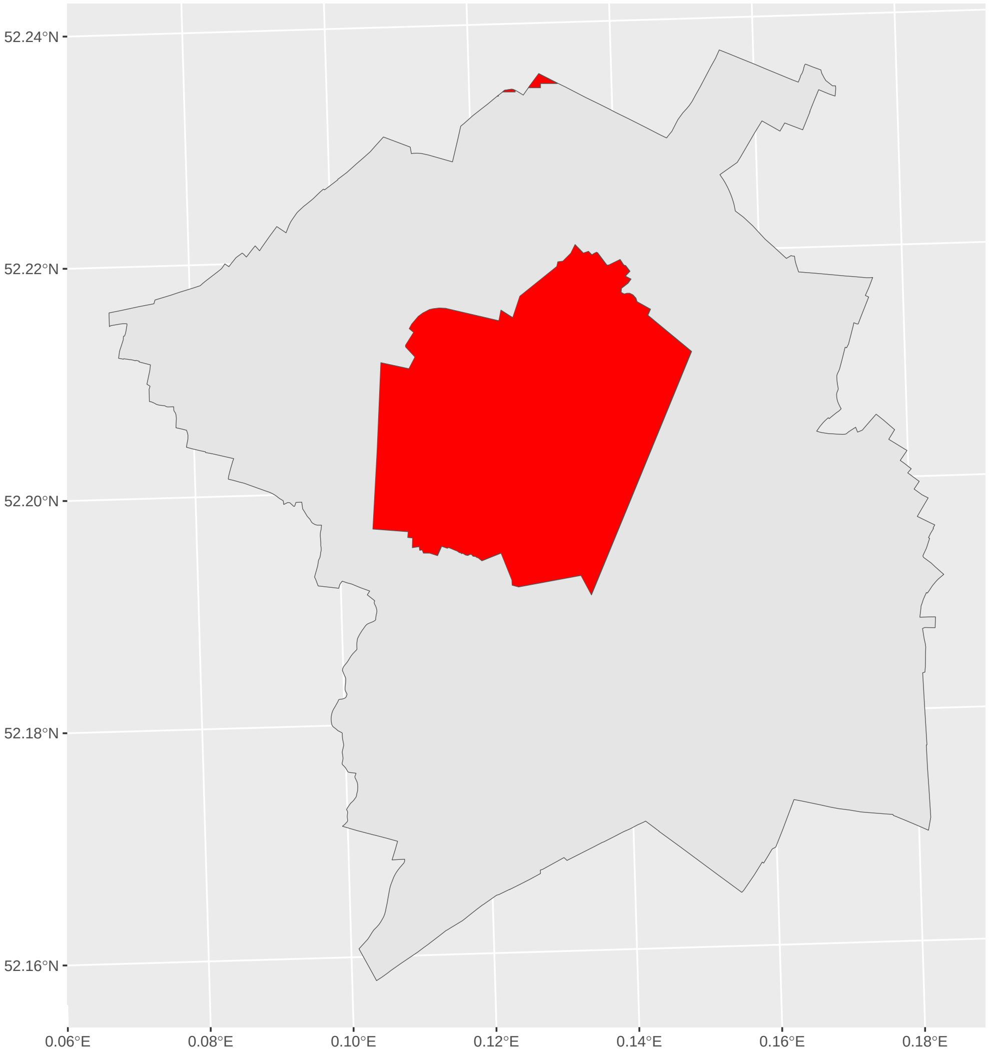


FIGURE 3.5: Plot of Air Quality Management Areas (AQMA) in Cambridge

- Income deprivation - The proportion of the population experiencing deprivation relating to low income.
- Employment deprivation - The proportion of the working age population who are involuntarily excluded from the labour market.
- Education, Skills and Training Deprivation - A measurement of the lack of attainment and skills in the population.
- Health, Deprivation and Disability - the risk of premature death and the impairment of quality of life through poor physical or mental health in a population.
- Crime - A measurement of the risk of personal and material victimisation at local level.
- Barriers to Housing and Services - A measurement of the physical and financial accessibility of housing and local services.

- Living Environment Deprivation - A measurement of the quality of the local environment.

Each domain is provided with a score ranging between 0 and 100, with higher values indicative of less deprived areas. A total IMD score is then calculated as a weighted combination of all the domains, using the domain weights provided in Table 3.6.

Table 3.6: Domain weights given to each individual domain when constructing the total Indices of Multiple Deprivation (IMD) score

Domain	**Domain weight (%)**
Income deprivation	22.5
Employment deprivation	22.5
Education, Skills and Training Deprivation	13.5
Health, Deprivation and Disability	13.5
Crime	9.3
Barriers to Housing and Services	9.3
Living Environment Deprivation	9.3

To interpret the IMD scores, it is suggested that deciles of the IMD scores be calculated at the UK level by ranking all LSOAs by their IMD scores and then dividing the LSOAs into ten equal groups. Deciles are interpreted as opposed to raw scores due to the scores not being easily interpretable on a continuous scale. For example a score of 60 does not necessarily indicate an area being twice as deprived as an area with a score of 30 (Noble et al., 2019), however modelling the IMD scores directly would assume that the difference observed is uniform for each unit increase. The IMD deciles are numbered between one and ten with one indicative of the most deprived areas and ten indicative of the least deprived areas.

Looking over the domain names it would appear that the Living Environment domain would be of particular interest when trying to establish environmental characteristics for modelling the observed and expected density of trees. The Living Environment domain is measured using an indoors sub-domain, based on the proportion of houses without central heating and the proportion of houses in poor condition, and an outdoors sub-domain, measuring the air quality based on emission rates for four pollutants and the number of road traffic accidents. For modelling tree densities only the air quality component would appear to be of particular interest, however air quality is only one component of the domain which is accounted for more directly using the AQMAs. We therefore consider the total IMD score when modelling as opposed to any of the individual IMD domains.

The IMD deciles for Cambridge suggest that there is generally little deprivation, with the exception of the North West of Cambridge which is largely more deprived in comparison to the rest of the city (Figure 3.6). Contrary to expectations, the rate of trees from the ProximiTREE dataset appear to be higher in the most deprived areas of Cambridge (Figure 3.7). As the most deprived Quintile consists of only 2% of the total area, we interpret this finding as an unexpected outlier resulting from such a small area being observed. The remaining quintiles behave approximately as expected with the second highest rate of trees per Ha being observed in the least deprived areas of Cambridge. As a result, the inclusion of an IMD decile variable has been considered when fitting statistical models.

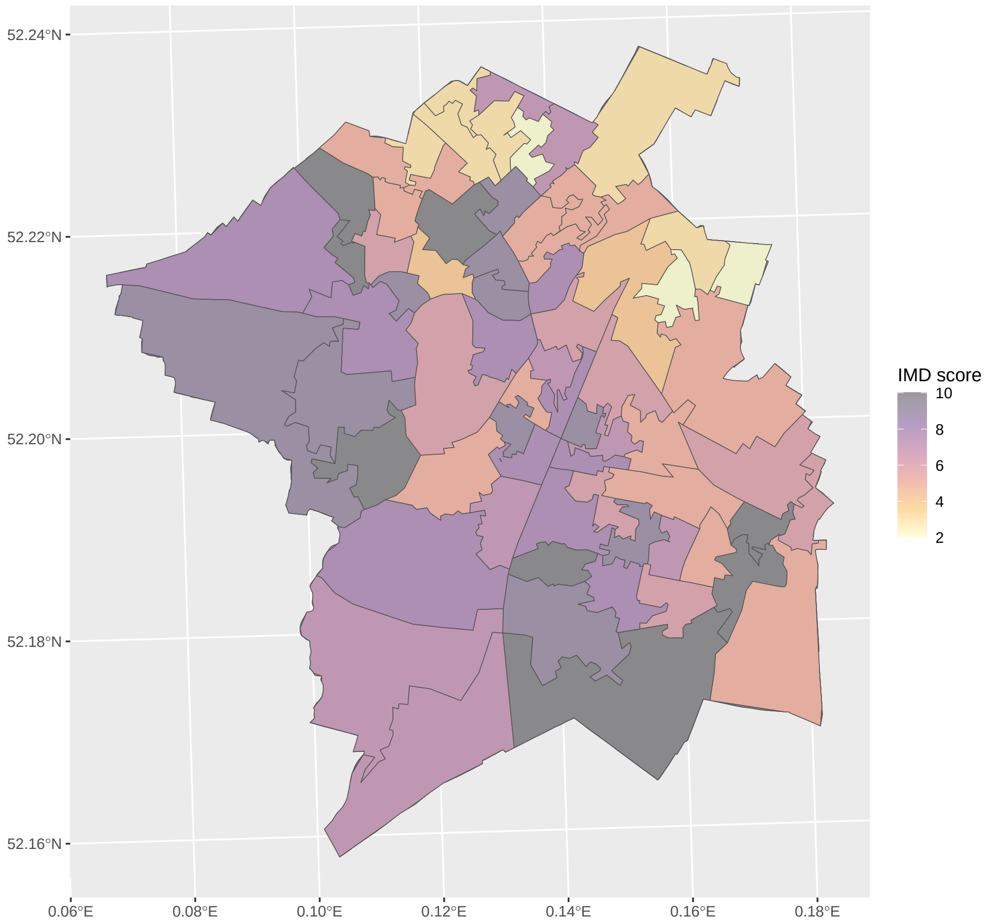


Figure 3.6: Plot of Indicies of Multiple Deprivation (IMD) deciles in Cambridge

Table 3.7: Summary of Indicies of Multiple Deprivation (IMD) quintiles for Cambridge. Numbers and rates of trees produced from ProximiTREE

	Indicies of Multiple Deprivation (IMD) quintile				
	1	**2**	**3**	**4**	**5**
Number of trees (%)	6,800 (2.0%)	53,491 (15.9%)	104,176 (31.0%)	117,529 (35.0%)	53,976 (16.1%)
Cambridge area coverage in Ha (%)	78.4 (1.9%)	398.3 (9.8%)	1152.8 (28.3%)	1293.1 (31.8%)	1147.4 (28.2%)
Number of trees (%)	8225 (2.4%)	31220 (9.3%)	98875 (29.4%)	93581 (27.9%)	104071 (31.0%)
Rate of trees per Ha	104.9	78.4	85.8	72.4	90.7

3.3 Full model definition

Using the model variables described so far as covariates, we consider approaches for fitting spatial models to the number of ProximiTREE trees observed in each cell. Relevant covariates are considered through the model selection technique, backwards selection, whereby parameters are systematically removed from the full model until they are deemed to be removing significant information. Spatial models have been fitted using the Leroux model in CARBayes as opposed to the BYM2 model in Stan. The Leroux model in CARBayes has been used as this provides a computationally efficient approach for fitting multiple spatial models as required for the backwards selection process. Furthermore, we consider the use of a scaled spatial component to be more beneficial when dealing with incomplete data that requires a consistent definition, than complete data where spatial components do not need to be established for missing areas. We note that CARBayes models have been fitted using the `Bcartime()` function in the R package bmstdr (Sahu, 2022), as discussed in Section 2.7.3.

Each of the ProximiTREE models were run for 100,000 samples following an initial burn in of 20,000 samples. After collecting the samples a thinning of 100 iterations was applied so as to make the MCMC samples more manageable. Due to a lack of additional information, default priors were used as defined in CARBayes. These priors are defined as:

$\beta \sim N(0, 1000)$ - A Normal distribution with mean 0 and variance 1000 for our regression parameters β

$\sigma^2 \sim IG(0.001, 0.001)$ - An Inverse-Gamma distribution with shape and scale equal to 0.001 for our overall standard deviation term, σ

$\tau^2 \sim IG(0.001, 0.001)$ - An Inverse-Gamma distribution with shape and scale equal to 0.001 for our precision term, τ

A summary of how the model selection process was conducted is included in Table 3.8. We note that the WAIC is treated with some caution due to the spatial nature of the data contradicting the assumption of independence between cell observations (Gelman et al., 2013). Therefore, if one model contains significant variables or does not present convergence issues, we have sometimes decided to select this model despite slightly higher WAIC values. For example, the natural and buildings variables sometimes have higher WAIC values, but were still included due to all of the parameters successfully converging to significant results.

As a final step of the model selection process, we consider the inclusion of an interaction parameter. Interaction parameters are not included in the full model at the beginning as this would make the model fitting impractical to perform, but are given consideration as part of the process. For fitting a model to the Cambridge ProximiTREE data, we found that an interaction term did not improve the model fit, due to a higher WAIC value and complications with parameter convergence.

The final model for the fully observed, Cambridge ProximiTREE data follows the Leroux model detailed in Equation 2.18 with the associated β values summarised in Table 3.9. We note that cases where cells did not contain any buildings was treated as the model baseline, meaning the category does not have an explicit associated β value. From Table 3.9 we observe that all β values have a significant effect, including each buildings coefficient. These results largely line up with the findings from our exploratory analysis, with more trees estimated in areas containing buildings and more trees generally observed in areas with more natural coverage.

From Figure 3.7 we observe that the mixing parameter, ρ, results are all close to one. This suggests strong evidence of a spatial effect present within our model. The presence of a strong spatial effect is in part used as justification for assuming a strong spatial effect when selecting our i-Tree Eco model priors in the following Chapter.

The accuracy of the model fit was assessed through comparison between the values provided by the MCMC samples and the observed Cambridge ProximiTREE values for each cell. Error values were calculated by subtracting the observed Y_j values from the mean of the expected Y_j MCMC sample values for each cell. From observing a density plot of the error values (Figure 3.8), the average expected values appear to accurately estimate the tree density for each cell, with estimates generally no more than four trees away from the observed value. The density appears symmetric around a centre of zero, suggesting the model is neither consistently over or underestimating the expected number of trees in the cells. We note that error values provided here are based on the mean expected number of trees for each cell and does not account for further variation observed within the cells. Generally all expected tree counts for the cells are relatively accurate, with 95% of all simulated expected tree values being within 15 trees of the observed tree counts.

Table 3.8: Summary of model selection process using backwards selection. The model components column is used to indicate which variables are included in the model at each stage of the backwards selection. The WAIC and an indicator of whether all variables are significant is presented for each model in the process. The addition of interaction terms is considered as the final stage of the selection process.

Model components (includes covariates and spatial component)	**Component removed**	**WAIC**	**Significant Variables?**
Spatial component, land use, NDVI, IMD Decile, Natural proportion and Buildings proportion (Categorical)	None	53764	No
	IMD	53771	No
	Land Use	53700	No
	NDVI	53761	Yes
	Natural	53705	No
	Buildings	53707	No
Land Use removed due to low WAIC			
Spatial component, NDVI, IMD Decile, Natural proportion and Buildings proportion (Categorical)	None	53764	No
	IMD	53701	No
	NDVI	53670	Yes
	Natural	53636	No
	Buildings	53640	No
NDVI removed due to low WAIC and removal resulting in significant variables			
Spatial component, IMD Decile, Natural proportion and Buildings proportion (Categorical)	IMD	53690	Yes
	Natural	53620	Yes
	Buildings	53630	Yes
IMD removed as convergence issues observed when removing other parameters			
Spatial component, Natural proportion and Buildings proportion (Categorical)	None	53690	Yes
	Interaction	53706	Yes
Interaction not included as WAIC is higher and the model includes convergence issues			

Table 3.9: Summary of Cambridge ProximiTREE model parameters

Name	**Symbol**	**Category**	**Mean (95% CI)**
Intercept	β_0	-	7.06 (6.95, 7.17)
Natural	β_1	-	1.14 (0.97, 1.29)
Buildings	β_2	$(0 < x \leq 0.11)$	0.92 (0.83, 1.02)
	β_3	$(0.11 < x \leq 0.18)$	1.19 (1.10, 1.32)
	β_4	$(0.18 < x \leq 0.26)$	1.17, (1.10 1.27)
	β_5	$(0.26 < x)$	0.96 (0.86, 1.11)

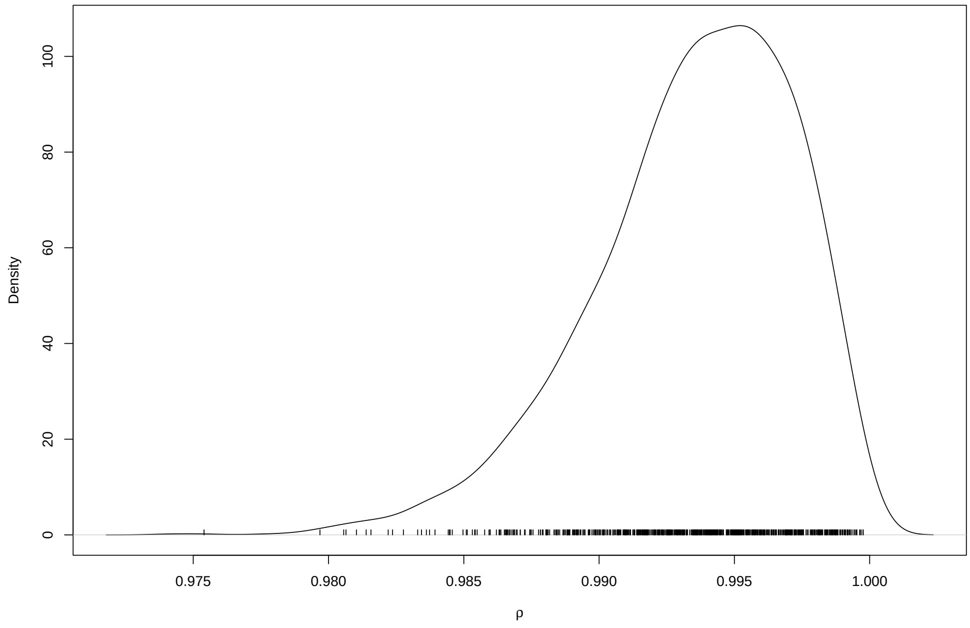


Figure 3.7: Plot of the Leroux mixing parameter, ρ, density for the Cambridge ProximiTREE model

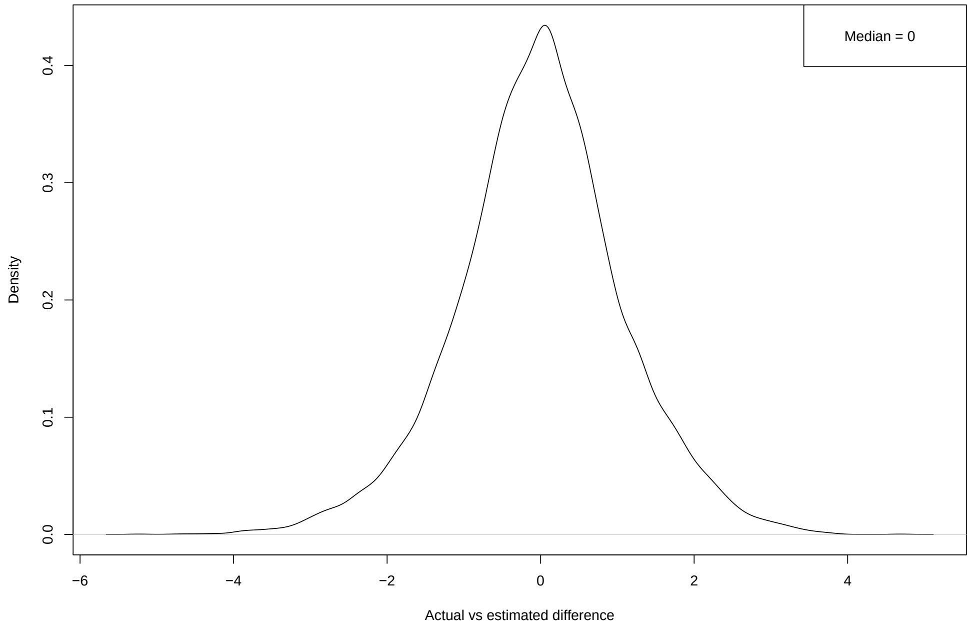


Figure 3.8: Plot of the mean Cambridge ProximiTREE model errors. Model error calculate by subtracting observed values from modelled values

Figure 3.9 provides an illustration of the mean expected number of trees in each cell for the Cambridge ProximiTREE model. From visual inspection, the model estimates appear to be accurately capturing the ProximiTREE data, with Figure 3.9 looking very similar to the plot of the ProximiTREE data presented in Figure 3.1.

The standard deviation in the expected number of trees, based on the ProximiTREE data, for each cell is displayed in Figure 3.10. In general, the standard deviation in the expected tree estimates appears to be higher in cells with higher expected tree estimates. We note that in cells where tree estimates are low, values cannot fall below zero which may, in part, be responsible for lower standard deviations in the modelled values.

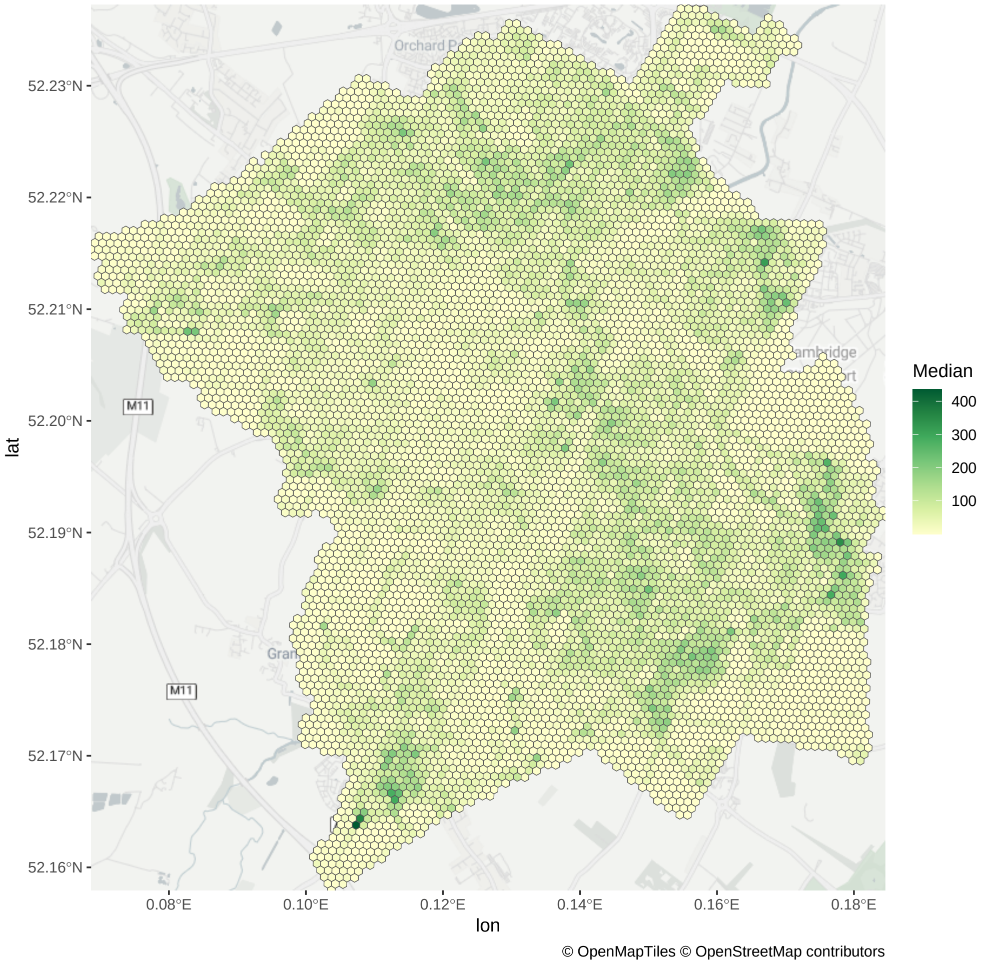


FIGURE 3.9: Median of the expected number of trees for each cell in Cambridge, as predicted from the Cambridge ProximiTREE data

3.4 Convergence and model fit diagnostics

Generally it appears that the model has converged as expected, with the traceplots for the β values, the precision parameters, τ^2_σ, and the mixing parameter, ρ, all suggesting convergence (Figure 3.11). Furthermore, the chains appear to have mixed well with

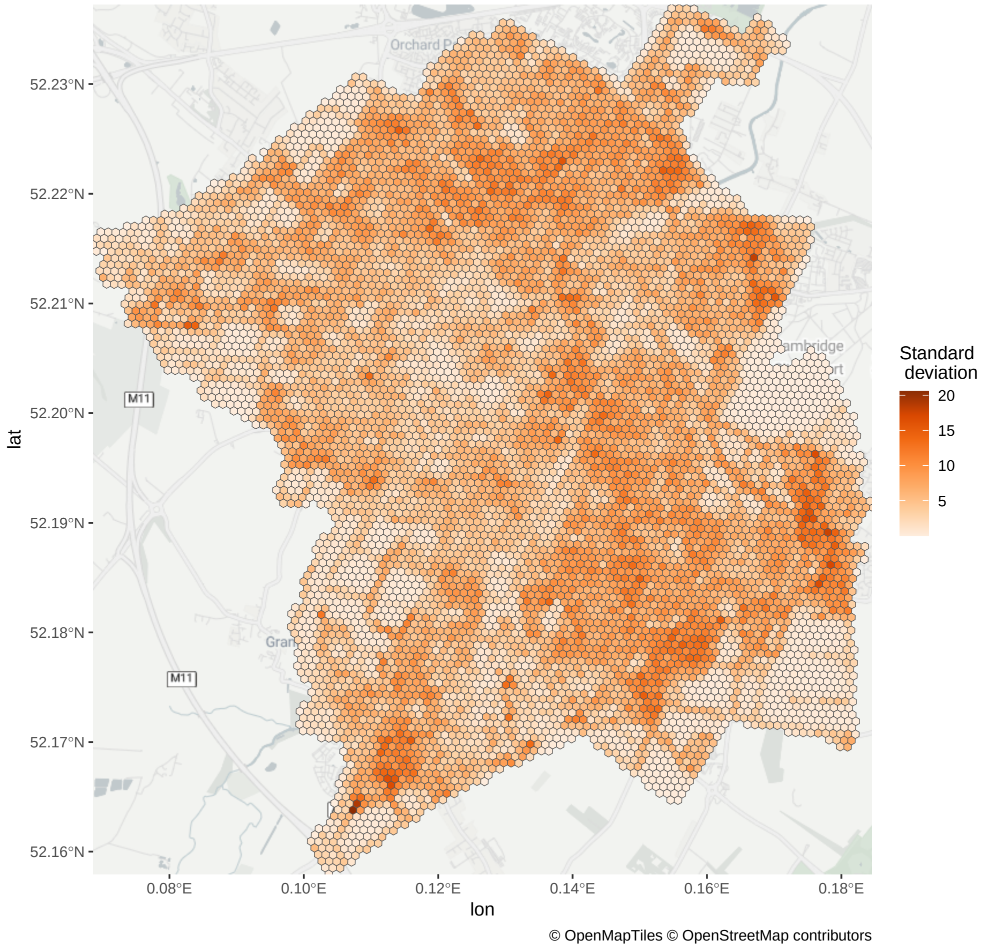


FIGURE 3.10: Standard deviation in the expected number of trees for each cell in Cambridge, as predicted from the Cambridge ProximiTREE data

few deviations, indicating an appropriate model fit. We note a much clearer presence of convergence and mixing amongst the uncertainty parameters, compared to the β parameters, but not to the level where this is of significant concern. The Geweke diagnostic values for each parameter were all below the 1.96 value, again suggesting the model is appropriate for our data.

3.5 Population densities and estimates

Total tree populations can be calculated for the model fitted to the Cambridge ProximiTREE data, by summing the number of trees estimated for each cell within each MCMC iteration. The expected tree populations appear roughly symmetrical (Figure 3.12), with a centre close to the total number of observed trees, 335,972. The range of populations observed by the model is relatively narrow, ranging between 332,700 trees 339,076 trees. The narrow range could be attributed to the relative accuracy of the

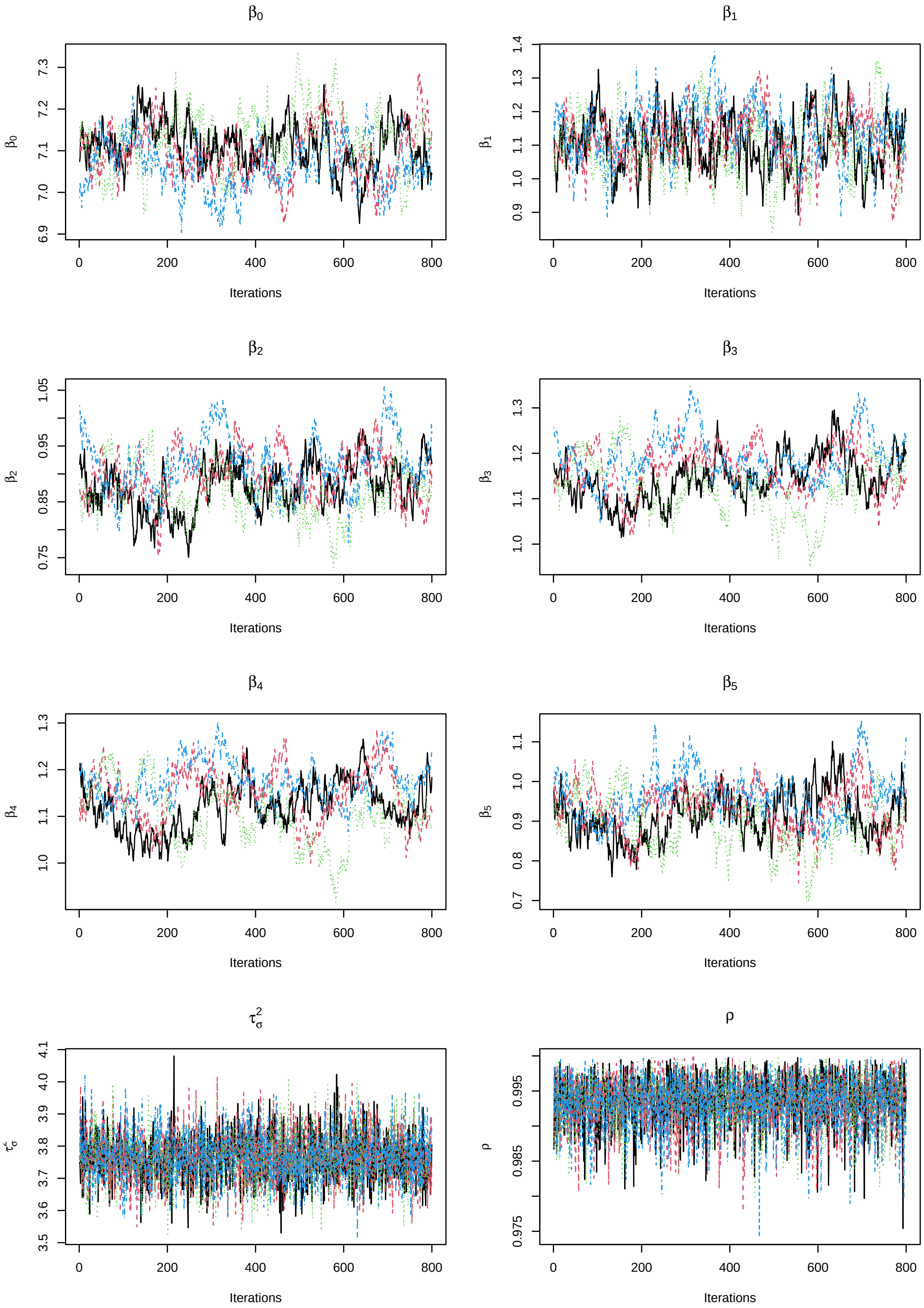


FIGURE 3.11: Parameter traceplots for the Cambridge ProximiTREE model

model and the lack of added variation from unobserved areas. We note that the narrow range of populations, suggests that using this model in simulations is also likely to result in a narrow range of simulated populations, however these populations appear to be representative of the observed data.

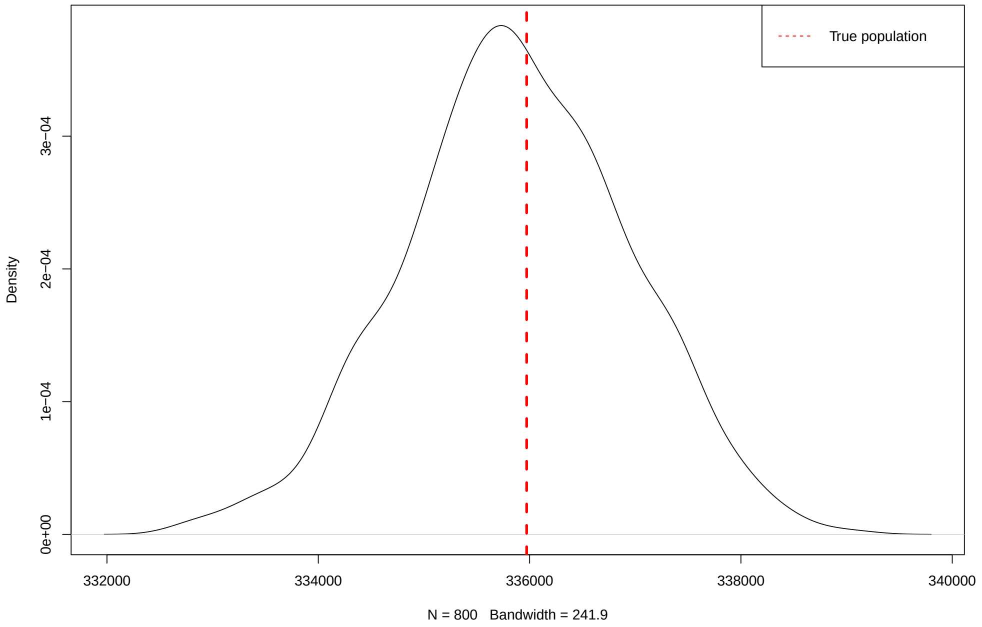


FIGURE 3.12: Plot of the population density as estimated from the Cambridge ProximiTREE data

3.6 Summary of findings in Petersfield

The approach outlined for modelling the Cambridge ProximiTREE dataset was replicated for National Tree Map (NTM) data located in the UK town of Petersfield. The Petersfield area is much smaller than that of Cambridge and Southampton, with a total size of approximately 801 hectares. The Petersfield area was therefore selected to explore how survey efficacy differs in smaller areas than Cambridge and Southampton. As Petersfield did not contain any AQMAs, we note that an AQMA variable was not considered for the Petersfield area.

Due to a smaller area size and more restrictive definition of what constitutes a tree than the ProximiTREE definition, the Petersfield NTM data was found to contain much less trees than the Cambridge ProximiTREE data. In total 25,689 trees were observed in the Petersfield NTM datasets, with a average of 32.1 trees per Ha across the entire area. Overlaying Hexagonal cells of size 0.1 Ha on the Petersfield NTM data, found a median of two trees in each cell, with a maximum of 21 trees in a cell. The use of smaller cells and the reduced rate of trees, resulted in cells generally containing less trees when compared to the Cambridge ProximiTREE data.

Fitting a Leroux model to the Petersfield NTM data reached similar conclusions to the Cambridge ProximiTREE model, albeit with the inclusion of an interaction term slightly improving the model. Cells with higher natural coverage were again estimated to contain more trees, however in contrast to the Cambridge ProximiTREE model, cells with higher building coverage were estimated to contain less trees. It would therefore appear that a relationship exists between the tree and building densities in both the Cambridge and Petersfield areas, however the nature of this relationship is not consistent. We note that the mixing parameter estimate is again close to one, suggesting heavy emphasis on the spatial error term over the independent error.

Conducting diagnostic checks, suggests the model fitted to the Petersfield NTM data is largely appropriate. Traceplots indicate that the coefficient and uncertainty parameters have converged, with chains mixing well. Convergence is also suggested by the Geweke diagnostics for the parameters, which are within the expected range for convergence. Analysis of the model residuals suggests a largely symmetric distribution centered around zero, with a little overestimation present, but not to the level where this is a concern. The populations estimated from the model are approximately centered around the observed tree population of 25,689 with all estimated populations no more than 1,000 trees away from the true populations.

Chapter 4

Spatial model application to partially observed survey data

For our analysis we explore a novel spatial modelling approach that fits a model based on the tree locations, whilst accounting for the fact that tree locations have only been observed within the survey plots. The spatial component of this model has been defined using an areal structure based on the tree locations found in an i-Tree Eco dataset. We begin this chapter by introducing the Southampton i-Tree Eco dataset, before proceeding to explain our approach for fitting a spatial model, containing a Conditional Autoregressive (CAR) component, to the data.

4.1 Initial data set introduction: Southampton i-Tree Eco

The area of Southampton is a city and port located along the Southern coast of the UK. Using Defra's 2011 Rural-Urban classification score, the Southampton area is classified as 'urban with city and town' ,due to less than 26% of the residential population living in areas classified as rural (Defra, 2014; Mutch et al., 2017). Based on the local authority district boundaries (ONS, 2020), Southampton is defined as a medium-sized city which encompasses an area of approximately 4,990 Ha, comparable to other UK cities such as Oxford and Exeter. Consideration towards ensuring Southampton is a green and environmentally sustainable city, have been explored and summarised in recent years, such as in the Southampton Green City Plan (Southampton City Council, 2020). The Southampton Green City Plan notes some of the environmental challenges Southampton is expected to face, including the expectation that Southampton's population will increase from 257,305 in 2020 to 292,505 by 2040.

In Summer 2016, a survey of trees in Southampton was carried out by the University of Southampton, working in partnership with Southampton City Council, Forest Research and Treeconomics, as a basis for investigating Southampton's urban forest structure. The survey was carried out by a group of students on the University of Southampton Excel Internship scheme, using the i-Tree methodology and following the i-Tree survey protocol (i-Tree, 2021a). For every observed tree in the survey plots, the location and characteristics of the tree, such as the species, crown condition and diameter at breast height were all recorded and entered into an i-Tree Eco dataset.

Commonly, i-Tree Eco surveys use approximately 200 survey plots of size 0.04 ha, as recommended by (Nowak et al., 2008b). However the Southampton i-Tree Eco survey instead aimed for a total of at least 400 accessible survey plots of size 0.04 ha, with 50 additional plots included so as to ensure that this target would be met. The Southampton i-Tree Eco report (Mutch et al., 2017) notes that the use of 400 survey plots results in a plot being observed on average every 12 ha, a much higher density of plots than any other UK based i-Tree Eco study, at the time the survey was conducted. We would expect that a higher density of plots should provide us with more information than comparable UK surveys, lending the Southampton i-Tree Eco survey to various model based simulations.

The locations of the survey plots were established through stratification, a process whereby we ensure that the observed area satisfies some predetermined strata criteria. For the i-Tree Eco Southampton dataset, plot locations were established through a stratification method whereby it was ensured that at least 30 plot centres were randomly located within each layer of the strata. Once this criteria was met, remaining plots were located randomly across Southampton. A circular boundary with radius 11.4m was placed around the randomly located centre points, providing an area of approximately 0.04 ha for observation. The strata were assigned using information on air quality zones, habitat, open spaces managed by the Southampton city council and index of multiple deprivation quintiles. These strata were selected as differences in the survey findings could potentially be observed between the different strata and because it was required for the strata to be sufficiently powered for intended subgroup analysis in the future. Once survey plot locations had been identified, it was then possible for the data to be observed and recorded for each of the survey plots locations.

Of the total number of survey plots, 8% of the plots were found to be inaccessible and resulted in missing data at these locations. These inaccessible plots are expected, with the i-Tree Eco user's manual (i-Tree, 2021b) recommending the addition of an extra 5-10% to the final number of plots used when setting up surveys. Common reasons for why survey plots are inaccessible in i-Tree Eco surveys include the area being too unsafe to approach and instances where the surveyors are denied access to the survey plot locations. In total, 38 plots could not be surveyed, leaving 412 plots containing 870 trees in the initial i-Tree Eco dataset. A visual inspection of the accessible and inaccessible plot locations (Figure 4.1) allowed us to conclude that the plot locations appear to be missing at random. We reached this conclusion as the missing plot locations do not appear to exhibit any indication of clustering, which could potentially be biasing any survey results. We therefore suggest that the number and location of inaccessible plots are not expected to affect the findings of our results.

Prior to any analysis, the dataset was first cleaned to remove any potential errors or inconsistencies in the data. As part of the data cleaning, the total number of trees rose to 876 as seven additional trees were found in a hard copy of the data and one tree, found to lie outside the plot radius, was removed. Distances and directions of the trees from the centre of the plots are provided in the i-Tree Eco dataset and were used to establish co-ordinates for each of the tree locations. In cases where either the direction or distance of the tree from the centre was missing, we instead estimated the tree location from areal images. While areal imaging would not be suitable for collecting all of the information required in the survey, it was deemed satisfactory for estimating tree locations given that we already knew that a tree was included inside the survey plot. In total, tree locations needed to be estimated for 28 trees across 13 different survey plots, a relatively small proportion of the overall data. Additional cleaning was carried out when prompted by the comments or the observation of unusual results in the data. For example, missing decimals and negative signs were inserted when appropriate for the plot centre latitudes and longitudes. Due to some discrepancies in the approaches taken to data cleaning, we would expect that the data analysed here will have slight differences when compared to the data used for the i-Tree Eco report.

From assessing the number of trees found in the plots, we note that in over half the survey plots no trees were observed. This high proportion of observed zeroes, sometimes referred to as zero inflation, should be addressed in our modelling procedure for providing accurate models. As zero inflated data is commonly observed in ecological data, we refer to the existing literature (Agarwal et al., 2002; Potts and Elith, 2006) on zero inflation when conducting our analysis. We have assessed whether the modelling approaches taken are currently compatible with existing approaches for dealing with zero inflated data. When not equal to zero, the number of trees observed in the survey plots tended to be relatively low, with a median value of three trees observed in survey plots containing at least one tree. In contrast, a few plots were found to be very densely populated, resulting in as many as 25 trees observed in a couple of plots. These findings are reflected in Table 4.1, which summarises the number of trees observed within the survey plots for six separate categories.

Table 4.1: Summary of the number of plots, whereby the observed number of trees lies within the given interval. Note that percentages are given as a proportion of the total number of plots

Number of trees	**0**	**1-5**	**6-10**	**11-15**	**16-20**	**21-25**
Number of plots (%)	231 (56%)	140 (34%)	21 (5%)	6 (1%)	5 (1%)	9 (2%)

Using the i-Tree Eco Southampton data, we are able to extract simple estimates of the total tree population by ignoring spatial and environmental effects.

By dividing through the total number of observed trees by the total observed area of the survey plots, an average of approximately 52.1 trees are observed per Hectare (Ha) within the survey plots. This value is similar to the 53 trees per Ha observed in the survey plots of an i-Tree Eco study which took place in inner and outer London, suggesting these results are typical of cities in the UK. Multiplying the average number of trees observed per Ha by the total area of Southampton provides us with a total tree population estimate of approximately 260,000 trees in Southampton. This is a little lower than the existing total population estimate of 267,000 trees found using the UFORE model in i-Tree Eco. This discrepancy in values can be attributed to differences in the data cleaning approaches and the use of a slightly tighter boundary for Southampton, that does not include the River Itchen. For a more accurate estimate with some associated error terms, we wish to account for characteristics of Southampton in our estimate. A more accurate estimate that also summarises the estimate error can be achieved through modelling techniques.

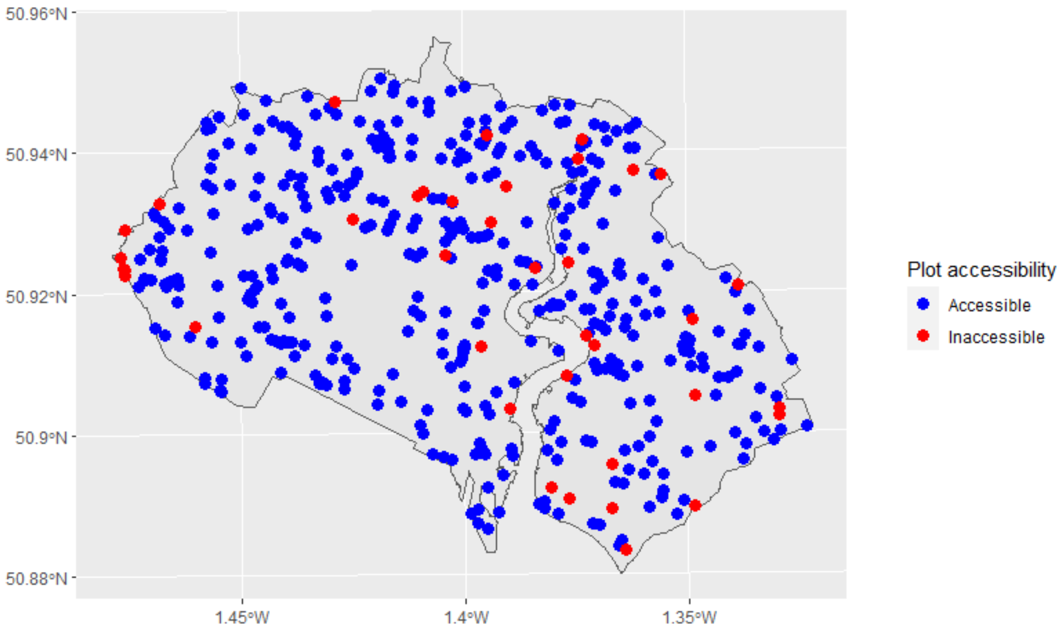


Figure 4.1: Accessible and inaccessible survey plots in the Southampton i-Tree Eco dataset. Note that the survey plots displayed here are not to scale

For the Southampton i-Tree data, point reference analysis could be carried out by assigning tree counts to each of the centre points of the sample plots, however this would not make use of information given by the tree locations within the plots. An example of how this looks can be seen in Figure 2.1.

4.2 Modelling setup for partially observed survey data

To provide a definition for our spatial component, we overlay an interlinking cell structure onto the Southampton area, as done in Chapter 3 for the ProximiTREE data. Following the findings discussed later in Section 4.8, the cells each have a hexagonal shape of size 0.5 Ha. Each of the covariates provided in Chapter 3 are calculated for the Southampton cells as previously discussed. Unlike with the ProximiTree data, the i-Tree Eco data only provides samples of an area as opposed to the entire tree population. We propose a novel approach whereby the number of observed trees from the i-Tree Eco data is calculated for each cell, along with the total area observed by the survey plots in each cell. The number of observed trees has been taken as our response variable when modelling and used to estimate the total number of trees we would expect to find within each cell. In cases where a cell bisects a survey plot, the number of trees in each cell is based on the given tree locations, whilst the survey area is calculated based on the amount of survey coverage in each cell. Commonly, cells do not contain any survey plot coverage, in which case we intend to predict the number present in the cell using our model. By applying a similar setup for our partially observed i-Tree Eco data to our fully observed ProximiTree data, we intend to allow for easy comparison between results. Comparison between i-Tree Eco and ProximiTree results is expected to be more advantageous for the Cambridge and Petersfield areas in which we have both i-Tree Eco and ProximiTree datasets available.

Using our modelling setup we establish how the number of observed trees in the overlaid cells can be used to provide an estimate of the total number of trees within each cell. Within this chapter we use $\mathcal{C}_i$ to refer to the survey plot area calculated for each cell, i, where $i = 1, 2, ..., n_c$ and n_c refers to the total number of cells within our area of interest. The term m_c is used for referring to the total number of cells in which at least some of the cell has been observed by the survey plots, i.e $C_i \neq 0$, and j as an indicator of length m_c for the cells in which $C_i \neq 0$, such that $C_j \in C_i$ and $m_c \leq n_c$.

We therefore fit a model using the number of trees in our observed cells, Y_j, as a response, where Y_j represents the trees observed within the survey plots for each cell j. From modelling the Y_j values we construct an approach for estimating the total number of trees inside all of the initial cells, indicated here using Z_i. We note that Z_i refers to the total number of trees both inside and outside of any survey plots for each cell, i, and that we expect $Y_j \leq Z_j$.

As C_j is not expected to be equal for all j, it must be ensured that the C_j values are accounted for when modelling Y_j. Because the C_j values arise from the design structure of the survey plot and cell locations, we include an offset in our model to account for the size of the survey plot area observed in each cell. By treating C_j as an offset, the survey plot areas in each cell are accounted for in our model without being treated as predictors. In practice, an offset is included by calculating the expected number of observed trees given the survey plot coverage, E_j, and including this in our model as discussed in Section 2.6.2. To calculate E_j a process know as internal standardisation was applied, whereby E_j is derived in part from our response variable, Y_j as follows:

$$E_j \equiv C_j \left(\frac{\sum_{j=1}^{G} Y_j}{\sum_{j=1}^{G} C_j} \right) \tag{4.1}$$

The use of internal standardisation to calculate the values of E_j has been described as 'correspond[ing] to a kind of null hypothesis' (Banerjee et al., 2014), whereby the observed number of trees are compared to the expected number of trees, given C_j.

By incorporating E_j, our model for Y_j takes the form:

$$Y_j|\eta_j \sim Poisson(E_j \eta_j) \tag{4.2}$$

where η_j is the relative risk defined using the BYM2 model as laid out in Section 2.6.5. From this equation we have a modelling setup that can be used for investigating the impact different variables have on tree densities, but which does not directly provide estimates for Z_i.

To provide estimates for Z_j we propose an approach whereby the expected number of trees, E_j^*, is calculated within the entire cell, rather than only the observed survey plots. The definition for our proposed value of E_j^* again uses internal standardisation and can formally be written as:

$$E_j^* \equiv A_j \left(\frac{\sum_{j=1}^{G} Y_j}{\sum_{j=1}^{G} C_j} \right) \tag{4.3}$$

where A_j is the total area of interest (i.e. Southampton) covered within each cell. A_j is calculated for each cell so as to adjust for cells placed around the border which are likely to contain land outside the area of interest.

We propose that estimates of Z_j are calculated from both the η values estimated by the BYM2 model in Equation 4.2 and the calculated E_j^* using:

$$Z_j|\eta_j \sim Poisson(E_j^* \eta_j) \tag{4.4}$$

By adjusting our offset value to use the expected rate of trees throughout the entire grid as a basis, our response is 'scaled up' so as to provide tree estimates for the entire cell. Under the approach presented in Equation 4.4 an assumption is made that the relative risk estimated from the partially observed survey plots in each hexagon is representative of the relative risk throughout the entire cell. Whilst this assumption may not fully hold in reality, we suggest that smaller cell sizes should often result in larger proportions of each cell being observed, minimising potential errors.

Using covariate information associated with each cell where $\mathcal{C}_i = 0$, estimates can be calculated for all E_i^*, η_i and using Equation 4.4, Z_i. For estimates of Z_i calculated outside the observed cells, extreme extrapolated values could be observed due to the use of an exponential scale in the BYM2 model, however this problem should be minimised through the use of survey locations representative of the underlying area.

When establishing η_i in areas where $C_i = 0$, we need to clearly define how we specify our spatial term, θ_i. We propose that θ_i be defined from the CAR model presented in Section 2.6.1, with θ_j values extracted from θ_i and included in the model provided in Equation 4.2. Through this setup we ensure that the spatial term is defined when fitting our model, whilst also ensuring spatial terms are estimated for all n_c cells. By scaling the θ_i values as part of the BYM2 model we ensure a consistent interpretation of the spatial component throughout the entire cell structure. A soft sum to zero constraint is place on the values θ_i, ensuring our spatial component is centered around zero and easily identifiable. An illustration of the median θ_i values modelled from the Southampton i-Tree Eco data is provided in Figure 4.2, in which we can observe how θ_i values are frequently clustered together.

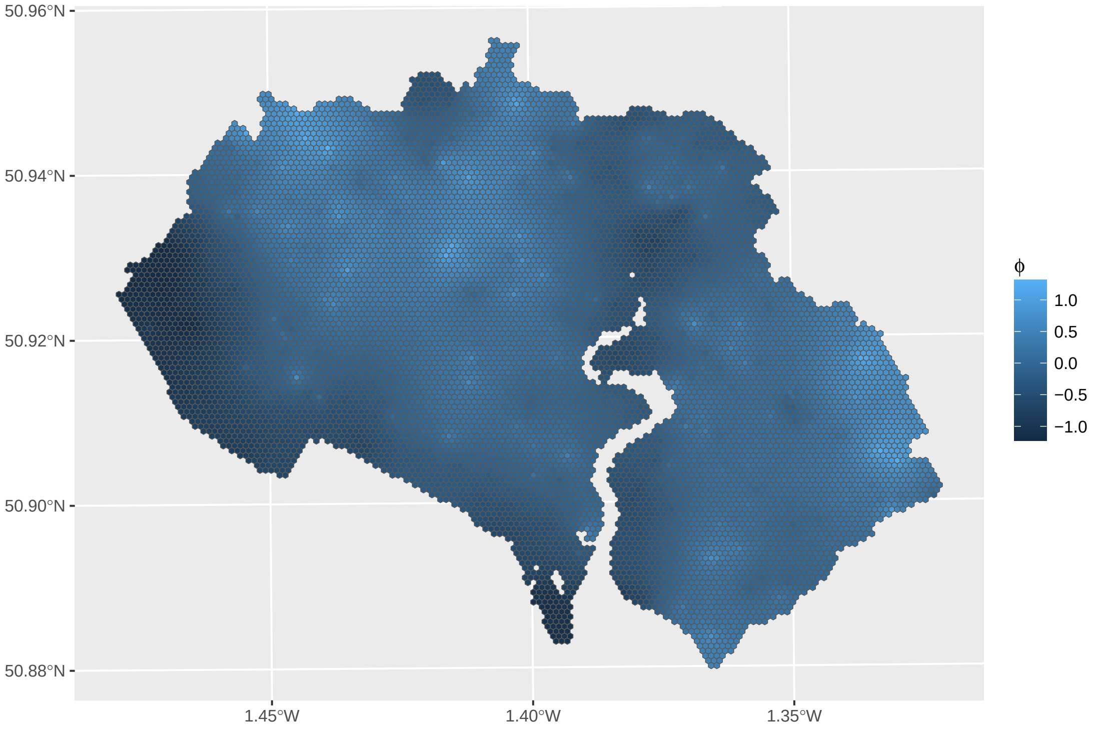


Figure 4.2: Plot of the median values of the spatial parameter, θ, by cell. Spatial model fitted to the i-Tree Eco Southampton dataset

To explore the efficacy of the modelling approach proposed in this section, we applied the methods presented here to partially observed ProximiTree data in Cambridge and

Petersfield. Generally we found that our estimated tree density values were accurate for a range of different cell sizes and shapes. The analysis conducted and the results found are presented in further detail in Section 4.8.

4.3 Zero-inflated data considerations

As the i-Tree Eco data is found to contain a large number of survey plots with no observations, alternative distributions to the Poisson that consider the presence of zero-inflation, should be considered. This is implemented by trading out the Poisson distribution used in Equations 4.2 and 4.3 with an alternative distribution. While the Hurdle and Zero-Inflated Poisson (ZIP) distributions are commonly used for modelling zero inflated data in ecology (Agarwal et al., 2002), both distributions present issues in predicting the values of Z_i. Under the Hurdle and ZIP distributions, the response is assumed to take the form of either 0 or some continuous value. While this assumption may hold when modelling within our observed responses, Y_j, the assumption is not expected to hold when estimating throughout the entire grid for Z_i. In short, while no trees are commonly observed within the survey plots, we would rarely expect no trees to be observed within the entire 0.5 Ha cells, as illustrated by the lack of empty cells when considering the Cambridge and Petersfield ProximiTree data. We instead propose using a negative binomial distribution as an alternative to the Poisson, when modelling our data. While still recommended as an approach for dealing with zero-inflated data within the ecology literature (Agarwal et al., 2002), the negative binomial distribution follows a similar definition to the Poisson with the addition of term controlling the level of variation observed. As opposed to the consistent variation assumed under the Poisson distribution, the negative binomial can assume less deviation around low values compared to higher values, ensuring relative risk estimates are close to zero but not constrained to zero.

4.4 Full model definition

Using the modelling approach outlined so far in this chapter, along with some of the covariates detailed in Chapter 3, a BYM2 model has been fitted to our Southampton i-Tree Eco data using the program Stan. The applied Stan code used existing approaches for modelling the BYM2 model in Stan (Morris et al., 2019) as a basis, but adapted the code to be suitable for our partially observed survey plot data approach, outlined in Sections 4.2 and 4.3. The final version of our Stan code is presented in full in Listing B.1 of the Appendix.

Each of the i-Tree Eco models were run for 10,000 samples following an initial burn in of 2,000 samples and thinning every tenth sample. The longer computational time associated with Stan modelling and the NUTS sampler resulted in a lower number of samples included in the model compared to Chapter 3. The number of samples used should still be large enough to produce simulations in Chapter 5 and tests are again conducted to ensure model parameters have converged.

Priors and hyperpriors were selected for the models as follows:

> $\beta \sim N(0, 5)$ - A Normal distribution with mean 0 and variance 5 for our regression parameters β. Default prior due to lack of specific information on covariate effect.
>
> $v \sim N(0, 1)$ - A Normal distribution with mean 0 and variance 1 for our independent uncertainty parameter, v. We expect the independent error to be centred around zero and have a standard deviation of one as a result of scaling the spatial component.
>
> $\frac{1}{\sqrt{\tau_\sigma}} \sim HalfCauchy(0, 25)$ - A half-Cauchy distribution with mean 0 and scale 25 for our inverse precision parameter $\frac{1}{\sqrt{\tau_\sigma}}$. A prior recommended in the literature for variance parameters in hierarchical models (Gelman, 2006).
>
> $\rho \sim Beta(0.5, 0.5)$ - A beta distribution with alpha and beta equal to 0.5 for our mixing parameter, ρ. A symmetric prior ranging between 0 and 1, which places more weight around the values 0 and 1. This prior was selected based on evidence of a strong spatial effect in the ProximiTree models.

Due to the computationally intensive nature of fitting multiple Stan models, a full model selection process is not presented here. Instead, the buildings and natural variables from the MasterMap data were included in the model to explore whether this provided a sufficient fit. The inclusion of the buildings and natural variables is based on our findings in Chapter 3, where both variables were included in the final model selection for models fitted to the Cambridge and Petersfield ProximiTree data. An interaction term between the natural and buildings variables has been included in the Southampton i-Tree Eco model, as including an interaction term was found to lower the LOOCV (1484.1 vs 1478.6) and the WAIC (1469 vs 1457.9).

Parameter information, related to the fitted Southampton i-Tree Eco model is provided in Table 4.2. From our results, we generally expect that as the proportion of natural coverage increases, the tree density will also increase. Furthermore, we expect the tree density to be lower in areas with a higher proportion of buildings, albeit with the tree density still larger in areas with higher natural coverage. In contrast, areas with low numbers of buildings are estimated to have lower tree densities as the amount of natural area increases. While there is no definite reason for this result, it could potentially be attributed to large natural areas generally containing less trees when a low number of buildings are present (e.g. golf course, school playing fields).

The uncertainty estimates provided in Table 4.2, appear to suggest that the spatial and non-spatial error is being accounted for appropriately. The mean of the uncertainty parameter suggests error terms are being accounted for in the model, whereas the mixing parameter being close to 1 suggests more emphasis being placed on the spatial random error term in comparison to the independent random error term.

TABLE 4.2: Summary of Southampton i-Tree Eco model parameters

Name	**Symbol**	**Category**	**Mean (SD)**	**95% CI**
Intercept	β_0	-	$-2.71\,(0.61)$	$(-3.96,\ -1.57)$
Natural	β_1	-	$2.3\,(0.52)$	$(1.33,\ 3.40)$
Buildings	β_2	$(0 < x \leq 0.12)$	$1.46\,(0.71)$	$(0.09,\ 2.86)$
	β_3	$(0.12 < x \leq 0.19)$	$1.05\,(1.00)$	$(-0.89,\ 3.00)$
	β_4	$(0.19 < x \leq 0.25)$	$-0.42\,(1.12)$	$(-2.61,\ 1.67)$
	β_5	$(0.25 < x)$	$-0.41\,(0.94)$	$(-2.26,\ 1.40)$
Interaction	β_6	$(0 < x \leq 0.12)$	$-1.12\,(0.75)$	$(-2.61,\ 0.31)$
	β_7	$(0.12 < x \leq 0.19)$	$-0.75\,(1.41)$	$(-3.56,\ 2.07)$
	β_8	$(0.19 < x \leq 0.25)$	$1.41\,(1.80)$	$(-2.10,\ 4.96)$
	β_9	$(0.25 < x)$	$0.61\,(1.74)$	$(-2.77,\ 4.03)$
Uncertainty parameter	$\frac{1}{\sqrt{\tau_\sigma}}$	-	$1.28\,(0.32)$	$(0.74,\ 1.97)$
Mixing parameter	ρ	-	$0.91\,(0.11)$	$(0.60,\ 1.00)$

Model accuracy was assessed through comparison of the observed Y_j to the expected Y_j values taken from the model. In Figure 4.3 we present a summary of the estimation error, calculated by subtracting the observed Y_j values from the mean of the expected Y_j MCMC sample values for each cell. Estimation errors close to zero should be indicative of an accurate model estimation, whereas values larger and smaller than zero are indicative of model over-estimation and underestimation respectively. Figure 4.3, largely suggests a high level of model accuracy, with the majority of the errors being close to zero. While larger errors are present, these appear to be infrequent and the symmetric nature of the density plot suggests the model is not consistently over-estimating or under-estimating the expected number of observed trees. These findings were largely found to hold when also considering the model error associated with non-zero Y_j values.

Using the fitted model, estimates can be provided for the total number of trees within each cell, as detailed in Section 4.2. Figure 4.4 illustrates the log of the expected number of trees for each cell covering Southampton. Log values are presented here for the purposes of illustration with median expected values from the MCMC samples taken for

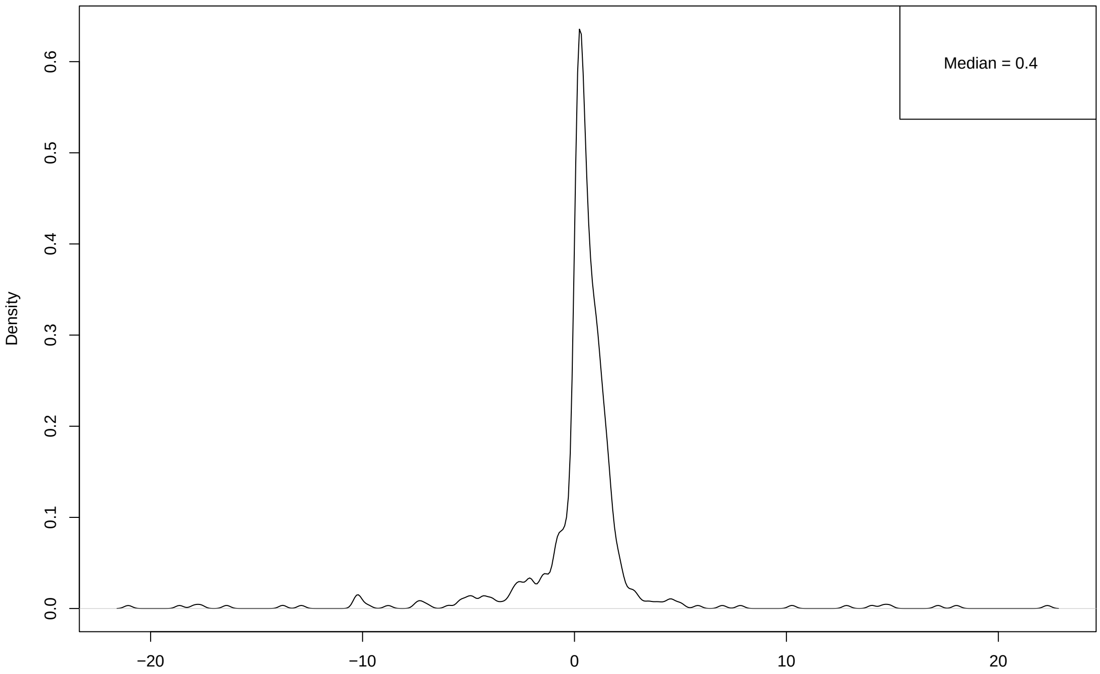


FIGURE 4.3: Plot of the mean Southampton i-Tree Eco model errors. Model error calculate by subtracting observed values from modelled values

each grid. From Figure 4.4, characteristics of the Southampton area are represented, such as the increased tree density in the Southampton common around the centre of the map and the low density of trees observed around the docks in the South-West. The median expected tree numbers for each cell range between 0.2 and 456.6, with a mean value of 17.2 trees in each cell. Comparing these results to the ProximiTREE and NTM data explored in Chapter 3 and tree density estimates from the i-Tree Eco model appear a little lower than expected, with mean estimates of 82.3 and 34.6 for the ProximiTREE and NTM data respectively. It is of note that the summarised expected tree numbers are averages and that much higher values have been observed for some cells. Furthermore, tree simulations will be generated from a negative binomial distribution, which could lead to more extreme tree density estimates in some of the cells.

The standard deviation in the expected number of trees was found to generally be higher in cells with higher tree density estimates, a point illustrated by comparing the log standard deviations (Figure 4.5) in each cell to the log medians(Figure 4.4). Generally the standard deviations appear to be high relative to the median estimates, however this is to be expected given how little of the Southampton area is observed by the survey plots. We note that the uncertainty associated with each cell estimate is likely to increase in the simulations, when generating samples from a negative binomial distribution.

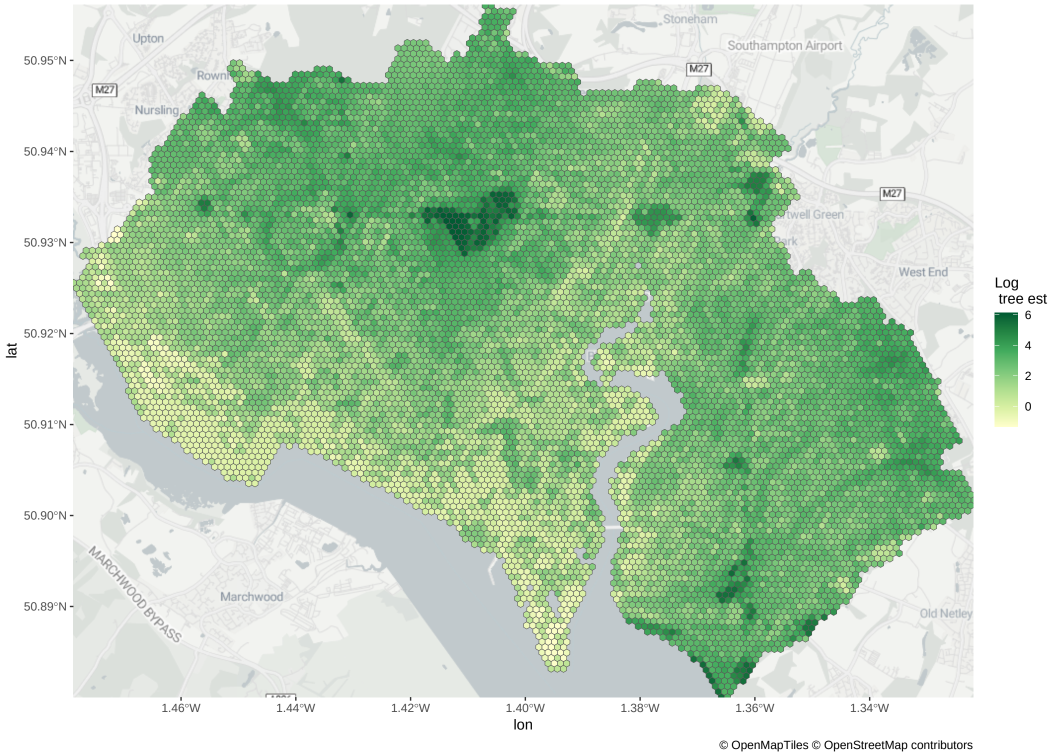


FIGURE 4.4: Log median of the expected number of trees for each cell in Southampton, as predicted from the Southampton i-Tree Eco data

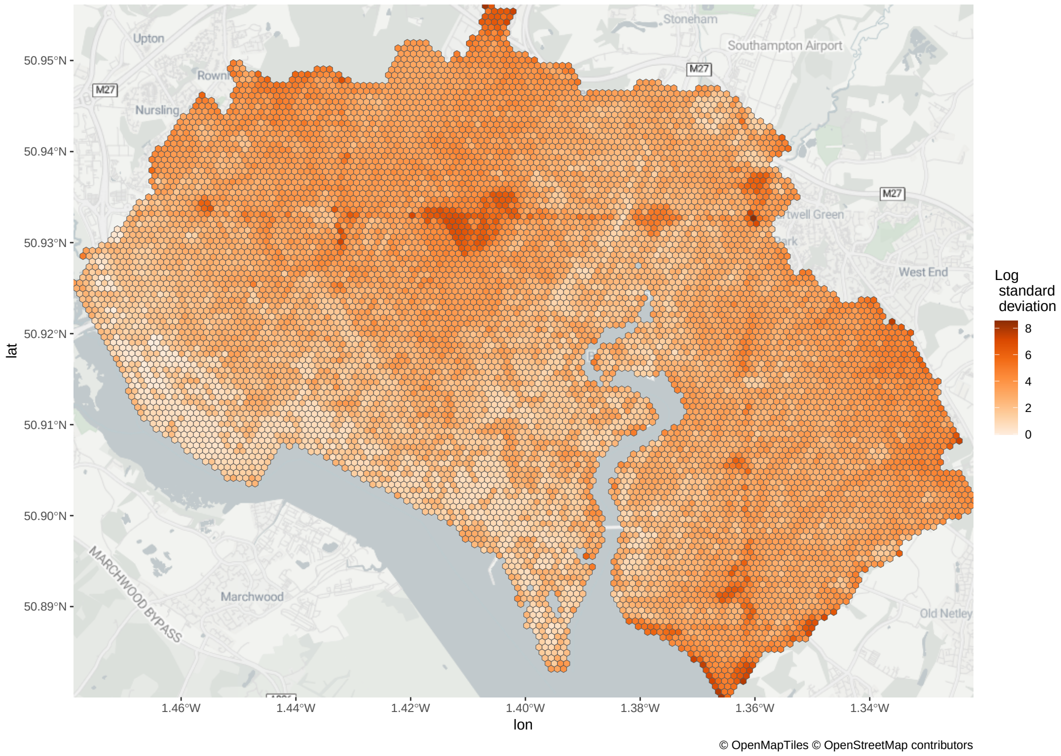


FIGURE 4.5: Log standard deviation in the expected number of trees for each cell in Southampton, as predicted from the Southampton i-Tree Eco data

4.5 Convergence and model fit diagnostics

The parameters included in the Southampton i-Tree Eco model appear to have converged as expected. Traceplots of important parameters (Figure 4.6) all illustrate parameters appearing to converge, with good mixing between the chains. Convergence is also suggested by the $\hat{R}$ values which provided to three significant figures are generally equal to 1.00. Exceptions are the inverse precision parameter and negative binomial variation parameter which have $\hat{R}$ values calculated as 1.02. While slightly higher, these $\hat{R}$ values only indicate the possibility of mild convergence issues, not considered high enough to be of concern. Density plots of parameters are as expected. From visual inspection of the density plots, each parameter appears to be normally distributed, with the exception of the mixing parameter which appears to have a half-normal distribution with a peak close to one. The modelled distribution of the mixing parameter is as expected and results from values being unable to exceed the maximum value of one.

Prior sensitivity was assessed by rerunning the model fitted to the Southampton i-Tree Eco data, with adjusted prior definitions for select parameters. Our sensitivity checks can be summarised as follows:

$\beta \sim N(0,1)$ - A Normal distribution with mean 0 and variance 1 for our regression parameters β. The adjusted prior has the same distribution as before, but with a smaller variance. The resulting model parameters were very similar to the original model, with no convergence issues observed.

$v \sim N(0,5)$ - A Normal distribution with mean 0 and variance 5 for our independent uncertainty parameter, v. The adjusted prior used here has an identical distribution, but with a larger variance. Resulting model parameters were very similar to the original model, albeit with the borderline significance observed for the β_2 parameter by the 95% CI now being just below, rather than just above, zero. No convergence issues were observed.

$\frac{1}{\sqrt{\tau_\sigma}} \sim Normal(0,1)$ - A Normal distribution with mean 0 and variance 1 for our inverse precision parameter $\frac{1}{\sqrt{\tau_\sigma}}$. The resulting model parameters were very similar to the original model, with no convergence issues observed.

$\rho \sim uniform(0,1)$ - A uniform distribution ranging between 0 and 1 for our mixing parameter, ρ. The resulting model parameters were very similar to the original model, with no convergence issues observed. Mixing parameter estimates were found to be a little lower, however generally remained close to one.

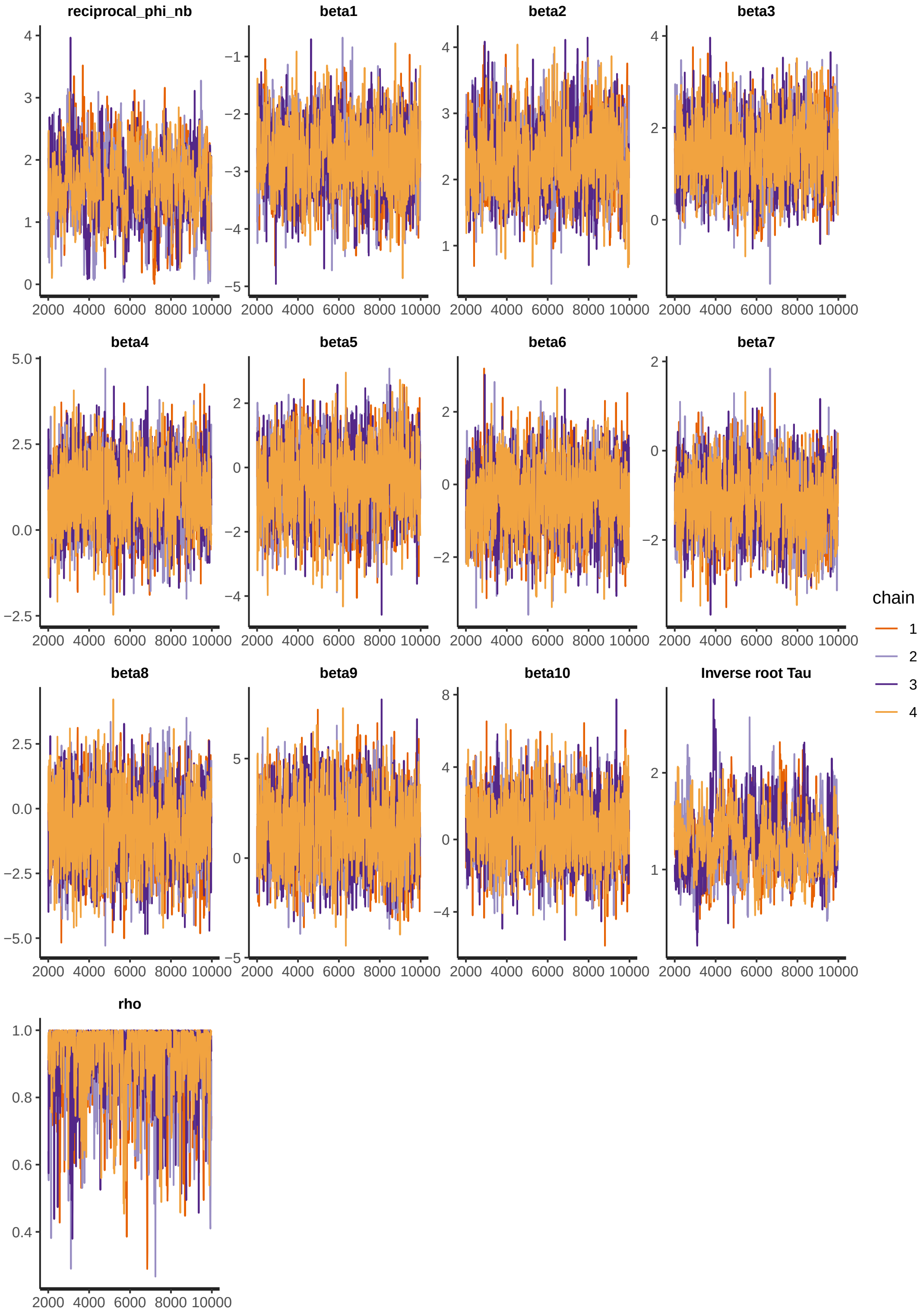


FIGURE 4.6: Parameter traceplots for the Southampton i-Tree Eco model

4.6 Population densities and estimates

Total tree population estimates for the Southampton area can be obtained by calculating the sum of the estimated number of trees in each cell, $\sum_{i=1}^{n_c} Z_i$ for cells $i = 1, 2, ..., n_c$. A density plot of the expected tree populations predicted from the Southampton i-Tree Eco model can be observed in Figure 4.7. The estimated tree population density peaks close to the original i-Tree Eco estimate of 267,000 trees, suggesting the modelled tree populations seem appropriate. A relatively high level of uncertainty is observed in the population estimates, with the 95% credible interval ranging between approximately 194,000 trees and 442,000 trees. This uncertainty can be attributed to the high levels of prediction required for estimating the total number of trees within each cell.

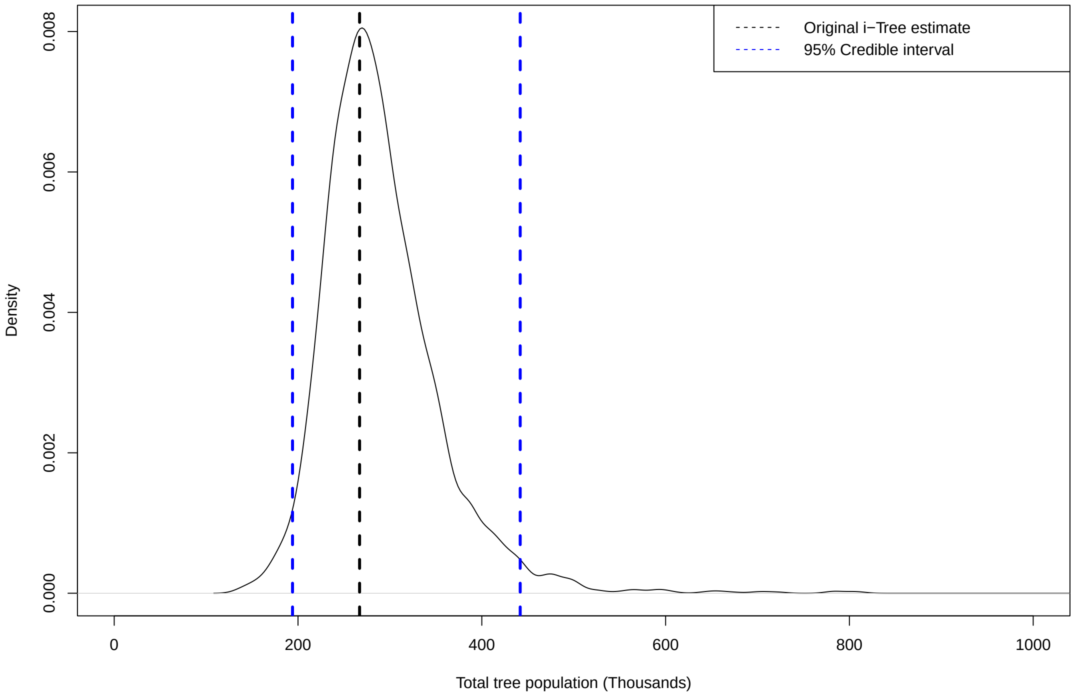


FIGURE 4.7: Plot of the population density as estimated from the Southampton i-Tree Eco data in thousands of trees

4.7 Summary of other areas

In addition to Southampton, BYM2 models were also fitted in Stan for i-Tree Eco data in Petersfield and Cambridge. By fitting spatial models to i-Tree Eco data in other areas we allow for tree densities to be simulated from a variety of locations. Through the inclusion of different locations, we can assess to what extent our findings on the efficacy of various survey plot designs has been observed across different areas. Plots and tables summarising the Petersfield and Cambridge i-Tree Eco models are included in Chapter B of the appendix.

4.7.1 Petersfield

The Petersfield i-Tree Eco data consists of 201 survey plots containing a total of 662 trees. The total number of trees corresponds to an overall rate of 82.1 trees per hectare within the survey plots resulting in an estimated total population of 65,805 trees throughout the area of Petersfield, providing the rate observed in the survey plots is representative of the entire area. We note that both the rate and estimated population of trees from the Petersfield i-Tree Eco data is much higher than the corresponding values in the NTM data. Like the Southampton data, a high number of survey plots (44%) were found to not contain any trees and of the plots containing trees, less than half contained more than five trees. Like the Southampton data, we therefore accounted for zero inflation in our dataset by using a negative binomial distribution, as opposed to a Poisson when model fitting.

In the Petersfield i-Tree Eco data, a low number of survey plots were found to include a high number of trees, with eight survey plots containing more than 20 trees and one plot containing as many as 39 trees. We note that the plot containing 39 trees corresponds to an average rate of 972.6 trees per hectare, much higher than is generally observed. These larger values present issues when modelling, due to the low number of plots providing limited information on when larger tree densities would be expected. Like in Southampton, the inclusion of an exponential component in the model and the reliance on extrapolating estimates, results in some high levels of uncertainty in areas where large numbers of trees are predicted.

Convergence and model fit diagnostics for the Petersfield model suggest a good model fit. Trace plots of the coefficient and uncertainty parameters indicate mixing and convergence, $\hat{R}$ values indicate convergence by not rising above 1.01, whilst density plot of the parameters distributions appear appropriate. Assessments of prior sensitivity all found very similar parameter values to the original model, without any indication of convergence issues.

As anticipated, the model estimated the expected number of trees to be very large in some areas. Whilst this follows the findings of the data, we suggest that the number of expected trees are often larger than could be observed in reality. Further consideration on how extreme values generated from the model can be considered is given when discussing the simulations in Chapter 5.

As observed in Southampton, analysis of the residuals largely suggest an accurate model, albeit with errors associated at extreme values. The 95% credible interval of the expected total population from the model is wide, however the inclusion of the i-Tree Eco population estimate from the overall rate of trees in the estimated population density, suggests model accuracy. We note that the NTM population estimate is below the lower interval of the calculated 95% credible interval, however this is expected due to the large discrepency in the tree rates found for the i-Tree Eco and NTM datasets.

4.7.2 Cambridge

The Cambridge i-Tree Eco data contains 202 survey plots in which 422 trees were observed, corresponding to an average rate of 52.2 trees per ha within the survey plots. The rate of trees in the Cambridge i-Tree Eco data is similar to that observed in Southampton but much lower than the rates in the Petersfield i-Tree Eco data and the Cambridge NTM data. Assuming the rate of trees in the survey plots is representative of the entire Cambridge area, we would predict an estimated tree population of 212,344 trees throughout Cambridge

Similarly to Southampton and Petersfield, no trees were observed in a large number of survey plots (50%) for the Cambridge i-Tree Eco data. Unlike Southampton and Petersfield, less plots contained high number of trees, with only three survey plots including more than 20 trees. We note that the lower maximum values within the survey plots generally corresponds to lower maximum tree estimates from the model, with unrealistically large tree densities being rarely estimated.

Generally the model fitted to the Cambridge i-Tree Eco data appears appropriate. Residuals were often observed to be close to zero, albeit with a low number of instances where the model underestimates the number of trees in areas with high tree densities. Total population estimates are significantly lower than the tree population provided by the ProximiTREE data, an expected result attributed to the loose definition of trees used for the ProximiTREE data. Based on the model results, population estimates were generally lower than the population estimate calculated by assuming a consistent rate of trees throughout the Cambridge area, however a significant difference was not observed.

Convergence and density plots appear appropriate for all coefficient and uncertainty parameters, with the exception of the mixing parameter. From the density plot of the mixing parameter, it appears that the MCMC samples contain a range of values, with higher estimates close to zero and close to one. The emphasis on either high levels of independent error or high levels of spatial error, can in part be attributed to the beta hyperprior placed on the mixing parameter. Consideration of alternative priors and hyperpriors, noted some changes in parameter estimates and similar convergence issues with the mixing parameter. Simulations from the Cambridge i-Tree Eco model are still explored in the following chapter, however we note the discrepancy in how the model error is accounted for, which may be attributed to a lack of information from the underlying survey plot locations.

4.8 Exploration of modelling accuracy using completely observed data

To further assess whether the modelling approach outlined in this chapter was suitable for the i-Tree Eco data, we employed the completely observed ProximiTREE and National Tree Map (NTM) data, modelled in Chapter 3. The survey plot locations from the Cambridge and Petersfield i-Tree Eco surveys, were overlaid on top of the ProximiTREE and NTM data and trees observed outside the survey plots were removed. We then fitted our spatial model, detailed in this chapter, to the survey plots and compared the modeled results to the initial ProximiTREE and NTM data. Comparisons were conducted by assessing whether the true values were within the 95% credible intervals estimated by the model, both for the total populations and by cell.

We assessed the accuracy of the model fits under a range of different conditions. To compare between different cell structures, models were fitted to cells of size 1.5, 1, 0.5 and 0.1 hectares using both hexagonal and gridded cell shapes. Cells of size 0.1 hectares were only explored for the smaller area of Petersfield, as cells of size 0.1 hectares were computationally infeasible for Cambridge. To ensure consistency of results, model accuracy was explored under different survey plot designs. In addition to the i-Tree Eco survey locations, we simulated an additional four complete and randomly located survey plot designs for both Cambridge and Petersfield. We note that two of the survey plot designs contained 200 plots, whilst the other two contained 400 plots. Plot locations were simulated according to the process outline in Section 5.1.1. Model covariates were obtained for each cell size and shape, by conducting a model selection process on the complete ProximiTREE and NTM datasets. While complete data was used to establish the included covariates, we note that only data inside the survey plots was accounted for in the model estimation.

Our results generally suggested that the outlined modelling process provided accurate summaries of the ProximiTREE and NTM datasets from the surveyed data. For all cell structures and survey designs considered, the true population was found to lie within the 95% credible intervals of the expected populations produced by the MCMC samples. These population results, suggest strong evidence of the models providing appropriate tree population estimates. We also considered the proportion of cells in which the observed cell values were found to lie within the 95% credible interval of the expected number of trees. Results by cell are summarised in Table 4.3, for hexagonal cells of all sizes. Our results generally suggest that the estimated tree densities were appropriate, with values above 95%, however a slight drop in accuracy is observed for the smaller cells.

We note that the drop could be attributed to smaller cells being more likely to contain no trees, whereas the expected numbers of trees cannot fall below zero. Furthermore, while a drop is observed, the values do not appear low enough to be of particular concern. Similar results to those illustrated in Table 4.3 were also found in each of the simulated survey plot designs (Tables B.4 and B.3 in Appendix B), further illustrating the suitability of our modelling approach. Due to the higher level of precision offered by smaller cell sizes, our final modelling process was conducted using cell sizes of 0.5 for Southampton and Cambridge and sizes of 0.1 for Petersfield.

Table 4.3: Proportion of cells in which the observed value is within the 95% credible interval for the modeled expected number of trees. Models based on i-Tree Eco survey locations containing trees from the ProximiTREE and National Tree Map data for Cambridge and Petersfield respectively

Location	**Hexagon size (Ha)**	**Proportion of hexagons in the 95% CI**
Petersfield	1.5	97.5
	1	97
	0.5	95.6
	0.1	90
Cambridge	1.5	99.3
	1	96.4
	0.5	93.8

Further examination of different cell structures and survey plot designs, suggested our modelling approach was effective for a range of different situations. Hexagonal and gridded cells were found to produce estimates with similar levels of accuracy. We selected the use of hexagonal cells, due to hexagonal cells usually allowing for a larger number of neighbours than grids. Similar levels of modelling efficacy were observed for 200 and 400 survey plots suggesting our modelling approach should be suitable for all of the i-Tree Eco datasets assessed in this thesis.

We note that our findings use the ProximiTREE and NTM definitions of a tree as a basis for modelling, as opposed to i-Tree Eco. While we believe that the results presented here demonstrate the efficacy of the modelling approach presented in this chapter, we cannot fully understand how accurate the modelling approach is for predicting i-Tree Eco data. Furthermore, a low number of fitted models contained convergence issues, similar to those observed in Section 4.7.2. Despite these concerns, the consistency of the modelling approach in producing reasonable estimates is encouraging.

Chapter 5

Simulation approach for assessing survey design efficacy

In this chapter an approach for using simulations to assess the efficacy of different survey plot design structures, is presented. The chapter begins by detailing the process used in the analysis, while the results of the analysis are summarised in the second half of the chapter.

5.1 Simulation methodology

To assess the efficacy of different survey plot designs, we propose an approach which employs simulating tree densities within cells covering some area of interest, using models fitted in Chapters 3 and 4. A variety of survey plot designs were placed over the simulated data, to assess how well the area contained within the survey plots summarised the full simulated data. The efficacy of a survey plot design was assessed by comparing the total tree population estimated by the survey plots to the total tree population simulated from our models. The accuracy of tree population estimates has been selected to summarise the efficacy of survey plot designs, as repeated accurate tree population estimates should be indicative of a survey design accurately representing the overall area. The use of tree populations for summarising survey design efficacy is discussed and considered in further detail in Section 6.1.

Over the following sections we detail the approach taken to applying the simulation approach described above. Each process has been written and run through the programming language R.

5.1.1 Simulating plot locations

The simulations conducted in our approach can be divided into simulating tree densities and simulating plot locations. The following details the broad approach taken to establishing survey plot locations.

STEP 1: Specify the number of plots, n_S, polygon representing the area being investigate, $\mathcal{D}$, and plot radius in metres, r.

STEP 2: Randomly locate n_S survey plots within the area, $\mathcal{D}$.

STEP 3: Calculate the distance between all survey plots and remove survey plots with a distance of less than $2r$ metres between each other.

STEP 4: Calculate the distance between the survey plots and the area boundary. Remove all survey plots less than r metres away from the boundary.

STEP 5: Repeat STEPS 2, 3 and 4 for the number of survey plots rejected. For STEP 3, the calculated distances include plots that have not been rejected, however only plots introduced in STEP 5 can be removed.

STEP 6: Repeat STEP 5 until all n_S survey plots have simulated locations.

To conduct the final step of the above process, a ‘for’ loop has been employed to repeat the step until the survey plot locations are satisfactory. The use of ‘for’ loops in R is computationally expensive, however allows for a process to be easily repeated whilst accounting for any results in the previous loop iteration. The number of loops required is generally low and rarely exceeds a maximum of four loops for the assessed areas.

Adaptations to the plot simulation process are required when incorporating stratification into the survey plot design. Under stratification, survey plots are located such that a set number of locations are placed within areas that meet some strata condition. For example, we have explored a stratification process which stratifies according to the natural and buildings categories detailed in Chapter 3. After establishing the number of survey plots to be assigned to each strata, the plot simulation process is repeatedly applied to the area satisfying the strata conditions, for the requisite number of survey plots. When applying the plot simulation process to stratified areas, a looser definition is applied whereby only the centre of the survey plot needs to lie in the strata rather than the entirety of the survey plot area. The looser definition for strata ensures that survey plots can be located in small or constrained strata areas, however are still set so that the survey plot cannot contain any area outside the overall area border.

Strata for the natural and buildings variables, consisting of five categories each, are applied by combining both variables into one single variable, consisting of 25 categories. The proportion of the total area meeting the strata criteria was then used to assess how many survey plots should be placed inside each strata level. For example, if 500 survey plots were being assigned in total and a strata level represented 1% of the total area, we would expect to place five survey plots within this strata. Strata numbers are selected based on the proportions of the strata in the overall area, as this is expected to result in the areas observed by the survey plots best representing the overall area.

5.1.2 Simulating tree densities

In addition to simulating survey plot locations, tree densities within overlaid hexagonal cells have also been simulated. The following process gives a brief overview of how the total number of trees and the number of trees within the survey plots have been simulated from some fitted Bayesian model with a spatial CAR component:

STEP 1: Generate random values from a N(0,1) distribution for our independent random error term, $\boldsymbol{v}$, in all cells

STEP 2: Select an MCMC iteration and use the simulated values, along with our independent random errors generated in STEP 1, to calculate the relative risk, $\boldsymbol{\eta}$,for each cell. Spatial samples taken from one MCMC iteration due to the interlinking nature of the fitted models.

STEP 3: Multiply the $\boldsymbol{\eta}$ values from STEP 2 by the expected number of trees within the simulated survey plot areas, $\boldsymbol{E}$, for each cell.

STEP 4: Multiply the $\boldsymbol{\eta}$ values from step 2 by the expected number of trees *within the entire cell*, $\boldsymbol{E}^*$.

STEP 5: Generate random tree densities from a Negative Binomial distribution with variance parameter taken from the MCMC iteration selected in STEP 2 and mean taken from the STEP 3 results. The resulting values represent the simulated number of trees observed in survey plot areas within each cell, $\boldsymbol{Y}$.

STEP 6: Repeat STEP 5 with the Negative Binomial mean taken from STEP 4. The resulting values represent the simulated number of trees observed within the entirety of each cell, $\boldsymbol{Z}$.

In the first step of the above process, an independent random error term is randomly simulated, to ensure that some non-spatial, independent random error is accounted for in the model. Simulations were drawn from a normal distribution with a mean of zero and standard deviation of one, in line with the prior distribution assigned to the independent random error term in the i-Tree Eco models. Due to the heavy emphasis on the spatial error over the independent error term in a lot of the observed models, it is of note that the non-spatial random error term is often having little effect on the simulated tree values.

The expected values used to simulate trees for both the cells and survey plots, were again calculated using internal standardisation. By keeping the rate of trees consistent with the definition from the observed data, it is ensured that interpretation of the expected values also remains consistent throughout the simulations.

To ensure the number of observed trees are whole numbers, random draws are taken from a negative binomial distribution with the mean taken as the expected number of trees and the dispersion parameter taken from the dispersion parameter used in the associated MCMC iteration. The negative binomial distribution has been selected for the i-Tree Eco models due to the inclusion of the dispersion parameter to deal with zero-inflated data. However, there is uncertainty on whether the total number of trees estimated throughout the entire cell also follows a negative binomial distribution. As the trees modelled in Chapter 3 follow a Poisson distribution, in which the variation is considered equal to the mean, generating tree density simulations from a Poisson distribution for the i-Tree Eco model simulations has also been considered. It is of note that the ProximiTREE and National Tree Map (NTM) model simulations are all calculated using Poisson distributions as opposed to negative binomial distributions.

As the tree density simulations for within the entire cell and within the survey plots are drawn from separate distributions, it is unlikely but not impossible for the trees observed within a survey plot to exceed the total number of trees simulated for the associated cell. Generally the plot simulations are more likely to exceed the cell simulations in areas where a large proportion of the cell has been observed, however the problem can also occur when drawing randomly from a negative binomial distribution with a high level of uncertainty. The decision was made to not adjust simulations such that plot simulations did not exceed cell simulations, as this could systematically lower population estimates calculated from the simulated survey plots.

An alternative approach, incorporating random draws from a Bernoulli distribution, was explored for simulating the number of trees found within the survey plots. As opposed to being directly drawn from a random distribution, simulations for the number of trees in the survey plots are instead calculated from the number of trees simulated for the entire cell. For each cell the proportion of the cell observed by the simulated survey plots was calculated and the result multiplied by the number of trees simulated for the entire cell. The decimal value was then used as the probability in a random draw from a Bernoulli distribution, to establish whether the estimate was rounded down or up to the nearest whole number. For example, if three trees were simulated for a cell of size two hectares, where one hectare of the cell was observed by survey plots, the number of simulated trees inside the survey area would be one with probability 0.5 or two with probability 0.5. Using this approach ensures that the number of trees simulated in the survey plots follows the rate of trees simulated, but cannot exceed the total number of trees simulated for the cell.

While using the Bernoulli approach results in simulated tree density values that logically make sense, the approach makes an implicit assumption that the rate of trees is largely consistent throughout a cell. In reality, we could expect the rate of trees to vary within a cell and by assuming a constant rate, the Bernoulli approach may be removing some of the variation in the plot estimates, and therefore the population estimates. In contrast, simulating the tree populations for the survey plots from either the Poisson or Negative Binomial distribution, ensures that variability in the tree density within the cells is maintained. In the simulations analysis, we therefore considered the impact of simulating the trees in the survey plots using both the Bernoulli approach and by taking draws from a random distribution.

5.1.3 Calculating population estimation accuracy from simulations

Using the approaches to simulating survey plots and trees densities, the following describes the process for assessing survey plot accuracy based on the number of survey plots in a survey plot design.

STEP 1: Simulate complete tree densities Z_i for each cell, i.

STEP 2: Calculate the simulated total tree population, by summing the results in STEP 2.

STEP 3: Select random plot locations and draw buffers around each survey plot.

STEP 4: Calculate the survey plot area included in each cell.

STEP 5: Simulate values for the number of trees observed within the survey plots, Y_j for cells containing survey plot areas.

STEP 6: Calculate an estimate of the tree population from the survey plot estimates simulated in STEP 5.

STEP 7: Compare the total population estimated from the survey plots in STEP 6 to the total population simulated in STEP 2.

In the final step of the approach detailed above, comparisons between the simulated and estimated total populations is conducted by calculating the relative error. The relative error is defined by dividing the populations estimated from the survey plots by the simulated populations. By calculating and presenting the relative error, survey design accuracy is assessed relative to the total populations, ensuring that areas with larger tree populations are not penalised. For comparison, raw errors are also presented and interpreted in the results. Consideration is given to how interpretation of both the raw and relative errors inform our results.

A key consideration of assessing survey plot designs is the number of survey plots included. However, repeatedly using the above process to explore population estimation accuracy for a range of different survey plot sizes is computationally expensive. An efficient way to make the process less computationally expensive, is to minimise the number of times survey plot locations are simulated in STEP 3. We propose selecting plot locations for only the largest survey plot designs, giving plot locations of $\boldsymbol{s} = (s_1, s_2, \ldots, s_n)$ where n represent the maximum number of survey plots being explored. Smaller plot designs were then obtained by randomly sampling plot locations from $\boldsymbol{s}$ without replacement, as opposed to simulating entirely new survey plot designs. We note that for stratified survey plot designs, locations must be randomly sampled so as to adhere to any outlined stratification criteria. STEPs 4 to 7 of the above simulation process were then conducted on both the full plot locations and the randomly sampled locations, providing accuracy estimates for plot designs across a range of sizes.

Running the above approach for estimating population accuracy multiple times across a number of iterations, provides results for a range of different plot location and tree density simulations. Due to survey plot designs being sampled from the largest design in each iteration, we note that independence will exist between, but not within, iterations. As interpretation of our results is provided across iterations this has not posed a problem when conducting our analysis. In our results, 5,000 iterations have been applied as this appears to always be sufficient for convergence without being computationally infeasible.

5.2 Simulation results

Using the simulations based approach outlined in this chapter, we now consider how the number of survey plots in a survey plot design affects the population estimate error for tree density values simulated from the spatial models fitted in Chapters 3 and 4. We begin by presenting the results taken from the National Tree Map (NTM) and ProximiTREE models, before presenting results from the i-Tree Eco models. Due to the large number of tables and figures assessed, additional tables and figures are provided in Appendix C .

5.2.1 Assessement of survey plot design error from models fitted to the ProximiTREE and National Tree Map (NTM) data

Generally, the relative errors associated with the ProximiTREE and NTM models appear to quickly converge to a low level of relative error as the the number of survey plots increases. A visual representation of the relative error against the number of survey plots is included in Figure 5.1, with solid lines representing the mean relative error from the simulations and the dashed lines representing the 90^{th} percentiles. The dashed 90^{th} percentile line indicates the higher relative error values that have been observed, with 90% of our relative errors found to lie below the 90^{th} percentile line. Generally, the mean relative error appears to reduce very little beyond 100 survey plots in both areas, with the 90^{th} percentiles reducing very little beyond 200 survey plots. From the 90^{th} percentiles, it would therefore appear that 200 survey plots would usually be sufficient, based on the ProximiTREE and NTM data.

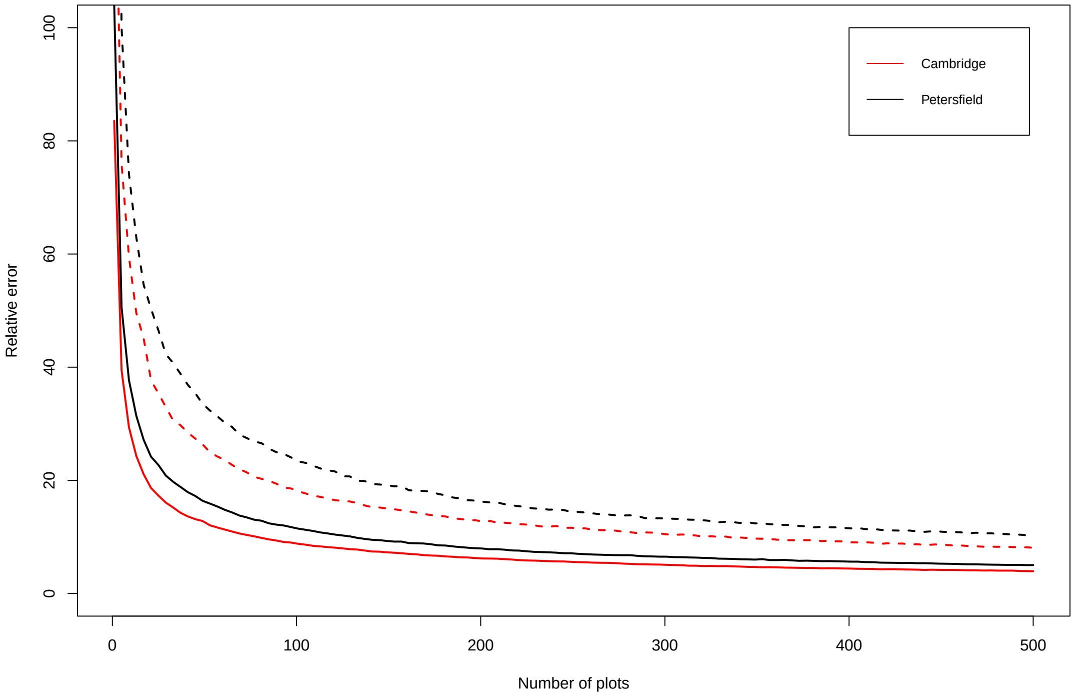


FIGURE 5.1: Plot of the relative error against the number of survey plots. Relative errors calculated using simulations from spatial models fitted to the Cambridge ProximiTREE and Petersfield National Tree Map (NTM) data. Solid lines represent mean values, whilst dashed lines represent 90^{th} percentiles

For ease of interpretation, relative errors are summarised numerically for the Cambridge ProximiTREE model and Petersfield NTM model in Tables 5.1 and 5.2 respectively. The tables present the relative errors associated with the total number of survey plots at the 50th, 75th, 90th and 95th percentiles, with the number of survey plots increasing in increments of 50 between 50 and 500. For example, in Table 5.1 we can observe that 90% of survey plot designs with 200 plots had a relative error of 12.8% or lower based on simulations from the Cambridge ProximiTREE model. Both Tables 5.1 and 5.2, again highlight how the relative error reduces at a slower rate after 200 survey plots. Based on the ProximiTREE and NTM data, it could therefore be suggested that sampling with more than 200 survey plots provides little benefit in reducing the observed relative error, however this conclusion is subject to personal interpretation of what constitutes an acceptably low relative error.

Table 5.1: Summary of the relative error by percentile against the number of survey plots. Relative errors calculated using simulations from modeled Cambridge ProximiTREE data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	10.5%	7.5%	6.1%	5.2%	4.8%	4.3%	4.0%	3.7%	3.5%	3.3%
75%	18.1%	12.8%	10.5%	8.9%	8.0%	7.3%	6.7%	6.3%	6.0%	5.6%
90%	25.7%	18.2%	15.1%	12.8%	11.5%	10.5%	9.7%	9.1%	8.7%	8.1%
95%	30.8%	21.6%	18.1%	15.4%	13.9%	12.6%	11.7%	10.9%	10.3%	9.7%

Table 5.2: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using simulations from modeled Petersfield National Tree Map (NTM) data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	13.8%	9.7%	7.9%	6.8%	6.0%	5.6%	5.0%	4.8%	4.5%	4.2%
75%	23.5%	16.6%	13.4%	11.5%	10.3%	9.4%	8.7%	8.1%	7.6%	7.3%
90%	33.1%	23.5%	19.2%	16.3%	14.6%	13.3%	12.4%	11.4%	10.9%	10.4%
95%	39.5%	27.8%	22.7%	19.2%	17.4%	15.7%	14.8%	13.6%	12.9%	12.4%

Despite differences in tree definitions and the underlying areas, the relative errors appear similar for the findings simulated from the Cambridge ProximiTREE and Petersfield NTM models, albeit with slightly lower relative errors observed for Cambridge. The ProximiTREE and NTM results would therefore appear to suggest that the accuracy of a survey plot design in estimating the underlying 'true' rate of trees, is not tied to the size of an area or even the total number of trees observed within an area. We note however that the findings presented in this section consider only the relative error, as opposed to the absolute error, and are using the definitions of trees provided for the ProximiTREE and NTM data, both of which differ from the definition of a tree given under i-Tree Eco. In the following section we explore the relationship between the number of survey plots and the relative error for simulations taken from i-Tree Eco data. Consideration of the absolute error, as opposed to relative error is provided in section 5.2.6.

5.2.2 Assessement of survey plot design error from models fitted to i-Tree Eco data

Relative errors calculated using i-Tree Eco model simulations were found to be much higher across all locations (Figure 5.2), compared to ProximiTREE and NTM simulations. In each location the mean relative error converges to a higher value of approximately 20% at a much slower rate than previously observed. Of particular concern are the high relative errors for the 90th percentiles, with 200 survey plots corresponding to a relative error of 46% at the 90th percentile in Southampton (Table 5.3). It is of note that an excess of 250 survey plots results in little reduction in the relative error, however the relative error remains larger than expected.

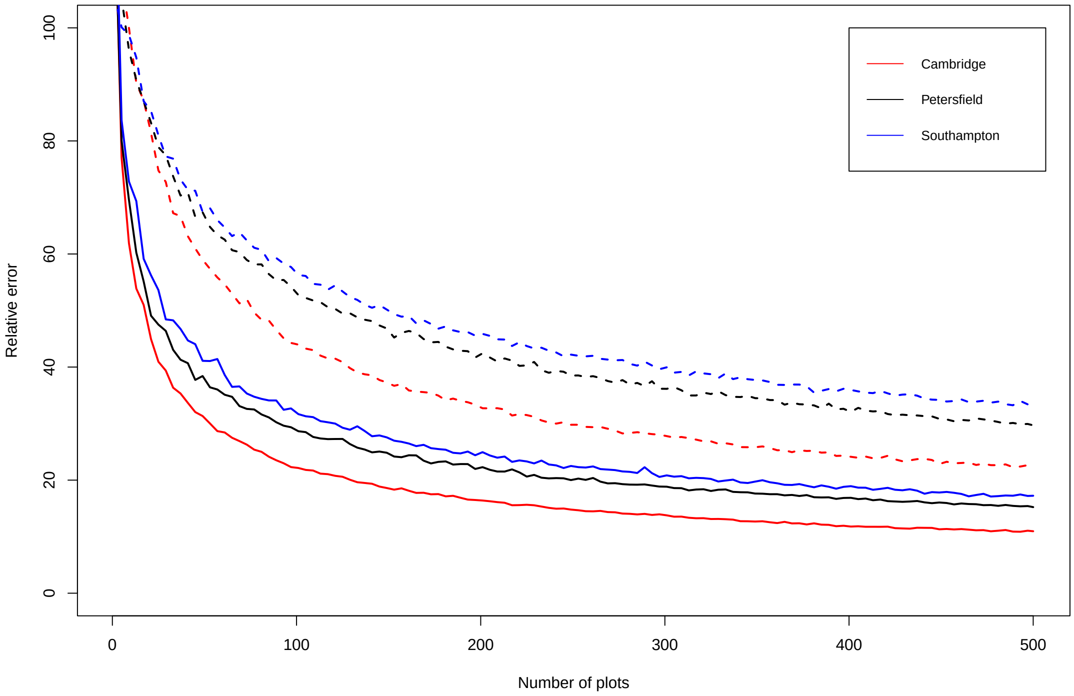


Figure 5.2: Plot of the relative error against the number of survey plots. Relative errors calculated using negative binomial simulations from modeled i-Tree Eco data. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

Table 5.3: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	31.5%	24.6%	20.8%	18.6%	16.9%	16.0%	15.3%	14.4%	13.6%	13.3%
75%	49.3%	39.7%	34.0%	31.3%	28.4%	27.2%	25.8%	24.3%	23.0%	22.4%
90%	68.3%	57.1%	50.1%	46.0%	41.8%	40.0%	37.8%	36.0%	34.3%	33.1%
95%	98.5%	79.7%	67.1%	61.0%	54.6%	51.7%	49.9%	46.1%	44.5%	43.1%

Simulating plot values from a Bernoulli distribution, as outlined in Section 5.1.2, was found to provide almost identical results to simulating plot values from a negative binomial distribution, throughout all areas. Furthermore, we note that applying the Bernoulli distribution approach does little to substantially effect the relative errors produced throughout all of the simulations we considered. The Bernoulli approach having little effect suggests that accounting for tree variation within the cells as part of our calculations has little effect on the overall accuracy of the survey plot designs. Due to offering very similar results, Bernoulli simulation results have been produced, but are generally not presented any further.

Unlike the ProximiTREE and NTM models, the relative errors simulated from the i-Tree Eco models were found to differ by location, with relative errors largely higher in Southampton and Petersfield than Cambridge. We suggest that the lower relative errors in Cambridge could be attributed to the maximum expected tree rate from the model being much lower in Cambridge, compared to Southampton and Petersfield. In cases where a small number of areas are accounting for a large proportion of the total tree population, survey plot error is expected to be much higher if cells with higher tree rates are not sufficiently accounted for in the survey plot design. Cells containing large number of trees for the i-Tree Eco models can be attributed to a number of reasons including, the uncertainty introduced in the prediction, the use of an exponential and the use of a negative binomial distribution. In the following sections, we consider how adaptations to the simulation process impact the population errors produced.

5.2.3 Simulations from a Poisson distribution

Relative errors calculated using i-Tree Eco model simulations, were generally found to be lower when tree densities for both cells and plots were drawn from a Poisson distribution, as opposed to a negative binomial (Figure 5.3). For example, in Southampton a relative error of of 33.9% was observed at the 90$^{\text{th}}$ percentile for 200 plots, approximately 12% lower than was observed when simulations were drawn from a negative binomial distribution. Relative errors being lower for Poisson simulations is not surprising given that the negative binomial is often providing additional uncertainty for the i-Tree Eco models. The use of a Poisson distribution can somewhat be justified by the distribution's efficacy in modelling the ProximiTREE and NTM data, which both suggest the number of trees in the cells follow a Poisson distribution. As i-Tree Eco data is provided at the survey plot level and often contains a large number of empty plots, it is possible that a negative binomial distribution is more appropriate for modelling survey plot data than simulating tree densities for entire cells.

We note that under the Poisson simulation approach, the relative errors is still lowest in Cambridge, however the Southampton relative error is now lower than Petersfield. The higher reduction in the relative error for Southampton could be attributed to higher uncertainty values for the negative binomial in the Southampton i-Tree Eco model, when compared to Petersfield. The rate at which the relative error reduces after 200 survey plots is comparable to that observed using a negative binomial distribution, however the relative error is lower due to a faster reduction in the relative error between 0 and 200 survey plots.

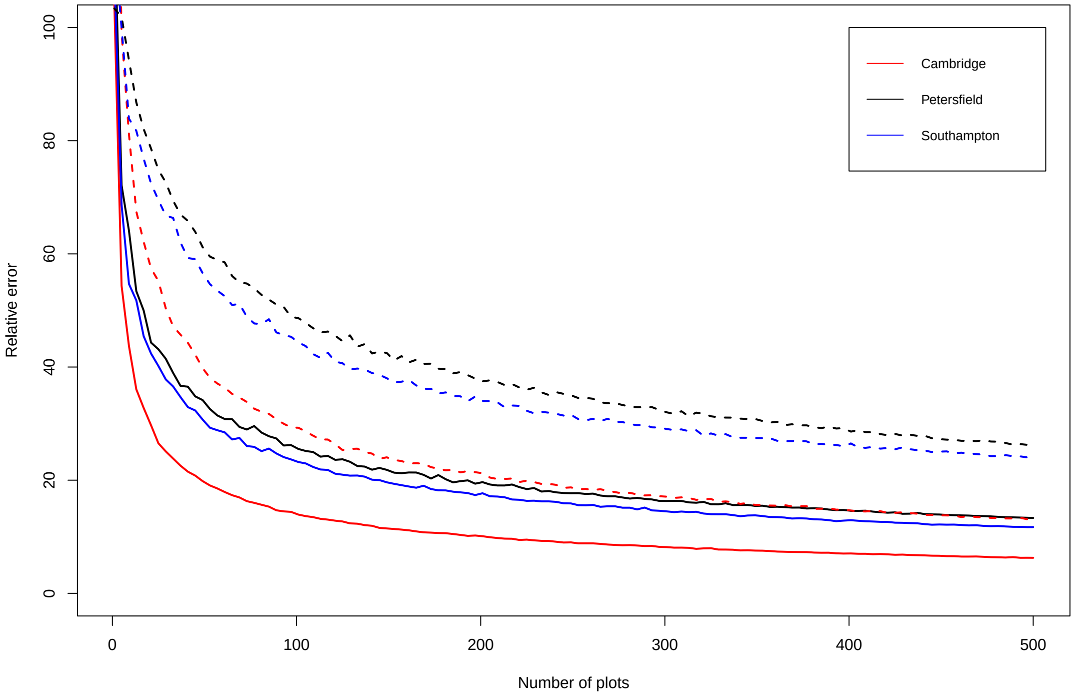


Figure 5.3: Plot of the relative error against the number of survey plots. Relative errors calculated using Poisson simulations from modeled i-Tree Eco data. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

Table 5.4: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	23.6%	18.1%	14.9%	13.4%	11.9%	11.1%	10.6%	9.8%	9.3%	9.0%
75%	39.2%	30.1%	25.5%	23.1%	20.7%	19.4%	18.4%	17.3%	16.4%	15.5%
90%	56.1%	44.5%	37.7%	33.9%	31.3%	29.2%	27.6%	26.5%	25.0%	24.1%
95%	75.1%	57.3%	48.8%	44.0%	39.8%	37.2%	35.0%	33.5%	31.4%	30.4%

5.2.4 Simulating from selected samples

For both the Southampton and Petersfield i-Tree Eco models, we observed extreme expected tree estimates that we do not believe would be observed in reality. We therefore consider how the population error would look if simulations drawn from the Southampton and Petersfield models did not include these extreme observations. For each MCMC sample we extracted the cell with the largest expected tree value in the cells, ordered the MCMC samples from smallest to largest by the largest expected tree value, removed the top half of the MCMC samples and assessed population error using our simulation approach from the lower half of the MCMC samples. By removing half the MCMC samples, we ensure that the largest simulated tree densities are lower than would be expected in the full model.

By removing half the MCMC samples, the relative error was reduced for both the Southampton (Figure 5.4) and Petersfield (Figure 5.5) i-Tree Eco models. Furthermore, consideration of only half the MCMC samples was found to reduce the relative error when sampling from both the negative binomial and Poission distributions. In Southampton a relative error of 28.4% was observed at the 90th percentile for 200 plots, when only including half the MCMC samples and simulating from a Poisson distribution (Table 5.5), much lower than the 46% observed when including all samples and simulating from a negative binomial distribution. The observed reduction in relative error, could suggest that extreme values produced from the models are inflating the relative error by considering simulations that would not exist in reality. A possible solution could be to constrain the expected number of trees in each cell at the modelling stage, preventing some of the larger tree estimates that were observed in Chapter 4.

We note again that the rate at which the relative error reduces after 200 survey plots is comparable to under the negative binomial distribution, however the relative error reduces at a much faster rate between 0 and 200 plots. An excess of 200 survey plots can therefore be viewed as providing little benefit in reducing the relative error further, however this is subject to personal interpretation of the results.

The total populations produced when considering only half the MCMC samples appear to be appropriate. For both Southampton and Petersfield, the population densities still peak close to the estimates provided in i-Tree Eco reports, however the long tails observed when using the full MCMC samples were missing. By sampling from only half the MCMC samples, the relative errors for Southampton and Petersfield were found to be much closer to those observed under the full MCMC samples in Cambridge, albeit with relative errors still lower in Cambridge.

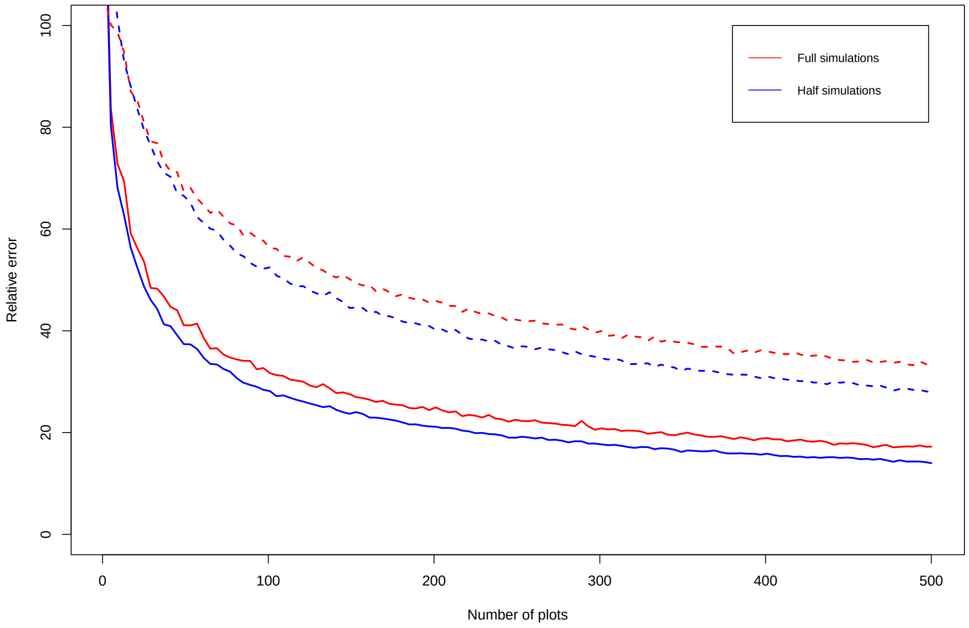


FIGURE 5.4: Plot of the relative error against the number of survey plots, for simulations produced from all and half of the MCMC samples. Relative errors calculated using negative binomial simulations from modeled i-Tree Eco data in Southampton. Solid lines represent mean values, whilst dashed lines represent 90$^{\text{th}}$ percentiles

TABLE 5.5: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson simulations from half the model samples for modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	21.0%	15.7%	12.9%	11.2%	10.4%	9.5%	8.8%	8.3%	7.9%	7.3%
75%	34.8%	26.3%	21.9%	19.3%	17.4%	16.1%	15.0%	14.1%	13.5%	12.6%
90%	51.1%	38.3%	31.8%	28.4%	25.6%	23.5%	21.8%	20.6%	19.5%	18.4%
95%	68.0%	48.5%	39.1%	35.1%	31.6%	28.8%	26.7%	25.2%	23.7%	22.5%

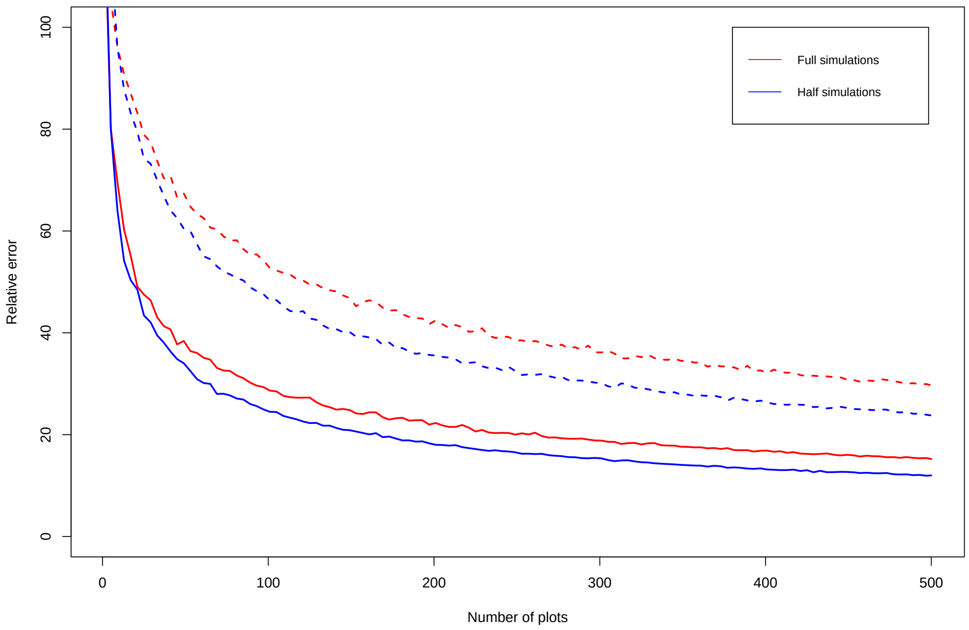


FIGURE 5.5: Plot of the relative error against the number of survey plots, for simulations produced from all and half of the MCMC samples. Relative errors calculated using negative binomial simulations from modeled i-Tree Eco data in Petersfield. Solid lines represent mean values, whilst dashed lines represent 90$^{\text{th}}$ percentiles

5.2.5 Stratification by variable

As an additional survey design consideration, we explored the impact of stratification on the relative error values. The stratification procedure used the proportion of natural and buildings coverage as strata, as outlined in Section 5.1.1. We explored applying stratification criteria to the survey plot locations for simulations drawn from the ProximiTREE, NTM and i-Tree Eco models. For i-Tree Eco models we considered drawing simulated tree values from both the negative binomial and Poisson distributions.

Generally, very little difference was observed in the relative errors between stratified and randomly located survey plot designs. We note however that stratification may be providing benefits, such as a more representative sample for other variables collected for i-Tree Eco surveys, despite having little impact on the population estimate. Furthermore, consideration of a stratification approach that accounts for spatial correlations in the area being observed, may result in lower relative errors than stratifying by covariates.

5.2.6 Absolute error

So far we have considered the relative error associated with population estimates, however it can also be important to consider the absolute error. As opposed to the relative error, absolute error provides the raw differences between the observed and the estimated populations in the simulations. By considering the relative error, we effectively assess how well the survey plots capture the overall rate of trees, whereas consideration of the absolute error, captures how well our estimated rate of trees 'scales up' to represent an entire area.

Absolute errors were found to generally reduce down quicker when simulating from the Petersfield i-Tree Eco model using a negative binomial distribution, as opposed to simulating from the Southampton and Cambridge i-Tree Eco models (Figure 5.6). Furthermore, the mean absolute error appears to generally reduce very little between 100 and 200 plots for Petersfield, whilst the 90^{th} percentile generally reduces very little beyond 200 survey plots. We note that our interpretation of the Petersfield absolute errors is provided within the context of the corresponding Southampton and Cambridge absolute errors.

Given the size of the area, the Cambridge absolute errors appear to also reduce down very quickly, albeit at a slower rate than observed for Petersfield. By 200 survey plots the mean absolute errors are very similar for both Petersfield and Cambridge, however the 90^{th} percentile remains higher in Cambridge. Relatively low absolute errors being observed in Cambridge, despite the Cambridge area being much larger than Petersfield, reflect the low relative errors observed when simulating from the Cambridge i-Tree Eco data earlier in the chapter.

Absolute errors produced from the Southampton i-Tree Eco data, were observed to be much higher than both Petersfield and Cambridge. We note that consideration of Poisson simulations and simulating from half the MCMC samples was found to reduce the absolute error in a comparable way to the relative error results presented earlier in the chapter. Tables and Figures summarising the absolute error for different simulation considerations in all areas, are provided in Appendix C.

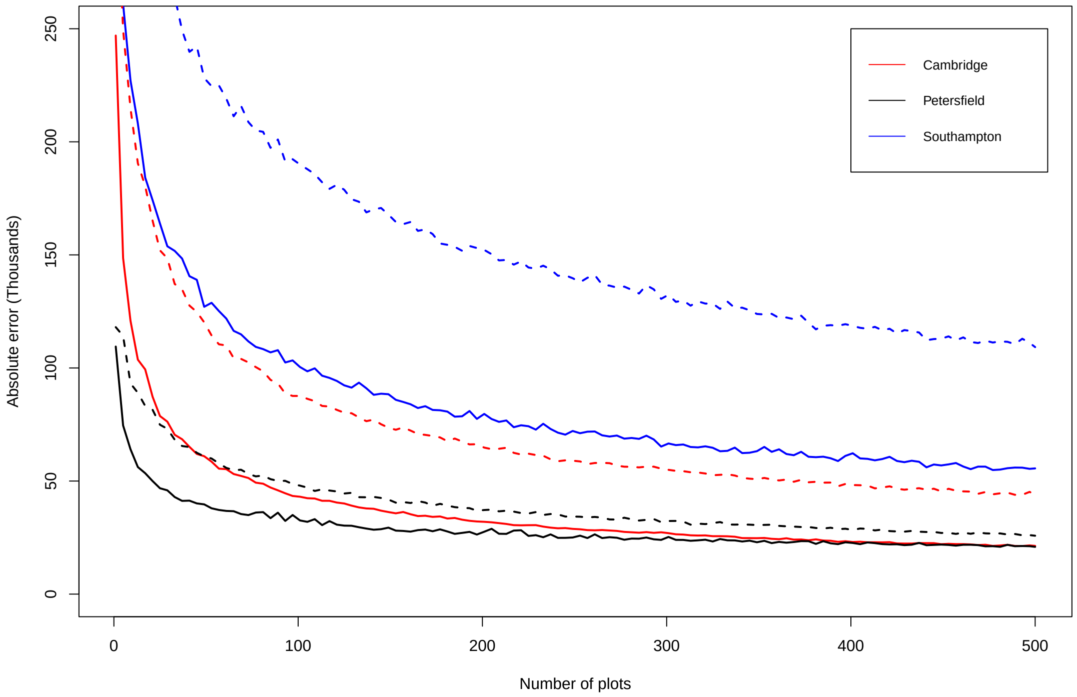


Figure 5.6: Plot of the absolute error against the number of survey plots. Absolute errors calculated using negative binomial simulations from modeled i-Tree Eco data. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

Chapter 6

Discussion and future work

6.1 Ecological conclusions and discussion

From our research, the suitability of using 200 survey plots in an i-Tree Eco survey design is dependent on a number of different conditions. More generally, we suggest that providing a representative survey is dependent on capturing the variety of information present within an urban area. When considering the accuracy of tree population estimates, we usually found much more accurate estimates when considering urban areas with lower variation in tree densities, compared to urban areas with higher variation in tree densities. The difference in relative population errors as a result of tree density variation, is perhaps best illustrated by considering the differences observed by generating from a negative binomial and a Poisson distribution for our i-Tree Eco models. For both distributions the expected number of trees in each cell was identical, however the additional variation introduced by the negative binomial distribution was found to result in higher relative errors from within the survey plots.

The conclusions reached on the suitability of using 200 survey plots in i-Tree Eco surveys, are also dependent on the definition of the term tree. Simulations from models fitted to ProximiTREE and National Tree Map(NTM) data found much lower relative errors under 200 survey plots, compared to simulations from models fitted to i-Tree Eco data. The ProximiTREE/NTM datasets differ from the i-Tree Eco datasets in two key ways, the first being that the i-Tree Eco data defines trees based on diameter at breast height, whilst the ProximiTREE/NTM data defines trees based on tree height. The second key difference is that the ProximiTREE/NTM data has been fully observed, whereas the i-Tree Eco data has only been observed at survey plot locations.

Due to their predictive nature, simulations of entire urban areas from i-Tree Eco models were found to contain high levels of uncertainty which could be inflating the relative error in population estimates. For Southampton and Petersfield, we observed that by removing simulations which contained larger tree density estimates, we reduced the relative population errors while maintaining reasonable estimates of the total population. This would suggest inflated relative population errors could be attributed to larger estimates in simulations, which result from high levels of modelling uncertainty. The completely observed nature of the ProximiTREE and NTM data result in less model uncertainty in our resulting simulations, suggesting that the results may be more reflective of tree densities that could be observed in reality.

However, differences in relative errors between the ProximiTREE/NTM and i-Tree Eco simulations, could also be attributed to differences in tree definitions. Comparisons between the ProximiTREE/NTM and i-Tree Eco data often found relatively low levels of agreement, which could suggest that the larger relative errors observed for the i-Tree Eco data are as a result of underlying differences in the ProximiTREE/NTM and i-Tree Eco datasets, as opposed to just modelling uncertainty. Furthermore, the simulation results from i-Tree Eco data are the only results which use the definition of trees as defined by i-Tree Eco, as a basis for exploring the relative errors. Despite the high levels of uncertainty, all simulations from the i-Tree Eco model are all predominantly based on findings observed within the i-Tree Eco data. We therefore conclude that whether or not the ProximiTREE/NTM models provide more representative population errors than the i-Tree Eco models, is largely reliant on whether or not the difference in tree definitions is expected to result in significant differences in tree densities across the observed areas.

A further consideration to whether 200 survey plots are suitable for an i-Tree Eco survey, is whether 200 survey plots is the most appropriate number of survey plots in all areas. Simulations from our i-Tree Eco data appear to indicate differences in the relative population errors, although these differences were somewhat mitigated by considering the use of a Poisson distribution and the use of only selected MCMC samples. Our finding that the accuracy of the relative error estimate is lower in areas with higher variation in tree densities, could suggest that less survey plots may be required in areas where the tree density is considered more homogeneous. However in practical terms, information on the presence of tree density homogeneity may not be present for an area prior to collecting information for the survey.

We note that when comparing survey accuracy between locations, we believe it is important to account for both the information being collected and consider the absolute errors produced in our results, as opposed to just the relative errors. By considering only the relative population error, we effectively consider how close the rate of trees estimated in the survey is to the actual rate. However considering the absolute error also accounts for the size of the area under investigation and considers in actual terms how close our population estimate from our surveys is to the simulated populations. For example, two areas of different sizes may have the same relative error, but we would expect the larger area to have a higher absolute error compared to the smaller area. In general, estimates provided by the UFORE model could be associated with higher absolute errors in larger areas if collected data is ‘scaled up’ to represent an entire urban area.

In contrast to other research in the literature (Jin and Yang, 2020), we found very little effect on the relative error as a result of stratifying survey plots. Our differing results could be attributed to a number of reasons, including differences in stratification approaches and differences in underlying areas. We note that while stratified survey plot designs did little to reduce the relative error, stratification should provide more representative survey data which may be accounted for in other variables collected in the i-Tree Eco surveys.

6.2 Appropriateness of model

Generally, we believe that the models fitted in Chapters 3 and 4 are appropriate for the simulation process outlined in Chapter 5, however we acknowledge the existence of some potential issues in the models. In particular, larger tree densities predicted in some cells for the Southampton and Petersfield i-Tree Eco models may be unrealistically high, despite following naturally from the fitted model. As model fit diagnostics and our findings in Section 4.8 suggest the models are otherwise appropriate, it may be of interest to explore constraining some of the model parameters for the Southampton and Petersfield i-Tree Eco models. Furthermore, some convergence issues were observed with the mixing parameter in the Cambridge i-Tree Eco model, which may require further investigation.

Despite all the hard work and effort that goes into planning, preparing and conducting every i-Tree Eco survey, it is only possible to observe a small proportion of the total area. For example, despite containing a larger number of plots than usual, the Southampton i-Tree Eco survey covered less than 0.4% of the total Southampton area. Models fitted from the survey plots are therefore based on relatively limited information, reflected in the high levels of uncertainty observed. We believe that unless the information collected in the survey plots is misleading in representing the underlying area, then the resulting simulations should provide logical tree densities for each cell, which follow naturally from the observed data. We note that the simulations are expected to provide a range of 'believable' tree densities across all of the cells, as opposed to accurately replicating what would be observed in reality.

When overlaying cells, the decision was made to bisect survey plots. Survey plots were bisected so as to use the exact tree locations given within the data, whilst the areas observed in each cell were accounted for by the inclusion of an offset term in the model. By bisecting survey plots, we risk complicating the spatial relationship due to trees potentially being separated into different cells despite being observed within the same survey plot. We note that the presence of a spatial component in the ProximiTREE and NTM models, along with exploratory spatial analysis of the trees included in the survey plots, suggest the high levels of spatial autocorrelation observed in our i-Tree Eco models are as expected. Furthermore, our results of Section 4.8, suggest that the cell structures applied should be sufficient for the purposes of the modelling conducted in Chapter 4. Alternative definitions for the cell structures were considered, such as Voronoi diagrams (Okabe et al., 2009), however cells of equal sizes were deemed more appropriate for summarising interaction between nearby areas. Additional consideration on the effect cell structures and definitions have on our model results, could further ensure that findings are attributed to the observations in our data as opposed to originating from the cell structure definition.

While our modelling process provides results to a reasonable level of accuracy (Section 4.8), the use of a negative binomial distribution to deal with zero inflation results in large levels of uncertainty in our simulations. Alternative distributions such as the Zero Inflated Poisson (ZIP) and hurdle models may provide a more appropriate modelling approach for the survey plots, but do not provide results conducive for estimating tree densities throughout entire cells. Further consideration of alternative distributions, such as the Tweedie distribution (Swallow et al., 2016), may provide an improved model fit compared to the negative binomial, however it must be ensured that the selected distribution is computationally feasible for both the model fitting in Chapter 4 and the simulations process in Chapter 5.

6.3 Further work

In this thesis, we have considered our results within the locations of Southampton, Cambridge and Petersfield. Whilst the range of simulations explored allowed us to consider the efficacy of survey plot design structures under a wide range of conditions, it may be beneficial to explore additional urban areas. In particular, our research does not directly address survey plot design efficacy in larger urban areas, such as the city of Manchester. While results from larger areas would enable further consideration on the relationship between the number of survey plots required and the area under investigation, model fitting and simulations for large areas could prove to be much more computationally expensive.

Modelling approaches for the ProximiTREE/National Tree Map data have been conducted separately from the i-Tree Eco data, due to the use of different tree definitions and resulting discrepancies in the datasets. An approach which considers employing the findings of the ProximiTREE and National Tree Map data for informing models fitted to tree locations in i-Tree Eco, may further reduce the variation observed in resulting simulations. Combining remote sensing and ground level data is a topic investigated within the ecology literature (Rattalino Edreira et al., 2020; Henrys and Jarvis, 2019; Uhl et al., 2021), however an approach suitable for our research could not be found.

More widely, we note that assessing the efficacy of survey plot designs through the design's population estimates has limitations. In general, we have made an assumption that by representing the 'true' rate of trees throughout an area, a survey plot design will be sufficient for conducting i-Tree Eco surveys. Whilst we suggest that survey plot designs with large population errors are unlikely to be representative of the wider area, consideration should be given to ensuring survey plot designs are representative for all information collected in i-Tree Eco studies.

Appendix A

Additional Chapter 3 plots and tables

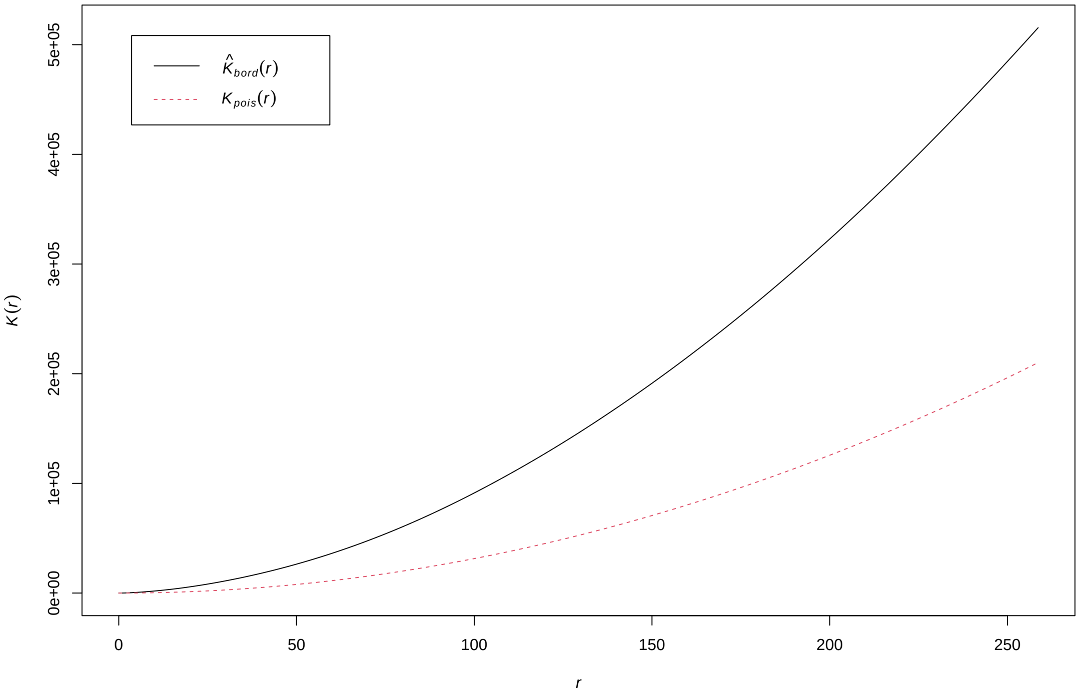


FIGURE A.1: Plot of the empirical and (completely random) Poisson G function for the Cambridge ProximiTREE dataset

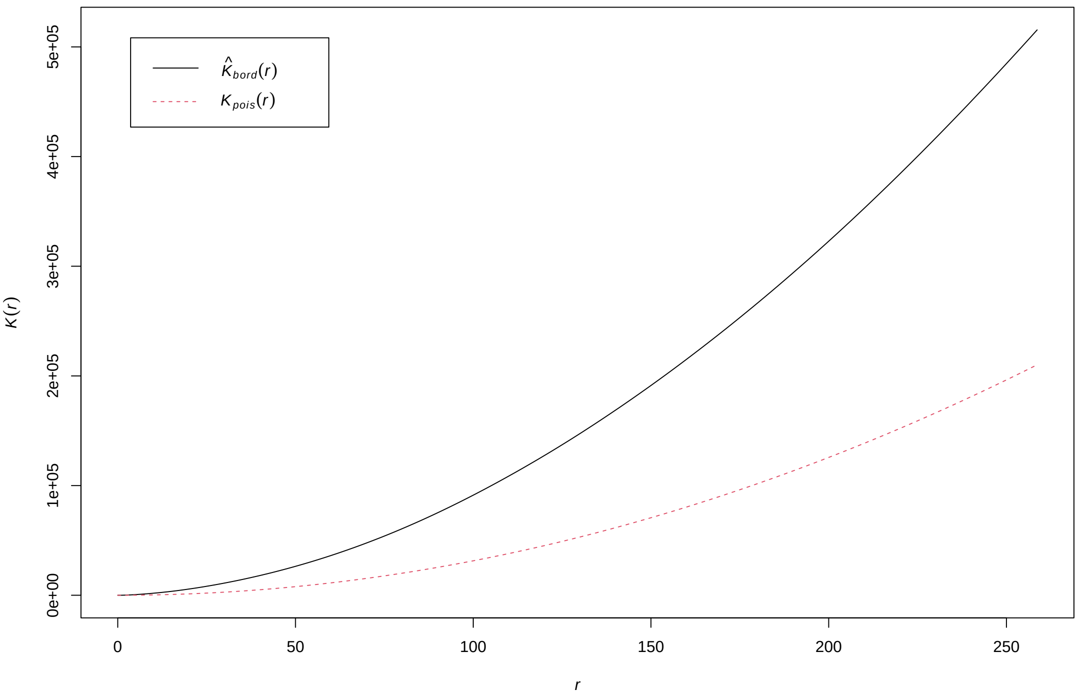


Figure A.2: Plot of the empirical and (completely random) Poisson Ripley's K function for the Cambridge ProximiTREE dataset

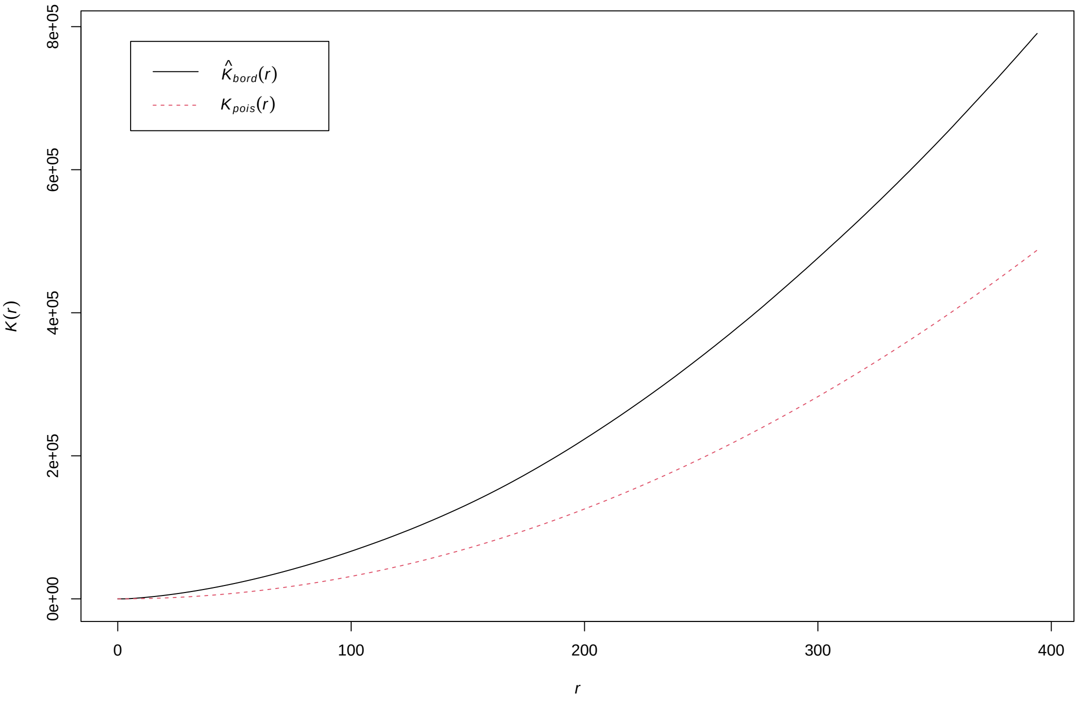


Figure A.3: Plot of the empirical and (completely random) Poisson G function for the Petersfield National Tree Map dataset

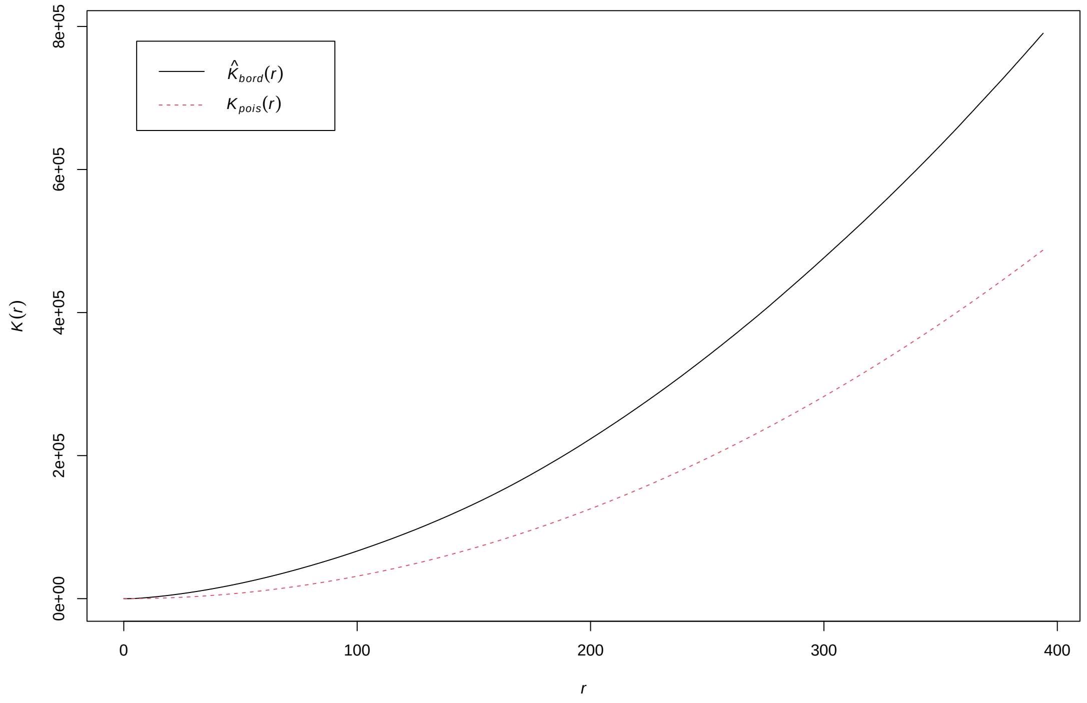


Figure A.4: Plot of the empirical and (completely random) Poisson Ripley's K function for the Petersfield National Tree Map dataset

Table A.1: Summary of land use categories for Petersfield, adapted from the 2015 Land Cover Map. Numbers and rates of trees produced from National Tree Map (NTM) data

	Land use category				
	Grassland	**Other**	**Suburban**	**Urban/ Urban industrial**	**Woodland**
Petersfield area coverage in Ha (%)	305.8 (38.2%)	64.2 (8.0%)	350 (43.7%)	45.0 (5.6%)	36.1 (4.5%)
Number of trees (%)	6,740 (26.2%)	767 (3.0%)	13,119 (51.1%)	902 (3.5%)	4,161 (16.2%)
Rate of trees per Ha	22.0	12.0	37.5	20.0	115.2

Table A.2: Summary of Indicies of Multiple Deprivation (IMD) quintiles. Numbers and rates of trees produced from National Tree Map (NTM) data

	Indicies of Multiple Deprivation (IMD) quintile				
	1	**2**	**3**	**4**	**5**
Petersfield area coverage in Ha (%)	0 (0%)	35.2 (4.4%)	2.6 (0.3%)	484.4 (60.5%)	279.0 (34.8%)
Number of trees (%)	0 (0%)	1143 (4.5%)	142 (0.6%)	14,115 (55.0%)	10,289 (40.1%)
Rate of trees per Ha	NA	32.5	54.6	29.1	36.9

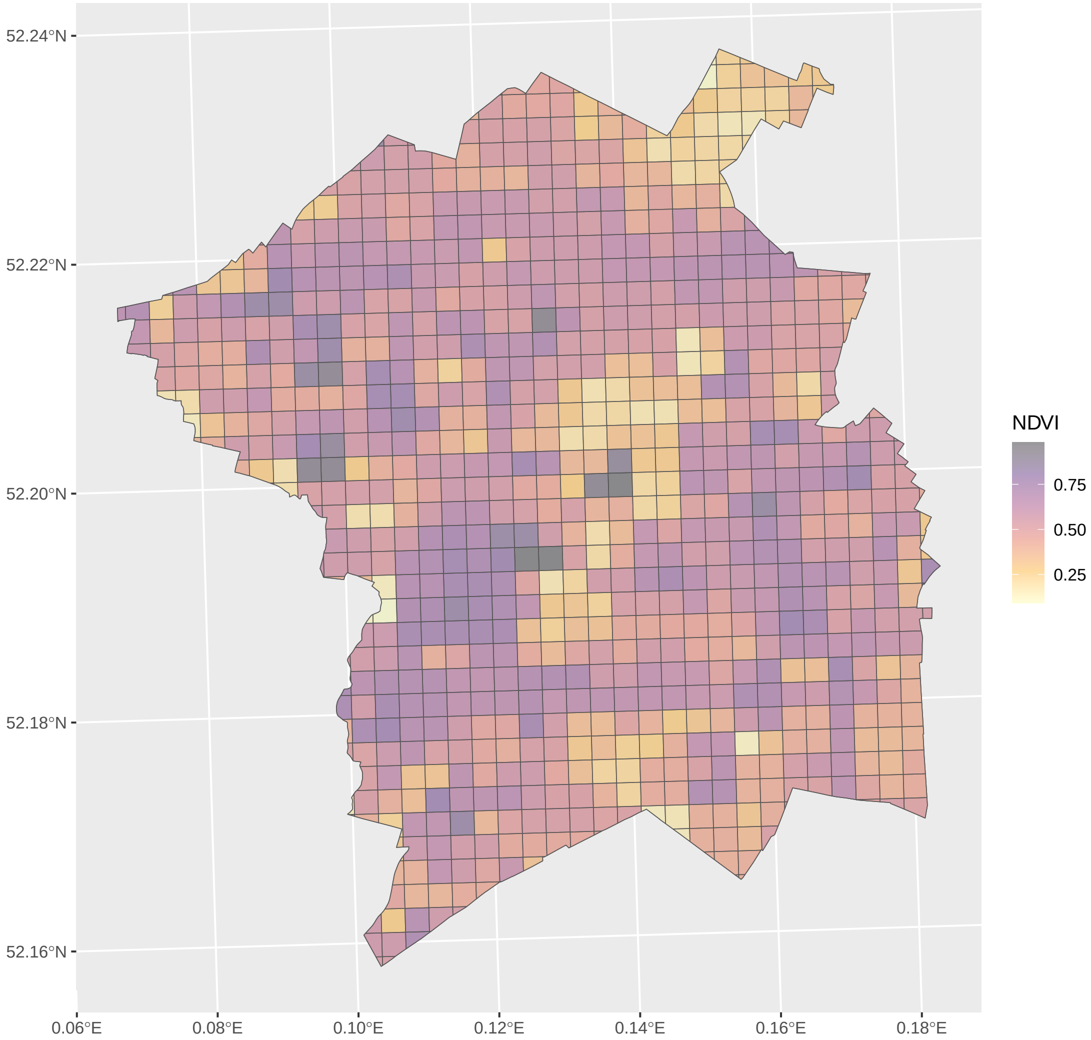


Figure A.5: Plot of Normalized difference vegetation index(NDVI) in Cambridge

Table A.3: Summary of OS MasterMap Green space data for Petersfield. Numbers and rates of trees produced from National Tree Map (NTM) data

	Natural coverage category				
	0	**1**	**2**	**3**	**4**
Number of trees (%)	3,076 (12.0%)	3,731 (14.5%)	4,700 (18.3%)	7,169 (27.9%)	7,013 (27.3%)
Rate of trees per Ha	14.5	25.5	31.5	48.3	48.2

Table A.4: Summary of OS MasterMap Buildings data for Petersfield. Numbers and rates of trees produced from National Tree Map (NTM) data

	Buildings coverage category				
	0	**1**	**2**	**3**	**4**
Number of trees (%)	14,505 (56.5%)	4,599 (17.9%)	3,103 (12.1%)	2,183 (8.5%)	1,299 (5.1%)
Rate of trees per Ha	33.0	50.9	34.2	24.1	14.3

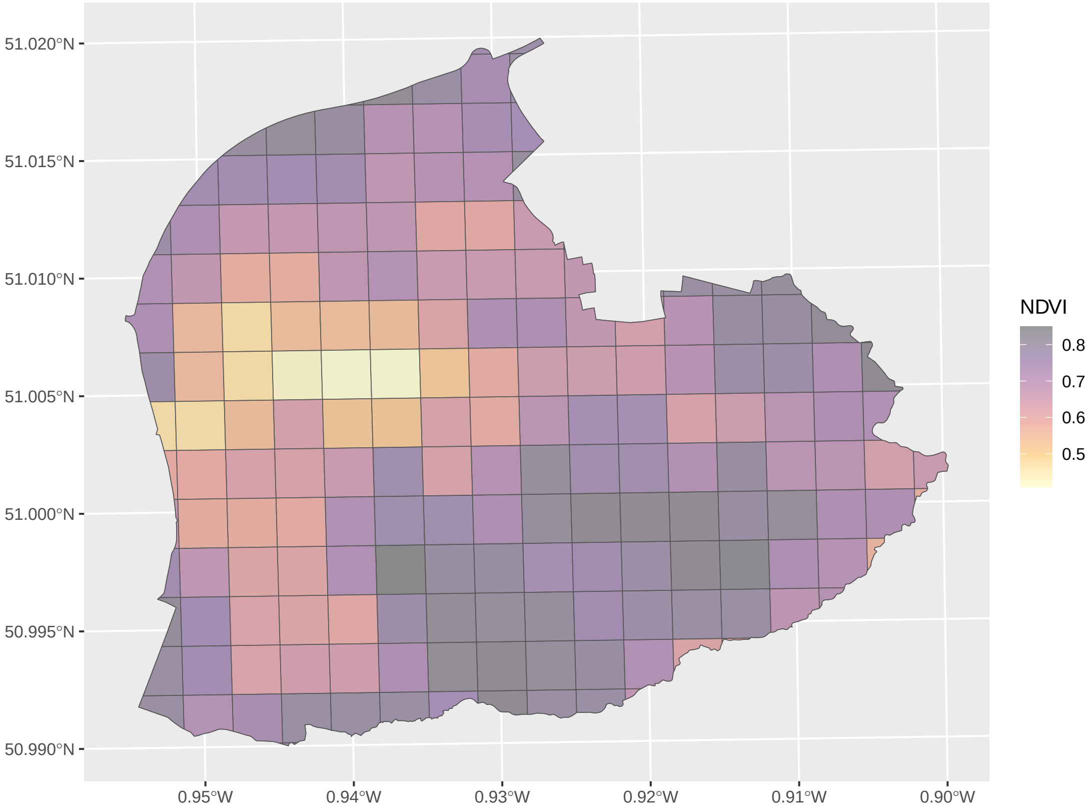


Figure A.6: Plot of Normalized difference vegetation index(NDVI) in Petersfield

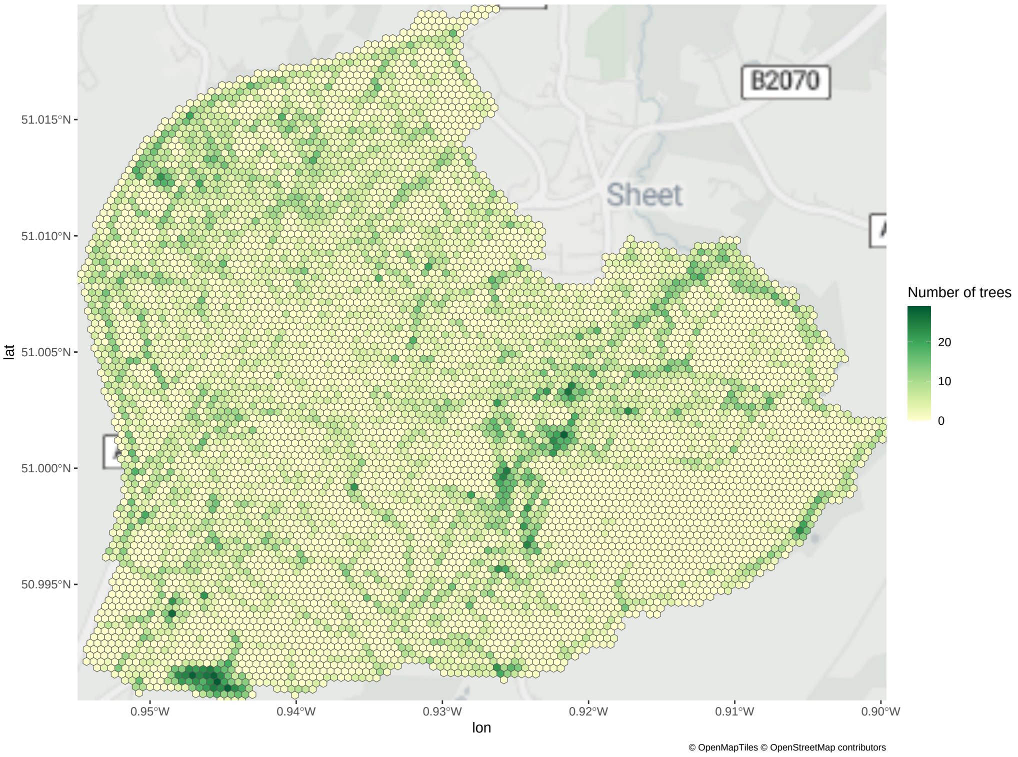


Figure A.7: Plot of the number of National Tree Map(NTM) trees located within overlaid hexagonal cells of size 0.1Ha in Petersfield

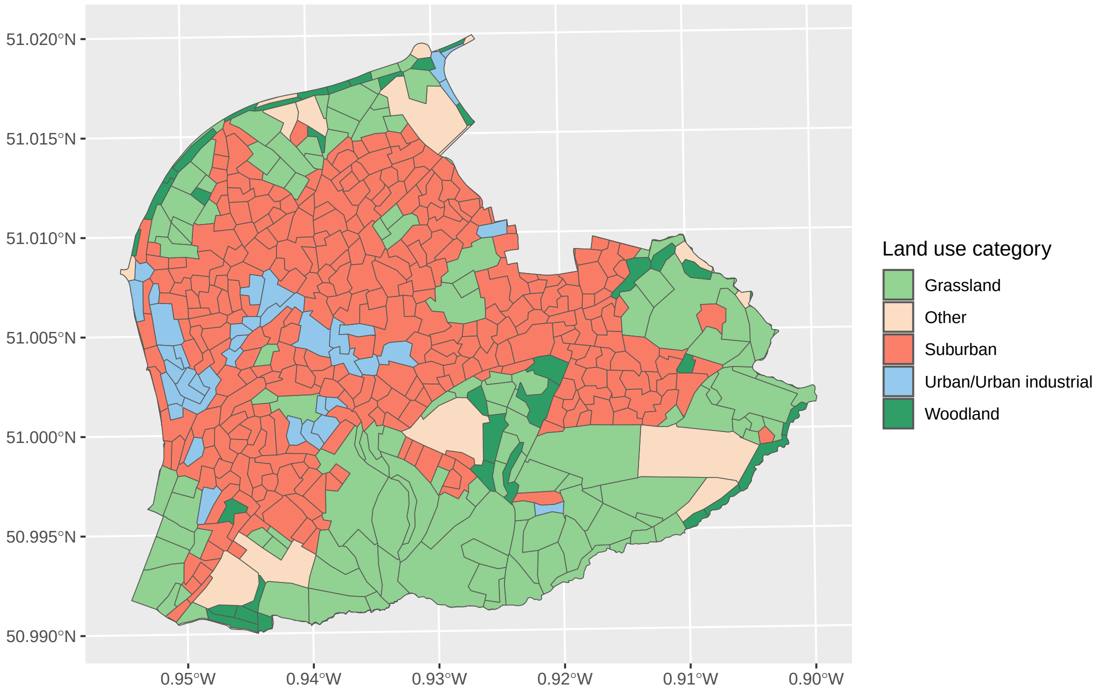


FIGURE A.8: Plot of Land Use categories in Petersfield

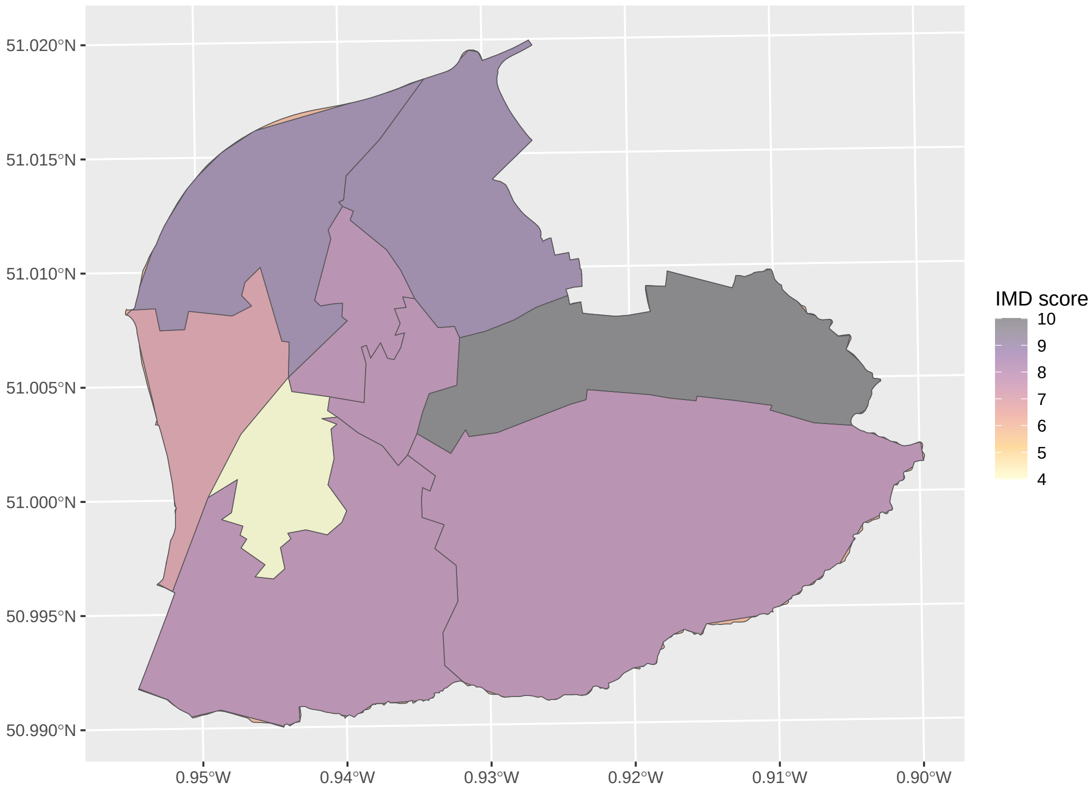


FIGURE A.9: Plot of Indicies of Multiple Deprivation(IMD) deciles in Petersfield

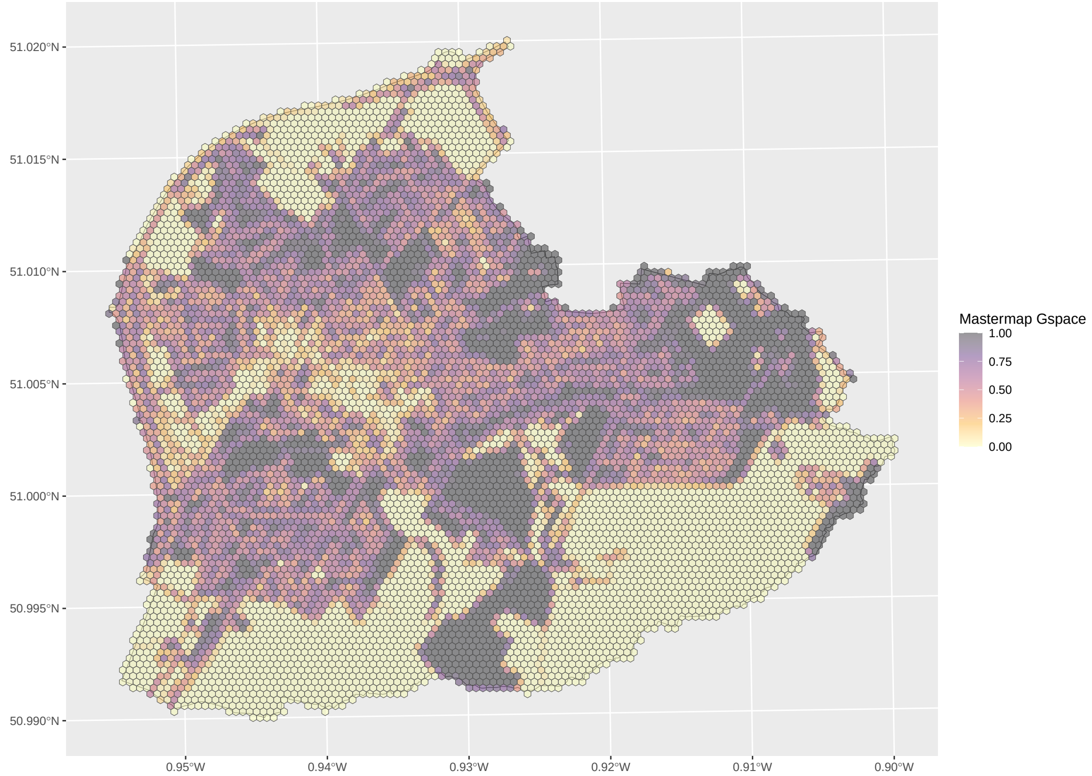


FIGURE A.10: Plot of Greenspace coverage within overlaid hexagonal cells of size 0.1Ha in Petersfield. Greenspace coverage defined from OS MasterMap Topology data

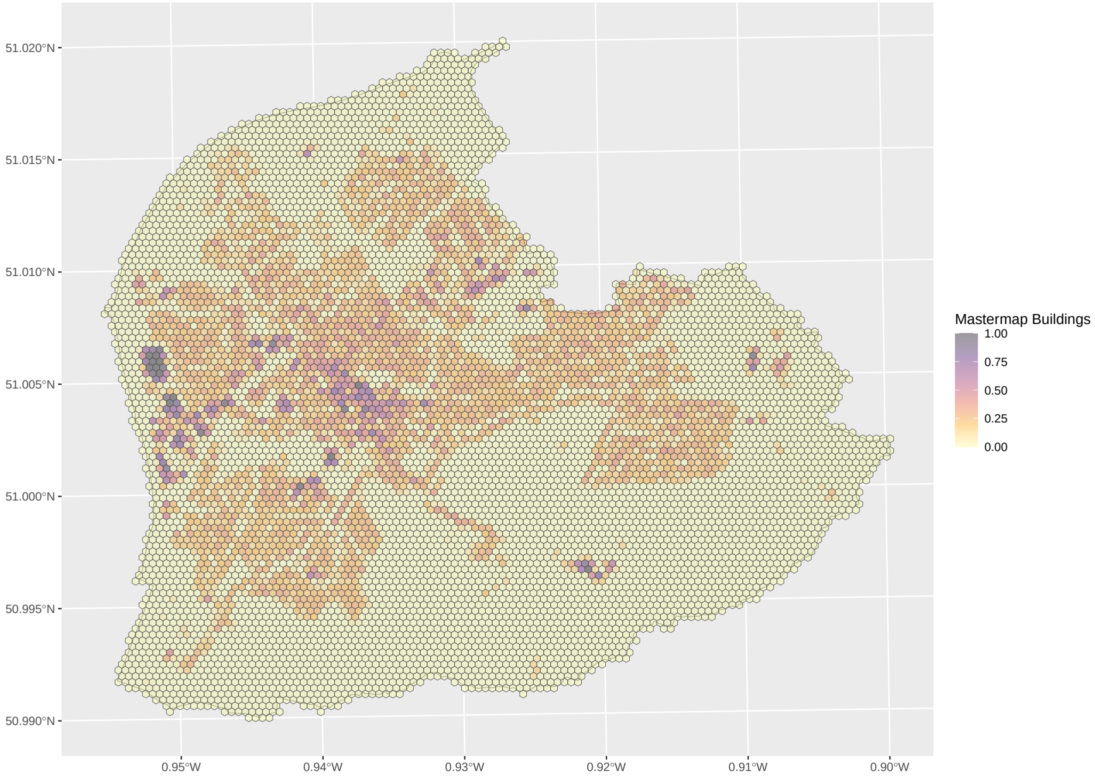


FIGURE A.11: Plot of Building coverage within overlaid hexagonal cells of size 0.1Ha in Petersfield. Building coverage defined from OS MasterMap Topology data

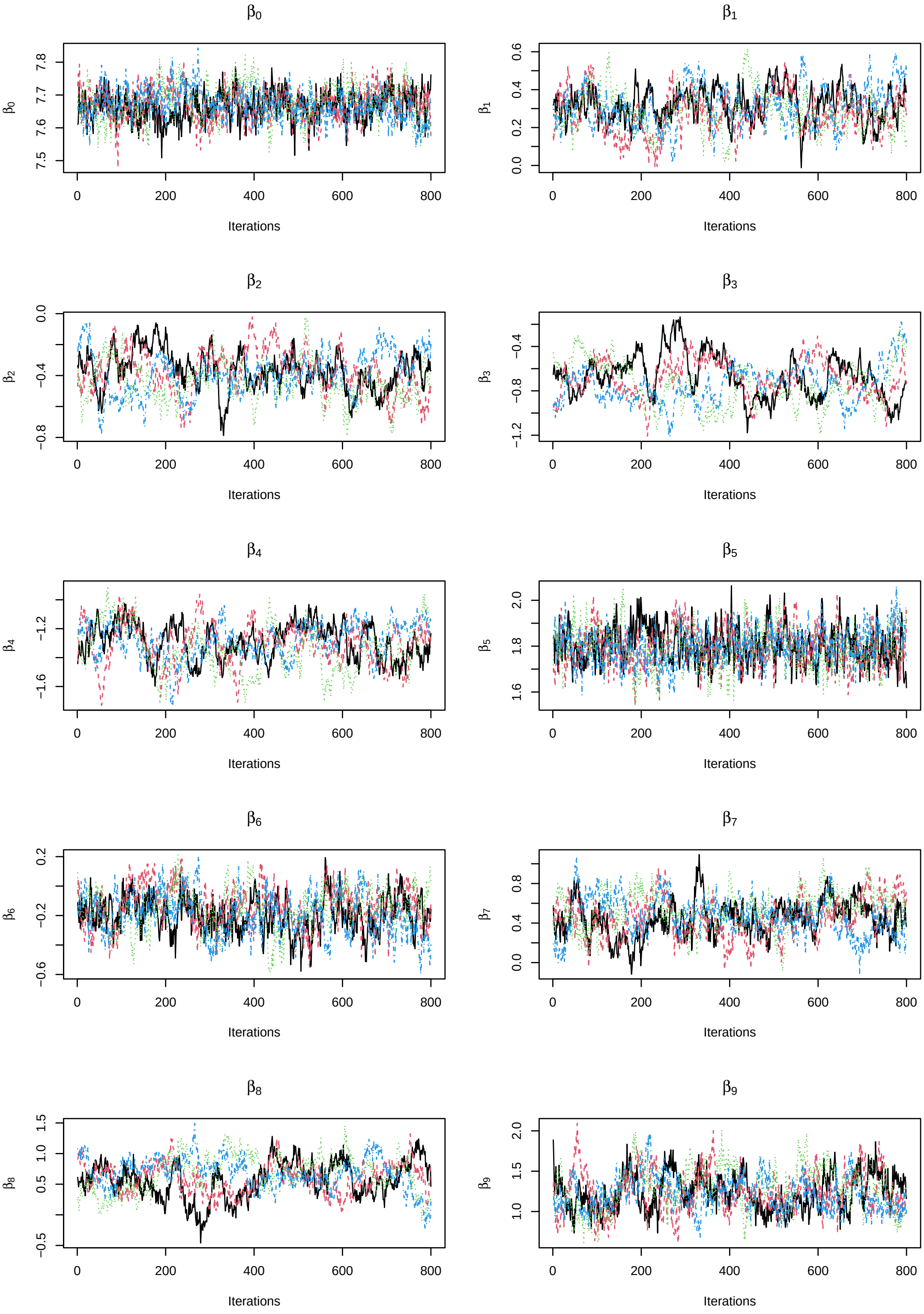


Figure A.12: Coefficient parameter traceplots for the Petersfield National Tree Map (NTM) model

Table A.5: Summary of the model parameters for the Petersfield National Tree Map (NTM) model

Name	Symbol	Category	Mean (95% CI)
Intercept	β_0	-	7.67 (7.58, 7.76)
Greenspace	β_1	-	1.80 (1.65, 1.95)
Buildings	β_2	$(0 < x \leq 0.08)$	0.32 (0.13, 0.53)
	β_3	$(0.08 < x \leq 0.18)$	$-0.39\,(-0.64,\ -0.13)$
	β_4	$(0.18 < x \leq 0.26)$	$-0.75\,(-1.04,\ -0.35)$
	β_5	$(0.26 < x)$	$-1.29\,(-1.53,\ -1.09)$
Interaction	β_6	$(0 < x \leq 0.08)$	$-0.20\,(-0.47,\ 0.08)$
	β_7	$(0.08 < x \leq 0.18)$	0.46 (0.10, 0.85)
	β_8	$(0.18 < x \leq 0.26)$	0.68 (0.16, 1.12)
	β_5	$(0.26 < x)$	1.21 (0.87, 1.65)

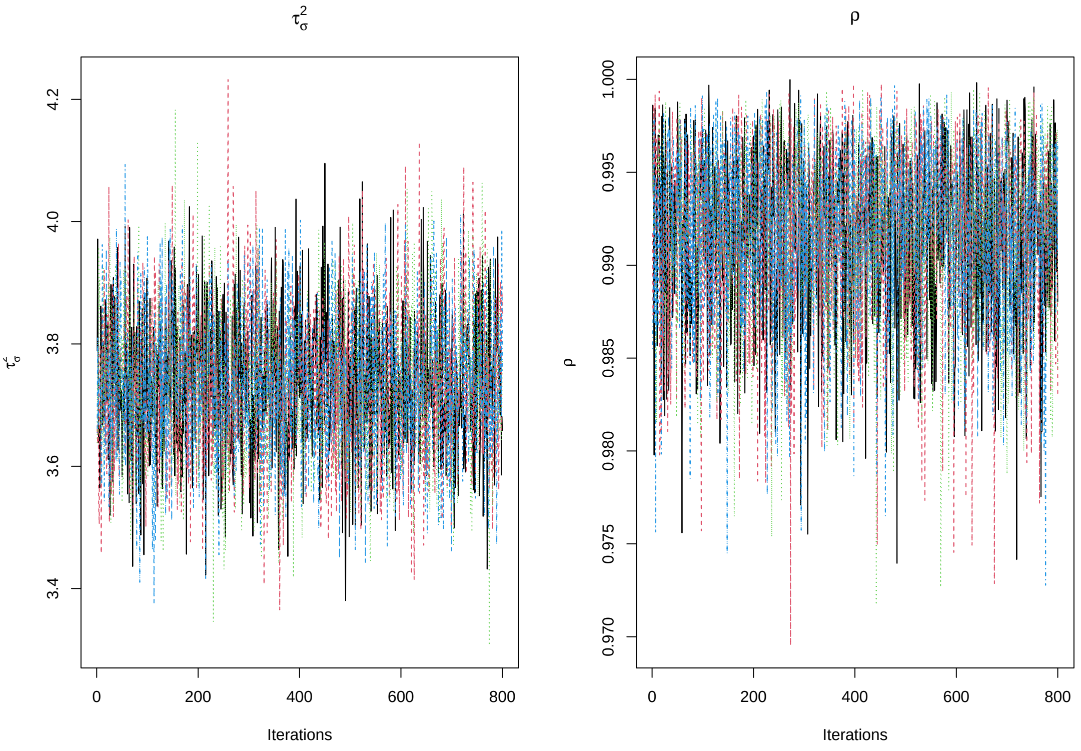


Figure A.13: Uncertainty parameter traceplots for the Petersfield National Tree Map (NTM) model

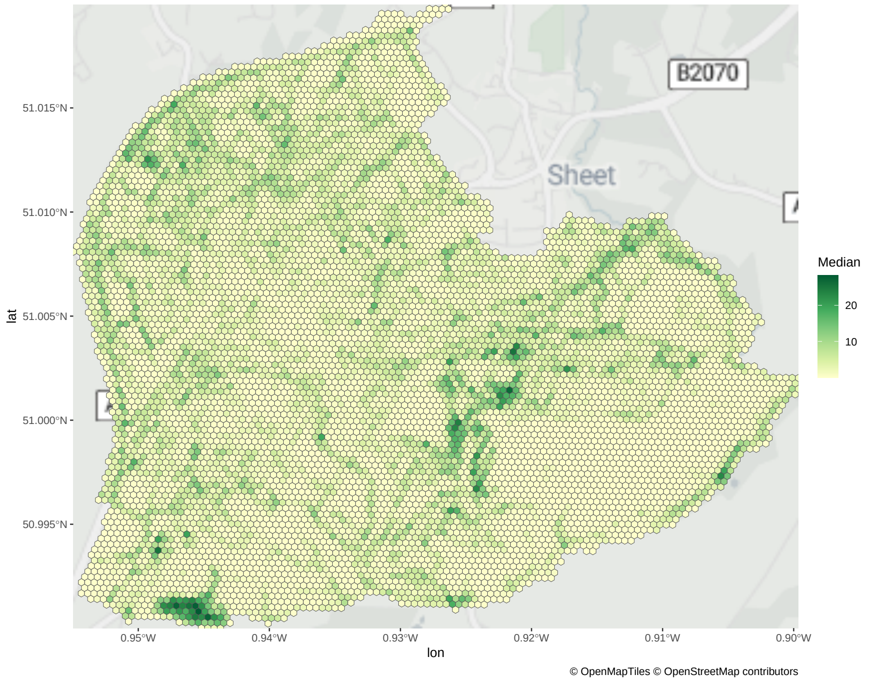


FIGURE A.14: Median of the expected number of trees for each hexagonal cell of size 0.1Ha in Petersfield, as predicted from the Petersfield National Tree Map (NTM) data

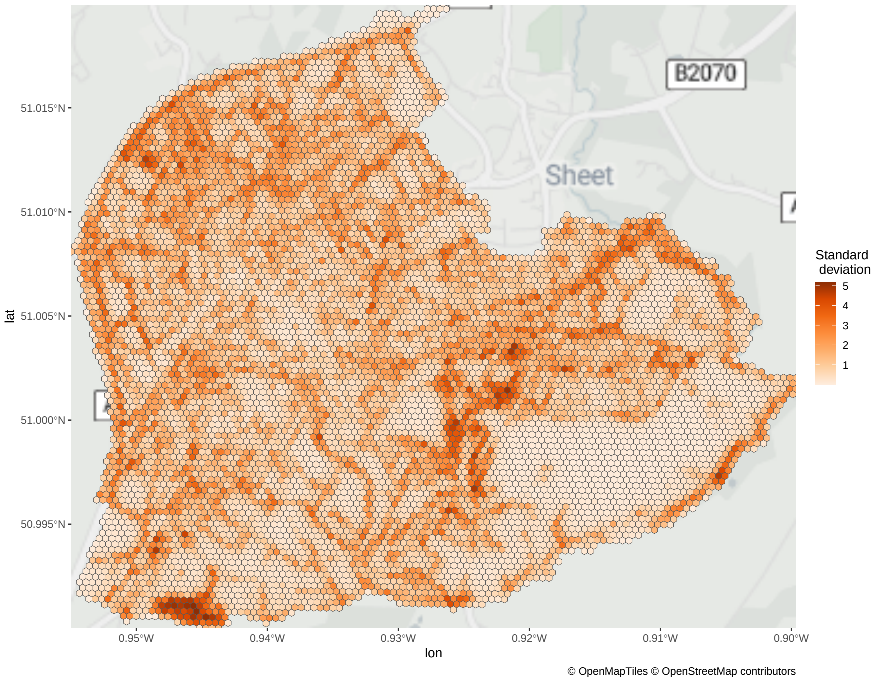


FIGURE A.15: Standard deviation in the expected number of trees for each hexagonal cell of size 0.1Ha in Petersfield, as predicted from the Petersfield National Tree Map (NTM) data

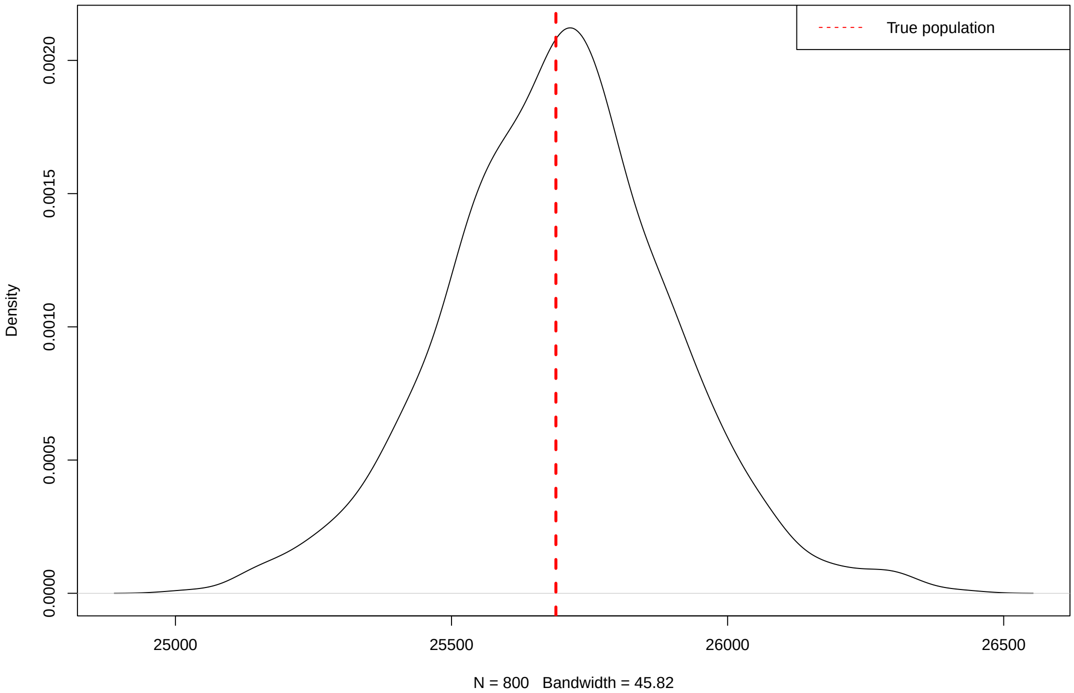


Figure A.16: Plot of the population density as estimated from the Petersfield National Tree Map (NTM) data

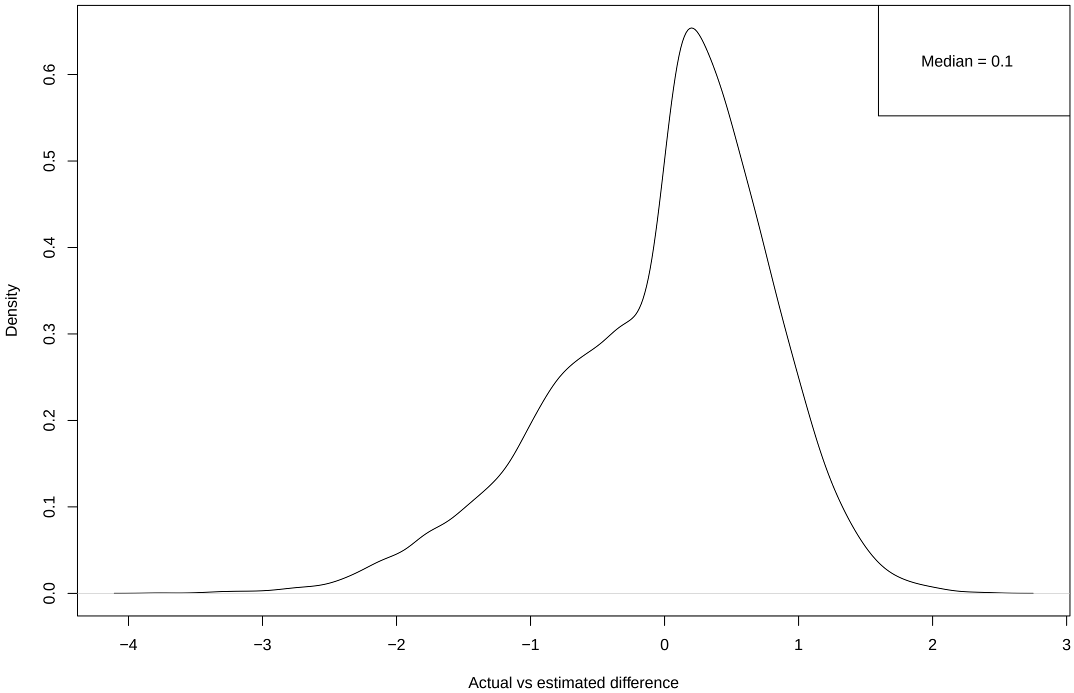


Figure A.17: Plot of the mean Petersfield National Tree Map (NTM) model errors. Model error calculate by subtracting observed values from modelled values

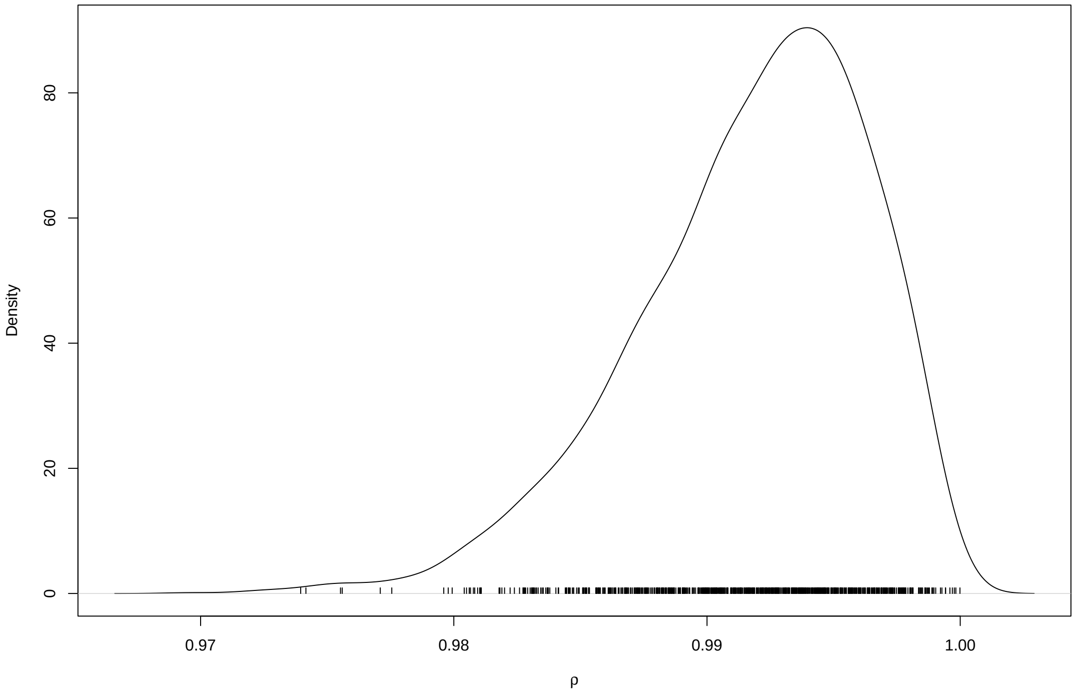


Figure A.18: Plot of the Leroux mixing parameter, ρ, density for the Petersfield National Tree Map (NTM) model

Appendix B

Additional Chapter 4 plots, code and tables

```
data {
  int<lower=0> N;
  int<lower=N> M;
  int<lower=0> N_edges;
  int<lower=1, upper=M> node1[N_edges];
  int<lower=1, upper=M> node2[N_edges];

  int<lower=0> y[N];
  int<lower=0> K;
  matrix[N, K] x;
  vector<lower=0>[N] E;
  real<lower=0> scaling_factor;
  real reciprocal_phi_nb_scale;

}

transformed data {
  vector[N] log_E = log(E);
}

parameters {
  real reciprocal_phi_nb;
  vector[K] beta ;
  real<lower=0> sigma;
  real<lower=0, upper=1> rho;
  vector[N] theta;
  vector[N] phi_fitted;
  vector[M-N] phi_missing;
}

transformed parameters {
  vector[N] convolved_re;
   real phi_nb;
   vector[N] eta;
  vector[M] phi_all;
  phi_all[1:N] = phi_fitted;
  phi_all[(N+1):M] = phi_missing;
  convolved_re =  sqrt(1 - rho) * theta + sqrt(rho / scaling_factor) * phi_fitted;

  phi_nb = 1. / reciprocal_phi_nb;
  eta = log_E + x * beta + convolved_re * sigma;
}

model {
  reciprocal_phi_nb ~ cauchy(0., reciprocal_phi_nb_scale);

  y ~ neg_binomial_2_log(eta, phi_nb);

  target += -0.5 * dot_self(phi_all[node1] - phi_all[node2]);
  sum(phi_fitted) ~ normal(0, 0.001 * N);

  beta ~ normal(0, 5);
  theta ~ normal(0, 1);
  sigma ~ cauchy(0, 25);
  rho ~ beta(0.5, 0.5);
}
```

Listing B.1: Stan code used to model the partially observed survey data

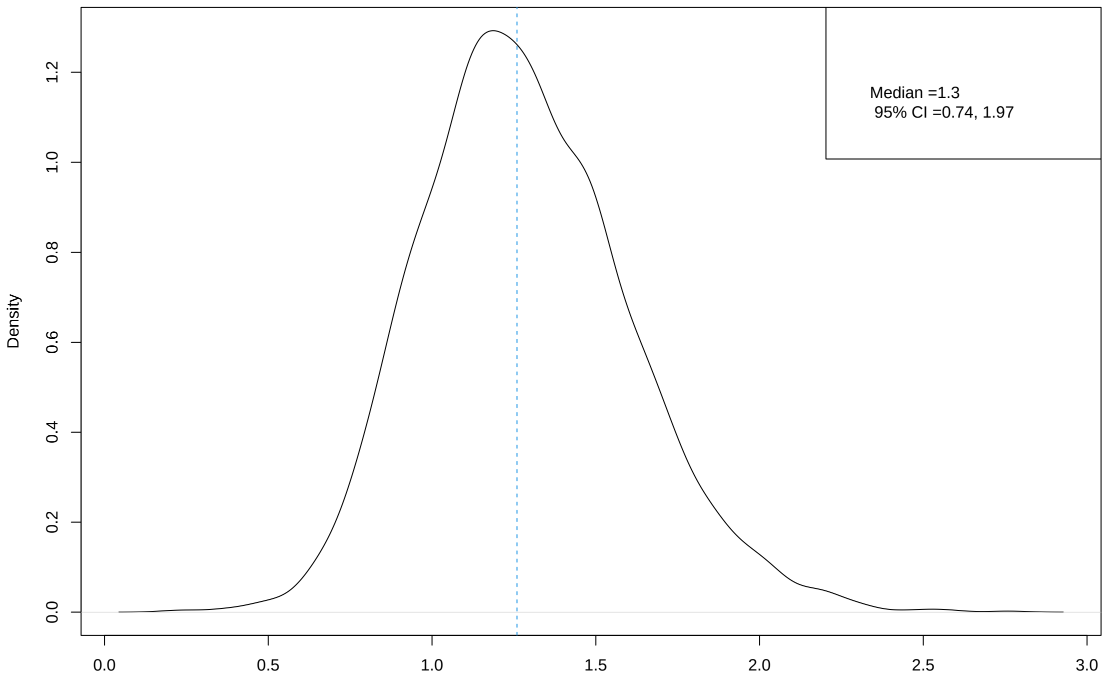


Figure B.1: Overall uncertainty distribution plot for the Southampton i-Tree Eco model

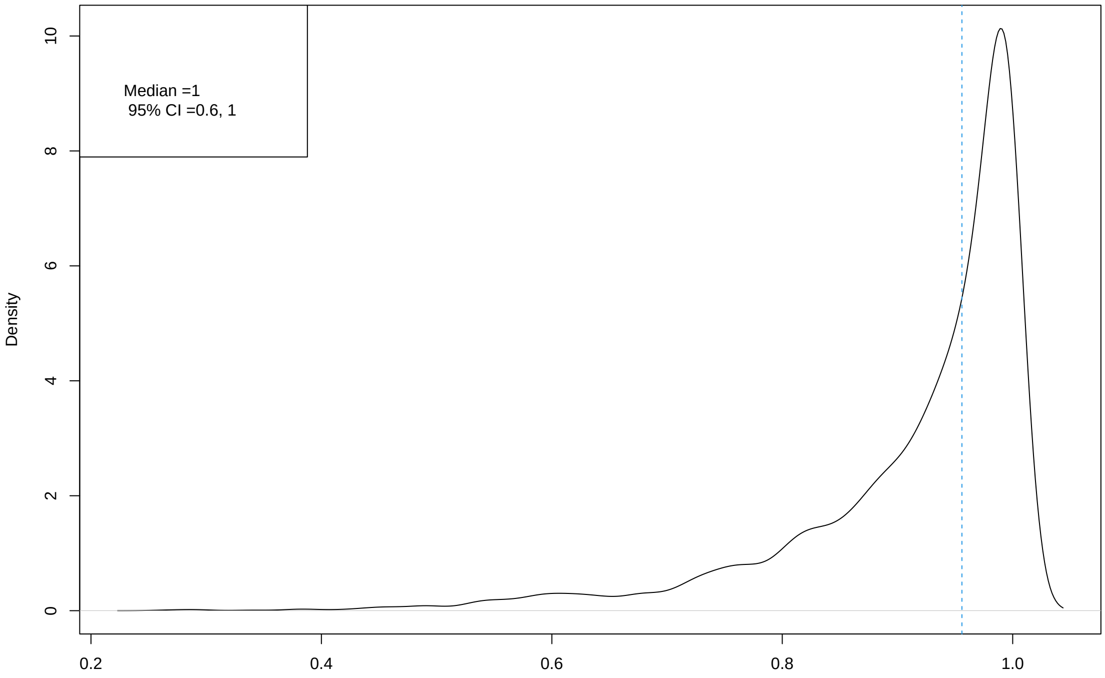


Figure B.2: Density plot of the mixing parameter, ρ, values for the Southampton i-Tree Eco model

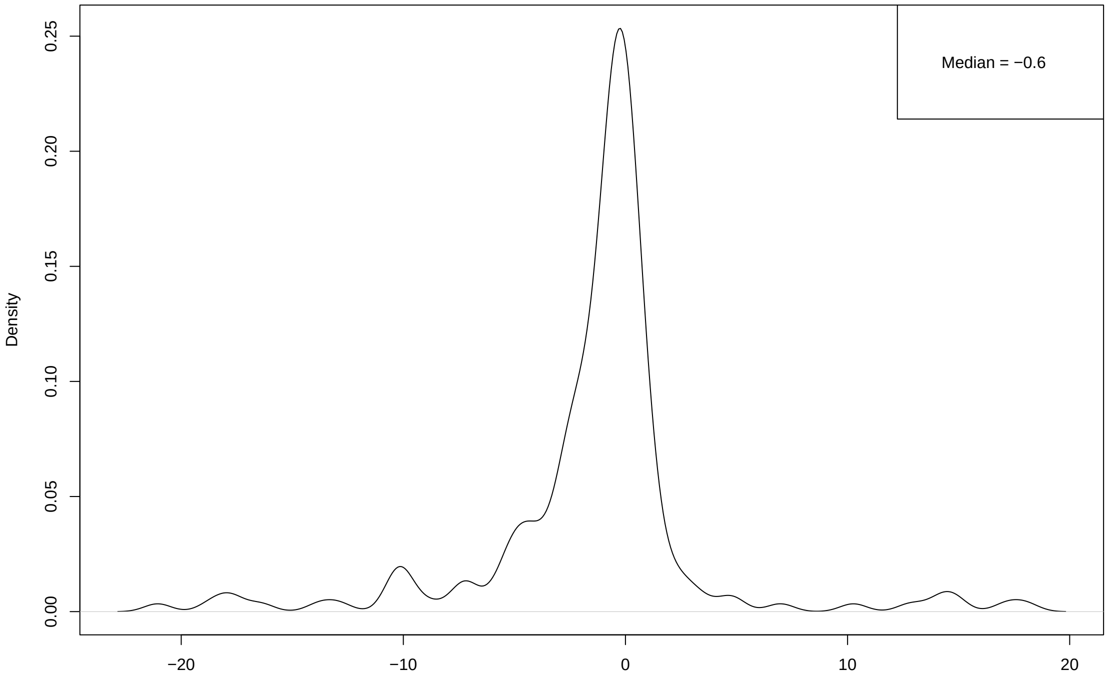
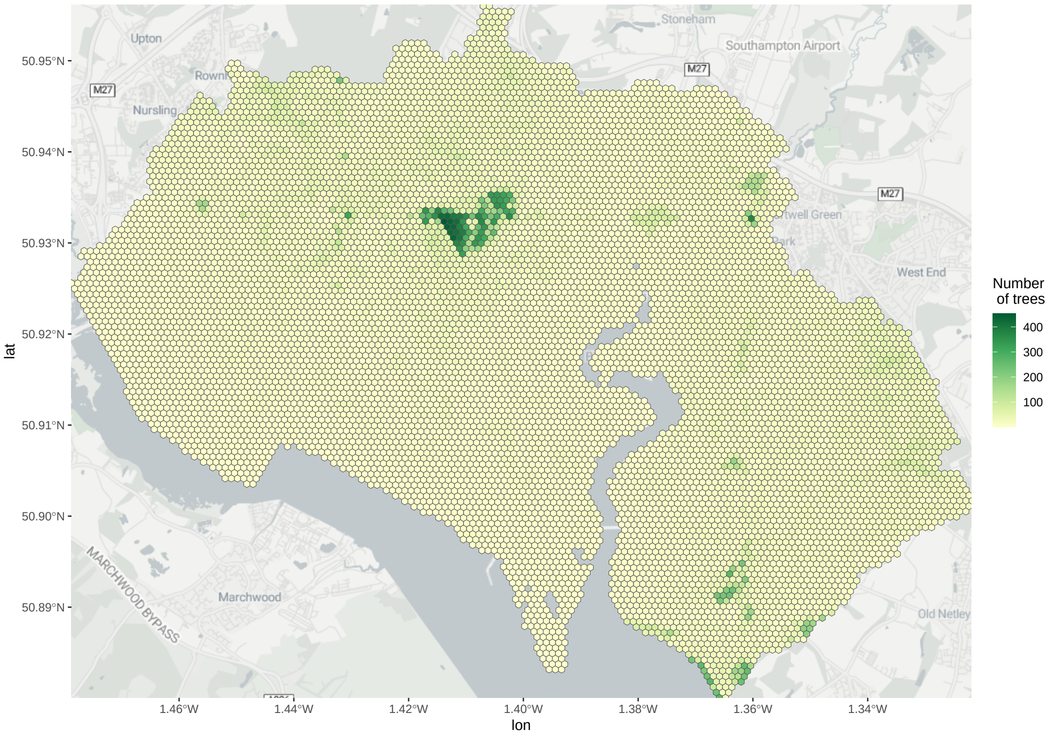


Figure B.3: Density plot of the model error associated with the Southampton i-Tree Eco model in observed cells where the observed number of trees was not equal to zero. Calculated by subtracting observed values from expected values



© OpenMapTiles © OpenStreetMap contributors

FIGURE B.4: Median of the expected number of trees for each cell in Southampton, as predicted from the Southampton i-Tree Eco data

TABLE B.1: Summary of the model parameters for the Cambridge i-Tree Eco model

Name	Symbol	Category	Mean (SD)	95% CI
Intercept	β_0	-	$-1.74\,(0.52)$	$(-2.84,\ -0.83)$
Greenspace	β_1	-	$0.84\,(0.38)$	$(0.13,\ 1.57)$
Buildings	β_2	$(0 < x \leq 0.11)$	$0.39\,(0.41)$	$(-0.40,\ 1.23)$
	β_3	$(0.11 < x \leq 0.18)$	$-0.20\,(0.47)$	$(-1.06,\ 0.74)$
	β_4	$(0.18 < x \leq 0.26)$	$0.36\,(0.42)$	$(-0.44,\ 1.21)$
	β_5	$(0.26 < x)$	$1.83\,(0.51)$	$(0.79,\ 2.80)$
Transformed precision	$\frac{1}{\sqrt{\tau_\sigma}}$	-	$0.86\,(0.41)$	$(0.09,\ 1.65)$
Mixing parameter	ρ	-	$0.48\,(0.34)$	$(0.00,\ 1.00)$

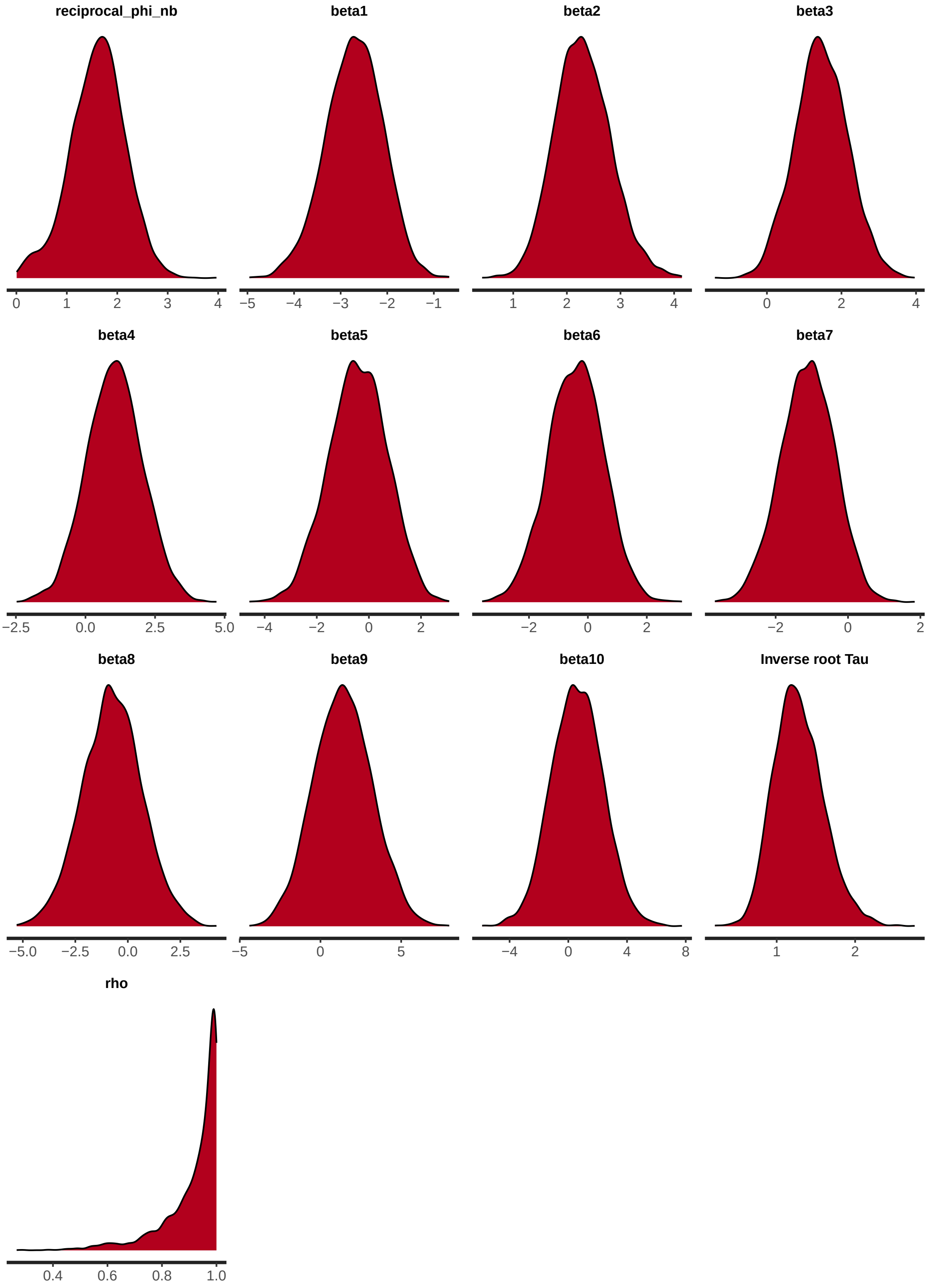


Figure B.5: Parameter density plots for the Southampton i-Tree Eco model

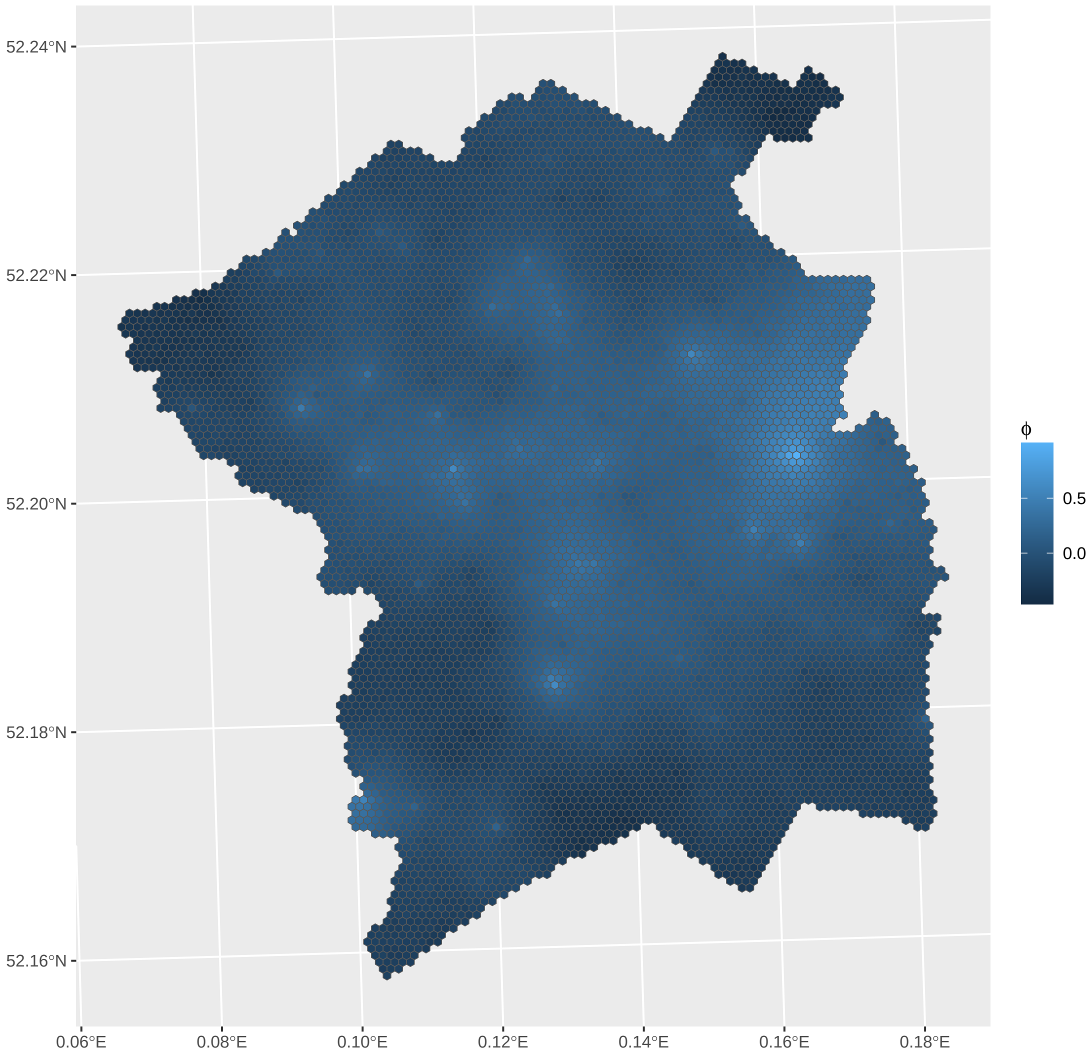


FIGURE B.6: Median of the spatial uncertainty, ϕ for each hexagonal cell of size 0.5Ha in Cambridge, as predicted from the Cambridge i-Tree Eco data

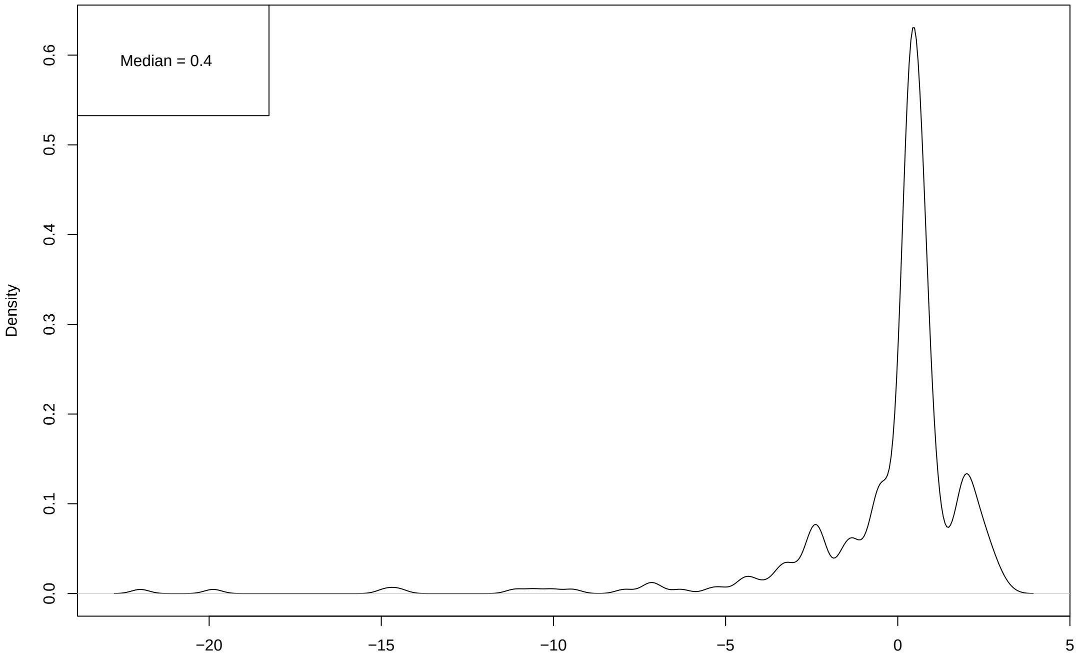


Figure B.7: Density plot of the model error associated with the Cambridge i-Tree Eco model in all observed cells. Calculated by subtracting observed values from expected values

Table B.2: Petersfield itree model parameters summary

Name	Symbol	Category	Mean (SD)	95% CI
Intercept	β_0	-	$-2.29\,(0.44)$	$(-3.18,\ -1.44)$
Greenspace	β_1	-	$1.64\,(0.55)$	$(0.54,\ 2.76)$
Buildings	β_2	$(0 < x \leq 0.08)$	$0.93\,(0.74)$	$(-0.58,\ 2.38)$
	β_3	$(0.08 < x \leq 0.18)$	$-0.56\,(0.96)$	$(-2.40,\ 1.32)$
	β_4	$(0.18 < x \leq 0.26)$	$0.08\,(0.89)$	$(-1, 67,\ 1.78)$
	β_5	$(0.26 < x)$	$-1.73\,(1.13)$	$(-4.06,\ 0.40)$
Interaction	β_6	$(0 < x \leq 0.08)$	$-0.36\,(1.01)$	$(-2.27,\ 1.67)$
	β_7	$(0.08 < x \leq 0.18)$	$1.47\,(1.38)$	$(-1.20,\ 4.23)$
	β_8	$(0.18 < x \leq 0.26)$	$0.50\,(1.45)$	$(-2.26,\ 3.40)$
	β_9	$(0.26 < x)$	$2.86\,(1.97)$	$(-0.99,\ 6.69)$
Transformed precision	$\frac{1}{\sqrt{\tau_\sigma}}$	-	$1.91\,(0.26)$	$(1.35,\ 2.39)$
Mixing parameter	ρ	-	$0.92\,(0.09)$	$(0.67,\ 1.00)$

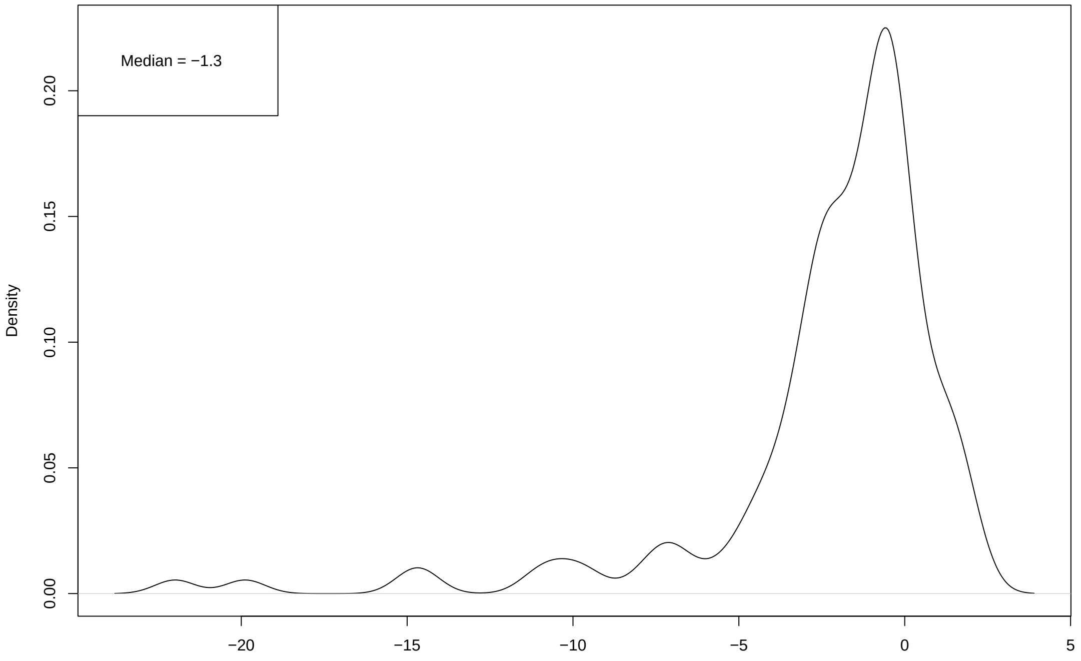


Figure B.8: Density plot of the model error associated with the Cambridge i-Tree Eco model in observed cells where the observed number of trees was not equal to zero. Calculated by subtracting observed values from expected values

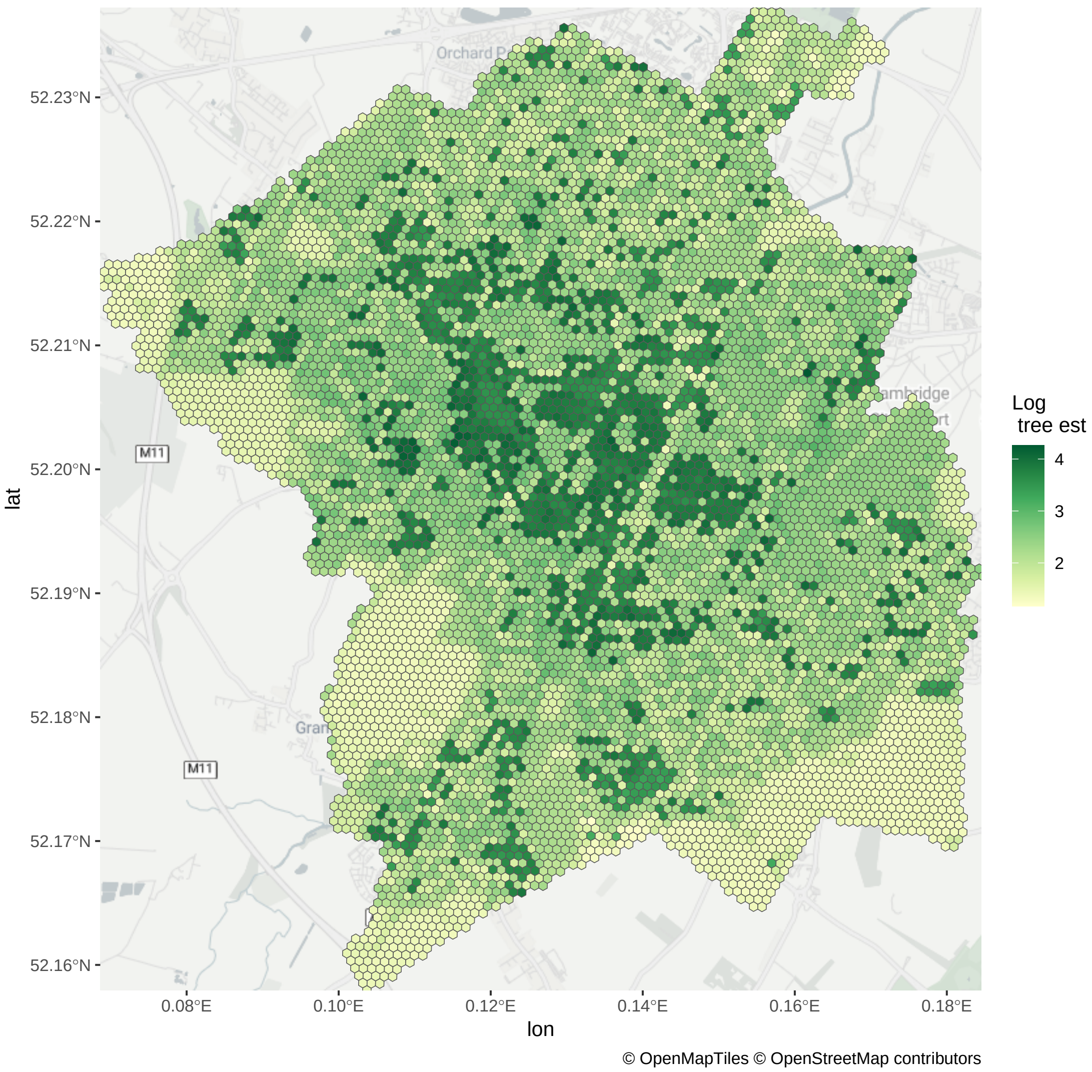


Figure B.9: Log median of the expected number of trees for each cell in Cambridge, as predicted from the Cambridge i-Tree Eco data

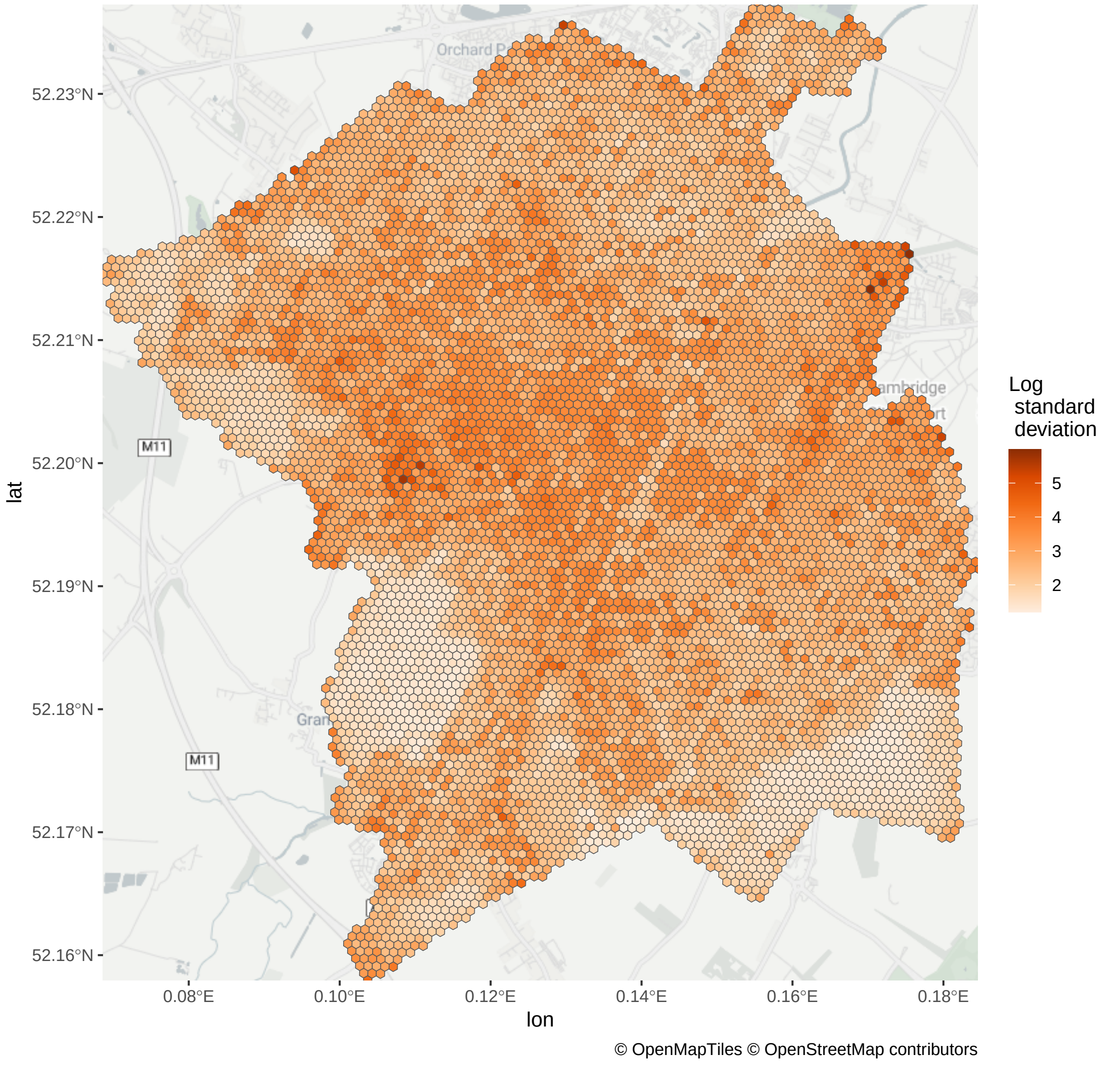


Figure B.10: Log standard deviation in the expected number of trees for each cell in Cambridge, as predicted from the Cambridge i-Tree Eco data

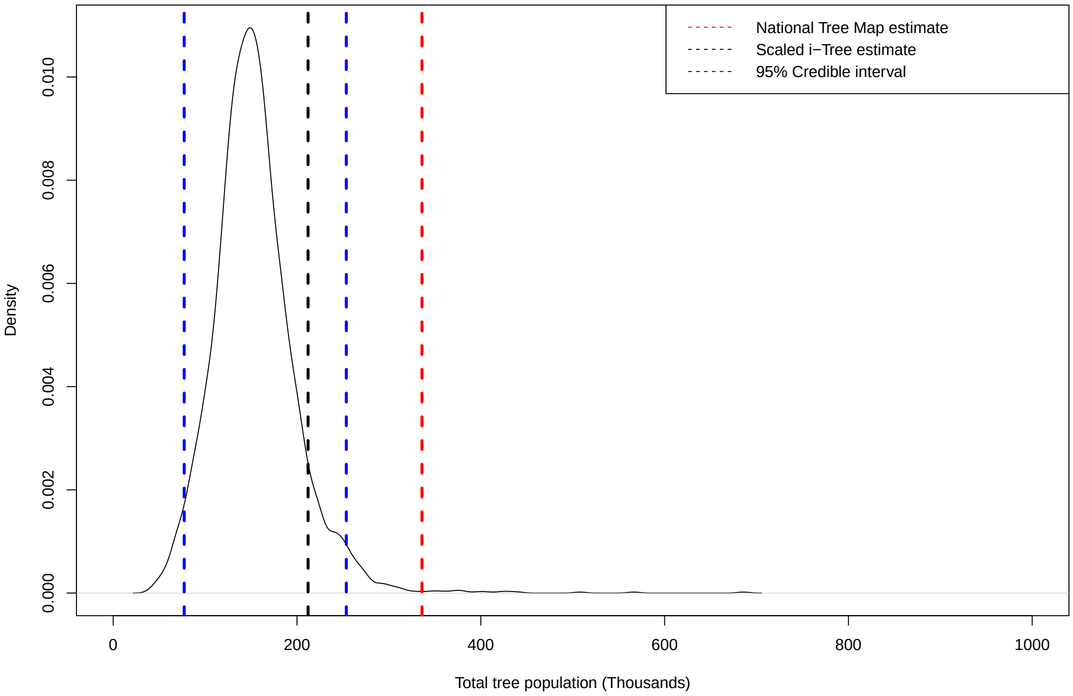


FIGURE B.11: Plot of the population density as estimated from the Cambridge i-Tree Eco data in thousands of trees

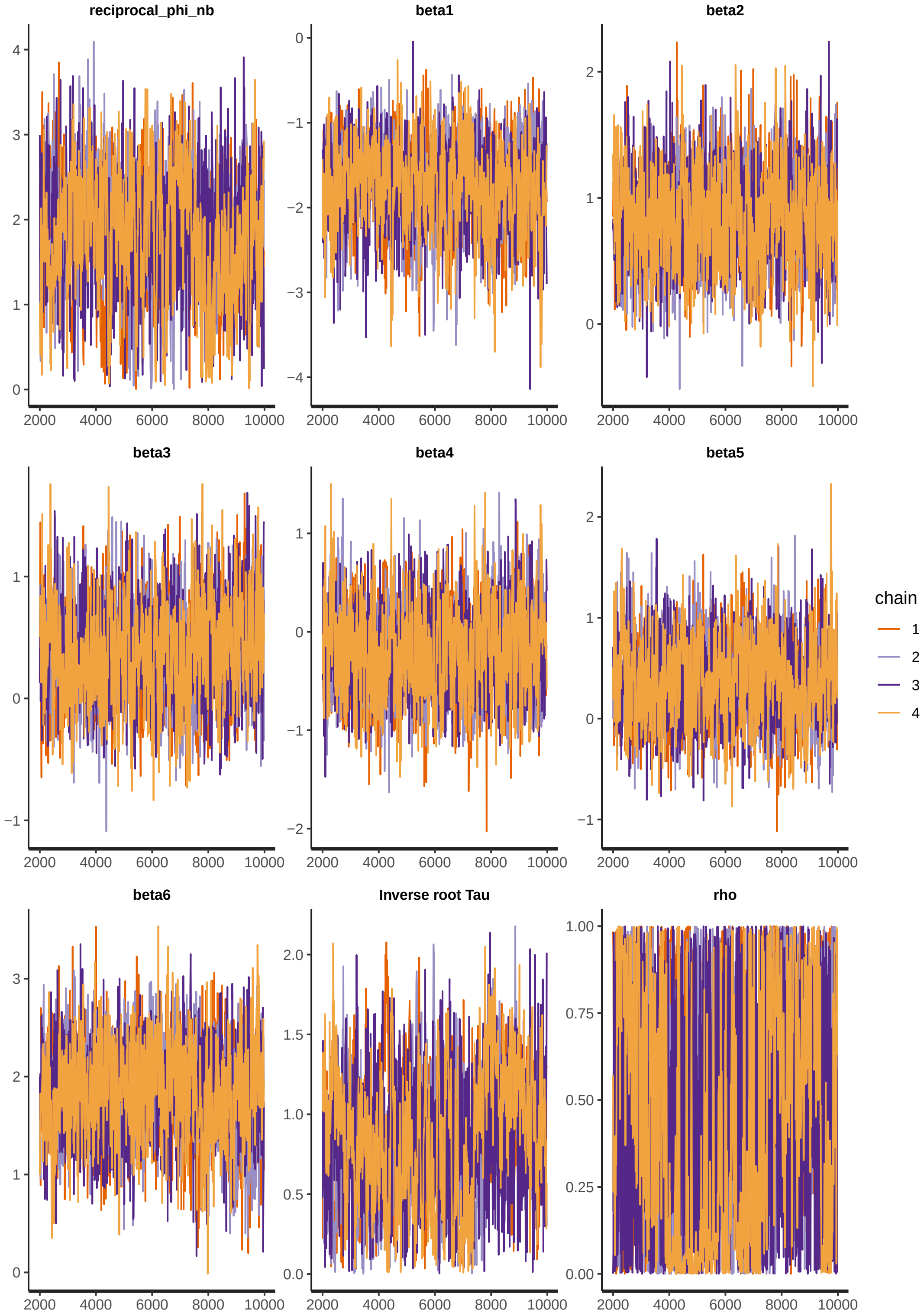


Figure B.12: Traceplots of the model coefficient parameters for the Cambridge i-Tree Eco model

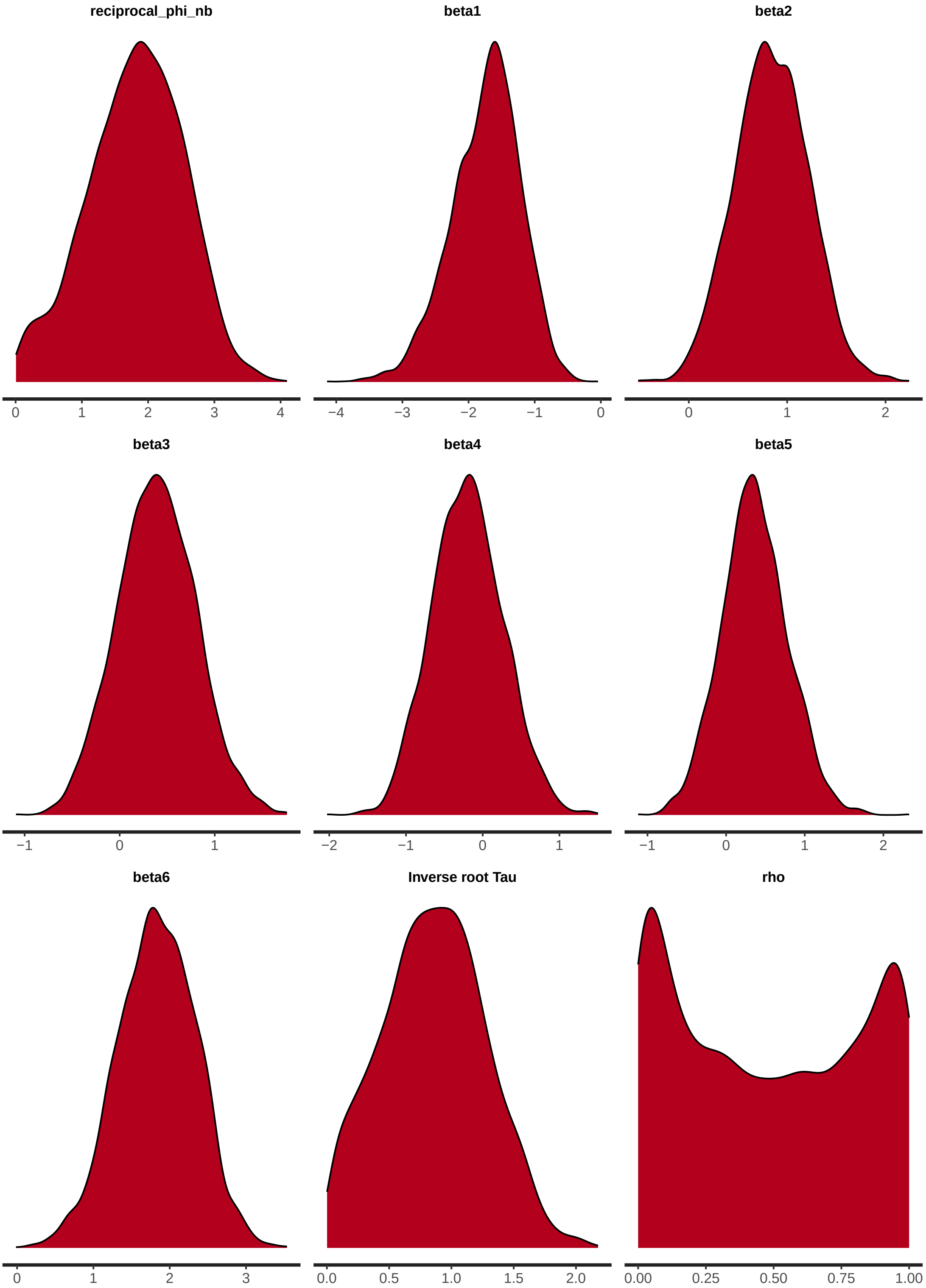


Figure B.13: Parameter density plots for the Cambridge i-Tree Eco model

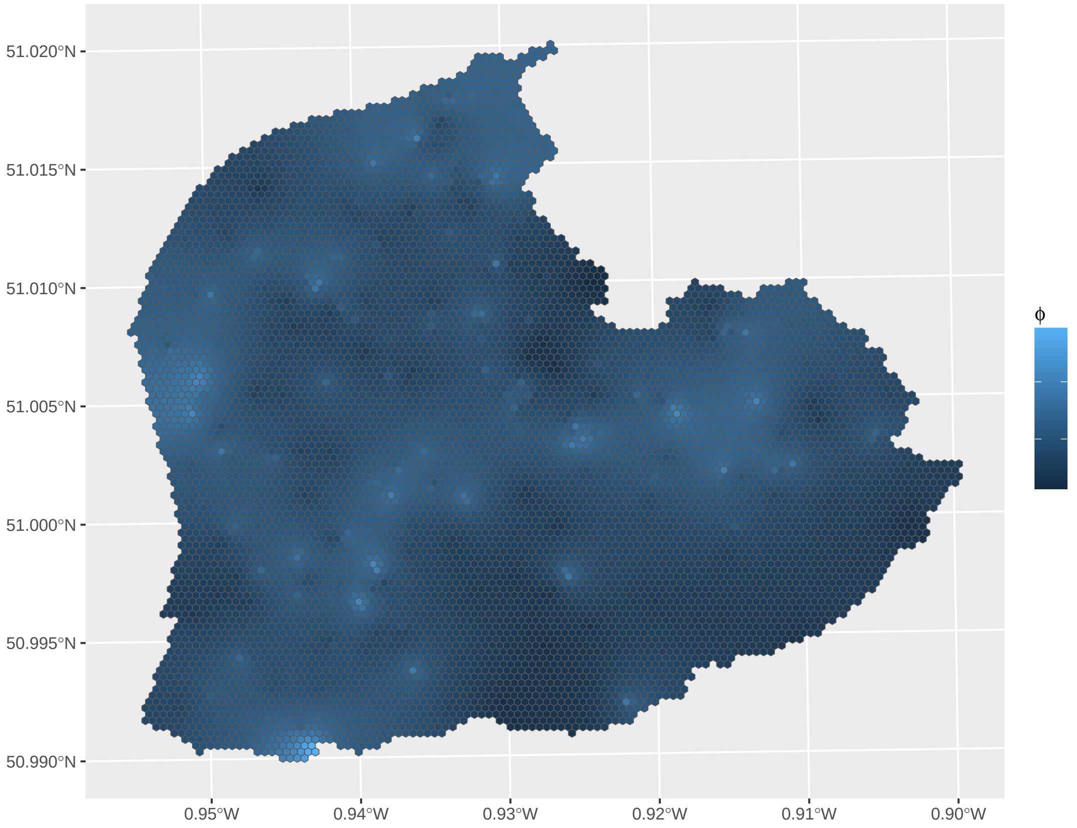


FIGURE B.14: Median of the spatial uncertainty, ϕ for each hexagonal cell of size 0.5Ha in Petersfield, as predicted from the Petersfield i-Tree Eco data

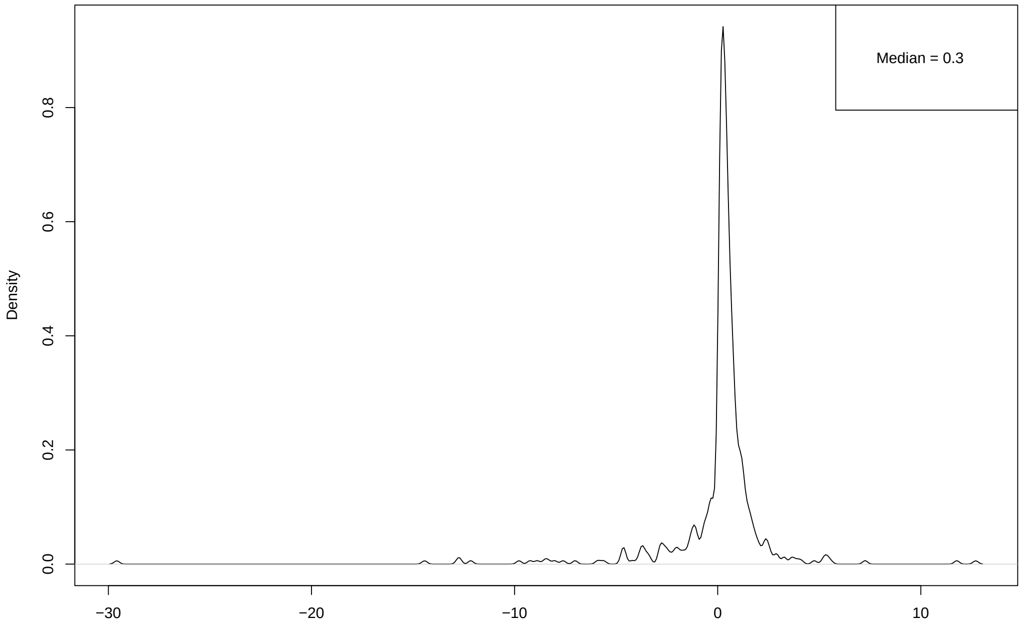


FIGURE B.15: Density plot of the model error associated with the Petersfield i-Tree Eco model in observed cells. Calculated by subtracting observed values from expected values

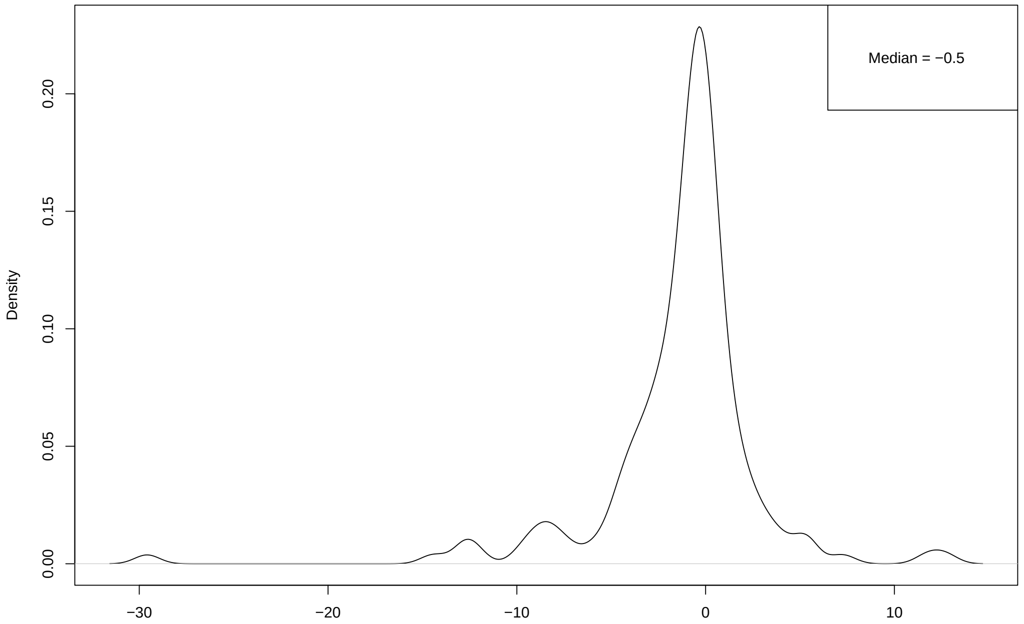


FIGURE B.16: Density plot of the model error associated with the Petersfield i-Tree Eco model in observed cells where the observed number of trees was not equal to zero. Calculated by subtracting observed values from expected values

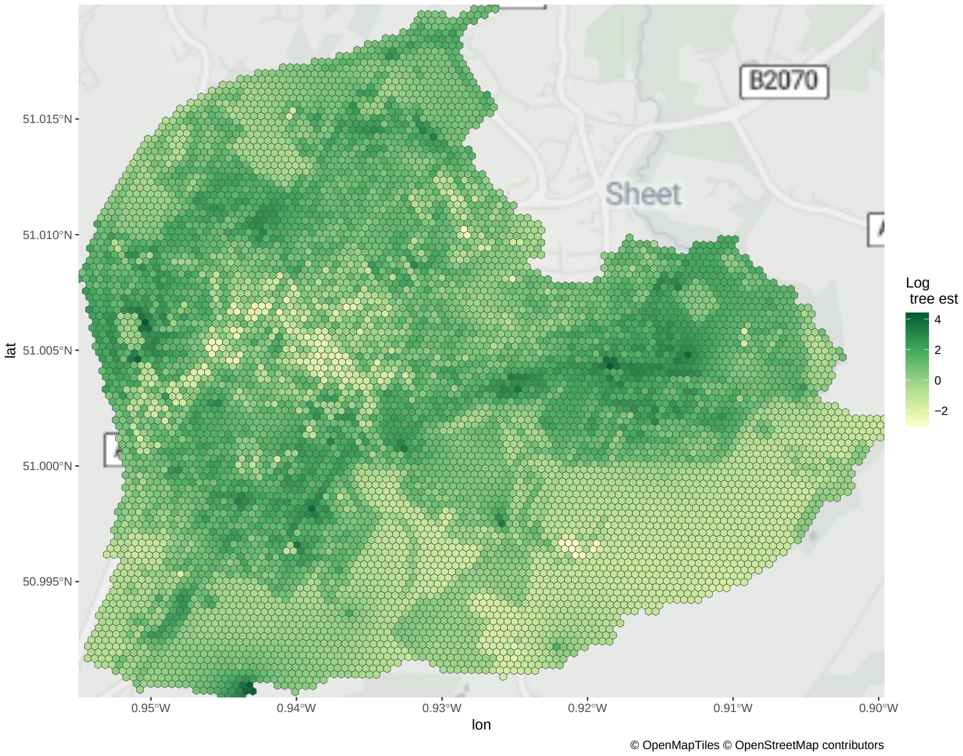


FIGURE B.17: Log median of the expected number of trees for each cell in Petersfield, as predicted from the Petersfield i-Tree Eco data

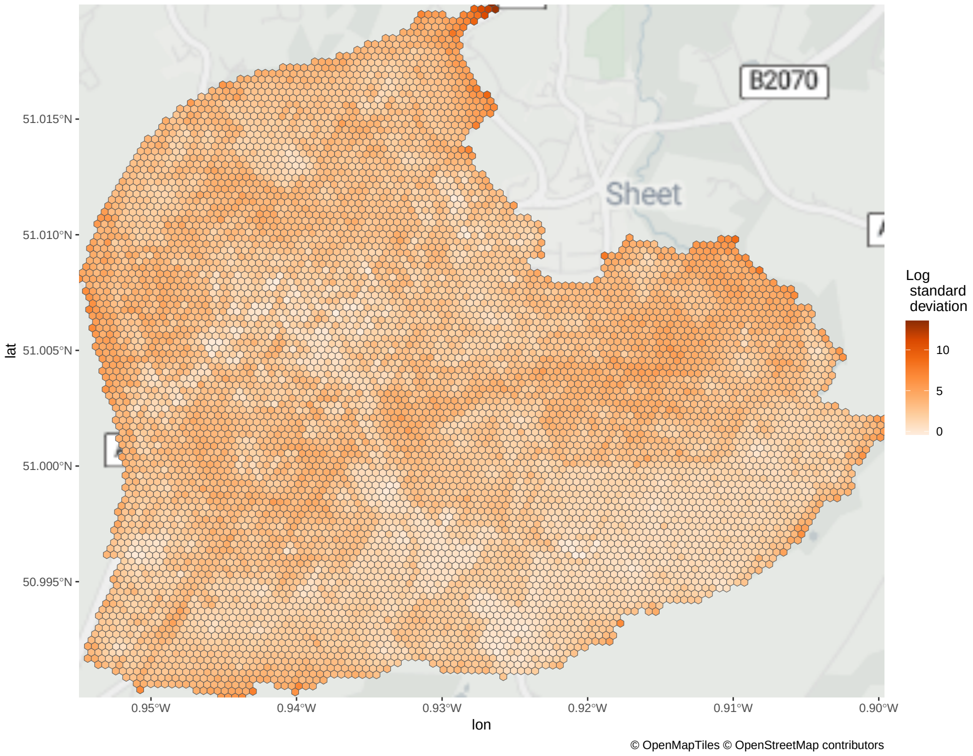


FIGURE B.18: Log standard deviation in the expected number of trees for each cell in Petersfield, as predicted from the Petersfield i-Tree Eco data

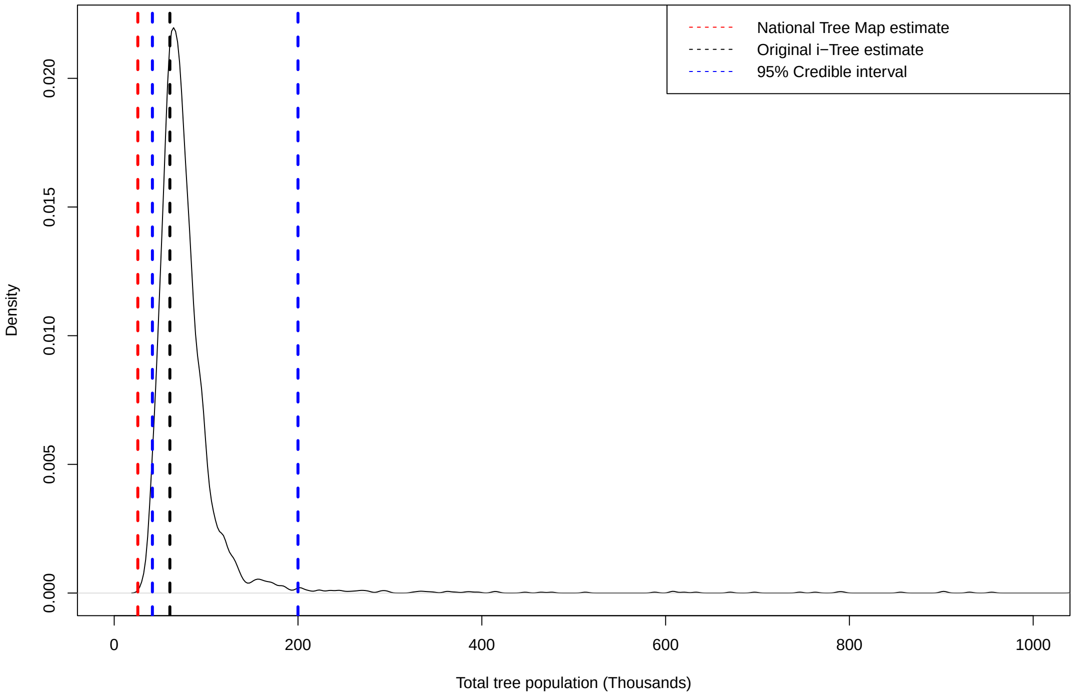


Figure B.19: Plot of the population density as estimated from the Petersfield i-Tree Eco data in thousands of trees

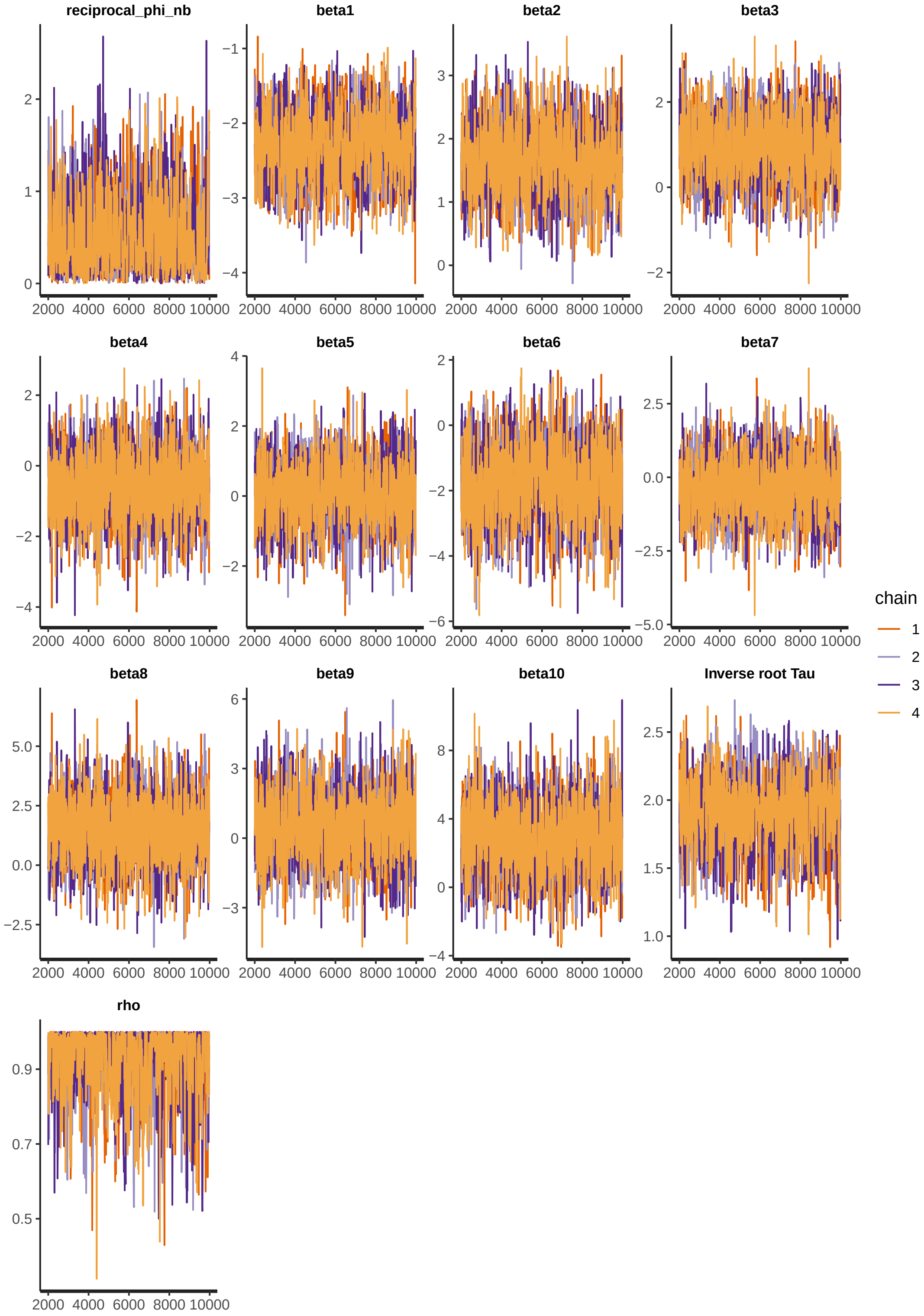


FIGURE B.20: Traceplots of the model coefficient parameters for the Petersfield i-Tree Eco model

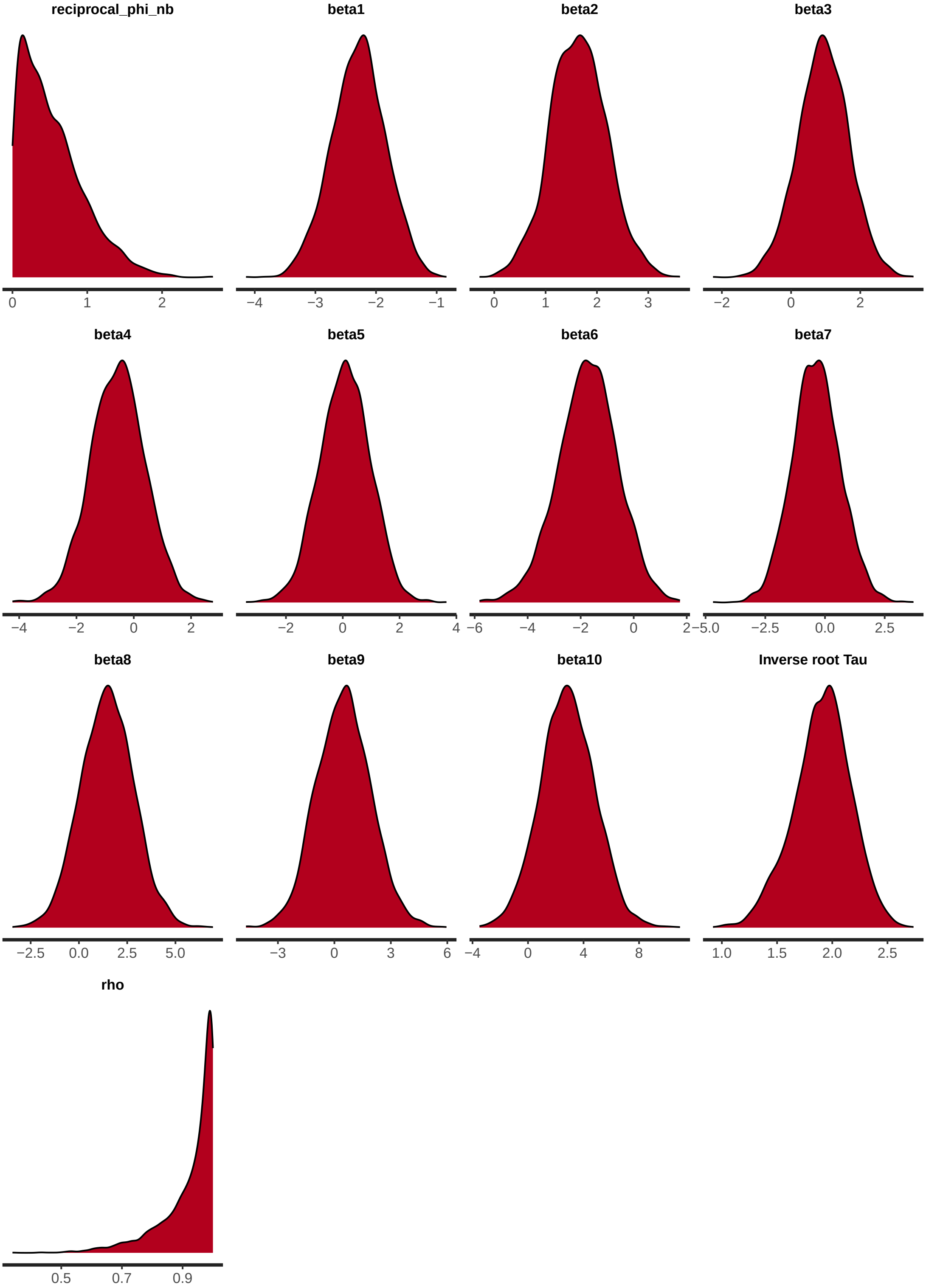


Figure B.21: Parameter density plots for the Southampton i-Tree Eco model

Table B.3: Proportion of hexagonal cells in which the observed value is within the 95% credible interval for the modeled expected number of trees. Models based on survey locations containing trees from the ProximiTREE data for Cambridge. Survey locations taken from the Cambridge i-Tree Eco data or randomly simulated from some random seed (Seed number one and Seed number two) for 200 and 400 survey plots. Results presented for cells of differing sizes

Seed	**Cell size (Ha)**	**Number of plots**	**Proportion of cells in the 95% CI**
Seed number one	1.5	200	96.8
		400	88.2
	1	200	90.7
		400	89.3
	0.5	200	84.9
		400	92.6
Seed number two	1.5	200	97.2
		400	92.5
	1	200	94
		400	84.6
	0.5	200	86.4
		400	83.3

Table B.4: Proportion of hexagonal cells in which the observed value is within the 95% credible interval for the modeled expected number of trees. Models based on survey locations containing trees from the National Tree Map data for Petersfield. Survey locations taken from the Petersfield i-Tree Eco data or randomly simulated from some random seed (Seed number one and Seed number two) for 200 and 400 survey plots. Results presented for cells of differing sizes

Seed	**Cell size (Ha)**	**Number of plots**	**Proportion of cells in the 95% CI**
Seed number one	1.5	200	99.2
		400	100
	1	200	99.2
		400	98.1
	0.5	200	96.8
		400	97.7
	0.1	200	91.2
		400	91.8
Seed number two	1.5	200	95.7
		400	86.7
	1	200	95
		400	90
	0.5	200	94.5
		400	94.3
	0.1	200	90.2
		400	90.1

Table B.5: Proportion of cells in which the observed value is within the 95% credible interval for the modeled expected number of trees. Models based on survey locations containing trees from the ProximiTREE data for Cambridge. Survey locations taken from the Cambridge i-Tree Eco data or randomly simulated from some random seed (Seed number one and Seed number two). Results presented for hexagonal and gridded cells of differing sizes

Seed	**Cell size (Ha)**	**Cell structure**	**Proportion of cells in the 95% CI**
i-Tree	1.5	Hexagon	96.4
		Grid	96.6
	1	Hexagon	93.8
		Grid	94.2
	0.5	Hexagon	99.3
		Grid	94.2
Seed number one	1.5	Hexagon	96.8
		Grid	91.9
	1	Hexagon	90.7
		Grid	81.2
	0.5	Hexagon	84.9
		Grid	81.2
Seed number two	1.5	Hexagon	92.5
		Grid	84.7
	1	Hexagon	84.6
		Grid	85.4
	0.5	Hexagon	83.3
		Grid	85.4

Table B.6: Proportion of cells in which the observed value is within the 95% credible interval for the modeled expected number of trees. Models based on survey locations containing trees from the National Tree Map data for Petersfield. Survey locations taken from the Petersfield i-Tree Eco data or randomly simulated from some random seed (Seed number one and Seed number two). Results presented for hexagonal and gridded cells of differing sizes

Seed	**Cell size (Ha)**	**Cell structure**	**Proportion of cells in the 95% CI**
i-Tree	1.5	Hexagon	97.5
		Grid	98.7
	1	Hexagon	97
		Grid	98.6
	0.5	Hexagon	95.6
		Grid	94.2
	0.1	Hexagon	90
		Grid	91.2
Seed number one	1.5	Hexagon	99.2
		Grid	98
	1	Hexagon	99.2
		Grid	99.3
	0.5	Hexagon	96.8
		Grid	98.1
	0.1	Hexagon	91
		Grid	91.2
Seed number two	1.5	Hexagon	95.7
		Grid	94.7
	1	Hexagon	95
		Grid	93
	0.5	Hexagon	94.5
		Grid	93.3
	0.1	Hexagon	90.2
		Grid	90.8

Appendix C

Additional Chapter 5 plots and tables

Table C.1: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial simulations from modeled Cambridge i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	24.5%	17.9%	14.8%	13.3%	11.8%	11.0%	10.2%	9.5%	9.2%	8.9%
75%	40.5%	30.7%	25.3%	22.7%	20.3%	19.0%	17.5%	16.6%	15.9%	15.4%
90%	57.9%	43.9%	37.1%	32.8%	29.7%	27.7%	25.9%	24.2%	23.0%	22.6%
95%	74.6%	54.5%	46.3%	41.0%	37.0%	34.1%	31.8%	30.2%	28.5%	27.7%

Table C.2: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial simulations from modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	29.0%	22.4%	18.9%	17.0%	15.3%	14.4%	13.5%	12.7%	11.9%	11.5%
75%	46.6%	36.6%	31.6%	28.5%	26.0%	24.2%	22.9%	21.7%	20.5%	19.5%
90%	66.6%	53.1%	46.6%	42.6%	38.5%	36.2%	34.6%	31.9%	30.7%	29.7%
95%	91.8%	71.2%	62.4%	57.4%	52.5%	48.5%	47.0%	42.6%	41.4%	40.3%

Table C.3: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial and Bernoulli simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	30.7%	24.0%	20.3%	18.2%	16.6%	15.4%	14.7%	13.8%	12.9%	12.1%
75%	48.1%	38.7%	33.4%	30.1%	27.6%	26.2%	24.6%	23.2%	22.1%	21.2%
90%	66.9%	55.1%	48.6%	44.1%	40.5%	38.9%	36.6%	35.0%	33.8%	32.4%
95%	101.1%	77.8%	66.2%	58.8%	54.3%	50.5%	48.7%	46.2%	43.4%	41.6%

Table C.4: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial and Bernoulli simulations from modeled Cambridge i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	23.7%	17.1%	14.4%	12.3%	11.0%	10.3%	9.3%	8.8%	8.5%	8.1%
75%	39.2%	29.3%	24.0%	21.2%	19.0%	17.5%	16.4%	15.4%	14.6%	13.9%
90%	56.4%	42.7%	35.3%	31.4%	28.1%	25.6%	24.2%	22.5%	21.1%	20.2%
95%	73.5%	53.4%	43.7%	38.7%	34.8%	31.9%	29.8%	27.6%	26.1%	24.9%

Table C.5: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial and Bernoulli simulations from modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	28.5%	22.1%	18.8%	16.8%	15.1%	14.3%	13.2%	12.5%	11.8%	11.3%
75%	45.7%	36.2%	31.6%	28.3%	25.8%	24.0%	22.5%	21.5%	20.3%	19.5%
90%	65.0%	53.2%	46.9%	42.3%	38.5%	35.5%	34.0%	32.3%	30.6%	29.3%
95%	88.6%	71.3%	62.8%	56.6%	51.9%	48.2%	45.8%	42.4%	41.1%	39.1%

Table C.6: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson and Bernoulli simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	23.0%	17.3%	14.5%	13.0%	11.4%	10.8%	10.3%	9.6%	9.1%	8.6%
75%	37.9%	29.4%	25.0%	22.4%	20.1%	19.0%	17.9%	16.8%	15.9%	15.2%
95%	74.2%	56.2%	47.7%	42.9%	39.4%	36.3%	34.7%	32.9%	31.1%	29.9%

Table C.7: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson simulations from modeled Cambridge i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	15.2%	11.1%	9.0%	7.8%	7.0%	6.3%	5.9%	5.5%	5.2%	4.9%
75%	26.7%	19.3%	15.8%	13.8%	12.4%	11.2%	10.3%	9.8%	9.1%	8.6%
90%	40.1%	28.9%	23.9%	21.3%	18.6%	17.1%	15.5%	14.8%	13.7%	13.1%
95%	49.4%	36.3%	29.6%	26.6%	23.4%	21.6%	19.7%	18.6%	17.4%	16.6%

Table C.8: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson and Bernoulli simulations from modeled Cambridge i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	13.9%	9.9%	8.1%	7.1%	6.3%	5.7%	5.3%	5.0%	4.6%	4.5%
75%	24.6%	17.6%	14.6%	12.7%	11.3%	10.2%	9.5%	8.9%	8.2%	7.8%
90%	37.2%	26.9%	22.2%	19.5%	17.4%	15.9%	14.5%	13.7%	12.8%	12.1%
95%	46.2%	34.0%	27.8%	24.8%	22.0%	20.3%	18.4%	17.6%	16.5%	15.9%

Table C.9: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson simulations from modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	26.2%	19.5%	16.5%	14.8%	13.1%	12.2%	11.3%	10.7%	10.1%	9.7%
75%	43.0%	33.0%	28.4%	25.1%	22.8%	21.3%	19.9%	18.7%	17.8%	16.9%
90%	60.7%	48.4%	42.5%	38.1%	34.9%	32.6%	30.9%	28.6%	27.2%	25.9%
95%	80.7%	63.2%	56.2%	51.6%	46.9%	43.0%	40.8%	38.2%	36.3%	35.2%

Table C.10: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson and Bernoulli simulations from modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	26.0%	19.3%	16.1%	14.6%	13.0%	12.1%	11.2%	10.6%	10.0%	9.5%
75%	42.6%	32.5%	27.9%	24.9%	22.7%	21.0%	19.6%	18.5%	17.7%	16.8%
90%	59.7%	47.8%	42.0%	37.8%	34.9%	32.4%	30.7%	28.5%	27.1%	25.8%
95%	80.2%	62.5%	55.1%	51.2%	46.4%	43.1%	41.1%	38.0%	36.1%	35.0%

TABLE C.11: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial simulations from half the model samples for modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	28.5%	22.2%	18.6%	16.8%	15.2%	14.3%	13.1%	12.7%	11.9%	11.4%
75%	44.8%	36.1%	30.6%	28.0%	25.3%	23.9%	22.3%	21.5%	20.1%	19.2%
90%	65.3%	51.5%	44.3%	40.7%	36.3%	34.8%	32.4%	31.5%	29.5%	27.9%
95%	96.8%	70.3%	57.8%	52.7%	46.4%	44.0%	40.5%	40.1%	36.7%	34.4%

TABLE C.12: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using negative binomial simulations from half the model samples for modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	26.9%	20.1%	16.6%	14.6%	13.2%	12.5%	11.4%	10.7%	10.2%	9.8%
75%	43.3%	32.7%	27.9%	24.7%	22.4%	20.9%	19.3%	18.2%	17.2%	16.4%
90%	60.0%	46.7%	40.3%	35.6%	32.2%	30.3%	28.0%	26.5%	25.1%	23.8%
95%	82.7%	61.0%	51.5%	44.9%	40.0%	37.7%	34.6%	32.5%	31.2%	29.8%

TABLE C.13: Summary of the relative error by percentiles against the number of survey plots. Relative errors calculated using Poisson simulations from half the model samples for modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	22.9%	16.8%	14.0%	12.2%	11.0%	10.1%	9.3%	8.8%	8.3%	7.8%
75%	37.9%	28.3%	24.0%	21.0%	18.8%	17.4%	16.0%	15.1%	14.2%	13.4%
90%	54.5%	41.4%	35.5%	30.9%	27.3%	25.4%	23.5%	21.9%	20.7%	19.5%
95%	69.7%	52.0%	43.9%	38.3%	33.6%	31.5%	28.8%	26.9%	25.6	24.1

TABLE C.14: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using negative binomial simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	92,727	72,144	60,678	54,553	49,474	46,731	44,775	41,909	39,409	38,879
75%	153,034	122,189	104,471	96,064	86,982	83,004	79,026	74,577	70,039	68,786
90%	232,285	192,674	167,779	152,535	137,738	132,662	125,000	117,571	113,675	109,381
95%	322,281	268,706	232,270	214,963	186,110	182,059	170,686	160,386	156,470	151,706

Table C.15: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using negative binomial simulations from modeled Cambridge i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	45,507	33,399	27,475	24,530	22,036	20,532	18,773	17,466	16,932	16,395
75%	76,556	57,774	47,892	42,417	38,313	36,227	32,998	31,226	30,097	28,300
90%	116,204	87,719	72,984	65,219	58,744	54,715	50,750	48,423	45,841	44,253
95%	154,124	113,000	95,026	85,168	76,234	68,971	65,447	62,210	57,769	57,366

Table C.16: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using negative binomial simulations from modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	21,728	16,900	14,242	12,717	11,466	10,793	10,025	9,471	8,970	8,577
75%	37,362	29,124	25,024	22,626	20,665	19,290	18,106	17,067	16,233	15,493
90%	60,366	48,458	41,505	37,609	34,535	32,571	31,015	28,318	27,211	25,828
95%	94,638	70,976	62,221	57,024	52,289	47,646	46,174	41,847	41,223	40,037

Table C.17: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using Poisson simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	68,736	52,241	43,684	39,197	34,562	32,408	31,179	28,659	27,333	26,180
75%	120,116	91,763	77,809	70,067	63,193	59,488	56,725	52,606	49,415	46,918
90%	187,601	149,768	125,430	113,188	103,332	96,727	90,941	86,355	82,121	79,011
95%	264,251	202,375	170,548	153,337	141,785	130,267	122,029	115,734	110,066	107,129

Table C.18: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using Poisson simulations from modeled Cambridge i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	28,091	20,472	16,710	14,371	12,841	11,636	10,832	10,043	9,547	8,999
75%	50,712	36,453	30,034	25,995	23,494	21,249	19,543	18,443	17,359	16,315
90%	79,189	57,834	47,147	42,434	36,969	34,547	31,208	29,567	27,773	26,375
95%	104,392	76,338	62,557	56,769	49,246	44,723	40,860	39,258	36,758	35,220

Table C.19: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using Poisson simulations from modeled Petersfield i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	21,702	16,939	14,216	12,653	11,482	10,817	10,014	9,491	8,985	8,574
75%	37,382	29,185	25,021	22,587	20,626	19,248	18,156	17,135	16,262	15,478
90%	60,141	48,310	41,622	37,812	34,583	32,272	31,049	28,178	27,139	26,012
95%	94,170	70,574	62,184	56,740	52,485	47,534	46,196	42,076	41,240	39,868

Table C.20: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using half the negative binomial simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	76,709	59,954	50,317	455,04	41,120	38,533	35,543	34,631	32,225	30,622
75%	126,048	99,355	84,631	77,233	69,703	65,961	61,398	59,386	55,423	52,640
90%	187,772	147,291	126,017	115,731	103,594	994,49	926,42	895,49	841,21	790,26
95%	267,593	197,419	165,036	151,340	134,550	128,817	116,543	115,976	105,772	99,434

Table C.21: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using half the Poisson simulations from modeled Southampton i-Tree Eco data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	64,265	48,935	40,584	36,597	32,289	30,317	28,827	26,690	25,372	24,387
75%	107,237	82,740	70,147	63,255	57,179	53,112	50,805	47,768	44,826	42,965
90%	157,822	123,248	104,742	94,187	88,006	81,174	76,912	73,357	69,379	66,195
95%	211,392	160,968	136,493	122,872	113,343	105,208	97,105	93,293	88,347	85,436

Table C.22: Summary of the absolute error by percentile against the number of survey plots. Relative errors calculated using simulations from modeled Cambridge ProximiTREE data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	35,130	25,110	20,427	17,426	15,990	14,309	13,283	12,548	11,833	11,206
75%	60,985	43,008	35,366	29,930	26,778	24,543	22,445	21,172	20,331	18,949
90%	86,286	61,166	50,553	42,904	38,765	35,171	32,416	30,536	29,060	27,192
95%	103,508	72,574	60,644	51,827	46,626	42,327	39,094	36,494	34,416	32,579

Table C.23: Summary of the absolute error by percentiles against the number of survey plots. Absolute errors calculated using simulations from modeled Petersfield National Tree Map (NTM) data

Percentile	**Number of survey plots**									
	50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
50%	3,533	2,497	2,018	1,736	1,550	1,431	1,297	1,243	1,143	1,079
75%	6,019	4,266	3,440	2,950	2,645	2,427	2,235	2,089	1,948	1,872
90%	8,507	6,056	4,915	4,184	3,746	3,411	3,183	2,938	2,808	2,658
95%	10,173	7,153	5,852	4,939	4,478	4,021	3,797	3,491	3,316	3,190

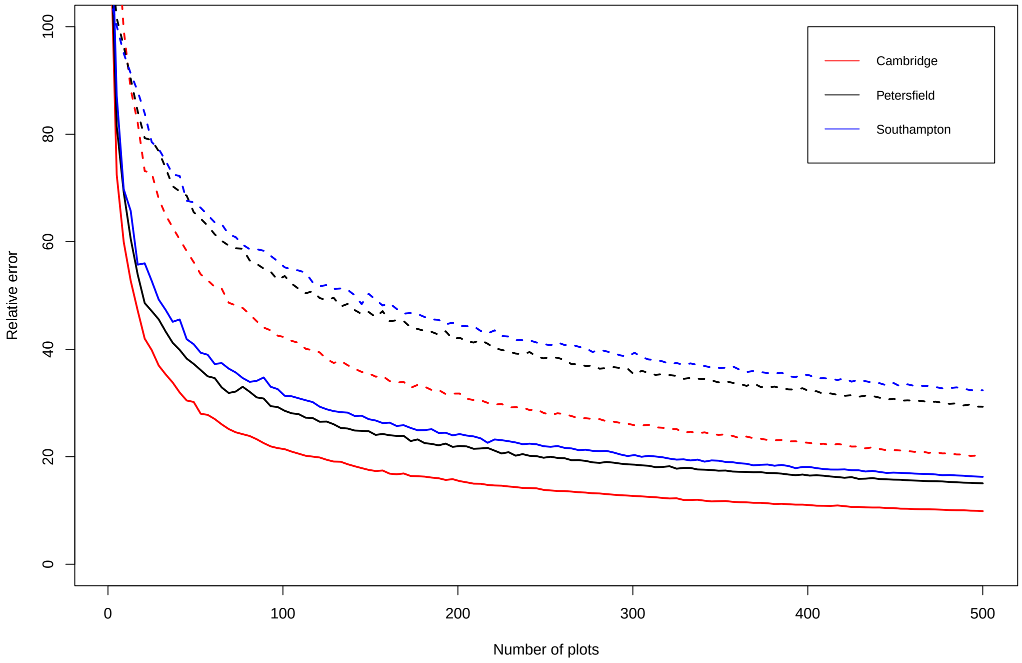


Figure C.1: Plot of the relative error against the number of survey plots. Relative errors calculated using negative binomial and Bernoulli simulations from modeled i-Tree Eco data. Solid lines represent mean values, whilst dashed lines represent 90$^{\text{th}}$ percentiles

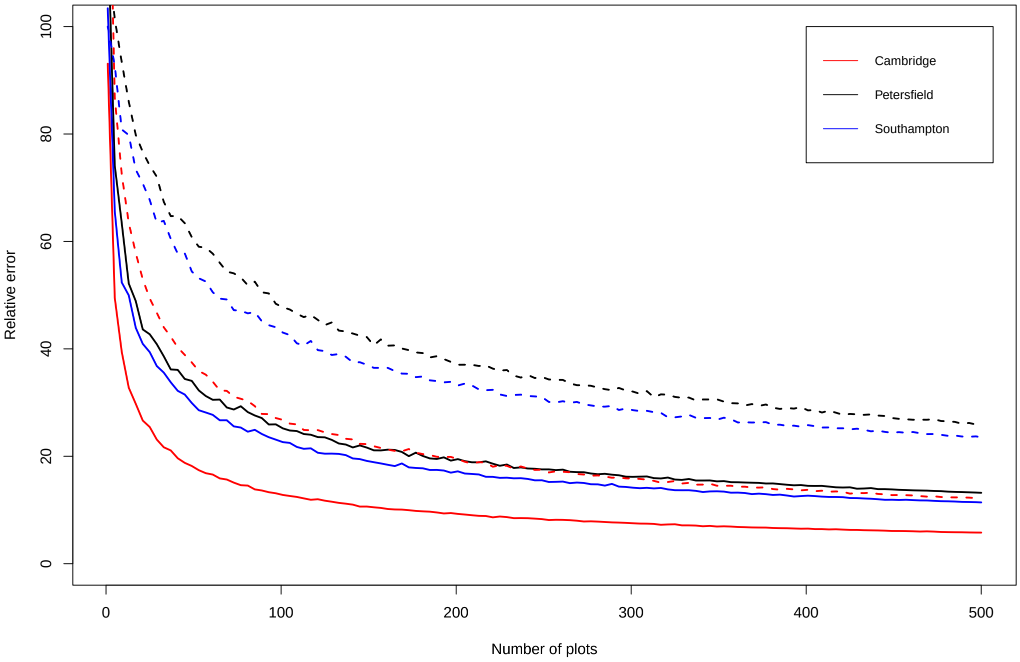


FIGURE C.2: Plot of the relative error against the number of survey plots. Relative errors calculated using Poisson and Bernoulli simulations from modeled i-Tree Eco data. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

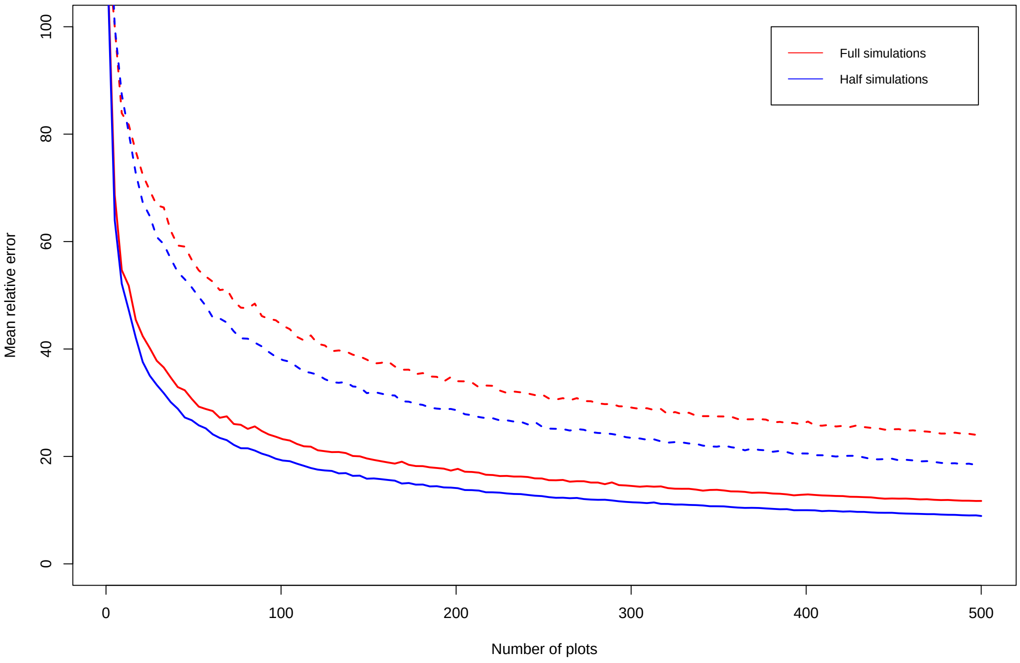


FIGURE C.3: Plot of the relative error against the number of survey plots, for simulations produced from all and half of the MCMC samples. Relative errors calculated using Poisson simulations from modeled i-Tree Eco data in Southampton. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

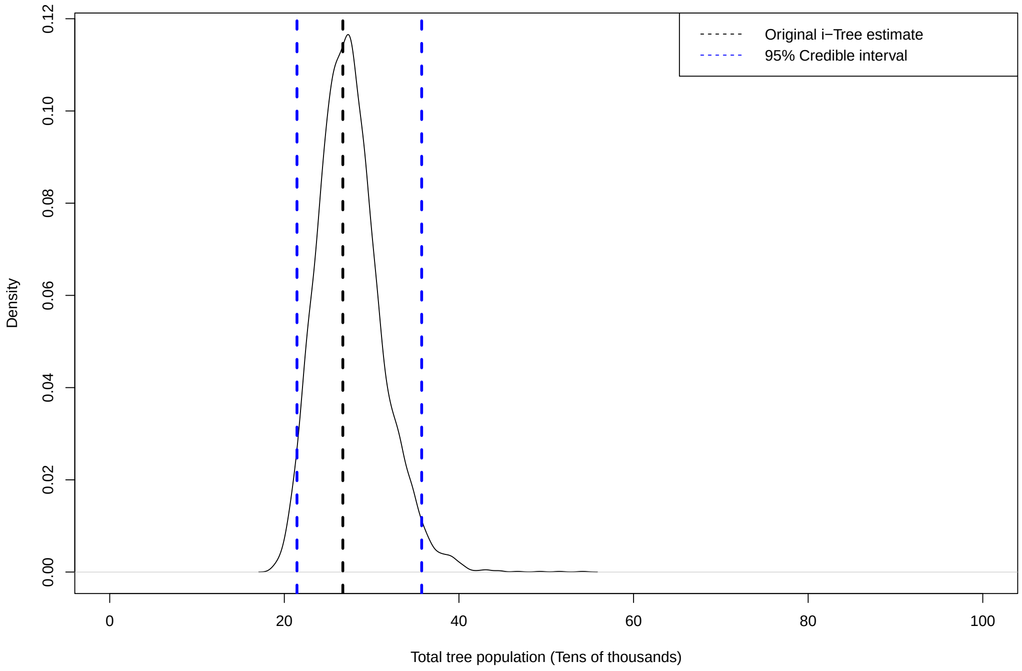


FIGURE C.4: Plot of tree populations simulated using a negative binomial distribution from half of the MCMC samples modeled using i-Tree Eco data in Southampton

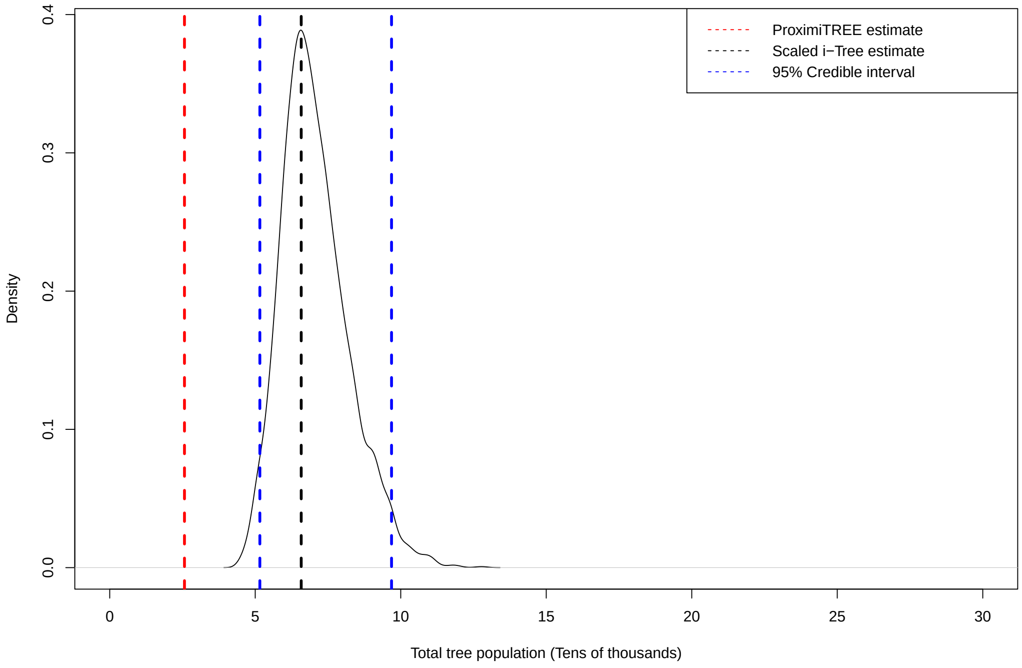


FIGURE C.5: Plot of tree populations simulated using a negative binomial distribution from half of the MCMC samples modeled using i-Tree Eco data in Petersfield

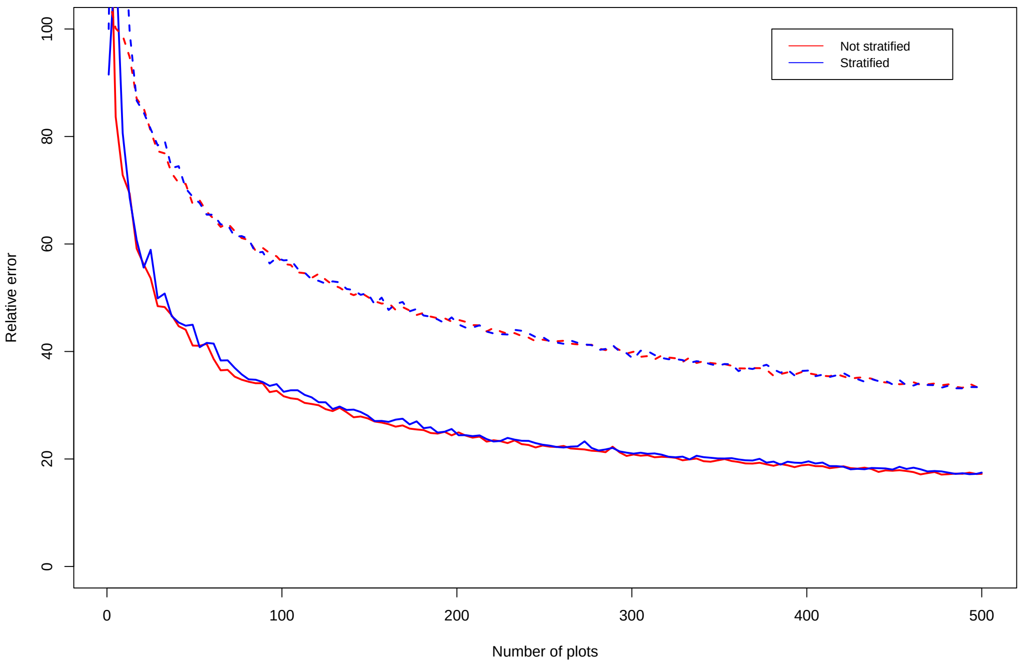


FIGURE C.6: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated using negative binomial simulations from modeled i-Tree Eco data in Southampton. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

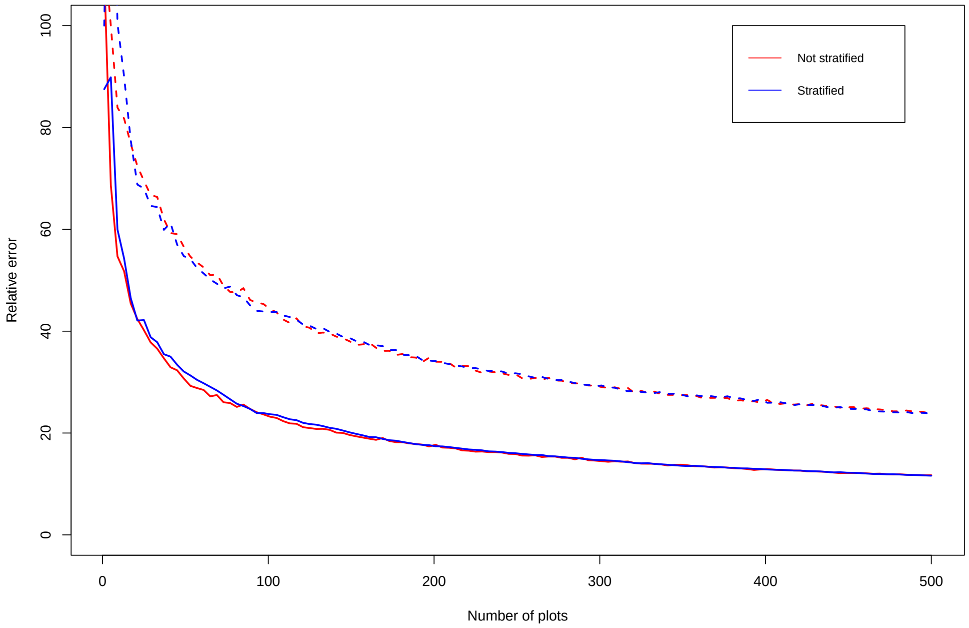


FIGURE C.7: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated using Poisson simulations from modeled i-Tree Eco data in Southampton. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

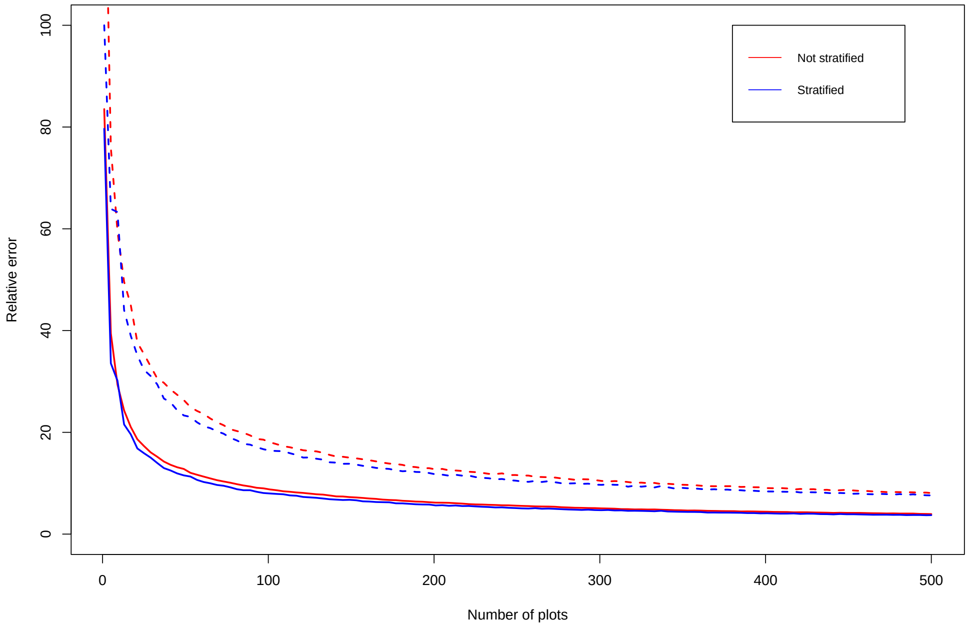


Figure C.8: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated from modeled ProximiTREE data in Cambridge. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

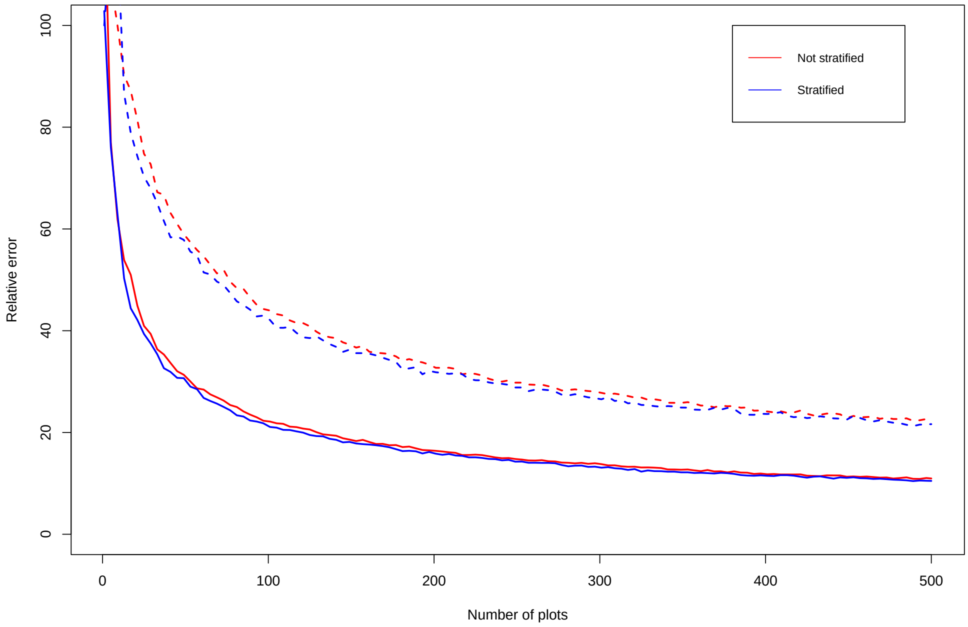


Figure C.9: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated using negative binomial simulations from modeled i-Tree Eco data in Cambridge. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

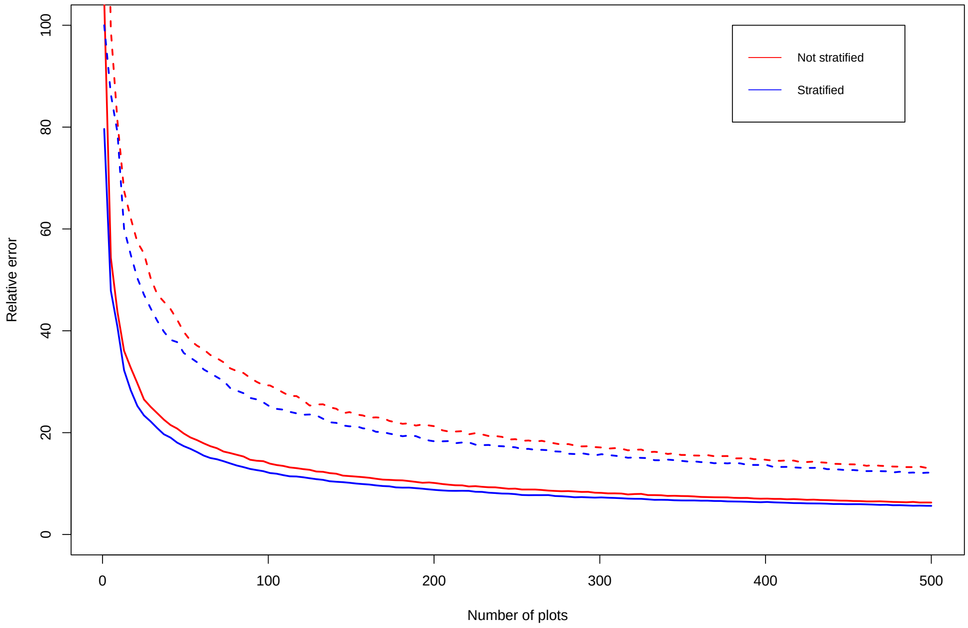


Figure C.10: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated using Poisson simulations from modeled i-Tree Eco data in Cambridge. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

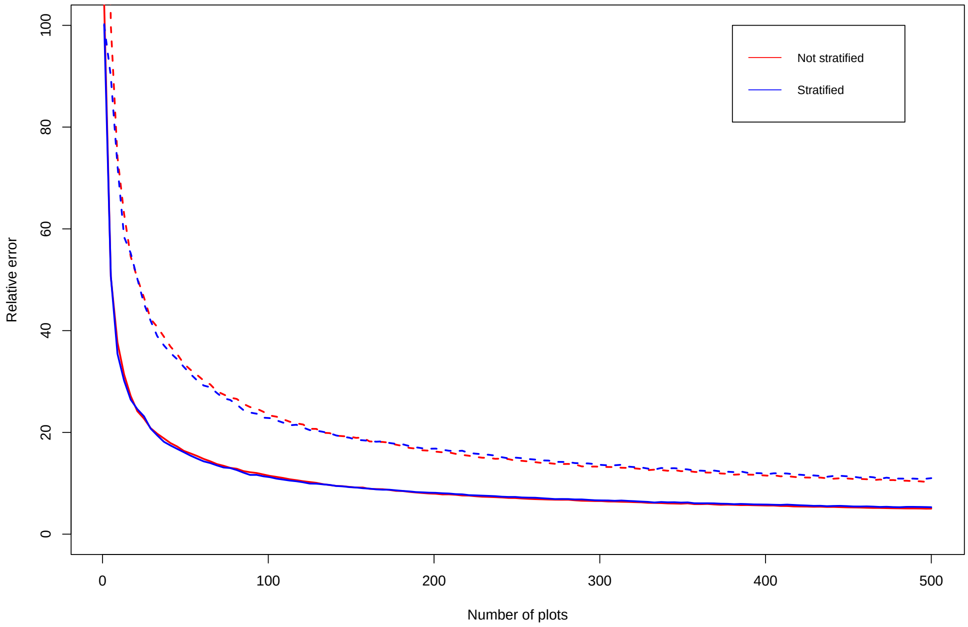


Figure C.11: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated from modeled National Tree Map(NTM) data in Petersfield. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

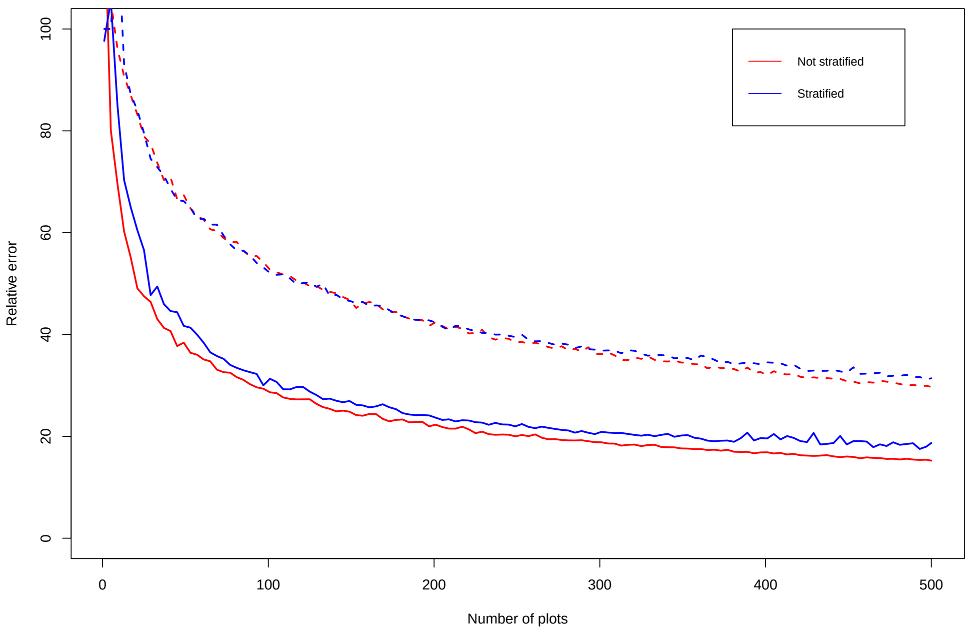


Figure C.12: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated using negative binomial simulations from modeled i-Tree Eco data in Petersfield. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

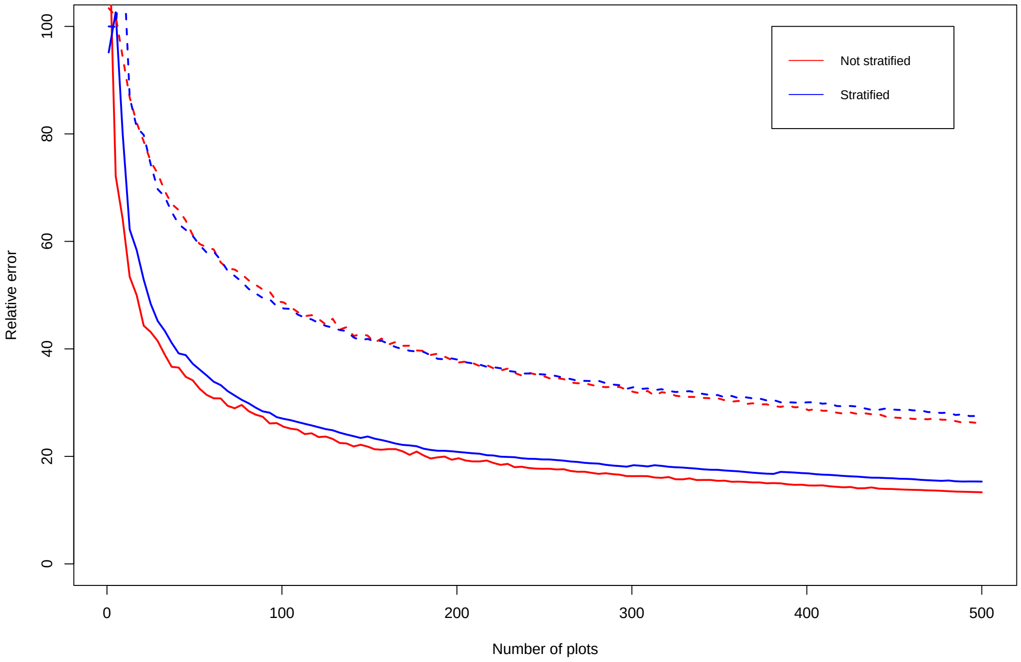


Figure C.13: Plot of the relative error against the number of survey plots, for random survey designs and survey designs stratified by natural and building coverage. Relative errors calculated using Poisson simulations from modeled i-Tree Eco data in Petersfield. Solid lines represent mean values, whilst dashed lines represent 90th percentiles

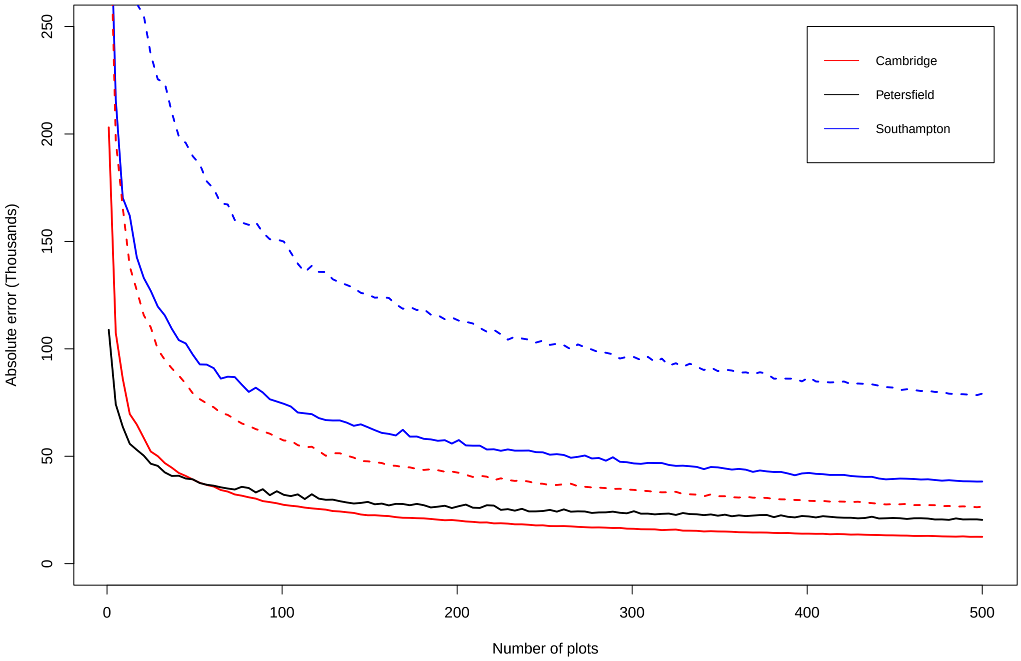


FIGURE C.14: Plot of the absolute error against the number of survey plots. Absolute errors calculated using Poisson simulations from modeled i-Tree Eco data. Solid lines represent mean values, whilst dashed lines represent 90$^{\text{th}}$ percentiles

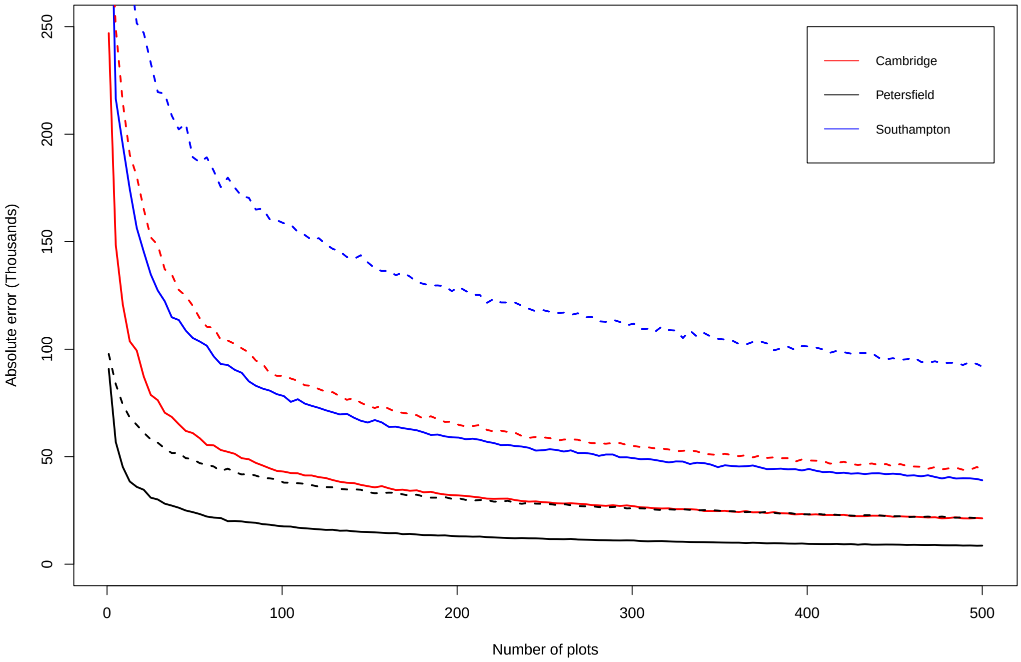


FIGURE C.15: Summary of the absolute error by percentiles against the number of survey. Absolute errors calculated using negative binomial simulations from modeled Cambridge i-Tree Eco data

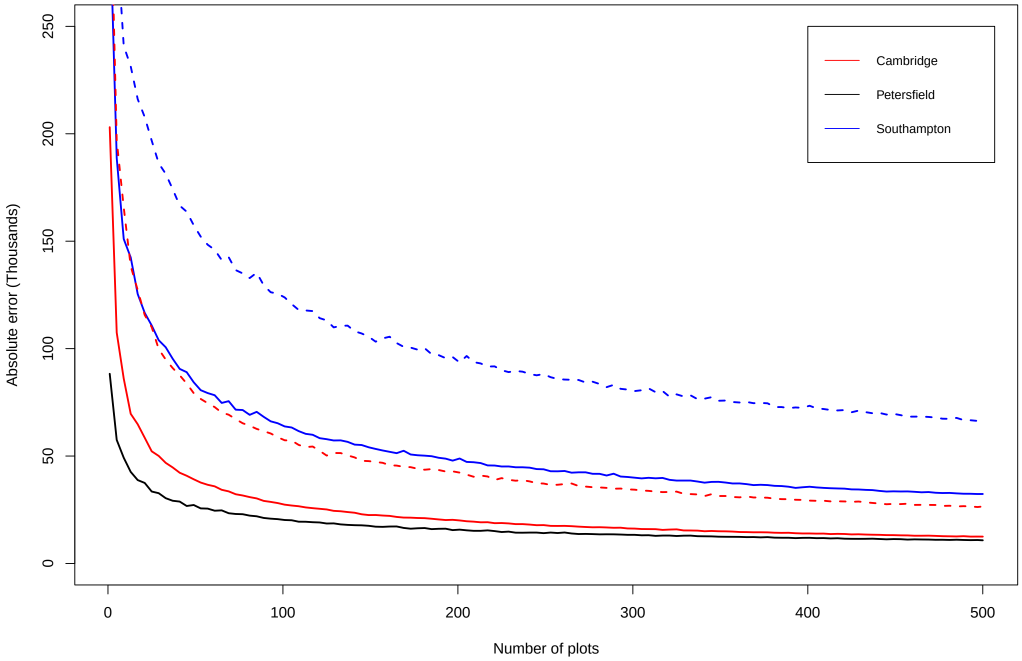


Figure C.16: Summary of the absolute error by percentiles against the number of survey plots. Absolute errors calculated using Poisson simulations from half the model samples for modeled Southampton i-Tree Eco data

References

Agarwal, D., Gelfand, A. and Citron-Pousty, S. (2002). Zero-Inflated Models with Application to Spatial Count Data, *Environmental and Ecological Statistics* **9**: 341–355.

Bacaro, G., Rocchini, D., Diekmann, M., Gasparini, P., Gioria, M., Maccherini, S., Marcantonio, M., Tordoni, E., Amici, V., Landi, S., Torri, D., Castello, M., Altobelli, A. and Chiarucci, A. (2015). Shape matters in sampling plant diversity: Evidence from the field, *Ecological Complexity* **24**: 37–45.

Baines, O., Wilkes, P. and Disney, M. (2020). Quantifying urban forest structure with open-access remote sensing data sets, *Urban Forestry & Urban Greening* **50**. Article 126653.

Banerjee, S., Carlin, B. and Gelfand, A. (2014). *Hierarchical Modeling and Analysis of Spatial Data*, 2 edn, Chapman & Hall/CRC Monographs on Statistical and Applied Probability, New York.

BCC (2021). Birmingham City Council Air Quality Action Plan. Birmingham City Council; URL: https://www.birmingham.gov.uk/downloads/file/19120/birmingham_city_council_air_quality_action_plan_2021-2026, Online, Accessed: 20-06-2024.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2): 192–236.

Besag, J. (1977). Efficiency of pseudolikelihood estimation for simple Gaussian fields, *Biometrika* **64**(3): 616–618.

Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics, *Annals of the Institute of Statistical Mathematics* **43**(1): 1–20.

Bivand, R. and Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association, *TEST* **27**(3): 716–748. URL: https://cran.r-project.org/web/packages/spdep/index.html.

BlueSky (2020a). National Tree Map. Data supplied by BlueSky; URL: https://www.bluesky-world.com/ntm.

BlueSky (2020b). ProximiTREE. Data supplied by Cambridge City Council and BlueSky; URL: https://www.bluesky-world.com.

Borck, R. and Schrauth, P. (2021). Population density and urban air quality, *Regional Science and Urban Economics* **86**. Article 103596.

Briggs, H. (2023). Everyone to live 15 minutes from green space or water in England under plans — BBC News. https://www.bbc.co.uk/news/science-environment-64456455 Online, Accessed 31-01-2023.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017). Stan: A probabilistic programming language, *Journal of statistical software* **76**(1): 1–32.

Carroll, R., Lawson, A. B., Faes, C., Kirby, R. S., Aregay, M. and Watjou, K. (2015). Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping, *Spatial and Spatio-temporal Epidemiology* **14-15**: 45–54.

Chave, J. (2013). The problem of pattern and scale in ecology: What have we learned in 20 years?, *Ecology letters* **16**(s1): 4–16.

Chiarucci, A. (2007). To sample or not to sample? That is the question... For the vegetation scientist, *Folia Geobotanica* **42**(2): 209–216.

Cressie, N. A. (1994). *Statistics for Spatial Data, Revised Edition.*, Wiley, New York.

Davies, H., Doick, K., Handley, P., O'Brien, L. and Wilson, J. (2017a). Delivery of ecosystem services by urban forests. Research Report, Forest Research, Forestry Commission: Edinburgh, URL: https://www.forestresearch.gov.uk/publications/delivery-of-ecosystem-services-by-urban-forests/.

Davies, H., Doick, K., Hudson, M. and Schreckenberg, K. (2017b). Challenges for tree officers to enhance the provision of regulating ecosystem services from urban forests, *Environmental Research* **156**: 97–107.

Dean, C. B., Ugarte, M. D. and Militino, A. F. (2001). Detecting Interaction Between Random Region and Fixed Age Effects in Disease Mapping, *Biometrics* **57**(1): 197–202.

Defra (2014). Official statistics: 2011 rural-urban classification of local authorities and other geographies. URL: https://www.gov.uk/government/statistics/2011-rural-urban-classification-of-local-authority-and-other-higher-level-geographies-for-statistical-purposes Online, Accessed 08-03-2022.

Defra (2020). Air Quality Management Areas. Data retrieved from the Department for Environment, Food and Rural affairs; URL: https://uk-air.defra.gov.uk/aqma/maps/,Accessedonvariousdates.

Dengler, J., Löbel, S. and Dolnik, C. (2009). Species constancy depends on plot size - A problem for vegetation classification and how it can be solved, *Journal of Vegetation Science* **20**(4): 754 – 766.

Di Zio, S., Fontanella, L. and Ippoliti, L. (2004). Optimal spatial sampling schemes for environmental surveys, *Environmental and Ecological Statistics* **11**: 397–414.

Doick, K., Davies, H., Handley, P., Vaz Monteiro, M., O'Brien, L. and Ashwood, F. (2016). Introducing England's urban forests: Definition, distribution, composition and benefits. UFWACN (Urban Forestry and Woodlands Advisory Committees (FWAC) Network).

Du, H., Hu, F., Zeng, F., Wang, K.-L., Peng, W., Zhang, H., Zeng, Z., Zhang, F. and Song, T. (2017). Spatial distribution of tree species in evergreen-deciduous broadleaf karst forests in southwest China, *Scientific Reports* **7**. Article 15664.

Fassnacht, F. E., White, J. C., Wulder, M. A. and Næsset, E. (2023). Remote sensing in forestry: current challenges, considerations and directions, *Forestry: An International Journal of Forest Research* **97**(1): 11–37.

Geary, R. C. (1954). The Contiguity Ratio and Statistical Mapping, *The Incorporated Statistician* **5**(3): 115–146.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper), *Bayesian Analysis* **1**(3): 515 – 534.

Gelman, A., Hwang, J. and Vehtari, A. (2013). Understanding predictive information criteria for Bayesian models, *Statistics and Computing* **24**: 997–1016.

Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Staff Report, Federal Reserve Bank of Minneapolis.

Gimond, M. (2023). Chapter 14. Spatial Interpolation in Intro to GIS and Spatial Analysis — GitHub. URL: https://mgimond.github.io/Spatial/spatial-autocorrelation.html Online, Accessed 03-10-2023.

GISGeography (2024). What is NDVI (Normalized Difference Vegetation Index)? URL: https://gisgeography.com/ndvi-normalized-difference-vegetation-index/ Online, Accessed 21-06-2024.

Güler, B., Jentsch, A., Apostolova, I., Bartha, S., Bloor, J., Campetella, G., Canullo, R., Házi, J., Kreyling, J., Pottier, J., Szabó, G., Terziyska, T., Ugurlu, E., Wellstein, C., Zimmermann, Z. and Dengler, J. (2016). How plot shape and spatial arrangement affect plant species richness counts: implications for sampling design and rarefaction analyses, *Journal of Vegetation Science* **27**(4): 692–703.

Hall, C., O'Brien, L., Hand, K. and Raum, S. (2018). Evaluation of i-Tree Eco surveys in Great Britain. Impacts and key lessons: The views of stakeholders. Farnham: Forest Research. Available from: https://www.researchgate.net/publication/324224060_Evaluation_of_iTree_Eco_surveys_in_Great_Britain_Impacts_and_key_lessons_The_views_of_stakeholders.

Henrys, P. A. and Jarvis, S. G. (2019). Intergration of ground survey and remote sensing derived data: Producing robust indicators of habitat extent and condition, *Ecology and Evolution* **9**(14): 8104 – 8112.

Hoffman, M. D. and Gelman, A. (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, *Journal of Machine Learning Research* **15**: 1593–1623.

Hoffmann, S., Steiner, L., Schweiger, A., Chiarucci, A. and Beierkuhnlein, C. (2019). Optimizing sampling effort and information content of biodiversity surveys: a case study of alpine grassland, *Ecological Informatics* **51**: 112–120.

i-Tree (2021a). i-Tree Eco field guide. User manual, URL: https://www.itreetools.org/support/resources-overview/i-tree-manuals-workbooks.

i-Tree (2021b). i-tree eco user guide. User manual, URL:https://www.itreetools.org/support/resources-overview/i-tree-manuals-workbooks.

i-Tree Eco (2021). i-Tree Eco v6 overview. URL: https://www.itreetools.org/tools/i-tree-eco Online, Accessed: 20-10-2021.

Jin, J. and Yang, J. (2020). Effects of sampling approaches on quantifying urban forest structure, *Landscape and Urban Planning* **195**. Article 103722.

Keeley, J. E. and Fotheringham, C. J. (2005). Plot Shape Effects on Plant Species Diversity Measurements, *Journal of Vegetation Science* **16**(2): 249–256.

Kenkel, N. C., Juhász-Nagy, P. and Podani, J. (1989). On Sampling Procedures in Population and Community Ecology, *Vegetatio* **83**: 195–207.

Khaemba, W. M. (2001). Spatial point pattern analysis of aerial survey data to assess clustering in wildlife distributions, *International Journal of Applied Earth Observation and Geoinformation* **3**(2): 139–145.

Kiskowski, M., Hancock, J. and Kenworthy, A. (2009). On the Use of Ripley's K-Function and Its Derivatives to Analyze Domain Size, *Biophysical journal* **97**(4): 1095–1103.

Konijnendijk, C. (2022). Evidence-based guidelines for greener, healthier, more resilient neighbourhoods: Introducing the 3-30-300 rule, *Journal of Forestry Research* **34**: 821–830.

Lauderdale, B. and Clark, T. (2016). Estimating Vote-Specific Preferences from Roll-Call Data Using Conditional Autoregressive Priors, *The Journal of Politics* **78**(4): 1153–1169.

Lavine, M. and Hodges, J. (2012). On rigorous specification of ICAR models, *American Statistician* **66**(1): 42–49.

Law, R., Illian, J., Burslem, D., Gratzer, G., Gunatilleke, C. and Gunatilleke, N. (2009). Ecological information from satial patterns of plants: Insights from point process theory, *Journal of Ecology* **97**(4): 616 – 628.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping, *Spatial and Spatio-temporal Epidemiology* **2**(2): 79–89.

Lee, D. (2013). CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors, *Journal of Statistical Software* **55**(13): 1–24.

Leitão, P., Moreira, F. and Osborne, P. (2011). Effects of geographical data sampling bias on habitat models of species distributions: A case study with steppe birds in southern Portugal, *International Journal of Geographical Information Science* **25**(3): 439–454.

Levin, S. A. (1992). The problem of pattern and scale in ecology, *Ecology* **73**(6): 1943–1967.

Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009). The BUGS project: Evolution, critique and future directions, *Statistics in Medicine* **28**(25): 3049–3067.

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modeling framework: Concepts, structure and extensibility, *Statistics and Computing* **10**: 325–337.

MacNab, Y. C. (2011). On Gaussian Markov random fields and Bayesian disease mapping, *Statistical Methods in Medical Research* **20**(1): 49–68.

Mell, I. C., Henneberry, J., Hehl-Lange, S. and Keskin, B. (2013). Promoting urban greening: Valuing the development of green infrastructure investments in the urban core of Manchester, UK, *Urban Forestry & Urban Greening* **12**(3): 296 – 306.

MHCLG (2019). Indicies of Multiple Deprivation. Data retrieved from the Ministry of Housing, Communities and Local Government; URL: `http://data-communities.opendata.arcgis.com/datasets/indices-of-multiple-deprivation-imd-2019-1`, Accessed on various dates.

Millennium Ecosystem Assessment (2005). *Ecosystems and Human Well-Being: Biodiversity Synthesis*, Synthesis Island press, Washington DC.

Moffat, A. J. and Doick, K. J. (2019). The Petersfield i-Tree Eco survey – an exercise in community ownership, *Arboricultural Journal* **41**(3): 153–171.

Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*, 1 edn, Chapman & Hall/CRC Biostatistics Series, New York.

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena, *Biometrika* **37**(1/2): 17–23.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A. and DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in Stan, *Spatial and spatio-temporal epidemiology* **31**. Article 100301.

Mutch, E., Davies, H., Hudson, M., Parks, K., Schreckenberg, K., Doick, K., Handley, P., Rogers, K., Kiss, S. and McCulloch, L. (2017). Understanding the value of Southampton's urban trees. Results of the 2016 i-Tree Eco survey. Technical Report. University of Southampton, Forest Research, Treeconomics and Southampton City Council, Southampton.

NASA (2016). Normalized Difference Vegetation Index (NDVI). Data retrieved from the Application for Extracting and Exploring Analysis Ready Samples (AppEEARS); URL: https://appeears.earthdatacloud.nasa.gov/, Accessed on 01-03-2023.

Neal, R. (2012). *Handbook of Markov Chain Monte Carlo*, 1 edn, Chapman & Hall, New York, chapter MCMC using Hamiltonian dynamics.

Noble, S., McLennan, D., Noble, M., Plunkett, E., Gutacker, N., Silk, M. and Gemma, W. (2019). The English Indices of Deprivation 2019 research report. Technical Report, Ministry of Housing, Communities and Local Government.

Nowak, D., Crane, D., Stevens, J., Hoehn, R., Walton, J. and Bond, J. (2008a). A ground-based method of assessing urban forest structure and ecosystem services, *Arboriculture & Urban Forestry* **34**(6): 347–358.

Nowak, D. J., Crane, D. E. and Stevens, J. C. (2006). Air pollution removal by urban trees and shrubs in the United States, *Urban Forestry & Urban Greening* **4**(3): 115–123.

Nowak, D., Walton, J., Stevens, J., Crane, D. and Hoehn, R. (2008b). Effect of Plot and Sample Size on Timing and Precision of Urban Forest Assessments, *Arboriculture and Urban Forestry* **34**(6): 386–390.

Obaromi, D. (2019). Spatial modelling of some Conditional Autoregressive priors in a disease mapping model: the Bayesian approach, *Biomedical journal of scientific and technical research* **14**(3): 10680–10686.

OED (2023). Tree, n.. — Oxford University Press. Oxford English Dictionary, URL: https://www.oed.com/dictionary/tree_n Online, Accessed 10-06-2024.

Okabe, A., Boots, B., Sugihara, K. and Chiu, S. N. (2009). *Spatial tessellations: concepts and applications of Voronoi diagrams*, John Wiley & Sons, Hoboken, NJ, United States.

ONS (2020). Local Authority Districts data. Data retrieved from the Office for National Statistics; URL: https://geoportal.statistics.gov.uk/, Accessed on various dates.

OS (2017). OS MasterMap Topography Layer Product Guide. Version 2, User manual, Ordinance Survey.

OS (2019). OS MasterMap® Topography Layer[GeoPackage geospatial data], Scale 1:1250, Tiles: GB. Data retrieved from the Ordnance Survey (GB), using the EDINA Digimap Ordnance Survey Service ; URL: https://digimap.edina.ac.uk, Accessed on various dates.

Ouchi, T. and Uekawa, T. (1986). Statistical analysis of the spatial distribution of earthquakes—variation of the spatial distribution of earthquakes before and after large earthquakes, *Physics of the Earth and Planetary Interiors* **44**(3): 211–225.

Palmer, M. W. and White, P. S. (1994). On the Existence of Ecological Communities, *Journal of Vegetation Science* **5**(2): 279–282.

Paradis, E. and Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R, *Bioinformatics* **35**: 526–528. URL: https://cran.r-project.org/web/packages/ape/index.html.

Paul, T. S. H., Kimberley, M. O. and Beets, P. (2019). Thinking outside the square: Evidence that plot shape and layout in forest inventories can bias estimates of stand metrics, *Methods in Ecology and Evolution* **10**(3): 381–388.

Potts, J. M. and Elith, J. (2006). Comparing species abundance models, *Ecological Modelling* **199**(2): 153–163.

Rattalino Edreira, J. I., Mourtzinis, S., Azzari, G., Andrade, J. F., Conley, S. P., Specht, J. E. and Grassini, P. (2020). Combining field-level data and remote sensing to understand impact of management practices on producer yields, *Field Crops Research* **257**. Article 107932.

Riebler, A., Sørbye, S. H., Simpson, D. and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling, *Statistical Methods in Medical Research* **25**(4): 1145–1165.

Ripley, B. D. (1977). Modelling spatial patterns, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(2): 172–212.

Rowland, C., Morton, R., Carrasco, L., McShane, G., A.W., O. and C.M., W. (2017). Land Cover Map 2015 (vector, GB). Data retrieved from the NERC Environmental Information Data Centre, Using: EDINA Digimap Service; URL: https://doi.org/10.5285/6c6c9203-7333-4d96-88ab-78925e7a4e73, Accessed on various dates.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2): 319–392.

Sahu, S. (2022). *Bayesian Modeling of Spatio-Temporal Data with R*, 1 edn, Chapman & Hall/CRC, New York. URL: https://cran.r-project.org/web/packages/bmstdr/index.html.

Sahu, S., Gelfand, A. and Holland, D. (2007). High-resolution space-time ozone modeling for assessing trends, *Journal of the American Statistical Association* **102**(480): 1221–1234.

Sales, K., Walker, H., Sparrow, K., Handley, P., Vaz Monteiro, M., Hand, K. L., Buckland, A., Chamber-Ostler, A. and Doick, K. J. (2023). The canopy cover Webmap of the United Kingdom's towns and cities, *Arboricultural Journal* **45**(4): 258–289.

Schweiger, A., Irl, S., Steinbauer, M., Dengler, J. and Beierkuhnlein, C. (2016). Optimizing sampling approaches along ecological gradients, *Methods in Ecology and Evolution* **7**(4): 463–471.

Smedt, T., Simons, K., Van Nieuwenhuyse, A. and Molenberghs, G. (2015). Comparing MCMC and INLA for disease mapping with Bayesian hierarchical models, *Archives of Public Health* **73**(Suppl 1). Article number: O2.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling, *Spatial Statistics* **8**: 39–51.

Southampton City Council (2020). Greener City Plan 2030. URL: https://www.southampton.gov.uk/our-green-city/council-commitments/plan-2030/ Online, Accessed 14-08-2022.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2003). *WinBUGS User Manual*, MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK. URL: http://www.mrc-bsu.cam.ac.uk/bugs.

Stan Development Team (2018). *Stan Modeling Language Users Guide and Reference Manual*. version 2.18, section 15.4. URL: https://mc-stan.org.

Steinbauer, M. J., Dolos, K., Reineking, B. and Beierkuhnlein, C. (2012). Current measures for distance decay in similarity of species composition are influenced by study extent and grain size, *Global Ecology and Biogeography* **21**(12): 1203–1212.

Sturtz, S., Ligges, U. and Gelman, A. (2005). R2WinBUGS: A Package for Running WinBUGS from R, *Journal of Statistical Software* **12**(3): 1–16.

Swallow, B., Buckland, S. T., King, R. and Toms, M. P. (2016). Bayesian hierarchical modelling of continuous non-negative longitudinal data with a spike at zero: An application to a study of birds visiting gardens in winter, *Biometrical Journal* **58**(2): 357–371.

Tiwary, A., Sinnett, D., Peachey, C., Chalabi, Z., Vardoulakis, S., Fletcher, T., Leonardi, G., Grundy, C., Azapagic, A. and Hutchings, T. R. (2009). An integrated tool to assess the role of new planting in PM10 capture and the human health benefits: A case study in London, *Environmental Pollution* **157**(10): 2645–2653.

Uhl, J. H., Leyk, S., Li, Z., Duan, W., Shbita, B., Chiang, Y.-Y. and Knoblock, C. A. (2021). Combining Remote-Sensing-Derived Data and Historical Maps for Long-Term Back-Casting of Urban Extents, *Remote Sensing* **13**(18). Article 3672.

UKCEH (2017). Land cover map dataset documentation. Version 1.2, User manual, UK Centre for Ecology and Hydrology.

UN (2018). World Urbanization Prospects: The 2018 Revision — Department of Economic and Social Affairs, Population Division. United Nations, URL: `https://population.un.org/wup/Country-Profiles/` Online, Accessed 11-03-2023.

Vehtari, A., Gelman, A. and Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Statistics and Computing* **27**(5): 1413–1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B. and Bürkner, P.-C. (2021). Rank-Normalization, Folding, and Localization: An Improved $\hat{R}$ for Assessing Convergence of MCMC (with Discussion), *Bayesian Analysis* **16**(2): 667–718.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y. and Gabry, J. (2022). Pareto Smoothed Importance Sampling v8, *arXiv preprint* . arXiv:1507.02646.

Ver Hoef, J., Peterson, E., Hooten, M., Hanks, E. and Fortin, M. J. (2017). Spatial Autoregressive Models for Statistical Inference from Ecological Data, *Ecological Monographs* **88**(1): 36–59.

Vranckx, M., Neyens, T. and Faes, C. (2019). Comparison of different software implementations for spatial disease mapping, *Spatial and Spatio-temporal Epidemiology* **31**. Article 100302.

Wang, X., Dallimer, M., Scott, C. E., Shi, W. and Gao, J. (2021). Tree species richness and diversity predicts the magnitude of urban heat island mitigation effects of greenspaces, *Science of The Total Environment* **770**. Article 145211.

Warhurst, J. R., Parks, K. E., McCulloch, L. and Hudson, M. D. (2014). Front gardens to car parks: Changes in garden permeability and effects on flood regulation, *Science of The Total Environment* **485-486**: 329 – 339.

Yoccoz, N. G., Nichols, J. D. and Boulinier, T. (2001). Monitoring of biological diversity in space and time, *Trends in Ecology & Evolution* **16**(8): 446–453.

Zhou, W., Wang, J. and Cadenasso, M. (2017). Effects of the spatial configuration of trees on urban heat mitigation: A comparative study, *Remote Sensing of Environment* **195**: 1–12.