Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2025:40
Faculty of Forest Sciences

Making better use of sample data: estimation of plant abundance and associated uncertainties

Léna Gozé

Making better use of sample data: estimation of plant abundance and associated uncertainties

Léna Gozé
Faculty of Forest Sciences
Department of Forest Resource Management
Umeå

SWEDISH UNIVERSITY
OF AGRICULTURAL
SCIENCES

DOCTORAL THESIS
Umeå 2025

Acta Universitatis Agriculturae Sueciae
2025:40

Cover: Arctic starflower (*Lysimachia europaea*) (photo by Jenny Svennås-Gillner, SLU, 2015)

ISSN 1652-6880
ISBN (print version) 978-91-8046-475-8
ISBN (electronic version) 978-91-8046-525-0
https://doi.org/10.54612/a.3a4d2oki20
© 2025 Léna Gozé, https://orcid.org/0000-0001-8974-1996
Swedish University of Agricultural Sciences, Department of Forest Resource Management, Umeå, Sweden

Print: SLU Grafisk service, Uppsala 2025

Errata for
Making better use of sample data: estimation of plant abundance and associated uncertainties

by Léna Gozé

ISBN (print version) 978-91-8046-475-8
ISBN (electronic version) 978-91-8046-525-0
Acta Universitatis Agriculturae Sueciae 2025:40
Umeå, 2025

Thesis summary

Page 11
Location: NFAM
Is now: National Forest Attribute Map
Should be: National Forest Attribute Maps

Page 18
Location: lines 30-31
Is now: Another consequence is that it is not possible to apply (…)
Should be: Another consequence is that it is typically not possible to apply (…)

Page 26
Location: line 9
Is now: National Forest Attribute Map
Should be: National Forest Attribute Maps

Page 32
Location: line 21
Is now: while both Matérn and Thomas are used as subcases in Paper I
Should be: while both Matérn and Thomas processes are used as subcases in Paper I

Page 52
Location: line 21

Is now: Baddeley et al. (2016) make the intensity of the parent process (…)
Should be: Baddeley et al. (2016) made the intensity of the parent process (…)

Page 52 Location: line 28
Is now: (Mrkvička 2014)
Should be: Mrkvička (2014)

Page 52 Location: line 33
Is now: Takashina et al. (2018) propose (…)
Should be: Takashina et al. (2018) proposed (…)

Page 53 Location: line 15
Is now: (…) to estimate the model bias for the cases studies in Papers I, II and III (…)
Should be: (…) to estimate the model bias for the cases studied in Papers I, II and III (…)

Paper I

Page 3 Location: line 22
Is now: Thus, with *M* denoting the total number of points in all the *N* cell cells (…)
Should be: Thus, with *M* denoting the total number of points in all the *N* cells (…)

Page 15 Location: line 14
Is now: meters
Should be: metres

Page 17 Location: line 5
Is now: corresponding estimates of variances
Should be: corresponding estimates of variance

Paper II

Page 3 Location: second column, lines 18-19
Is now: $\lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in U$

Should be: $\lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \notin U$

Page 4 Location: Page 4, equation (15)

Is now: $\widehat{\mathrm{Var}}\left(\hat{R}(\widehat{\boldsymbol{\beta}})\right) = \frac{\widehat{\mathrm{Var}}(\hat{\Lambda}(\widehat{\boldsymbol{\beta}}))}{a_U}$

Should be: $\widehat{\mathrm{Var}}\left(\hat{R}(\widehat{\boldsymbol{\beta}})\right) = \frac{\widehat{\mathrm{Var}}(\hat{\Lambda}(\widehat{\boldsymbol{\beta}}))}{a_U^2}$

Page 4 Location: Page 4, equation (11), second row

Is now: $+\sum_{k\_1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\hat{\beta}_k\hat{\beta}_l)\hat{v}_k\hat{v}_l$

Should be: $+\sum_{k\ 1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\hat{\beta}_k,\hat{\beta}_l)\hat{v}_k\hat{v}_l$

Paper III

Page 17 Location: Page 31

Is now: National 148 Inventories of Landscapes

Should be: National Inventories of Landscapes

Paper IV

Page 9 Location: Footnote 1, line 3

Is now: Thus, the reasononing is valid

Should be: Thus, the reasoning is valid

Page 11 Location: lines 25-26

Is now: the 95 percentile

Should be: the 95th percentile

Page 12 Location: lines 5-6

Is now: of 0 percentiles

Should be: of the 0th percentiles

Making better use of sample data: estimation of plant abundance and associated uncertainties

Abstract

Environmental monitoring has become increasingly important in the current context of global ecological change. More directives and reporting guidelines are issued, hence the need for additional methods for exploiting data from environmental monitoring programmes in order to obtain relevant information about the current state of forests and landscapes. National monitoring programmes, such as the Swedish National Forest Inventory and the National Inventory of Landscapes in Sweden, are core infrastructures for describing and analysing state and change in terrestrial ecosystems. These programmes have large, but not fully exploited potential as a basis for basic and applied research. This thesis aims to develop and apply novel tools for analysing presence/absence (P/A) data from environmental monitoring programmes. Although the area of spatial statistics has been extensively studied, the issue of relating P/A data to plant abundance is an underdeveloped field that needs further attention. The primary goal of this thesis is thus to estimate plant abundance both locally and across large regions for various species. Such plant abundance estimators are derived through models for spatial distribution of plants, by using inhomogeneous point process models that are capable of modelling various categories of point patterns across the landscape, taking geographical covariate information into account. The methods are applied to data collected in the field as well as simulated data to assess the performance of the estimators of plant abundance and associated estimators of uncertainty. The results are promising and show the potential of P/A data in environmental analyses. Another objective of this thesis is to provide reliable estimators of uncertainty in different contexts, with a particular study that takes into account several sources of uncertainty when applying model-based inference (Paper IV). That study shows that the variance of a predictor is a fairly good approximation of uncertainty in large-area surveys, whereas other components come into play when the study area is decreased.

Keywords: Presence/absence data, plant density, model-based inference, generalised linear models, forest inventory data, spatial point processes, uncertainty analysis

Effektivare användning av stickprovsdata: skattning av planttäthet och tillhörande osäkerhet

Sammanfattning

Miljöövervakning har fått allt större betydelse i samband med globala miljöförändringar. Ett ökande antal direktiv och rapporteringskrav utfärdas, vilket kräver utveckling av metoder som stärker användandet av data från miljöövervakningsprogram och som ger relevant information om statusen för skogar och landskap. Nationella övervakningsprogram, som Riksskogstaxeringen och Nationell Inventering av Landskapet i Sverige (NILS), fungerar som viktiga infrastrukturer för att beskriva och analysera tillstånd och förändringar som sker i miljön. Dessa program representerar stora men underutnyttjade möjligheter som bas för grundläggande och avancerad forskning. Det primära målet med denna avhandling är att skapa och implementera nya verktyg för att analysera närvaro-/frånvaro-data (N/F-data) som härrör från miljöövervakningsprogram. Även om området för rumslig statistik har utforskats i stor utsträckning, kvarstår utmaningar med att koppla N/F-data till planttäthet, vilket motiverar ytterligare studier. Ett syfte med denna avhandling är därför att skatta planttäthet både lokalt och över större geografiska regioner för olika arter baserat på N/F-data. Planttäthet skattas via modeller för den rumsliga fördelningen av växter, med ickehomogena punktprocessmodeller som kan ta hänsyn till olika typer av punktmönster över landskapet, samt genom att integrera geografisk kovariatinformation i beräkningarna. För att utvärdera skattningar av planttäthet och tillhörande osäkerhet i skattningar tillämpas metoderna på såväl faktiska fältdata som simulerade data. Resultaten är lovande och belyser potentialen hos N/F-data inom miljöanalys. Ett annat mål med avhandlingen är att ta fram tillförlitliga skattningar av osäkerhet i olika sammanhang, med en specifik studie som behandlar olika osäkerhetskällor inom modellbaserad inferens (papper IV). Den studien visar att för t.ex. en predikterad mängd biomassa ger variansen en bra approximation av osäkerhet vid storskaliga undersökningar, medan andra komponenter kan få större betydelse ifall studieområdet är mindre.

Nyckelord: Närvaro- och frånvarodata, planttäthet, modellbaserad inferens, generaliserade linjära modeller, skoglig inventering, punktprocesser, osäkerhetsanalys

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Gozé, L., Ekström, M., Wallerman, J., Dahlgren, J., Jonsson, B. G., Sandring, S., & Ståhl, G. Large-area estimation of plant density using presence/absence data and binary regression, and a correlation test of the binary regression model (manuscript).

II. Gozé, L., Ekström, M., Sandring, S., Jonsson, B. G., Wallerman, J., & Ståhl, G. (2024). Estimation of plant density based on presence/absence data using hybrid inference. Ecological Informatics, 80, 102377. https://doi.org/10.1016/j.ecoinf.2023.102377

III. Ekström, M., Gozé, L., Sandring, S., Jonsson, B. G., Wallerman, J., & Ståhl, G. Estimation of parameters in inhomogeneous Neyman-Scott processes using presence/absence data (submitted).

IV. Ståhl, G., Gozé, L., Papucci, E., Gobakken, T., Saarela, S., Healey, S. P., Yang, Z., Ekström, M., Kellner, J., Hou, Z., Xu, Q., Ørka, H. O., Næsset E. , McRoberts, R. E. A closer look at uncertainties in forest ecosystem surveys using remotely sensed data and model-based inference (submitted).

Paper II is published open access.

The contribution of Léna Gozé to the papers included in this thesis was as follows:

I. Wrote a major part of the main draft, performed the analyses and simulations and contributed to the theoretical developments.

II. Wrote the main draft, performed the analyses and simulations and contributed to the theoretical developments.

III. Performed analyses and simulations, contributed critically to the main draft and contributed to the theoretical developments.

IV. Performed the simulation study and produced the results, contributed critically to the main draft.

List of figures

Abbreviations

ACL	Actual Confidence Level
AGB	Aboveground Biomass
ALS	Airborne Laser Scanning
EU	European Union
GEDI	Global Ecosystem Dynamics Investigation
GLM	Generalised Linear Model
i.i.d.	independent and identically distributed
LGCP	Log-Gaussian Cox Process
MSE	Mean Squared Error
NFAM	National Forest Attribute Map
NFI	National Forest Inventory
NILS	National Inventory of Landscapes in Sweden
NSP	Neyman-Scott Process
p.d.f.	Probability distribution function
P/A	Presence/Absence
PPP	Poisson Point Process
RS	Remote Sensing
SRS	Simple Random Sampling
USA	United States of America

1. Introduction

1.1 Motivation

Vegetation monitoring has long played a central role in the study of ecosystem processes (Elzinga et al. 1998; Bonham 2013). In the context of the current environmental crisis, the study of plant populations is more important than ever as part of ecological assessments, including biodiversity quantification, tracking of threatened or invasive species, and evaluation of restoration effectiveness. In particular, it is of interest to understand the current state of plant populations and communities and how they evolve with time. Plant occurrence and abundance are good indicators of biodiversity status and can be used to evaluate state and change that are relevant for ecosystem function and resilience. In this thesis, emphasis is made on non-tree vegetation in Papers I, I and III, and on vegetation in a broader sense (in the form of biomass) in Paper IV.

Legal directives (e.g., the EU's Habitats Directive (Commission of the European Communities 2003); or the EU's Biodiversity and Forestry strategies (The European Commission 2020, 2021)) were instituted and require the regularity of reports of vegetation characteristics. Hence, one of the primary objectives is to estimate the number of plants (i.e., plant abundance) in forests.

Environmental monitoring programs such as the Swedish National Forest Inventory (NFI, Fridman et al. 2014) and the National Inventory of Landscapes in Sweden (NILS, Ståhl et al. 2011a) perform different types of inventories in order to collect data on the current state of forests and landscapes in the entirety of Sweden. Similar programs exist in other countries (Tomppo et al. 2010). These infrastructures have large, but not fully exploited potential for performing environmental analyses. Therefore, this thesis aims at developing new methods to make better use of the largely untapped data in the environmental monitoring programmes' databases, with an emphasis on presence/absence (P/A) data.

Monitoring data can be gathered in several ways. For instance, using sample plots is a widespread method to register information on plants in the field (Gregoire & Valentine 2007). Sample plots, generally of circular shape (although they can also be of, e.g., quadratic shape), are placed in a region of interest according to some sampling design. Then, registrations and

measurements are made at individual plot level. Sampling designs can be quite simple, like simple random sampling (SRS) or systematic sampling (Thompson 2012), or take more aspects into consideration, like spatially balanced sampling (Grafström et al. 2012; Grafström & Matei 2018). Other methods to make measurements in the field include distance and line transect methods (see Bonham (2013) for an overview).

1.2 How to register plant information?

The basis for vegetation monitoring is well established, but fundamental problems remain. Trees are relatively easy to count and monitor, but this is far from being the case for ground vegetation (Elzinga et al. 1998). How to correctly define a plant individual? The definition differs depending on the kind of plant under study (Bonham 2013). In addition, some plants grow in numbers so high that it is no simple task to count them accurately, not to mention species with clonal growth pattern, or clustering. All these reasons could explain why so many inventories rely on alternatives to counting plant individuals.

There are several alternatives to simple count data. Vegetation cover estimation is relatively common in vegetation studies and inventory programmes (Godínez-Alvarez et al. 2009; Bonham 2013). However, this method is impacted by a phenomenon called observer judgement bias (Gallegos Torell & Glimskär 2009). Surveyors might interpret cover percentage differently, and it is difficult to derive reasonably accurate estimates of plant cover with only the naked eye. This is similar to what could happen with counting plants. Some surveyors could miss individuals, especially if the plants are numerous in a given sample plot. Traditional cover estimates through ocular assessment are prone to significant observer errors, and hence estimates may vary between surveys. This entails a serious risk that results from monitoring are misinterpreted and may result in reported changes that have not occurred in reality. Potentially equally problematic is the fact that the observer-generated variation in cover estimates can become so large that trends are missed due to low power in analyses – in theory, even when there is no bias. As highlighted by Ståhl (2003), these spurious conclusions need to be avoided, stressing the importance of developing improved methods that limit observer bias.

An alternative to vegetation cover estimation is the point intercept method, where a thin device is used to assess what species cover randomly selected points (Rochefort et al. 2013). The proportion of points where the species occurs is used as a measure of plant cover, but since the device cannot be made infinitesimally thin, the methods will usually overestimate cover (Ståhl 2003). Point intercept methods are less prone to judgement bias than ocular assessments, but require large samples and are time-consuming and costly to conduct (Ringvall et al. 2005).

A simpler, cost-effective alternative to all the methods mentioned above is P/A sampling (Elzinga et al. 1998). All that is required with this kind of survey is to verify whether a specific species is present in a given sample plot. In case the species of interest is present, the surveyor enters "1", otherwise they enter "0". No accurate counts or cover estimation are necessary, hence the advantage of the method cost- and time-wise (Ståhl et al. 2017). In addition, the method is less susceptible to surveyor judgement bias compared to the aforementioned types of survey (Ringvall et al. 2005). All the surveyor needs to know is how the species looks like. On the other hand, P/A sampling presents several challenges. For example, P/A data usually do not provide direct information on plant density (defined as the mean number of individuals per unit area, usually square metres or hectares), and plant occurrence frequencies are difficult to interpret due to their dependence on spatial occurrence patterns and plot size (Ståhl et al. 2017). Hence, one of the main motivations for the present thesis is to find new ways to make use of P/A data to obtain information related to plant density, based on model assumptions regarding the spatial distribution of plant individuals. The very same question has been continually studied, recently by, e.g., Fithian et al. (2015), Ståhl et al. (2017, 2020), Gelfand & Shirota (2019) and Ekström et al. (2020). Earlier references are presented in the next subsection.

1.3 Modelling plant locations and estimating plant density: earlier developments

The concept of frequency, which is closely linked to the concept of density, was first used by Raunkiaer in 1909 (English translation in Raunkiaer (1934)). That researcher observed the presence of plant species in a number of sample plots placed in an area of interest. Plant frequency is then defined

as the number of sample plots where the species is present divided by the total number of sample plots.

However, plant frequency estimates are dependent on the spatial distribution of individuals. Poisson point process (PPP) models are often used, explicitly or implicitly, to model locations of plants and other species. PPP models consider the plant locations as randomly distributed and independent of each other. A closely related point process is the binomial point process, where the point locations are considered random but the process will always generate a fixed number of points (Baddeley et al. 2016).

Works that attempt to find suitable ways to estimate plant density from P/A data assuming random populations date back from at least the start of the twentieth century, with the pioneering article by Arrhenius (1921). There, the author considered P/A data under a binomial point process and estimated the number of species in an area of interest. Later, Kylin (1926) derived a formula for the expected proportion of sampling plots in which the species would be absent, assuming individuals are randomly distributed in the study area. Under the same assumption, Blackman (1935) stated that if the percentage of absence is known then the density can be deduced. In the discussion of the same article by Bartlett, an estimate of the probability of absence in a plot was derived, followed by an estimator for the density as well as a corresponding estimate of variance, based on plant occurrence proportions. Bartlett further stated that the most efficient plot size for density estimation corresponds to around 20% absence of the species under study (Bartlett 1948). Aberdeen (1958) developed a formula that links sample plot size, plant size, plant density and frequency under the assumption of a PPP for plant positions and SRS for the sampling design. Later, Greig-Smith (1983) proposed a model that describes the relationship between frequency and density when plant individuals are randomly distributed. Swindel (1983) determined the optimal size and number of plots to estimate density from P/A data when the plant locations are supposed to be at random. Later, Ståhl et al. (2020) estimated plant density from P/A data with an explicit assumption about a homogeneous PPP in the special case where sample plot sizes vary.

However, plant individuals are rarely randomly distributed (Bonham 2013). Additional assumptions might be needed, as some plant patterns exhibit spatial dependence. Clustered patterns began to be studied around the second half of the twentieth century, although the following studies did not

necessarily handle P/A data. Thomas (1949) proposed a method to estimate density for clustered plant populations based on abundance data. The (generalised) Thomas process was named in recognition of Thomas' contribution (Diggle et al. 1976). A little later, Neyman & Scott (1952) studied clustering of galaxies, although the model developed therein has also been applied to model clustered plant populations (e.g., Batista & Maguire 1998; Ogata 2020). Neyman and Scott also gave their names to a type of clustered point process, and the generalised Thomas process is actually a special case of Neyman-Scott process (NSP). A list of all special subcases of Neyman-Scott processes that have been studied around that time is provided by Guttorp & Thorarinsdottir (2012). During the same decade, Pielou (1957) studied the effect of plot size when estimating parameters from a Neyman-Scott and a Thomas process.

Another type of Neyman-Scott cluster process that is sometimes used for the modelling of clustering in plant populations is the Matérn cluster process (Matérn 1960, 1986), that differs from the generalised Thomas process regarding the distribution of plants in the clusters. Matérn focused on applications of point processes in forestry. Applications of Matérn cluster processes in forest studies include Fleischer et al. (2006), Eichhorn (2010) and Ekström et al. (2020). Out of these references, only the latter presented estimators of expected plant density using P/A data.

Other models for plant populations have also been applied for studying associated properties. One of the most popular models for plant abundance is the negative binomial model, especially when plants are known to exhibit clustering. He & Gaston (2000, 2007) proposed a method for estimating plant abundance based on occurrence data based on the assumption that plant abundance follows a negative binomial distribution. Similarly, Hwang & He (2011) showed how to estimate plant abundance based on P/A maps using a Gamma-Poisson model, which is a generalisation of the negative binomial model. However, the method presented in He & Gaston (2000) tends to overestimate species abundance (Conlisk et al. 2007). In addition, the negative binomial model does not appear to be very suitable for studying plant populations (Holt et al. 2002; Gaston et al. 2011), especially since only two known homogeneous point processes produce the negative binomial distribution for plot abundances, and both of them are extreme cases (Daley & Vere-Jones 2008). For an overview of applications of the negative binomial model in ecology, see Stoklosa et al. (2022). Some other studies,

such as Chang & Huang (2024), used techniques such as kernel estimation to estimate plant abundance from P/A data under several population assumptions. Holt et al. (2002) provided an overview of other models to model plant abundance and density from P/A data.

1.4 Use of additional data in the modelling

In for example Ståhl et al. (2020) or Ekström et al. (2020), the intensity of the point process, defined as the expected number of points per unit area (Baddeley et al. 2016), was supposed to be constant at every point of the region of interest (i.e., homogeneous). However, this is an oversimplification of the reality in most cases. Plant density is known to vary depending on environmental factors such as soil moisture, ambient humidity, chemical composition in the soil, tree cover, and many more (Schulze et al. 2019). As a consequence, it would be an advantage to take such factors into account when modelling abundance and deriving estimates of plant density, in order to make the latter more accurate. Thus, it is suitable to introduce explanatory variables, also called covariates, in the point process models (that are thus called inhomogeneous point process models) and the generalised linear models (GLMs, McCullagh 1989) implied by the point processes. Based on available knowledge, no previous study has suggested large-area estimators of plant density from P/A data and inhomogeneous PPPs, including corresponding variance estimators; nor has any study presented estimators of plant density based on P/A data and inhomogeneous NSPs.

There exists several possible sources for obtaining covariates, either from field surveys or from remote sensing (RS). Covariates from field surveys are collected by surveyors when they visit the different sample plots in a specific region of interest. On the one hand, the main advantage of field covariate registrations is that they usually are very thorough since a lot of environmental aspects are taken into account. On the other hand, a major drawback of auxiliary data taken locally at plot level is that they are generally not available in the whole region of interest. Another consequence is that it is not possible to apply the standard model-based framework if only field-based covariate data are used (Ståhl et al. 2016). This drawback can be counteracted by using RS covariate data, also called “wall-to-wall” data, since they are available for any point of the region of interest. Nowadays, such remotely-sensed covariate information for modelling is currently

increasing in amount, resolution and quality (e.g., Kangas et al. 2018; Dubayah et al. 2022). Moreover, some of them are being made available at short intervals of time (Lindgren et al. 2021). Remotely-sensed covariate data such as airborne laser scanning (ALS) are made less frequently but contain a plethora of useful information on vegetation sites that can be used for modelling (Lidberg et al. 2020). The availability and ready-to-use nature of this kind of covariate information has made the modelling of species abundance depending on environmental factors easier compared to yesteryear.

Another factor that could potentially contribute to the improvement of species modelling is the increasing availability of presence data offered by citizen science data, i.e. spontaneous, voluntary species registrations made outside of structured monitoring programmes and research projects. Most of citizen data are in the form of presence-only data, i.e. only the presence of species is registered. Such data can be used in connection with P/A data from planned surveys to create a more incorporating framework (Fithian et al. 2015; Bradter et al. 2018; Gelfand & Shirota 2019; Mäkinen et al. 2024). However, one should keep in mind that such modelling offers substantial challenges, mostly because of preferential bias, where observers tend to focus on the species they know or appreciate and usually make observations at easily accessible sites (Robinson et al. 2018; Johnston et al. 2020, 2023; Cretois et al. 2021).

The maximum entropy method (shortened as MaxEnt, Phillips et al. (2017)), first mentioned by Jaynes (1957) in a more global context, is equivalent to a regression model based on an inhomogeneous PPP (Renner & Warton 2013), except for the intercept. It has become extremely popular in ecology (e.g., Dudík et al. 2005), although it is rather used with presence-only data instead of P/A data and is often applied uncritically (Royle et al. 2012).

For modelling binary response data, GLMs can be used with an appropriate choice of link function (Mehtätalo & Lappi 2020). In ecology, the most commonly chosen link function is the logit link, which leads to the logistic regression model (see, e.g., Wintle et al. 2005; Foody 2008; Pellissier et al. 2013; Sipek et al. 2022). However, Baddeley et al. (2010) warn that the logistic regression model might not be appropriate when tessellating the region of interest if plant locations are considered as a realisation of an inhomogeneous PPP. On the other hand, binary regression models with

complementary log-log link are not affected by this issue (Baddeley et al. 2010), hence their use in Paper I where a cell grid is used to tessellate the area of interest. See Fortin et al. (2008), Lindenmayer et al. (2009), Yee & Dirnböck (2009), Fithian et al. (2015) or Fiorentin et al. (2019) for other examples of ecological studies based on P/A data and a complementary log-log link function.

1.5 Other considerations

In papers I, II and III, model-based inference is used, which implies that the variable under study is considered as random. Hence, the plant density is considered as a random variable. However, to facilitate the derivations, the expected value of the plant density, which is fixed, is estimated instead of the actual plant density being predicted. Not much is lost by making this adjustment, since the relative difference between the actual and expected values for the plant density are small in large-area surveys if the model used is approximately correct (Ståhl et al. 2016).

Instead of focusing on specific plant species and P/A data, one could register continuous variables, for example aboveground biomass (AGB). Biomass is the variable under study in Paper IV. AGB, or its density, can be estimated by field data (Næsset et al. 2016) or coupled with ALS methods, for example via the Global Ecosystem Dynamics Investigation (GEDI, Dubayah et al. 2022).

Whenever an estimation is made, uncertainty comes into play and must be taken into account. Reporting of estimates usually comprises associated measures of uncertainty, for example variance estimates, mean squared error (MSE) values, or confidence intervals. Indeed, errors can come from different sources, such as the mathematical modelling, the map products produced by remote sensing tools, or the measurements done in the field. In this thesis, it is supposed that the covariate and field data (including the P/A registrations) are devoid of errors, although this is a simplification of reality.

A large number of studies utilising model-based inference use solely the variance of the predictor of the target variable as a measure of uncertainty. Some studies (e.g., McRoberts et al. 2018) suggest that the variance estimator alone can be a sufficient estimator to quantify uncertainty in large-area surveys. However, the variance estimator alone does not take into account all the potential sources of error, largely because it does not directly

factor in the fact that the true value of the target variable is a random variable. A more thorough measure to estimate uncertainty is the MSE, which takes into account the variance and model bias of the predictor, the variance of the true value and the covariance between the predicted and true value. Thus, in-depth uncertainty assessments in model-based inference should use the MSE rather than the variance of the predictor (Cassel et al. 1977). Estimating different components in a broadened uncertainty analysis is the main subject of Paper IV.

2. Aims and objectives

The main objective of this thesis is to propose new ways to make better use of sample data in assessing characteristics of plant populations. A particular emphasis is made on P/A data, which are believed to have an underexploited potential. By proposing new methods and models to facilitate the use of this kind of data, it is believed that researchers and monitoring programmes will be able to get more extensive understanding of the data they are working with, and increase the possibilities of interpretation as well as analytic capacities.

More specifically, the use of model-based and hybrid inference in relation to inhomogeneous spatial point processes and P/A data is investigated in order to estimate plant density taking environmental factors into account. Corresponding variance estimators are also derived.

Additionally, a widened uncertainty analysis is performed in one of the articles, where the variance is one of the components of the MSE formula when model-based inference is used and AGB is the target variable. The extent of the error components in the MSE formula is studied in different subcases.

The specific objectives for each of the papers were as follows:

1. To develop a model-based method in combination with P/A data collected in the field, remotely-sensed covariate data and inhomogeneous PPP in order to derive large-area estimates of expected plant density for a selection of plant species, as well as corresponding estimators of variance, and to apply a residual-based test to test whether the derived GLM implied by the inhomogeneous PPP model has independent response variables given the covariates (**Paper I**);
2. To develop a method that makes use of hybrid-based inference together with P/A data and covariate data collected in the field and inhomogeneous PPP in order to obtain large-area estimates of expected plant density for a selected species, as well as corresponding estimators of variance (**Paper II**);
3. To develop a method that makes use of model-based inference jointly with P/A data collected in the field, remotely-sensed covariate data and inhomogeneous NSPs in order to estimate parameters from the process and to get local estimates of

expected plant density for some selected species, as well as corresponding estimators of variance (**Paper III**);

4. To widen the uncertainty analysis in a model-based framework, and thus to investigate the extent of the different components of the MSE formula in a model-based inference framework, using simulated data and biomass as the target variable (**Paper IV**).

3. Material and Methods

3.1 Data

3.1.1 Field data

Data collected from four different locations were used in the papers contained in this thesis, including a field study that was performed especially for Paper III. For the other papers, data were collected for other purposes, such as registrations as part of forest inventories and monitoring.

For Papers I and II, the data were collected in Northern Sweden, more precisely in the Lappland region of Norrbotten County (Fig. 1). In Paper I, the data, collected during the years 2011 to 2013, originated from the permanent plots of the Swedish NFI (Fridman et al. 2014). The P/A data in the study were originally available for 293 plots. Other covariates related to field and terrain properties were also registered as part of the survey. P/A data for *Luzula pilosa* (L.) Willd. (hairy woodrush), and *Lysimachia europaea* (L.) U. Manns & Anderb. (arctic starflower) were used.

In Paper II, the variable of interest was P/A data of *Vaccinium vitis-idaea* L. (lingonberry). Two samples, both from the Swedish NFI, were considered in the study. The first sample, called S_1, consisted of the centres of the small vegetation plots included in permanent plots in the Lappland region of Norrbotten during the years 2008 to 2012. Sample S_1 had a size n_1 equal to 724 plots. Cluster sampling was used to obtain the second sample, called S_2. This sample had a size of $n_2 = 111$ tract centres, which corresponds to 1132 temporary sample plots in total.

In Paper III, the data were collected in the forest connected to the SLU field station at Kulbäcksliden (Västerbotten County, Sweden, Fig. 2) during the month of September 2022. In total, P/A data for 559 sets of concentric circular plots were obtained (Fig. 3). P/A data for the following plant species were recorded: *L. pilosa, Maianthemum bifolium* (L.) F.W. Schmidt (false lily of the valley), and *L. europaea*.

Concerning Paper IV, the data were collected in the field in Liwale District in Southeast Tanzania, originally as part of another study (Næsset et al. 2016). The sample survey in Tanzania was conducted according to a systematic single-stage cluster design. Each of the 11 clusters consisted of

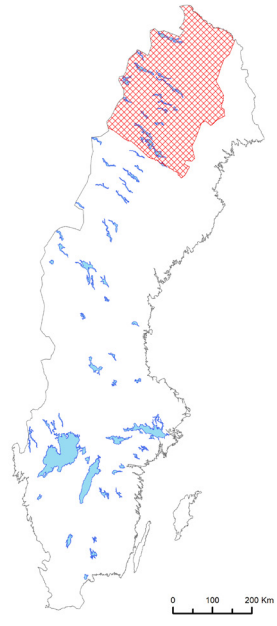


Figure 1. Position of the Lappmark region of Norrbotten County in Sweden.

eight plots, forming an L-shape. The circular plots had a radius of 15 metres, and the data were collected in 2014. AGB was calculated for each tree and then summed at sample plot level. Several tree measurements (for example diameter at breast height) were also conducted on the plots. Data that mimic conditions in Western USA (Saarela et al. 2025) were also used in Paper IV.

3.1.2 Remote sensing data

For papers I, II and III, covariates were obtained from several forest raster map products: the SLU Forest Map (Reese et al. 2003; Wallerman et al. 2021), the National Forest Attribute Map (NFAM, Nilsson et al. 2017), and a soil moisture map produced by Ågren et al. (2021).The SLU Forest Map is made of raster maps of the Swedish forest state, generated from satellite images using the Swedish NFI sample plots as reference data. A similar method was used in the NFAM to create forest raster maps from airborne laser scanning data collected in Sweden between 2009 and 2016. The soil moisture map by Ågren et al. (2021) was derived from terrain indices generated from a national ALS digital elevation model and environmental features.

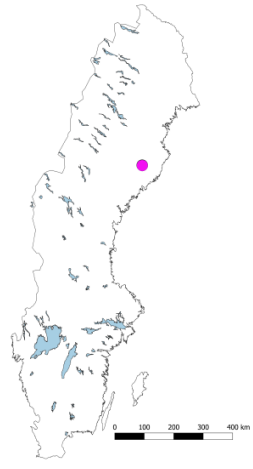


Figure 2. Map of Sweden showing the position of the Kulbäcksliden research park.

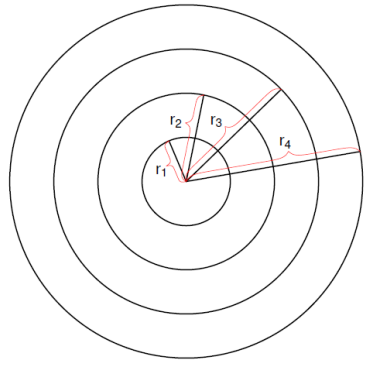


Figure 3. Example of a plot design with four concentric circular sample plots.

For Paper IV, the remotely sensed data came from different sources, the first one being ALS (or GEDI (Dubayah et al. 2022) in the case with simulated USA data), and the other one being satellite image data from the Landsat 8 sensor. The Western USA data were obtained through copula modelling (Ene et al. 2013) based on RS data available from a previous study (Saarela et al. 2025).

3.2 Estimation frameworks

3.2.1 Model-based inference

Model-based inference relies on model assumptions rather than sampling designs. The values that are linked to the elements in the population of interest are realisations of random variables (Ståhl et al. 2016). The realisations come from a so-called superpopulation model that attributes new values to every population unit (in particular values of the response variable) every time it is run. When model-based inference is applied to real data, it is assumed that the real population is a realisation of an invisible, unknown superpopulation model (Cassel et al. 1977). Alternatively, if model-based inference is applied as part of a simulation study, it is relatively easy to make realisations out of a superpopulation model since the latter is assumed to be known. Covariates that originate from, e.g., RS can be used as auxiliary data when creating a model based on a realisation of the superpopulation model. Inference is then based on this model, and estimators of fixed quantities are constructed (alternatively, predictions of random variables are made). In this thesis, model-based inference is used in Papers I, III and IV, as well as partially in Paper II.

Examples of application of model-based inference in forestry include Askne et al. (2013), in which AGB was estimated with models that include RS auxiliary data; Hou et al. (2017), that estimated firewood volume while relying on auxiliary data; Saarela et al. (2018), that developed a hierarchical model-based framework for the estimation of biomass based on ALS and GEDI data; and Mukhopadhyay et al. (2024), that predicted AGB based on GLMs and computed associated prediction intervals. Contrary to the aforementioned studies, that focus on AGB or volume estimation, Papers I, II and III are among the rare ones (including, e.g., Ekström et al. (2020)) that used model-based inference and plant P/A data as a response variable.

3.2.2 Design-based inference

Design-based inference relies on specific sampling designs. Contrary to model-based inference, the randomness in design-based inference comes from the sampling process, while the population is assumed to be fixed. The sample is random for each realisation, while the population parameters (for example the population total or population mean) do not vary. Estimations of population parameters are then made based on these samples, using so-

called design-based estimators. The probability of inclusion of each unit in the sample must be known. In this thesis, design-based inference is used partially in Paper II.

Examples of applications of design-based inference in forestry include Fattorini et al. (2019), Marcelli et al. (2019) and Di Biase et al. (2022), that concentrate on estimating diverse forest attributes. The latter study focused on the estimation of several plant indicators, including plant presence.

3.2.3 Hybrid inference

Hybrid inference is a mixture of both model-based and design-based inference. Two samples are involved in the process. The first sample is used to fit a model, and thus covariate data (that do not need to be wall-to-wall but only available at field plot level) as well as response data are needed. The second sample, which is a sample of covariate data and on which the model fitted on the first sample is applied, is used exclusively to estimate a population parameter, for example expected plant density or biomass per hectare in an entire region. Thus, while covariate data at plot level are required even for the second sample, no response data are needed. Inclusion probabilities for all sampling units in the second sample need to be known, since most of the design-based estimators, for example the Horvitz-Thompson estimator (Horvitz & Thompson 1952), involve inclusion probabilities of some sort.

The term "hybrid" inference was first introduced in an article by Corona et al. (2014) about estimation of standing wood volume in Italy, even if the method itself was already in use before (for example in Ståhl et al. (2011)). Most applications of hybrid inference in forestry concern biomass estimation and prediction. For example, Ståhl et al. (2011) and Gobakken et al. (2012) used a hybrid inference framework to estimate biomass based on ALS sample data in a Norwegian county. A similar study, based in North America, was performed by Margolis et al. (2015). Likewise, Bullock et al. (2023) used GEDI data in combination with NFI data to estimate biomass in Paraguay. Prediction of biomass based on hybrid inference has also been done for larger regions, for example by Saarela et al. (2022), using GEDI data. McRoberts et al. (2019) used that type of inference to compare forest biomass estimates at different resolutions in the USA. Hybrid inference has also been applied to estimate growing stock volume in Finland (Saarela et al. 2015) and Spain (Condés & McRoberts 2017). Hybrid inference can also be

used with mixed-effects models (e.g., Fortin et al. 2016). Note that all the aforementioned studies used hybrid inference through models with a continuous response variable, contrary to Paper II where the response variable was binary. Based on available knowledge, Paper II is the first study that involves GLMs in hybrid inference.

3.3 Spatial point processes

Spatial point processes are, as the name indicates, processes that randomly generate points in space (Møller & Waagepetersen 2003). Spatial point processes are generally denoted by capital letters, such as $\mathbf{X}$. In the following, a point pattern generated by a spatial point process $\mathbf{X}$ will be denoted by $\mathbf{x}$. $\mathbf{x}$ is a set of points $\mathbf{x}_i$, $i = 1, \dots, n$, in the two-dimensional space $\mathbb{R}^2$. It can be written as

$$\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}.$$

Let U be a region of $\mathbb{R}^2$. The subset that consists of the points generated by $\mathbf{x}$ that fall in U is denoted by $\mathbf{x} \cap U$.

The intensity of a point process, which can be defined as the average number of points per unit area (Baddeley et al. 2016), will be denoted by λ. Examples of spatial point processes studied in this thesis are presented below.

3.3.1 Poisson point processes

Homogeneous PPPs are sometimes called Complete Spatial Randomness (Baddeley et al. 2016). They fulfil two main properties: first, the homogeneity property, which states that the points have no preference for any spatial location, and thus that the intensity of the process is constant in the whole region of interest. A homogeneous PPP verifies the independence property as well. The latter states that the positions of points do not impact each other. In short, point locations are independent of each other. When the process is inhomogeneous, the homogeneity property is not respected anymore and λ varies, for example, depending on external factors such as environmental covariates. However, the process still fulfils the independence property.

Let $\boldsymbol{\beta}$ denote a vector of model coefficients and denote its transpose by $\boldsymbol{\beta}^{\mathrm{T}}$. Let $\boldsymbol{z}(\boldsymbol{u})$ denote the vector of covariate data, of size q, at location $\boldsymbol{u}$. The intensity of a PPP (Baddeley et al. 2010) can be modelled as

$$\lambda(\boldsymbol{u}) = \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}(\boldsymbol{u})\right), \boldsymbol{u} \in U \subset \mathbb{R}^2, \quad (1)$$

as in Paper II. This implies that the expected total number of plants in region U is

$$\Lambda(\boldsymbol{\beta}) = \int_U \lambda(\boldsymbol{u}) d\boldsymbol{u} = \int_U \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}(\boldsymbol{u})\right). \quad (2)$$

In case a grid tessellation over U is used, as in Paper I, the intensity of the PPP in cell $i, i = 1, \ldots, N$, can be expressed as

$$\lambda_i = \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i\right), \quad (3)$$

where $\boldsymbol{z}_i$ denotes the covariate vector in grid cell i, assuming that the covariate vector inside a cell is constant. Then, the expected total number of plants in U can be expressed as

$$\Lambda = a_P \sum_{i=1}^{N} \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i\right), \quad (4)$$

where a_P is the size of the cells tessellating U.

3.3.2 Neyman-Scott processes

NSPs are part of a larger family of point processes called cluster processes (Baddeley et al. 2016). Cluster processes involve two mechanisms: the first one generates points (called parent points) according to a given process (e.g., a PPP) in the region of interest. Subsequently, a second process creates points (also called offspring points) around each and every parent point. Parent points are then removed to produce the realisation of the cluster process. Clustered point processes have the potential to mimic reality pretty well when it comes to plant locations, since offspring plants usually grow in clusters around some parent plant (Schulze et al. 2019). In fact, this kind of model might be more appropriate than PPP models to model plant populations since random populations seldom exist in nature, although individual species in most plant communities may be randomly scattered (Bonham 2013).

The unobserved parent points in NSPs follow a PPP with a given intensity τ. Each parent point will produce a cluster of offspring points, and these clusters will be independent and identically distributed (i.i.d.) (Baddeley et al. 2016). Offspring points are generated by another point process and are independent within a cluster. The point process that generates the offspring points has an average number of points per cluster μ. The intensity of the NSP as a whole thus becomes $\tau\mu$.

Some NSPs are characterised by a positive third parameter, γ, that refers to the size of the clusters (e.g., their radius). In that case, for each parent point $\mathbf{x}_i \in \mathbb{R}^2$, the offspring points $\boldsymbol{y}_{ij} \in \mathbb{R}^2$ are i.i.d. with a spatial offspring probability density $f_\gamma(\boldsymbol{y} - \mathbf{x}_i)$ that depends on γ.

There are several ways to make the above process inhomogeneous. In this thesis (as well as in Paper III), the method proposed by Waagepetersen (2007) is applied. For a parent point at location $\mathbf{x}_i$, the offspring follow a PPP with intensity $\mu(\boldsymbol{u})f_\gamma(\boldsymbol{u} - \mathbf{x}_i),\ \boldsymbol{u} \in \mathbb{R}^2$. The mean number of points in the clusters μ varies according to environmental covariates, while the intensity of the parent process τ stays constant. More precisely,

$$\mu(\boldsymbol{u}) = \exp\left(\beta_0 + \sum_{i=1}^{q-1} \beta_i z_i(\boldsymbol{u})\right), \tag{5}$$

where each $z_i(\boldsymbol{u})$ denotes an environmental spatial covariate from vector $z(\boldsymbol{u})$, $i = 1, \ldots, q-1$. Other methods to make a NSP inhomogeneous are mentioned in subsection 7.2.

If $f_\gamma(\boldsymbol{u})$ is a uniform density function in a disc of radius γ, then the point process can be seen as an inhomogeneous Matérn cluster process. If, instead, $f_\gamma(\boldsymbol{u})$ is the density function of an isotropic Gaussian distribution $N(0, \gamma^2 I)$, where γ is a standard deviation parameter and I is the identity matrix, then the point process can be seen as an inhomogeneous generalised Thomas cluster process. The inhomogeneous Matérn cluster process is studied in Paper III, while both Matérn and Thomas are used as subcases in Paper I, where non-Poisson processes were generated to investigate the power of the correlation tests.

3.3.3 Other cluster processes

NSPs are not the only examples of processes that generate clustered patterns. For example, the log-Gaussian Cox process (LGCP), which is also employed in Paper I during the simulation study, is part of this family. Cox processes are basically PPPs with a random intensity function (Baddeley et al. 2016), the latter varying depending on unobservable external factors (and possible environmental covariates). No offspring and parent points are involved in the process, contrary to Neyman-Scott cluster processes. A LGCP is a Cox process whose driving intensity is of the form

$$\Lambda(\boldsymbol{u}) = \exp G(\boldsymbol{u}),$$

where $G(\boldsymbol{u})$ is a Gaussian random field.

4. Estimation of plant density based on spatial point processes and P/A data

4.1 Using inhomogeneous Poisson point processes

Both Paper I and Paper II make use of inhomogeneous PPP and P/A data for obtaining estimates of expected plant density and associated variance estimators within large regions. The difference between these two papers is that a tessellation of the study area U is used in conjunction with wall-to-wall covariate date for Paper I, while Paper II makes use of hybrid inference where the covariate data come from sampling plots in U. Paper I makes use of tessellation cells, while circular field plots are the main sampling units in Paper II (cluster sampling is also studied in that paper). For the sake of simplicity, it is supposed henceforth that all grid cells or circular plots are entirely inside the region of interest U. In Paper II, this is not necessarily the case and a buffer is used around U to mitigate potential edge effects (Gregoire & Valentine 2007).

Let N_i denote the number of plants in grid cell i or in circular plot i (called unit i henceforth) in U. Assume that the area of every unit is constant and equal to a_P. If the values of the covariate vector $\boldsymbol{z}_i$ are assumed to be constant within unit i, then the expected number of plants in unit i can be expressed as

$$E(N_i) = a_P \lambda_i,$$

with λ_i as defined in (3). N_i is Poisson distributed, which implies that the probability of presence in unit i can be expressed as

$$p_i = 1 - P(N_i = 0) = 1 - \exp\left(-a_P \exp\left(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i\right)\right).$$

It follows that the loglikelihood for the binary response variables Y_i (that denote presence or absence of plant individuals in unit i) becomes the loglikelihood of a complementary log-log regression model (Baddeley et al. 2010) with an offset equal to the log of the area of unit i

$$\log(-\log(1 - p_i)) = \log(a_P) + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{z}_i. \qquad (6)$$

The estimator of the model parameter vector $\boldsymbol{\beta}$, denoted $\widehat{\boldsymbol{\beta}}$, is obtained from the model above.

In Paper I, there is no need for a second sample since the covariates are available in the entirety of the region of interest. Thus, the expected plant density in U can be expressed as

$$R_U = \frac{\Lambda}{Na_P},$$

with Λ defined as in (4). Note that R_U is the expected total number of plants in U divided by the total area of U, which follows from the definition of the density. R_U can be estimated by

$$\hat{R}_U = \frac{\widehat{\Lambda}}{Na_P} = \frac{1}{N}\sum_{i=1}^{N} \exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_i\right), \tag{7}$$

where $\widehat{\Lambda}$ is the estimator of Λ, obtained by replacing $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$ in (4).

The next step is to derive the associated variance of $\hat{R}_U$ and its estimator. Using the computation rules of the variance, it follows that $\mathrm{Var}\left(\hat{R}_U\right) = (Na_P)^{-2}\mathrm{Var}(\widehat{\Lambda})$. Let n denote the sample size. For estimating the variance of Λ, one may use the fact that for large samples and under mild conditions, $\widehat{\boldsymbol{\beta}}$ is asymptotically normally distributed, i.e.

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{D} N_q(0, \boldsymbol{\Sigma}),$$

where N_q designates the q-variate normal distribution. The covariance matrix $\boldsymbol{\Sigma}$ (whose definition can be seen in Paper I) is assumed to be positive definite (Sen & Singer 1993). This implies that $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_i$ is approximately normally distributed with mean $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i$ and variance $n^{-1}\boldsymbol{z}_i^{\mathrm{T}}\Sigma\boldsymbol{z}_i$. It follows that

$$\mathrm{Var}(\widehat{\Lambda}) \approx a_P^2 \sum_{i=1}^{N}\sum_{j=1}^{N}\left[\exp\left(\frac{\boldsymbol{z}_i^T\boldsymbol{\Sigma}\boldsymbol{z}_j}{n} - 1\right)\right]\exp\left(\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{z}_i + \boldsymbol{z}_j) + \frac{\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{z}_i + \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{z}_j}{2n}\right),$$

which is estimated by replacing $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}$ by $\widehat{\boldsymbol{\Sigma}}$ (see Paper I for details).

Now, the variance of $\hat{R}_U$ can be estimated by

$$\hat{\sigma}_U^2 = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[\exp\left(\frac{\boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{z}_j}{n} - 1\right)\right]\exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{z}_i + \boldsymbol{z}_j) + \frac{\boldsymbol{z}_i^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{z}_i + \boldsymbol{z}_j^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{z}_j}{2n}\right).$$

The above method can be applied when one has access to wall-to-wall covariate data that cover the entire region of interest U. However, this might not be the case in every study. In such cases, hybrid inference may be preferred (as in Paper II), and additional steps are required to obtain estimators of expected plant density and associated variance estimators.

Let f be the joint probability density function (p.d.f.) for the plot centres in sample S_2, and $f_i(\boldsymbol{u})$ the marginal p.d.f. for plot centre $\boldsymbol{u}_i$ in S_2, $i = 1, \ldots, n_2$. The inclusion density function is

$$\pi(\boldsymbol{u}) = \sum_{i=1}^{n_2} f_i(\boldsymbol{u}), \tag{8}$$

and can be considered as a local measure of the number of sample points to be selected per unit area (Cordy 1993). The generalised Horvitz-Thompson estimator of the expected number of plants in U is then given by

$$\widehat{\Lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n_2} \frac{\lambda(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} = \sum_{i=1}^{n_2} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i)}{\pi(\boldsymbol{u}_i)}, \tag{9}$$

where $\pi(\boldsymbol{u})$ is given by (8) and $\lambda(\boldsymbol{u}_i)$ is the average expected intensity in sample plot i with centre $\boldsymbol{u}_i$. If the covariate data are supposed to be constant in sample plot i, then $\lambda(\boldsymbol{u}_i)$ is the same as $\lambda(\boldsymbol{u})$ as defined in (1), with $\boldsymbol{z}(\boldsymbol{u}) = \boldsymbol{z}_i$, $\boldsymbol{z}_i$ being equal to the value of the covariate vector in sample plot i. Since the parameter vector $\boldsymbol{\beta}$ is usually unknown, $\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})$ is used as an estimator of the expected number of plants instead.

It is of necessity to know the area of U, or to estimate that area, to get an estimate of the expected plant density. In case the area of U is known (denote this area by a_U), the expected density in U is expressed as

$$R(\boldsymbol{\beta}) = \frac{\Lambda(\boldsymbol{\beta})}{a_U},$$

where $\Lambda(\boldsymbol{\beta})$ is defined in (2). This expected density can be estimated by

$$\hat{R}(\widehat{\boldsymbol{\beta}}) = \frac{\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})}{a_U},$$

where $\widehat{\Lambda}(\boldsymbol{\beta})$ is defined in (9). The case where a_U is unknown and estimated is presented in detail in Paper II.

In order to get an estimate of the variance of $\widehat{\Lambda}(\boldsymbol{\beta})$, the Sen-Yates-Grundy variance formula defined in Cordy (1993) is applied (see Paper II). Let

$$\Delta(\boldsymbol{u}, \boldsymbol{u}') = \pi(\boldsymbol{u})\pi(\boldsymbol{u}') - \pi(\boldsymbol{u}, \boldsymbol{u}') \text{ and } \pi(\boldsymbol{u}, \boldsymbol{u}') = \sum_{i \in I_n} \sum_{j \in J_{n,i}} f_{ij}(\boldsymbol{u}, \boldsymbol{u}'),$$

the latter being the pairwise inclusion density function with $I_n = \{1, \dots, n_2\}, J_{n,i} = \{1, \dots, n_2\} \backslash \{i\}$, and f_{ij} is the joint p.d.f. of $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$. According to Cordy (1993), if $\pi(\boldsymbol{u})$ and $\pi(\boldsymbol{u}, \boldsymbol{u}')$ are strictly positive for all $(\boldsymbol{u}, \boldsymbol{u}') \in U$, an unbiased estimator of $\mathrm{Var}(\widehat{\Lambda}(\boldsymbol{\beta}))$ is given by

$$\widehat{\mathrm{Var}}\left(\widehat{\Lambda}(\boldsymbol{\beta})\right) = \frac{1}{2} \sum_{i \in I_n} \sum_{j \in J_{n,i}} \frac{\Delta(\boldsymbol{u}_i, \boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i, \boldsymbol{u}_j)} \left(\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_i)}{\pi(\boldsymbol{u}_i)} - \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}_j)}{\pi(\boldsymbol{u}_j)} \right)^2.$$

When the model coefficients are unknown, $\boldsymbol{\beta}$ is estimated by $\widehat{\boldsymbol{\beta}}$ and an estimate of the variance of $\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})$ can be written as

$$\widehat{\mathrm{Var}}\left(\hat{\Lambda}(\hat{\boldsymbol{\beta}})\right) = \frac{1}{2}\sum_{i\in I_n}\sum_{\substack{j\in J_{n,i}\\ q\quad q}} \frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_i)}{\pi(\boldsymbol{u}_i)} - \frac{\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_j)}{\pi(\boldsymbol{u}_j)}\right)^2 + \sum_{k=1}^{q}\sum_{l=1}^{q} \widehat{\mathrm{Cov}}_{S_1}(\hat{\beta}_k,\hat{\beta}_l)\,\hat{v}_k\hat{v}_l,$$

with

$$\hat{v}_k = \sum_{i=1}^{n_2} \frac{1}{\pi(\boldsymbol{u}_i)}\lambda^{(k)}(\mathbf{u}_i)\,,$$

where $\hat{\beta}_k$ denotes the kth component of the $\hat{\boldsymbol{\beta}}$ vector and

$$\lambda^{(k)}(\boldsymbol{u}_i) = \frac{\partial\lambda(\boldsymbol{u}_i)}{\partial\hat{\beta}_k} = z_{ik}\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{z}_i),$$

with z_{ik} denoting the kth component of $\boldsymbol{z}_i$.

Then, it follows that the variance estimator of the estimator of expected plant density with known area a_U is

$$\widehat{Var}\left(\hat{R}(\hat{\boldsymbol{\beta}})\right) = \frac{\widehat{Var}(\hat{\Lambda}(\hat{\boldsymbol{\beta}}))}{a_U^2}.$$

A variance estimator in the case where the area a_U is unknown is given in Paper II.

4.2 Using inhomogeneous Neyman-Scott processes

In Paper III, plant locations are supposed to be generated by cluster point processes, in particular Matérn cluster processes. A design with sampling plots is applied, and the inference is based on the plant registrations (more specifically, P/A data of plants) done within these sampling plots. The estimation of the parameters of the inhomogeneous NSP cannot be done in the same way as in Papers I and II, i.e. a binary regression model cannot be used. It will be shown below that a multinomial regression model is used to estimate the parameter vector $\boldsymbol{\theta} = (\tau, \beta_0, \dots, \beta_{q-1}, \gamma)$. For a Matérn cluster processes, τ denotes the intensity of the parent process, $\mu(\boldsymbol{u})$ is defined as in (5) as a function of a parameter vector $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{q-1})$, and γ denotes the cluster radius parameter.

It is necessary to know about the disposition of plots in a given sampling design in order to calculate the probabilities of presence in the different plots (or in the different configurations of plots). The following design with concentric circular plots, used in Paper III, can be taken as an example (see

Fig. 3). The use of sampling designs with concentric plots when spatial patterns are to be expected is recommended by Morrison et al. (1995). Assume that there are n sets of (concentric) plots $C_{i,j}, i = 1, \dots, n,\ j = 1, \dots, k$. Suppose that the n sets of concentric plots are so far apart that the point patterns within these sets can reasonably be considered as independent. In main plot i, the jth innermost circle $C_{i,j}$ has a radius $r_j, j = 1, \dots, k$. Let $B_{i,1} = C_{i,1}$ and $B_{i,j} = C_{i,j} \backslash C_{i,j-1}, i = 1, \dots, n, j = 2, \dots, k$, and let $N_{C_{i,k}}$ denote the number of plants in plot $C_{i,k}$. A survey of such set of plots is done from the innermost plot outwards, and is over as soon as a plant is encountered on one of the $B_{i,j}$s or if no plants at all are registered in the k concentric circular plots. Thus, the events corresponding to this survey are the following:

$$A_{i,0} = \{\text{absence in } C_{i,k}\} = \{N_{C_{i,k}} = 0\},$$
$$A_{i,1} = \{\text{presence in } C_{i,1}\} = \{N_{C_{i,1}} > 0\},$$
$$A_{i,j} = \{\text{presence in } B_{i,j} \text{ but not in } C_{i,j-1}\} = \{N_{C_{i,j-1}} = 0 \text{ and } N_{B_{i,j}} > 0\},$$

for $i = 1, \dots, n$ and $j = 2, \dots, k$.

Let $I_{i,j}$ be the indicator of the event $A_{i,j}$ and $\pi_{i,j}(\boldsymbol{\theta})$ the associated probabilities for every $A_{i,j}$, computed using Theorem 1 in Paper III. Set Y_i equal to j if the event $A_{i,j}$ occurs, $i = 1, \dots, n$. The variable Y_i then becomes the dependent variable in a multinomial regression model (cf. Amemiya 1985), for which

$$P\{Y_i = j\} = P\{A_{i,j}\} = \pi_{i,j}(\boldsymbol{\theta}).$$

Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. This parameter vector is estimated by maximum likelihood. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is any parameter vector in $\Theta = \{\boldsymbol{\theta} = (\tau, \beta_0, \dots, \beta_{q-1}, \gamma)^{\mathrm{T}} : \tau, \gamma > 0\}$ that maximises the log-likelihood function

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{k} I_{i,j} \log \pi_{i,j}(\boldsymbol{\theta}).$$

By standard arguments (cf. Rao 1973; Amemiya 1985),

$$i_{rsn}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_r} \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_s}\right] = \sum_{i=1}^{n} \sum_{j=0}^{k} \frac{1}{\pi_{i,j}(\boldsymbol{\theta})} \frac{\partial \pi_{i,j}(\boldsymbol{\theta})}{\partial \theta_r} \frac{\partial \pi_{i,j}(\boldsymbol{\theta})}{\partial \theta_s},$$

where $\theta_1 = \tau, \theta_k = \beta_{k-2}, k = 2, \dots, q+1, \theta_{q+2} = \gamma$. Let $I_n(\boldsymbol{\theta}) = (i_{rsn}(\boldsymbol{\theta}))$. It is assumed that the limiting matrix $\lim_{n \to \infty} n^{-1} I_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ exists, is finite and positive definite (Sen & Singer 1993). The maximum likelihood

estimator $\widehat{\boldsymbol{\theta}}$ is supposed to be consistent and asymptotically normally distributed, i.e.

$$n^{\frac{1}{2}}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \xrightarrow{D} N(0,[I(\boldsymbol{\theta}_0)]^{-1}). \tag{10}$$

Let $i^{rsn}(\boldsymbol{\theta}), r, s = 1,2,\dots,q+2$, denote the elements of the inverse of the matrix $I_n(\boldsymbol{\theta})$. An approximate 95% confidence interval for the individual parameters θ_r is given by

$$\hat{\theta}_r \pm 1.96\sqrt{i^{rrn}\left(\widehat{\boldsymbol{\theta}}\right)}, r = 1,2,\dots,q+2. \tag{11}$$

An approximate confidence interval can also be constructed for the intensity $\lambda_{\boldsymbol{\theta}}(\boldsymbol{u}) = \tau\mu(\boldsymbol{u}) = \tau\exp(\beta_0 + \sum_{i=1}^{q-1}\beta_i z_i(\boldsymbol{u}))$ of the point process at a spatial location $\boldsymbol{u}$. By (10) and the delta-method (e.g., Shao 2003),

$$n^{\frac{1}{2}}\left(\lambda_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{u}) - \lambda_{\boldsymbol{\theta}_0}(\boldsymbol{u})\right) \xrightarrow{D} N\left(0, \nabla\lambda_{\boldsymbol{\theta}_0}(\boldsymbol{u})^{\mathrm{T}}[I(\boldsymbol{\theta}_0)]^{-1}\nabla\lambda_{\boldsymbol{\theta}_0}(\boldsymbol{u})\right),$$

where $\nabla\lambda_{\boldsymbol{\theta}}(\boldsymbol{u})$ denotes the gradient, i.e., the vector formed by the partial derivatives $\frac{\partial\lambda_{\boldsymbol{\theta}}(\boldsymbol{u})}{\partial\theta_r}, r = 1,\dots,q+2$. Then, an approximate 95% confidence interval for the intensity at a given spatial location $\boldsymbol{u}$ is given by

$$\lambda_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{u}) \pm 1.96\sqrt{\sum_{r=1}^{q+1}\sum_{s=1}^{q+1} i^{rsn}\left(\widehat{\boldsymbol{\theta}}\right)\left.\frac{\partial\lambda_{\boldsymbol{\theta}}(\boldsymbol{u})}{\partial\theta_r}\right|_{\widehat{\boldsymbol{\theta}}}\left.\frac{\partial\lambda_{\boldsymbol{\theta}}(\boldsymbol{u})}{\partial\theta_s}\right|_{\widehat{\boldsymbol{\theta}}}}. \tag{12}$$

The estimated standard errors for the estimators $\hat{\theta}_r$ and $\lambda_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{u})$ are provided by the square root expressions of (11) and (12), respectively.

4.3 Assessing the models

There is no guarantee that the regression models presented in 4.1 and 4.2 are directly applicable. These models, that are implied by their respective types of point processes, need to be evaluated for suitability. First, standard methods for model selection and validation of the regression models were applied. In addition, in Paper I, the correlation between the residuals in the fitted model (6) was investigated. If spatial correlation was found between the paired residuals (obtained from the paired observations in the sampling design in Fig. 4), this indicated that the assumption of independence of the response variables—given the covariates—in the binary regression model (6), as implied by the inhomogeneous PPP model, was violated. This would then make the underlying hypothesis of inhomogeneous PPP invalid. The

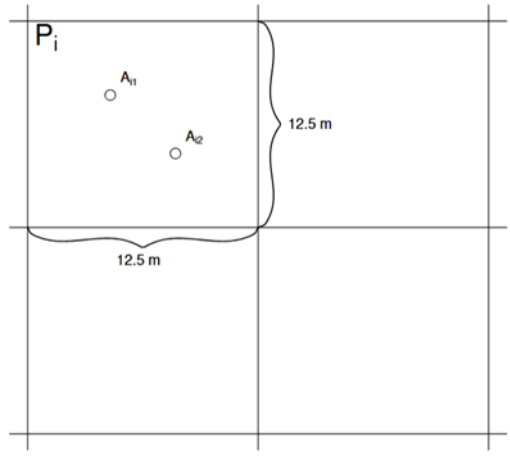


Figure 4. Disposition of the paired vegetation plots in a pixel, based on the design used in the Swedish NFI.

regression model residuals that were studied were the Pearson, working and randomised quantile residuals (Dunn & Smyth 1996, 2018). Additionally, the Pearson and Spearman correlation coefficients were the coefficients that were studied.

In Paper III, randomised quantile residuals as defined in Trijoulet et al. (2023) were used to evaluate the fitted multinomial regression models.

4.4 Simulation studies

Simulation studies are an essential part of this thesis. Indeed, Monte Carlo simulations are convenient in order to evaluate the performance of the developed estimators and their estimators of variance. They are also practical to investigate the actual significance levels (i.e., the percentage of times the null hypothesis is rejected when the data are generated according to the null hypothesis) and the power (i.e., the percentage of times the null hypothesis is rejected when the data are generated according to the alternative hypothesis) of the tests. All the simulations in this thesis were performed in R (R Core Team 2025).

In Paper I, simulation studies were conducted in order to investigate the power and actual significance levels of the correlation tests. The actual significance levels should ideally be close to the nominal significance level,

i.e. 0.05 as it was set, and the power should be close to 1 the more the studied point processes differ from an inhomogeneous PPP (e.g., when investigating processes that produce strong clustering). Different point processes were generated: log-Gaussian Cox, Matérn cluster and Thomas processes. The parameters for these processes were adjusted to express different strength of clustering. A GLM was then fitted on the P/A data coming from the generated point data and the covariates associated with each sample plot under the assumption that the process was generated by an inhomogeneous PPP, which was then used to produce residuals to use for the considered tests.

In Paper II, simulations were performed in order to evaluate the performance of the derived density estimators and their associated variance estimators. This was done for two cases, both when the area of the region of interest was known and when it was unknown. Covariates were created artificially based on the ones that were included in the *V. vitis-idaea* model. P/A data were generated from an inhomogeneous PPP for each replicate. The plot centres for samples S_1 and S_2 were randomly selected for each replicate. A model was then created on S_1 and model coefficients $\widehat{\boldsymbol{\beta}}$ were estimated. Then, the expected plant density and corresponding variance were estimated based on S_2.

In Paper III, the main objective of the simulation study was to verify the performance of the parameter estimators and their associated variance estimators. This involved computing the actual confidence level (ACL) of the derived confidence intervals for the individual parameters included in vector $\boldsymbol{\theta}$. The ACL should ideally be close to the nominal confidence level, set to 95% for this study. The design used during the Monte Carlo study was the design with concentric circles presented in subsection 3.1, with 6 concentric circular plots. Realisations of Matérn cluster processes were generated and maximum likelihood estimates of $\boldsymbol{\theta}_0$ were produced. The models investigated were the ones for the 3 plant species introduced in subsection 3.1, with covariates based on the actual covariates in the study area. The intensity of the process when the covariate was equal to its ith sample quartile, $i = 1,2,3$, was estimated, and the performance of these intensity estimators was also evaluated.

5. Estimation of the components of the MSE based on simulations

In Papers I to III, the measure of uncertainty was uniquely the variance estimator for the estimator of expected plant density. However, it can be of interest to investigate a more thorough measure of uncertainty in some cases, if each of the components of said measure of uncertainty can be estimated. The MSE can be applied as a measurement of uncertainty when model-based inference is used. The MSE gives an expression for the expected squared deviation between predicted and actual values, and takes several relevant uncertainty factors into consideration, in addition to the variance of the estimator/predictor.

Let Y be the (random) targeted response variable (e.g., AGB measurement in ton/ha) and $\hat{Y}$ be its predictor. In the model-based inference case, the MSE of the predictor is expressed as

$$\mathrm{MSE}(\hat{Y}) = \mathrm{Var}(\hat{Y}) + \mathrm{Bias}^2(\hat{Y}) + \mathrm{Var}(Y) - 2\mathrm{Cov}(\hat{Y}, Y),$$

where $\mathrm{Var}(\hat{Y})$ is the variance of the predictor, $\mathrm{Bias}(\hat{Y}) = \mathrm{E}(\hat{Y} - Y)$ is the model bias of the predictor, $\mathrm{Var}(Y)$ is the variance of the target variable and $\mathrm{Cov}(\hat{Y}, Y)$ is the covariance of the target variable and its predictor.

In a number of model-based studies that make use of remotely sensed data, the formula of the MSE is believed to be used wrongly (Ståhl et al. 2024). Indeed, some studies use the design-based version of the MSE formula instead of the model-based version. In other words, the variance of the target variable, the model bias and the covariance term are omitted. However, these terms can be non-negligible in some cases. Looking at the magnitude of each term of the model-based MSE in different cases is the main objective of Paper IV.

For large areas, the MSE can reasonably be approximated by the variance of the predictor solely (McRoberts et al. 2018). This is not the case anymore when the area of study is small, as can be seen in the results of Paper IV.

Paper IV relied entirely on simulations. For each combination of study area (Tanzania or USA) and RS data source (ALS/GEDI or Landsat 8), a proxy of the superpopulation model was estimated and applied to generate a large number of population realisations. For each population and field data sample size, a new prediction model was created based on a fixed systematic sample, and the biomass density predicted for the target area. Thus, for each

replicate, a true and a predicted biomass density values were obtained. Based on these, all the terms constituting the MSE formula could be empirically estimated.

In order to make the simulations more realistic, a spatial autocorrelation structure was introduced. It was applied to the error terms in the superpopulation model. The strength of the spatial autocorrelation could be controlled (no autocorrelation, mild autocorrelation, moderate autocorrelation and strong autocorrelation), and several subcases were examined. The benchmark case in the simulations had moderate autocorrelation and a sample size of 150 cells.

6. Results from the empirical data studies and simulations

6.1 Large-area estimation of plant density using presence/absence data and binary regression, and correlation tests of the binary regression model (Paper I)

The results were separated into two categories for this paper. First, an application with field data was conducted, and then a simulation study was performed to evaluate the performance of the proposed estimators and associated variance estimators.

A model was first constructed on field data, i.e. P/A data of *L. europaea* and *L. pilosa*, in the Lappland region of Norrbotten county in Sweden. Both binary regression models based on data from these species passed the correlation test at the 5% significance level. Both models contained two explanatory variables: the first model, for *L. europaea*, comprised the proportion of tree stem volume of deciduous trees and basal area; whereas the second model, for *L. pilosa*, comprised the basal area-weighted mean tree DBH and an index of soil moisture. The expected plant density was estimated for both species in the region of interest. The estimate of expected plant density, as defined in (7), was 0.111 individuals per square metres for *L. europaea* and 0.054 individuals per square metres for *L. pilosa*. The associated estimated variances were respectively 0.00125 and 0.00052.

In the Monte Carlo simulation study, the variant with the Pearson residuals and Pearson correlation test was the one that produced the largest power, whereas the variant with randomised quantile residuals was the one that produced actual significance levels closer to the nominal level of 5%. It was clear that the larger the sample, the higher the power. The power increased monotonously for the LGCP with increasing standard deviation parameter from 0 to 3, whereas it increased and then decreased with the two types of NSPs (i.e., Matérn cluster and generalised Thomas) with increasing cluster size. It was also apparent that when the distance between the vegetation plots in a same pair was reduced, the power of the test increased (as can be seen in Fig. 5).

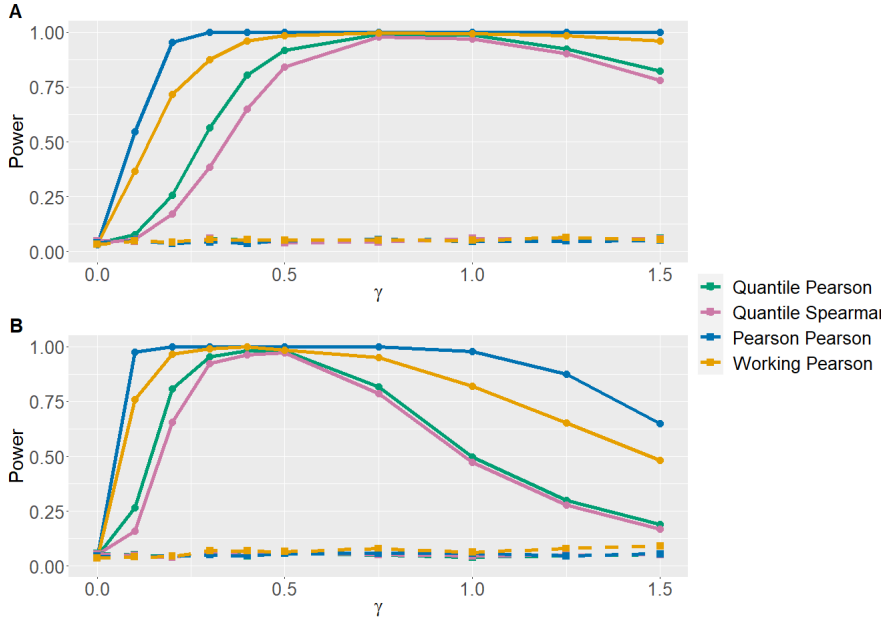


Figure 5. Examples of power curves, with Matérn and Thomas processes, for different types of residuals (quantile, Pearson, working) and correlation coefficients (Pearson, Spearman), with varying γ. The curves with solid lines represent the cases with a distance of 0.62 metres between the plot centres, and the curves with dashed lines represent the cases with a distance of 5 metres between the plot centres.

6.2 Estimation of plant density based on presence/absence data using hybrid inference (Paper II)

The results were separated into two categories for this paper as well, in the same way as in Paper I. A model was first constructed based on real data, i.e. P/A data of *V. vitis-idaea* in the Lappland region of Norrbotten county in Sweden. The model comprised two explanatory variables: the number of tree stems per hectare, and an indicator variable indicating whether the soil was humid or wet. The density of *V. vitis-idaea* was estimated to be approximately 7.5 individuals per square metre in the whole Lappland region of Norrbotten county. If only the areas consisting of productive forests were considered, this estimated density increased to approximately 9.7 individuals per square metre. The associated estimated variances were respectively 0.209 and 0.411.

A Monte Carlo study was also performed to see whether these estimators and associated variance estimators performed satisfactorily. In both cases with known and unknown area, the estimator of expected density presented slight negative bias and the associate variance estimators showed small bias (see Paper II).

6.3 Estimation of parameters in inhomogeneous Neyman-Scott processes using presence/absence data (Paper III)

There were two parts in the results section in this paper as well. Models were first created on field data collected in the Kulbäcksliden research park in Västerbotten County, Sweden. Three species were selected to apply the proposed method: *M. bifolium*, *L. pilosa* and *L. europaea*. Only one covariate variable was used for each model. All the models passed the Shapiro-Wilk test of normality based on the model residuals, and the parameters for the underlying NSPs were estimated. Local estimates of plant density when the covariate was equal to specific, meaningful values were also provided in the article. The estimated process intensities (as well as the associated 95% confidence interval) when the covariates equal their sample median values were respectively 0.052 (0.034, 0.071), 0.11 (0.063, 0.17) and 0.026 (0.018, 0.035) for *M. bifolium*, *L. pilosa* and *L. europaea*.

In the Monte Carlo study, it was apparent that the (mean and median) biases and standard deviations of the estimators decreased as the sample size increased. For the largest sample size considered, the biases were very small and the ACLs were pretty close to the nominal level of 95%. However, for the smallest sample size considered, the ACLs were notably lower than the nominal level 95% for several parameters and estimates of plant density. The standard error estimators exhibited some biases, but they were not severe for any of the sample sizes considered.

6.4 A closer look at uncertainties in forest ecosystem surveys using remotely sensed data and model-based inference (Paper IV)

In the simulation study of Paper IV, several cases were investigated. In the first case, variability due to study area size was studied. The sample size was

held constant at 150 cells, while the study area (i.e. the area where the models are applied) varied between very small area, small area, intermediate area and very large area (the whole grid). The main observation from that subcase was that the variance of the true value decreased with study area size, whereas the other components of the MSE did not appear to be particularly affected (Fig. 6). A particularly interesting observation when the area of application was the entire grid was that the MSE was smaller than the variance of the predictor due to the non-negligible covariance term.

The second case was variability due to field sample size. The study area was the whole grid in both regions, and the field sample size varied between 50, 150 and 500 sample units. This change in sample size principally affected the variance of the predictor (Fig. 7). More specifically, the larger the sample, the smaller the variance of the predictor.

The third case was variability due to spatial autocorrelation of the error terms. The sample size was set to 150 sample units, the study area was set to the whole grid, and the strength of the autocorrelation varied between no autocorrelation, mild autocorrelation, moderate autocorrelation and strong autocorrelation. The results showed that both the variance of the true value and the covariance term increased significantly with the strength of spatial autocorrelation (Fig. 8). A similar change occurred for the variance of the predictor, albeit less obviously. However, this had the effect of decreasing the MSE with autocorrelation strength, mostly because of the covariance term.

The fourth case was the effect of model transfer on the components of the MSE. The model constructed on Tanzania data was applied on USA data and vice versa. Significant model bias occurred in both cases, and the squared model bias component made up for approximately 95% of the total MSE. This can be explained by strong extrapolation that occurs when applying a model from one region to another, since these regions might not have the same range for the target variable.

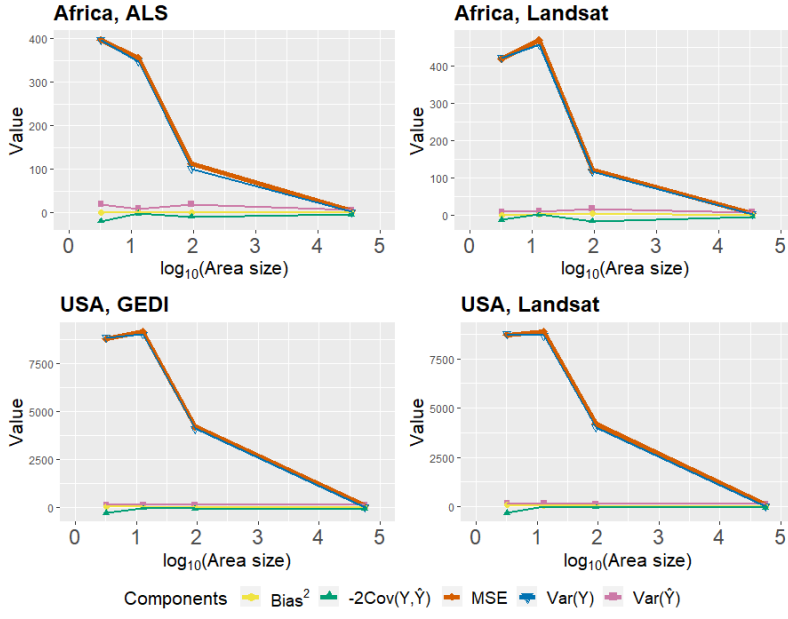


Figure 6. Effect of area size on the different MSE components in the different areas of study and for each kind of RS data.

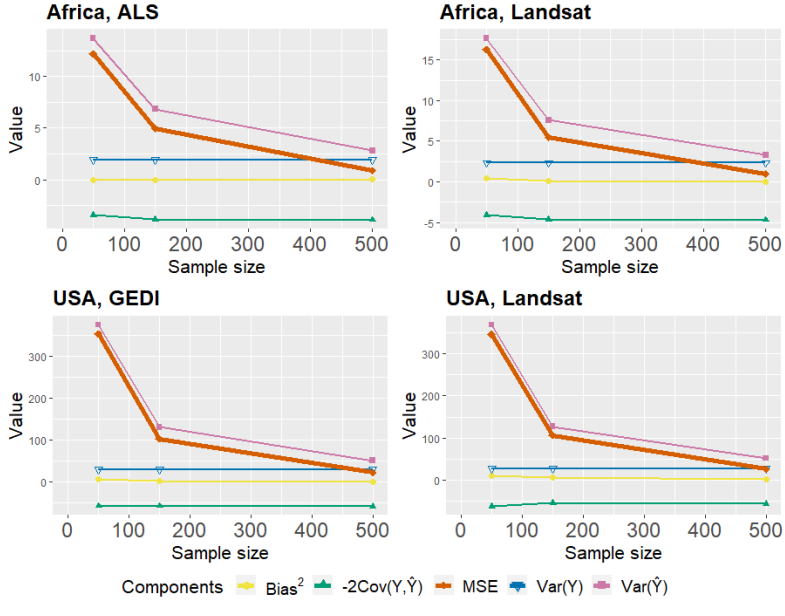


Figure 7. Effect of sample size on the different MSE components in the different areas of study and for each kind of RS data.

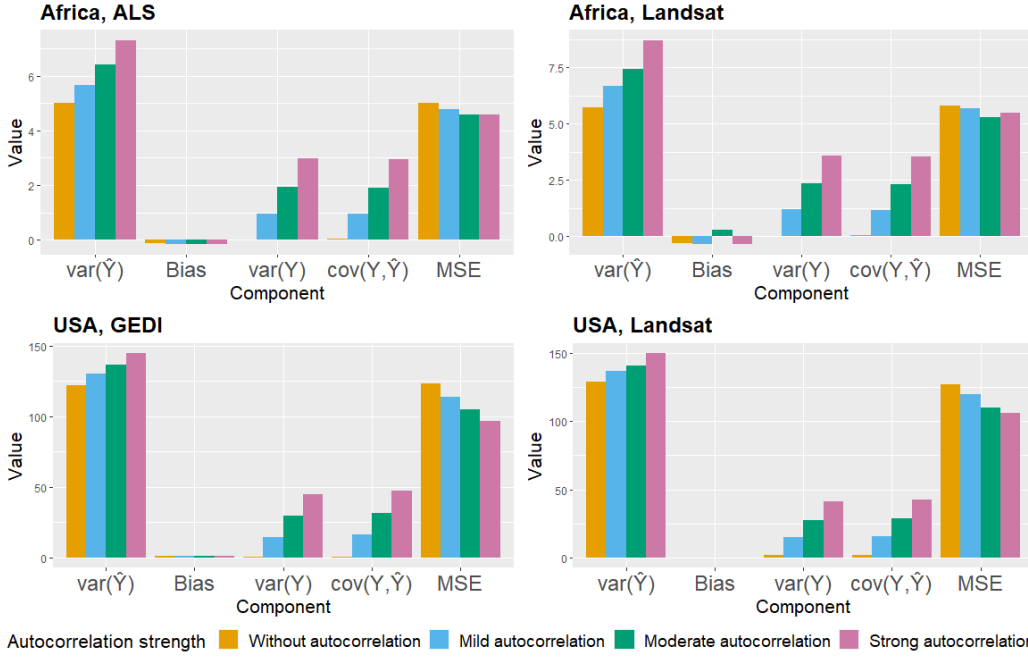


Figure 8. Effect of autocorrelation strength on the different MSE components in the different areas of study and for each kind of RS data.

7. Discussion and future research

7.1 Some reflections and conclusions

In this thesis, several methods intended for application in ecological research and environmental monitoring programmes were developed. A particular emphasis was made on P/A data, which are oftentimes registered as part of these programmes but are only occasionally used in analyses. The methods presented in this thesis for estimating plant density from P/A data show versatility, since they can be adapted to different assumptions regarding plant populations (more precisely, plant locations) and different types of covariate information (wall-to-wall or at plot sample level). The results were promising but additional research is needed.

P/A data are known for being hard to interpret (Ståhl et al. 2017). It is difficult to get ecologically valuable measurements directly from P/A data (for example total cover area or number of plants per hectare), and the occurrence proportions obtained from P/A data analyses are usually comparable only within monitoring inventories as long as the sample plot sizes are the same. In this thesis, several methods that mitigate these aspects have been presented. Inhomogeneous point process models, i.e. point process models that take environmental covariates into account, were used to model plant positions. These point process models imply GLMs that create a link between these environmental covariates and P/A data of plants. Based on these GLMs, estimates of plant density can be obtained locally and in large regions when P/A data are registered at sample plot level. Thus, if the assumed point process model is approximately correct, the proposed methodology provides estimates of plant density that are more easily interpretable than, for example, occurrence proportions. Such estimates can be used for state and trend analyses, as well as for reporting by environmental monitoring programmes within the scope of the habitat directives mentioned in Section 1.

In a way, this thesis can be seen as an argument for the more frequent use of P/A data in environmental monitoring analyses. The cost- and time-effectiveness of P/A sampling, in combination with the readily-available nature of RS data, can be reasonable arguments in favour of the wide-spreading of GLMs with binary response variables in ecology, especially for environmental analyses. This thesis has hopefully demonstrated that model-

based and hybrid inference can be used efficiently to derive meaningful and construable characteristics of plant populations – in particular, plant density and abundance – using spatial models to represent the locations of plants as well as P/A data in connection with auxiliary environmental data.

One should keep in mind that inference from model-based estimators always rely on some model assumptions, and these should be tested. Different ways to assess the GLMs implied by the underlying spatial point process models have been suggested in this thesis. The tests presented herein, as well as in Papers I and III, are based on model residuals. Randomised quantile residuals appear to be useful for assessing binary regression models, although their randomness could cause problems in the correlation tests. Indeed, as stated in Paper I, the test statistic and associated p-value will vary each time the test is run, even for the exact same values of response variable and covariate data. Moreover, it has been shown in the simulation study in Paper I that the test variant with Pearson residuals is more powerful than the one making use of randomised quantile residuals, although the actual significance levels for the former variant were a little less satisfactory.

One relevant observation with regard to environmental monitoring programmes based on the analyses in Paper I is that the method developed within would detect deviations from random populations better if the distance between the plots in a same pair of vegetation plots would be decreased to its minimum possible without any overlap occurring (in case such a sampling design is used, for example in the Swedish NFI). This aspect could be taken into account for future planning in the monitoring programmes.

The derived estimators and corresponding variance estimators in Papers I, II and III were evaluated through Monte Carlo simulations. The results indicated that these estimators performed relatively well, with only slight bias and reasonable orders of magnitude for the variance estimates, particularly when the sample size increased. In addition, the confidence intervals in Paper III were relatively narrow for the larger sample sizes considered, which shows that the method used therein is rather precise.

In Papers I and II, some estimates of expected density for several plant species in an entire region of Sweden were produced. However, it is difficult to see if these values are reasonably accurate, since there exists no reference data to compare with. In spite of that, these values do not appear to be too far-fetched, since for example *V. vitis-idaea* is a relatively common plant

species in the whole of Sweden, and thus relatively high densities are to be expected.

An important aspect to take into consideration when performing plant surveys is to decide what constitutes a presence registration for a given species. In other words, it has to be decided what part of the plant is sufficient for the plant to be considered as present on a sample plot. In Paper III, presence is recorded if a predetermined part of the plant is located on the vegetation plot (i.e., their root points), as advised by, e.g., Cain & Castro (1959), whereas in Papers I and II, presence was recorded if any part of the plant was located in the vegetation plot. As a consequence, the plot radii needed to be adjusted in the calculations by adding a presumed average radius for the plant, in accordance with Ståhl et al. (2017).

In Paper IV, the MSE and its different components were estimated in a model-based context. The results showed that the variance of the estimator (or predictor) is not the only component that can be significant, especially in small-area surveys. In the latter case, the variance of the true value of the variable under study can be largely influential. Moreover, if a spatial autocorrelation structure exists, the covariance between the predicted value and the true value can impact the MSE and contribute to it negatively. Model bias can be an issue if one applies a model to another region compared to where said model was constructed. Thus, the conclusion is that one should be aware of other uncertainty components than the variance of the predictor when performing model-based inference. This conclusion has consequences for Papers I, II and III, as will be explained in subsection 7.2.

The models presented in Papers I, II and III contain at most two explanatory variables. As a consequence, these models can be considered too simplistic, and can be criticised for failing to capture the complexity of the environment and the different processes within it that can affect the presence of plants at a given location. It would have been interesting to consider models with more explanatory variables, although this might necessitate an increase of the sample size as well. However, the methods presented in these papers are computationally demanding, and the complexity rises when the number of explanatory variables increases, especially in Paper III. Hence, future attempts at such modelling should take computer efficiency as a factor.

7.2 Ideas for future research

An interesting area to look further into is the estimation of change of plant density between two time points in case plant data are assumed to be generated from inhomogeneous point processes. Ståhl et al. (2017) have already studied this issue for homogeneous PPP. Modelling of P/A data from two different point processes are needed: one from the process at the first time point and another one at the second time point. If the time between the two considered time points is long enough, these two processes can be considered independent, and then the extension from the methods in, e.g., Paper I can be relatively straightforward. However, if this time interval is relatively short, then the two point processes will probably be dependent, and adjustments would be needed in the modelling.

In Paper II, hybrid inference is applied to data that are supposed to be realisations of an inhomogeneous PPP. It could be feasible to develop the method for other kinds of point processes as well, not least Neyman-Scott cluster processes, that is to expand the methods presented in Paper III to obtain an estimator of expected plant density in an entire region.

In Paper III, inhomogeneity in the NSPs was introduced by letting μ vary by environmental covariates. The idea followed from Waagepetersen (2007). Nevertheless, this is far from being the only way to make a NSP inhomogeneous. Baddeley et al. (2016) make the intensity of the parent process, τ, vary with covariates. Other methods to include non-homogeneity in cluster processes include second order intensity reweighted stationarity (Baddeley et al. 2000), which is another way to make μ vary through thinning the offspring points. This is similar to dependent thinning (Prokešová 2010). Other methods make γ vary. Such examples include inhomogeneous space-location dependent scaling (in which τ varies as well (Hahn et al. 2003)) or the "growing clusters" approach developed in (Mrkvička 2014). In the latter, μ varies as well. It would be interesting to expand the methods derived in Paper III by applying these other techniques to make NSPs inhomogeneous, as well as different other types of NSPs such as Thomas processes, although this would complicate the numerical calculations.

Takashina et al. (2018) propose a framework to estimate abundance from count data using SRS and cluster sampling while making assumptions about an underlying homogeneous Thomas process. An area of interest could be to develop the method presented therein so that it takes into account inhomogeneous point processes as well. Several other models already exist

to estimate population abundance or density based on abundance data. Design-based methods that make use of, e.g., adaptive cluster sampling have been used to estimate abundance for rare plant species that are known to grow in clusters, using, for example, a Horvitz-Thompson estimator as in Philippi (2005). Abrahamson et al. (2011) compared the performance of several sampling designs, including adaptive cluster sampling, to estimate understory plant abundance.

The measure of uncertainty in Papers I, II and III is solely the variance of the estimator of the variable of interest. However, it has been shown in Paper IV that other uncertainty components (that constitute the formula for the MSE) can be significant in model-based studies. Moreover, the MSE is more suitable as a measurement of uncertainty compared to the variance of the estimator/predictor alone, since it takes into account the deviation between the true and estimated/predicted value. A problem is that it seems to be difficult to estimate the model bias for the cases studies in Papers I, II and III, since there are no registrations of plant abundance available at sample plot level. Instead, the bias is estimated by a simulation study. Likewise, a possible further development of the studies regarding plant density can be to specify predictors for actual plant density at regional level and perform simulations similar to the ones conducted in Paper IV for a wider uncertainty analysis of the estimations of plant density.

In the simulation study in Paper IV, the biomass values in each cell were supposed to be strictly positive. This is not always the case in reality, where some areas, even inside forests, can have an absence of biomass (for example in Næsset et al. (2016)). Further simulations would be needed where zero values can be accepted, for example by adopting a two-part modelling approach (Duan et al. 1983), where the 0-data are first generated from a specific model (e.g., a logistic regression model), and then the non-0 data are generated according to another model, e.g., a GLM (cf. Min & Agresti 2002).

There is a growing need for methods that integrate multiple data types into a single analytical framework. Over the past decade, initial efforts have been made to combine unbiased P/A data from structured monitoring programs with abundant presence-only data from citizen science sources; see, e.g., Fletcher et al. (2019) for a review. Using an inhomogeneous PPP model for multispecies data, Fithian et al. (2015) propose pooling presence-only and P/A data to simultaneously estimate and correct for the sampling bias affecting the presence-only data. A key assumption of their model is that

the sampling bias is consistent across all species considered. Under these assumptions, they argue that unbiased or nearly unbiased estimates of plant density can be obtained for all species, including rare ones. Pacifici et al. (2017), Gelfand & Shirota (2019), and Ahmad Suhaimi et al. (2021) developed similar frameworks that also account for spatial dependence. However, because these models do not incorporate a multispecies setting, they are less suitable for rare species.

References

Aberdeen, J. (1958). The effect of quadrat size, plant size, and plant distribution on frequency estimates in plant ecology. *Australian Journal of Botany*, 6 (1), 47–58. https://doi.org/10.1071/BT9580047

Abrahamson, I.L., Nelson, C.R. & Affleck, D.L. (2011). Assessing the performance of sampling designs for measuring the abundance of understory plants. *Ecological Applications*, 21 (2), 452–64. https://doi.org/10.1890/09-2296.1

Ågren, A.M., Larson, J., Paul, S.S., Laudon, H. & Lidberg, W. (2021). Use of multiple LIDAR-derived digital terrain indices and machine learning for high-resolution national-scale soil moisture mapping of the Swedish forest landscape. *Geoderma*, 404. https://doi.org/10.1016/j.geoderma.2021.115280

Ahmad Suhaimi, S.S., Blair, G.S. & Jarvis, S.G. (2021). Integrated species distribution models: A comparison of approaches under different data quality scenarios. *Diversity and Distributions*, 27 (6), 1066–1075. https://doi.org/10.1111/ddi.13255

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press, Cambridge.

Arrhenius, O. (1921). Species and area. *Journal of Ecology*, 9, 95–99. https://doi.org/10.2307/2255763

Askne, J., Fransson, J., Santoro, M., Soja, M. & Ulander, L. (2013). Model-Based Biomass Estimation of a Hemi-Boreal Forest from Multitemporal TanDEM-X Acquisitions. *Remote Sensing*, 5 (11), 5574–5597. https://doi.org/10.3390/rs5115574

Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D. & Shah, R. (2010). Spatial logistic regression and change-of-support in poisson point processes. *Electronic Journal of Statistics*, 4, 1151–1201. https://doi.org/10.1214/10-EJS581

Baddeley, A., Rubak, E. & Turner, R. (2016). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press, Boca Raton. https://doi.org/10.1201/b19708

Baddeley, A.J., Møller, J. & Waagepetersen, R. (2000). Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica*, 54 (3), 329–350. https://doi.org/10.1111/1467-9574.00144

Bartlett, M.S. (1948). Determination of Plant Densities. *Nature*, 162 (4120), 621. https://doi.org/10.1038/162621a0

Batista, J.L.F. & Maguire, D.A. (1998). Modeling the spatial structure of topical forests. *Forest Ecology and Management*, 110 (1–3), 293–314. https://doi.org/10.1016/S0378-1127(98)00296-5

Blackman, G.E. (1935). A study by statistical methods of the distribution of species in grassland associations. *Annals of Botany*, os-49 (4), 749–777. https://doi.org/10.1093/oxfordjournals.aob.a090534

Bonham, C.D. (2013). *Measurements for Terrestrial Vegetation: Second Edition*. Wiley, New York. https://doi.org/10.1002/9781118534540

Bradter, U., Mair, L., Jönsson, M., Knape, J., Singer, A. & Snäll, T. (2018). Can opportunistically collected Citizen Science data fill a data gap for habitat suitability models of less common species? *Methods in Ecology and Evolution*, 9 (7), 1667–1678. https://doi.org/10.1111/2041-210X.13012

Bullock, E., Healey, S., Yang, Z., Acosta, R., Villalba, H., Insfran, K., Melo, J., Wilson, S., Duncanson, L., Naesset, E., Armston, J., Saarela, S., Stahl, G., Patterson, P. & Dubayah, R. (2023). Estimating aboveground biomass density using hybrid statistical inference with GEDI lidar data and Paraguay's national forest inventory. *Environmental Research Letters*, 18 (8). https://doi.org/10.1088/1748-9326/acdf03

Cain, S.A. & Castro, G.M.O. (1959). *Manual of Vegetation Analysis*. Harper, New York.

Cassel, C.M., Särndal, C.E. & Wretman, J.H. (1977). *Foundations of inference in survey sampling*. John Wiley & Sons.

Chang, Y.-M. & Huang, Y.-C. (2024). Estimating Species Abundance from Presence–Absence Maps by Kernel Estimation. *Journal of Agricultural, Biological and Environmental Statistics*, 29 (4), 812–830. https://doi.org/10.1007/s13253-023-00589-4

Commission of the European Communities (2003). Council directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora.

Condés, S. & McRoberts, R.E. (2017). Updating national forest inventory estimates of growing stock volume using hybrid inference. *Forest Ecology and Management*, 400, 48–57. https://doi.org/10.1016/j.foreco.2017.04.046

Conlisk, E., Conlisk, J. & Harte, J. (2007). The impossibility of estimating a negative binomial clustering parameter from presence-absence data: A comment on He and Gaston. *American Naturalist*, 170 (4), 651–654. https://doi.org/10.1086/521339

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics and Probability*

Letters, 18 (5), 353–362. https://doi.org/10.1016/0167-7152(93)90028-H

Corona, P., Fattorini, L., Franceschi, S., Scrinzi, G. & Torresan, C. (2014). Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: Model-based, Design-based, And hybrid perspectives. *Canadian Journal of Forest Research*, 44 (11), 1303–1311. https://doi.org/10.1139/cjfr-2014-0203

Cretois, B., Simmonds, E., Linnell, J., van Moorter, B., Rolandsen, C., Solberg, E., Strand, O., Gundersen, V., Roer, O. & Rod, J. (2021). Identifying and correcting spatial bias in opportunistic citizen science data for wild ungulates in Norway. *Ecology and Evolution*, 11 (21), 15191–15204. https://doi.org/10.1002/ece3.8200

Daley, D.J. & Vere-Jones, D. (2008). *An introduction to the theory of point processes (2nd ed.).* Springer Verlag, New York.

Di Biase, R., Marcelli, A., Franceschi, S., Bartolini, A. & Fattorini, L. (2022). Design-based mapping of plant species presence, association, and richness by nearest-neighbour interpolation. *Spatial Statistics*, 51. https://doi.org/10.1016/j.spasta.2022.100660

Diggle, P., Besag, J. & Gleaves, J. (1976). Statistical Analysis of Spatial Point Patterns by Means of Distance Methods. *Biometrics*, 32 (3), 659–667. https://doi.org/10.2307/2529754

Duan, N., Manning, W.G.Jr., Morris, C.N. & Newhouse, J.P. (1983). A comparison of alternative models for the demand for medical care (Corr: V2 P413). *Journal of Business and Economic Statistics*, 1, 115–126. https://doi.org/10.2307/1391852

Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Blair, J.B., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G. & Luthcke, S. (2022). GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, 17 (9). https://doi.org/10.1088/1748-9326/ac8694

Dudík, M., Schapire, R.E. & Phillips, S.J. (2005). Correcting sample selection bias in maximum entropy density estimation. *Proceedings of Advances in Neural Information Processing Systems*, 2005. 323–330

Eichhorn, M. (2010). Pattern reveals process: spatial organisation of a Kamchatkan stone birch forest. *Plant Ecology & Diversity*, 3 (3), 281–288. https://doi.org/10.1080/17550874.2010.528804

Ekström, M., Sandring, S., Grafström, A., Esseen, P.A., Jonsson, B.G. & Ståhl, G. (2020). Estimating density from presence/absence data in

clustered populations. *Methods in Ecology and Evolution*, 11 (3), 390–402. https://doi.org/10.1111/2041-210X.13347

Elzinga, C.L., Salzer, D.W. & Willoughby, J.W. (1998). *Measuring and Monitoring Plant Populations. BLM Technical Reference 1730-1.* BLM National Applied Resource Sciences Center, Denver.

Ene, L.T., Næsset, E. & Gobakken, T. (2013). Model-based inference for k-nearest neighbours predictions using a canonical vine copula. *Scandinavian Journal of Forest Research*, 28 (3), 266–281. https://doi.org/10.1080/02827581.2012.723743

Fattorini, L., Di Biase, R., Giuliarelli, D., Marcheselli, M., Pisani, C. & Corona, P. (2019). Mapping the diversity of forest attributes: a design-based approach. *Canadian Journal of Forest Research*, 49 (2), 190–197. https://doi.org/10.1139/cjfr-2018-0204

Fiorentin, L., Bonat, W., Pelissari, A., Machado, S., Téo, S. & Orso, G. (2019). Generalized Linear Models for Tree Survival in Lobolly Pine Plantations. *Cerne*, 25 (4), 347–356. https://doi.org/10.1590/01047760201925042649

Fithian, W., Elith, J., Hastie, T. & Keith, D.A. (2015). Bias correction in species distribution models: Pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6 (4), 424–438. https://doi.org/10.1111/2041-210X.12242

Fleischer, F., Eckel, S., Schmid, I. & Kazda, M. (2006). Point process modelling of root distribution in pure stands of Fagus sylvatica and Picea abies. *Canadian Journal of Forest Research*, 36 (1), 227–237. https://doi.org/10.1139/X05-232

Fletcher, R.J., Hefley, T.J., Robertson, E.P., Zuckerberg, B., McCleery, R.A. & Dorazio, R.M. (2019). A practical guide for combining data to model species distributions. *Ecology*, 100 (6). https://doi.org/10.1002/ecy.2710

Foody, G.M. (2008). Refining predictions of climate change impacts on plant species distribution through the use of local statistics. *Ecological Informatics*, 3 (3), 228–236. https://doi.org/10.1016/j.ecoinf.2008.02.002

Fortin, M., Bédard, S., DeBlois, J. & Meunier, S. (2008). Predicting individual tree mortality in northern hardwood stands under uneven-aged management in southern Quebec, Canada. *Annals of Forest Science*, 65 (2). https://doi.org/10.1051/forest:2007088

Fortin, M., Manso, R. & Calama, R. (2016). Hybrid estimation based on mixed-effects models in forest inventories. *Canadian Journal of Forest Research*, 46 (11), 1310–1319. https://doi.org/10.1139/cjfr-2016-0298

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H. & Ståhl, G. (2014). Adapting National Forest Inventories to changing requirements - The case of the Swedish National Forest Inventory at the turn of the 20th century. *Silva Fennica*, 48 (3). https://doi.org/10.14214/sf.1095

Gallegos Torell, Å. & Glimskär, A. (2009). Computer-aided calibration for visual estimation of vegetation cover. *Journal of Vegetation Science*, 20 (6), 973–983. https://doi.org/10.1111/j.1654-1103.2009.01111.x

Gaston, K.J., He, F., Maguran, A. & McGill, B. (2011). Species occurrence and occupancy. *Biological diversity: frontiers in measurement and assessment*, 141–151

Gelfand, A.E. & Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 89 (3). https://doi.org/10.1002/ecm.1372

Gobakken, T., Næsset, E., Nelson, R., Bollandsås, O., Gregoire, T., Ståhl, G., Holm, S., Orka, H. & Astrup, R. (2012). Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning. *Remote Sensing of Environment*, 123, 443–456. https://doi.org/10.1016/j.rse.2012.01.025

Godínez-Alvarez, H., Herrick, J.E., Mattocks, M., Toledo, D. & Van Zee, J. (2009). Comparison of three vegetation monitoring methods: Their relative utility for ecological assessment and monitoring. *Ecological Indicators*, 9 (5), 1001–1008. https://doi.org/10.1016/j.ecolind.2008.11.011

Grafström, A., Lundström, N.L.P. & Schelin, L. (2012). Spatially Balanced Sampling through the Pivotal Method. *Biometrics*, 68 (2), 514–520. https://doi.org/10.1111/j.1541-0420.2011.01699.x

Grafström, A. & Matei, A. (2018). Spatially Balanced Sampling of Continuous Populations. *Scandinavian Journal of Statistics*, 45 (3), 792–805. https://doi.org/10.1111/sjos.12322

Gregoire, T.G. & Valentine, H.T. (2007). *Sampling strategies for natural resources and the environment*. Chapman & Hall/CRC Press, Boca Raton.

Greig-Smith, P. (1983). *Quantitative Plant Ecology, 3rd Edition*. University of California Press.

Guttorp, P. & Thorarinsdottir, T.L. (2012). What Happened to Discrete Chaos, the Quenouille Process, and the Sharp Markov Property? Some History of Stochastic Point Processes. *International Statistical Review*, 80 (2), 253–268. https://doi.org/10.1111/j.1751-5823.2012.00181.x

Hahn, U., Jensen, E., van Lieshout, M. & Nielsen, L. (2003). Inhomogeneous spatial point processes by location-dependent scaling. *Advanced in Applied Probability*, 35 (2), 319–336. https://doi.org/10.1239/aap/1051201648

He, F. & Gaston, K.J. (2000). Estimating species abundance from occurrence. *American Naturalist*, 156 (5), 553–559. https://doi.org/10.1086/303403

He, F. & Gaston, K.J. (2007). Estimating abundance from occurrence: An underdetermined problem. *American Naturalist*, 170 (4), 655–659. https://doi.org/10.1086/521340

Holt, A.R., Gaston, K.J. & He, F. (2002). Occupancy-abundance relationships and spatial distribution: A review. *Basic and Applied Ecology*, 3 (1), 1–13. https://doi.org/10.1078/1439-1791-00083

Horvitz, D.G. & Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47 (260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Hou, Z., Xu, Q., McRoberts, R., Greenberg, J., Liu, J., Heiskanen, J., Pitkänen, S. & Packalen, P. (2017). Effects of temporally external auxiliary data on model-based inference. *Remote Sensing of Environment*, 198, 150–159. https://doi.org/10.1016/j.rse.2017.06.013

Hwang, W.H. & He, F. (2011). Estimating abundance from presence/absence maps. *Methods in Ecology and Evolution*, 2 (5), 550–559. https://doi.org/10.1111/j.2041-210X.2011.00105.x

Jaynes, E.T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, 106 (4), 620–630. https://doi.org/10.1103/PhysRev.106.620

Johnston, A., Matechou, E. & Dennis, E. (2023). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*, 14 (1), 103–116. https://doi.org/10.1111/2041-210X.13834

Johnston, A., Moran, N., Musgrove, A., Fink, D. & Baillie, S. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 422. https://doi.org/10.1016/j.ecolmodel.2019.108927

Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K.T., Maltamo, M., Nilsson, M., Nord-Larsen, T., Næsset, E. & Olsson, H. (2018). Remote sensing and forest inventories in Nordic countries – roadmap for the future. *Scandinavian Journal of Forest Research*, 33 (4), 397–412. https://doi.org/10.1080/02827581.2017.1416666

Kylin, H. (1926). Uber Begriffsbildung und Statistik in der Pflanzensociologie. *Botaniska notiser*, ii. (81)

Lidberg, W., Nilsson, M. & Agren, A. (2020). Using machine learning to generate high-resolution wet area maps for planning forest management: A study in a boreal forest landscape. *Ambio*, 49 (2), 475–486. https://doi.org/10.1007/s13280-019-01196-9

Lindenmayer, D.B., Welsh, A., Donnelly, C., Crane, M., Michael, D., Macgregor, C., McBurney, L., Montague-Drake, R. & Gibbons, P. (2009). Are nest boxes a viable alternative source of cavities for hollow-dependent animals? Long-term monitoring of nest box occupancy, pest use and attrition. *Biological Conservation*, 142 (1), 33–42. https://doi.org/10.1016/j.biocon.2008.09.026

Lindgren, N., Olsson, H., Nyström, K., Nyström, M. & Ståhl, G. (2021). Data Assimilation of Growing Stock Volume Using a Sequence of Remote Sensing Data from Different Sensors. *Canadian Journal of Remote Sensing*, 48 (2), 127–143. https://doi.org/10.1080/07038992.2021.1988542

Mäkinen, J., Merow, C. & Jetz, W. (2024). Integrated species distribution models to account for sampling biases and improve range-wide occurrence predictions. *Global Ecology and Biogeography*, 33 (3), 356–370. https://doi.org/10.1111/geb.13792

Marcelli, A., Corona, P. & Fattorini, L. (2019). Design-based estimation of mark variograms in forest ecosystem surveys. *Spatial Statistics*, 30, 27–38. https://doi.org/10.1016/j.spasta.2019.02.002

Matérn, B. (1960). Spatial variation. *Meddelanden Fran Statens Skogsforskningsinstitut*, 49 (5), 1–144. https://res.slu.se/id/publ/125179

Matérn, B. (1986). *Spatial variation. Lecture notes in statistics 36*. Springer Verlag.

McCullagh, P. (1989). *Generalized Linear Models*. 2nd ed. Springer US.

McRoberts, R., Næsset, E., Gobakken, T., Chirici, G., Condés, S., Hou, Z., Saarela, S., Chen, Q., Ståhl, G. & Walters, B. (2018). Assessing components of the model-based mean square error estimator for remote sensing assisted forest applications. *Canadian Journal of Forest Research*, 48 (6), 642–649. https://doi.org/10.1139/cjfr-2017-0396

McRoberts, R.E., Næsset, E., Liknes, G.C., Chen, Q., Walters, B.F., Saatchi, S. & Herold, M. (2019). Using a Finer Resolution Biomass Map to Assess the Accuracy of a Regional, Map-Based Estimate of Forest Biomass. *Surveys in Geophysics*, 40 (4), 1001–1015. https://doi.org/10.1007/s10712-019-09507-1

Mehtätalo, L. & Lappi, J. (2020). *Biometry for Forestry and Environmental Data : With Examples in R*. Chapman and Hall/CRC, Boca Raton.

Min, Y. & Agresti, A. (2002). Modeling Nonnegative Data with Clumping at Zero: A Survey. *Journal of the Iranian Statistical Society*, 1 (1–2), 7–33

Møller, J. & Waagepetersen, R. P., R. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. CRC Press.

Morrison, D.A., Le Brocque, A.F. & Clarke, P.J. (1995). An assessment of some improved techniques for estimating the abundance (frequency) of sedentary organisms. *Vegetatio*, 120 (2), 131–145. https://doi.org/10.1007/BF00034343

Mrkvička, T. (2014). Distinguishing Different Types of Inhomogeneity in Neyman–Scott Point Processes. *Methodology and Computing in Applied Probability*, 16 (2), 385–395. https://doi.org/10.1007/s11009-013-9365-4

Mukhopadhyay, R., Ekström, M., Lindberg, E., Persson, H., Saarela, S. & Nilsson, M. (2024). Computation of prediction intervals for forest aboveground biomass predictions using generalized linear models in a large-extent boreal forest region. *Forestry*, cpae006, 1–11. https://doi.org/10.1093/forestry/cpae006

Næsset, E., Ørka, H.O., Solberg, S., Bollandsås, O.M., Hansen, E.H., Mauya, E., Zahabu, E., Malimbwi, R., Chamuya, N., Olsson, H. & Gobakken, T. (2016). Mapping and estimating forest area and aboveground biomass in miombo woodlands in Tanzania using data from airborne laser scanning, TanDEM-X, RapidEye, and global forest maps: A comparison of estimated precision. *Remote Sensing of Environment*, 175, 282–300. https://doi.org/10.1016/j.rse.2016.01.006

Neyman, J. & Scott, E. (1952). A Theory of the Spatial Distribution of Galaxies. *Astrophysical Journal*, 116 (1), 144–163. https://doi.org/10.1086/145599

Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J. & Olsson, H. (2017). A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. *Remote Sensing of Environment*, 194, 447–454. https://doi.org/10.1016/j.rse.2016.10.022

Ogata, Y. (2020). Cluster analysis of spatial point patterns: posterior distribution of parents inferred from offspring. *Japanese Journal of Statistics and Data Science*, 3 (1), 367–390. https://doi.org/10.1007/s42081-019-00065-9

Pacifici, K., Reich, B.J., Miller, D.A.W., Gardner, B., Stauffer, G., Singh, S., McKerrow, A. & Collazo, J.A. (2017). Integrating multiple data sources in species distribution modeling: A framework for data fusion. *Ecology*, 98 (3), 840–850. https://doi.org/10.1002/ecy.1710

Pellissier, V., Bergès, L., Nedeltcheva, T., Schmitt, M., Avon, C., Cluzeau, C. & Dupouey, J. (2013). Understorey plant species show long-range spatial patterns in forest patches according to distance-to-edge. *Journal of Vegetation Science*, 24 (1), 9–24. https://doi.org/10.1111/j.1654-1103.2012.01435.x

Philippi, T. (2005). Adaptive Cluster Sampling for Estimation of Abundances within Local Populations of Low-Abundance Plants. *Ecology*, 86 (5), 1091–1100. https://doi.org/10.1890/04-0621

Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E. & Blair, M.E. (2017). Opening the black box: an open-source release of Maxent. *Ecography*, 40 (7), 887–893. https://doi.org/10.1111/ecog.03049

Pielou, E. (1957). The Effect of Quadrat Size on the Estimation of the Parameters of Neyman and Thomas Distributions. *Journal of Ecology*, 45 (1), 31–47. https://doi.org/10.2307/2257075

Prokešová, M. (2010). Inhomogeneity in Spatial Cox Point Processes – Location Dependent Thinning Is Not the Only Option. *Image Analysis & Stereology*, 29 (3). https://doi.org/10.5566/ias.v29.p133-141

R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rao, C.R. (1973). *Linear statistical Inference and its Applications*. Wiley, New York.

Raunkiaer, C. (1934). *The life forms of plants and statistical plant geography; being the collected papers of C. Raunkiaer. The life forms of plants and statistical plant geography; being the collected papers of C. Raunkiaer.*. Oxford: Clarendon Press.

Reese, H., Nilsson, M., Pahlén, T.G., Hagner, O., Joyce, S., Tingelöf, U., Egberth, M. & Olsson, H. (2003). Countrywide Estimates of Forest Variables Using Satellite Data and Field Data from the National Forest Inventory. *Ambio*, 32 (8), 542–548. https://doi.org/10.1579/0044-7447-32.8.542

Renner, I.W. & Warton, D.I. (2013). Equivalence of MAXENT and Poisson Point Process Models for Species Distribution Modeling in Ecology. *Biometrics*, 69 (1), 274–281. https://doi.org/10.1111/j.1541-0420.2012.01824.x

Ringvall, A., Petersson, H., Ståhl, G. & Lämås, T. (2005). Surveyor consistency in presence/absence sampling for monitoring vegetation in a boreal forest. *Forest Ecology and Management*, 212 (1–3), 109–117. https://doi.org/10.1016/j.foreco.2005.03.002

Robinson, O., Ruiz-Gutierrez, V. & Fink, D. (2018). Correcting for bias in distribution modelling for rare species using citizen science data. *Diversity and Distributions*, 24 (4), 460–472. https://doi.org/10.1111/ddi.12698

Rochefort, L., Isselin-Nondedeu, F., Boudreau, S. & Poulin, M. (2013). Comparing survey methods for monitoring vegetation change through time in a restored peatland. *Wetlands Ecology and Management*, 21 (1), 71–85. https://doi.org/10.1007/s11273-012-9280-4

Royle, J., Chandler, R., Yackulic, C. & Nichols, J. (2012). Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, 3 (3), 545–554. https://doi.org/10.1111/j.2041-210X.2011.00182.x

Saarela, S., Grafström, A., Ståhl, G., Kangas, A., Holopainen, M., Tuominen, S., Nordkvist, K. & Hyyppä, J. (2015). Model-assisted estimation of growing stock volume using different combinations of LiDAR and Landsat data as auxiliary information. *Remote Sensing of Environment*, 158, 431–440. https://doi.org/10.1016/j.rse.2014.11.020

Saarela, S., Healey, S.P., Yang, Z., Roald, B.-E., Patterson, P.L., Gobakken, T., Næsset, E., Hou, Z., McRoberts, R.E. & Ståhl, G. (2025). A Separable Bootstrap Variance Estimation Algorithm for Hierarchical Model-Based Inference of Forest Aboveground Biomass Using Data From NASA's GEDI and Landsat Missions. *Environmetrics*, 36 (1). https://doi.org/10.1002/env.2883

Saarela, S., Holm, S., Healey, S., Andersen, H., Petersson, H., Prentius, W., Patterson, P., Næsset, E., Gregoire, T. & Ståhl, G. (2018). Generalized Hierarchical Model-Based Estimation for Aboveground Biomass Assessment Using GEDI and Landsat Data. *Remote Sensing*, 10 (11). https://doi.org/10.3390/rs10111832

Saarela, S., Holm, S., Healey, S.P., Patterson, P.L., Yang, Z., Andersen, H.E., Dubayah, R.O., Qi, W., Duncanson, L.I., Armston, J.D., Gobakken, T., Næsset, E., Ekström, M. & Ståhl, G. (2022). Comparing frameworks for biomass prediction for the Global Ecosystem Dynamics Investigation. *Remote Sensing of Environment*, 278. https://doi.org/10.1016/j.rse.2022.113074

Schulze, E.-D., Beck, E., Buchmann, N., Clemens, S., Müller-Hohenstein, K. & Scherer-Lorenzen, M. (2019). *Plant Ecology, Second Edition*. 926. https://doi.org/10.1007/978-3-662-56233-8

Sen, P.K. & Singer, J.M. (1993). *Large sample methods in statistics: An introduction with applications*. Chapman & Hall, New York.

Shao, J. (2003). *Mathematical Statistics*. Springer Verlag, New York.

Sipek, M., Horvat, E., Kosic, I. & Sajna, N. (2022). Presence of Alien Prunus Serotina and Impatiens Parviflora in Lowland Forest Fragments in NE Slovenia. *Sumarski List*, 146 (5–6), 215–224. https://doi.org/10.31298/sl.146.5-6.4

Ståhl, G. (2003). Presence/absence sampling as a substitute for cover assessment in vegetation monitoring. *Advances in Forest Inventory for Sustainable Forest Management and Biodiversity Monitoring*, 137–142

Ståhl, G., Allard, A., Esseen, P.A., Glimskär, A., Ringvall, A., Svensson, J., Sundquist, S., Christensen, P., Torell, Å.G., Högström, M., Lagerqvist, K., Marklund, L., Nilsson, B. & Inghe, O. (2011a). National Inventory of Landscapes in Sweden (NILS)-scope, design, and experiences from establishing a multiscale biodiversity monitoring system. *Environmental Monitoring and Assessment*, 173 (1–4), 579–595. https://doi.org/10.1007/s10661-010-1406-7

Ståhl, G., Ekström, M., Dahlgren, J., Esseen, P.A., Grafström, A. & Jonsson, B.G. (2017). Informative plot sizes in presence-absence sampling of forest floor vegetation. *Methods in Ecology and Evolution*, 8 (10), 1284–1291. https://doi.org/10.1111/2041-210X.12749

Ståhl, G., Ekström, M., Dahlgren, J., Esseen, P.-A., Grafström, A., Jonsson, B.-G. & Molofsky, J. (2020). Presence–absence sampling for estimating plant density using survey data with variable plot size. *Methods in Ecology and Evolution*, 11 (4), 580–590. https://doi.org/10.1111/2041-210x.13348

Ståhl, G., Gobakken, T., Saarela, S., Persson, H.J., Ekström, M., Healey, S.P., Yang, Z., Holmgren, J., Lindberg, E., Nyström, K., Papucci, E., Ulvdal, P., Ørka, H.O., Næsset, E., Hou, Z., Olsson, H. & McRoberts, R.E. (2024). Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time – and how this affects applications. *Forest Ecosystems*, 11. https://doi.org/10.1016/j.fecs.2023.100164

Ståhl, G., Holm, S., Gregoire, T.G., Gobakken, T., Næsset, E. & Nelson, R. (2011b). Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark county, Norway. *Canadian Journal of Forest Research*, 41 (1), 96–107. https://doi.org/10.1139/X10-161

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E. & Gregoire, T.G. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems*, 3 (1). https://doi.org/10.1186/s40663-016-0064-9

Stoklosa, J., Blakey, R.V. & Hui, F.K.C. (2022). An Overview of Modern Applications of Negative Binomial Modelling in Ecology and Biodiversity. *Diversity*, 14 (5). https://doi.org/10.3390/d14050320

Swindel, B.F. (1983). Choice of size and number of quadrats to estimate density from frequency in Poisson and binomially dispersed populations. *Biometrics*, 39 (2), 455–464. https://doi.org/10.2307/2531016

Takashina, N., Kusumoto, B., Beger, M., Rathnayake, S. & Possingham, H.P. (2018). Spatially explicit approach to estimation of total population abundance in field surveys. *J Theor Biol*, 453, 88–95. https://doi.org/10.1016/j.jtbi.2018.05.013

The European Commission (2020). EU Biodiversity Strategy for 2030 (COM/2020/380)

The European Commission (2021). New EU Forest Strategy for 2030 (COM/2021/572)

Thomas, M. (1949). A generalization of Poisson's binomial limit for use in ecology. *Biometrika*, 36 (Pt. 1-2), 18–25. https://doi.org/10.1093/biomet/36.1-2.18

Thompson, S.K. (2012). *Sampling, 3rd Edition*. Wiley.

Tomppo, E., Gschwantner, T., Lawrence, M. & McRoberts, R.E. (2010). Preface. *National Forest Inventories: Pathways for Common Reporting*, v–vi. https://doi.org/10.1007/978-90-481-3233-1

Trijoulet, V., Albertsen, C., Kristensen, K., Legault, C., Miller, T. & Nielsen, A. (2023). Model validation for compositional data in stock assessment models: Calculating residuals with correct properties. *Fisheries Research*, 257. https://doi.org/10.1016/j.fishres.2022.106487

Waagepetersen, R.P. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics*, 63 (1), 252–258. https://doi.org/10.1111/j.1541-0420.2006.00667.x

Wallerman, J., Axensten, P., Egberth, M., Janzen, J., Sandstrom, E., Fransson, J.E.S. & Nilsson, M. (2021). SLU Forest Map - Mapping Swedish Forests Since Year 2000 In: Proceedings of IGARSS 2021, Crossing Borders, Virtual Symposium, Brussels, Belgium, 11-16 July, 2021, pp. 6056–6059., 2021.

Wintle, B.A., Elith, J. & Potts, J.M. (2005). Fauna habitat modelling and mapping: A review and case study in the Lower Hunter Central Coast region of NSW. *Austral Ecology*, 30 (7), 719–738. https://doi.org/10.1111/j.1442-9993.2005.01514.x

Yee, T. & Dirnböck, T. (2009). Models for analysing species' presence/absence data at two time points. *Journal of Theoretical Biology*, 259 (4), 684–694. https://doi.org/10.1016/j.jtbi.2009.05.004

Popular science summary

As the environment continues to change due to global warming, land use, and other human impacts, keeping an eye on plant populations has become increasingly important. Environmental monitoring programmes, such as the Swedish national forest inventory (NFI) and the national inventory of landscapes in Sweden (NILS), regularly collect large amounts of data pertaining to forests and landscapes that can be used in meaningful environmental analyses.

A particular sampling method that is used in such programmes and that has a large, but not fully exploited potential is the one where only presence or absence of a specific plant species is registered in each sample plot. That method is called presence/absence (P/A) sampling. This method is simple, time- and cost-effective, and easier to carry out compared to many traditional survey methods.

By combining mathematical models with P/A data and environmental covariate data from the NFI or remote sensing, plant density (defined as the number of plants per unit area) is estimated for several forest species, both locally and for entire regions. Different assumptions about the plant populations are considered (whether the plant individuals are randomly scattered or grow in groups). The developed approaches can also account for how environmental factors, such as surrounding trees or soil moisture, might influence the abundance of plants. In short, these techniques help turn simple yes-or-no data into valuable insights about where plants are likely to grow and how they interact with their environment. The methods were applied to both real-world and simulated data and showed promising results.

Nevertheless, the proposed estimates of plant density cannot be trusted blindly. There is some uncertainty at play whenever such values are presented. Errors can emerge from the mathematical models, the remote sensing products, the field measurements, and many more. That is why it is important that a measure of uncertainty is presented in connection with these estimates. In this thesis, it is supposed that the P/A and covariate data are error-free, for instance that there were no measurement errors and that the presences or absences of plants were registered correctly, although this is a simplification of reality.

In most studies, as in Papers I, I and III in this thesis, uncertainty is presented by means of the variance of the estimator. The variance expresses

how an estimator can vary. However, more uncertainty sources can arise when applying mathematical models. The study of the extent of these uncertainty components is the main objective of Paper IV, with a case study focused on biomass based on simulations. The results show that the variance can be used as a suitable approximation of uncertainty when the studies occur on a large area, whereas additional measures of uncertainty need to be taken into account when the study area is small.

Populärvetenskaplig sammanfattning

När miljön förändras på grund av global uppvärmning, markanvändning och annan mänsklig påverkan har övervakning av växtpopulationer blivit allt viktigare. Miljöövervakningsprogram, som Riksskogstaxeringen och Nationella Inventeringar av Landskapet i Sverige (NILS), samlar rutinmässigt in omfattande datamängder relaterade till skogar och landskap som kan användas för värdefulla miljöanalyser.

En specifik inventeringsmetod som används i sådana program och som har stora men relativt outnyttjade fördelar kallas närvaro/frånvaro. Den innebär att det i varje provyta endast registreras om en viss växtart registreras förekommer eller inte. Metoden är enkel, tidseffektiv och med mindre risk för att olika personer gör sinsemellan olika bedömningar än för många andra konventionella inventeringsmetoder.

Genom att kombinera matematiska modeller med data om närvaro/frånvaro samt andra data som beskriver miljön uppskattas planttätheten (definierad som antalet plantor per ytenhet) för flera skogsarter, både lokalt och för stora områden. Miljödata kommer från olika källor som t.ex. Riksskogstaxeringen eller fjärranalys. Olika förhållanden avseende växtpopulationer beaktas, t.ex. om växterna är slumpmässigt utspridda eller grupperade. De utvecklade metoderna kan också ta hänsyn till hur miljöfaktorer, såsom omgivande träd eller markfuktighet, påverkar planttätheten. I korthet kan dessa metoder, baserade på inventeringar som registrerar närvaro/frånvaro av arter, ge värdefulla insikter om planttäthet och hur den påverkas av sin omgivning. Metoderna har tillämpats på både faktiska fältdata och simulerade data och har gett lovande resultat.

De erhållna skattningarna av planttäthet bör dock inte ses som sanningar utan vidare granskning. Det finns osäkerheter i skattningarna och fel kan uppstå från matematiska modeller, kartprodukter framtagna med hjälp av fjärranalysdata, fältmätningar och från andra källor. Det är därför det är viktigt skattningarna också presenteras tillsammans med en uppskattning av osäkerheten, vilket också är en viktig del av avhandlingen. I denna avhandling antas dock att utnyttjade data är utan fel, exempelvis att de inte är behäftade med mätfel och att det inte finns några felregistreringar av frånvaro och närvaro av växtarter, vilket är en förenkling av verkligheten.

För många studier, inklusive papper I, II och III i denna avhandling, presenteras osäkerheten genom skattningars varians. En varians ger ett mått

på hur en skattning kan variera. Dock kan fler osäkerhetskällor förekomma när matematiska modeller tillämpas. Att undersöka omfattningen av dessa är i huvudfokus i papper IV, som är en fristående studie med fokus på biomassa och som bygger på simuleringar. Den studien visar att variansen för t.ex. en uppskattad mängd biomassa ger en rätt god approximation av osäkerhet vid storskaliga undersökningar, medan andra osäkerhetskomponenter kan få större betydelse ifall studieområdet är mindre.

Acknowledgements

To start with, a big thank you to my main supervisor Magnus Ekström. I really think you've been the best supervisor I could have thought of. Thanks for your patience, your help, your understanding, your amazing pedagogical skills, but also for the fascinating non-work related discussions. Of course, this big thank you includes my other supervisors, Göran Ståhl, Saskia Sandring and Bege Jonsson, for the constant help with my studies. My thanks go also to Jörgen Wallerman, who wasn't officially part of the supervising team but was definitely considered as part of the team. This wouldn't have been possible without you all! You were always available for me, gave me lots of insights (and comments) and I gained invaluable knowledge thanks to you all. We also had fun during the field excursions. Thanks as well to Lucy, my cat supervisor, for being present at (almost) all our supervision meetings.

There have been many people helping me at SLU, with my studies as well as morally, along the way. I'd like to thank some people in particular. Thanks to Hilda, for the tremendous musical discussions we've had (and discussions about our favourite musical genres… you still haven't converted me to "1-minute" punk music, sorry!); thanks to the "metal club" at SRH, especially Sabina and Mateusz (let's go to more gigs together, and expand the metal culture at SLU and worldwide!); thanks to Cornelia, for the possibility to speak my mother tongue once in a while; thanks to Olivia, for making me play the greatest video game ever (you know which one I'm referring to ☺); thanks to Emanuele, for being always so funny; thanks to Mariana for the varied discussions (and for being the first person that likes Visual Kei I meet in real life!); thanks to Felix, for being my only queer ally at the department; thanks to Kohsuke, for the practice in Japanese; thanks to Svetlana, for keeping in touch even after you left SLU; thanks to the remainder of the statistics team: Wilmer, Fabio and Anton (Oh, by the way, did you know that there's a strong correlation between being a statistician and playing the bass? Hilda, Felix, Wilmer and I proved it). Thanks to Carl for the last-minute help with the kappa (aaah, these cursed programmes that are always up to mischief at the worst moments…). Thanks to my co-authors that aren't included in my supervision group. And I guess thanks to some of my other PhD student colleagues. Thanks to all the people that have helped me with my studies, the logistics (for the field study), the courses, the programming, the administration (I promise I'll settle on a first name soon enough… I

already decided on my pronoun)… The list goes on and on, it's difficult to name everyone…

Luckily, I also got moral support from people outside of SLU. If it weren't for them, I probably wouldn't have survived in this dark, snowy, remote place (just joking… or not?). A gigantic thank you to my band mates Jakob, Nicklas, Olof, Jonathan, Karo, Lisa, Gena, Natalia and Tony. Music is my life, and playing with you guys has been a blast. I hope our adventures keep on going forever! Look out for our next gigs and releases (shameless self-promotion detected). Thanks to solitude, for giving me inspiration… I also thank my family for the support and for having put up with me during all these years…

Finally, thanks to the fuel of my life, music (wasn't it clear already?). Thanks to (old-school) Visual Kei, Japanese Rock/Metal, but also non-Japanese Metal (not Nu-metal tho). Special mention to Black Metal, which I started listening to while staying here (the environment sure helped: endless forests, the darkness, the harsh and cold winters…). It's an acquired taste…

I

Large-area estimation of plant density using presence/absence data and binary regression, and correlation tests of the binary regression model

Léna Gozé[a,1], Magnus Ekström[a,b], Jonas Dahlgren[a], Bengt Gunnar Jonsson[c,d], Saskia Sandring[a], Göran Ståhl[a]

[a] Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

[b] Department of Statistics, USBE, Umeå University, SE-901 87 Umeå, Sweden

[c] Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden

[d] Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

Abstract. Inventories of populations are of great importance in ecological research. For monitoring plants, presence/absence (P/A) sampling is simple to conduct, since only presence or absence of a species on a plot needs to be registered. Estimates of expected plant density may be obtained from P/A data but need to be based on model assumptions about the spatial distribution of plants. Assuming that plant locations follow an inhomogeneous Poisson point process, we used binary regression modelling of P/A data to estimate expected plant density across large regions and subregions. This binary regression model implied by the inhomogeneous Poisson point process model assumes that the response variables are independent given the covariates. To validate this assumption, we compared standard correlation tests using different types of residuals from the fitted model. Using empirical data from the Swedish National Forest Inventory as well as simulated plant population data, we evaluated the performance of the suggested estimators of expected plant density and the correlation tests, respectively. The estimator of expected plant density performed relatively well, and the simulation results suggested that the correlation test using the Pearson correlation coefficient and Pearson residuals is preferable. The formal estimates could be useful for analysis of state and trends, as well as for reporting, for example in connection with the EU's habitats directive.

KEYWORDS: inhomogeneous Poisson point process, plant monitoring, remote sensing, vegetation survey, correlation test, residual study

1 | Introduction

Ground vegetation, being part of biodiversity and playing many functional roles, is a major component of forest ecosystems. Hence, it is important to correctly estimate plant species abundance and distribution. However, in contrast to trees, for many plant species the individual plants are difficult to count visually due to clonal growth patterns or the sheer number of individuals (Ståhl et al. 2017). Instead, inventories use vegetation cover or proportion of presences in sample plots (Godínez-Alvarez et al.

[1]Corresponding author: Léna Gozé (lena.goze@slu.se)

2009; Bonham 2013). Assessment of cover through visual inspection is a commonly used method, but is sensitive to observer judgement bias (Gallegos-Torell & Glimskär 2009).

Presence/absence (P/A) sampling is an alternative where only presence or absence of a species on a plot is registered (Elzinga et al. 1998; Bonham 2013), and is therefore simple and relatively inexpensive to conduct (Ståhl et al. 2017). When fairly small plots are used, studies suggest that the method is less influenced by surveyor judgement than is cover assessment (Ringvall et al. 2005). A drawback is that it does not generally provide information on plant density (mean number of plants per unit area) or plant cover, and plant occurrence frequencies are difficult to interpret due to their dependence on spatial occurrence patterns and plot size. Several studies stress the need for linking P/A data with plant density (e.g. Royle & Nichols 2003; Hwang & He 2011). Estimates of (expected) plant density may be obtained from P/A data, but such outputs need to be based on model assumptions regarding the spatial distribution of plants (Fithian et al. 2015; Ståhl et al. 2017, 2020; Gelfand & Shirota 2019; Ekström et al. 2020).

To take into account that the expected density of plants can be spatially varying, covariate data from remote sensing can be used in the modelling. The amount and quality of auxiliary covariate data available for modelling is currently increasing. Remote sensing data of different kinds are nowadays being made available at short intervals (Lindgren et al. 2021). For example, optical satellite data from sensors such as Sentinel-2 can be obtained several times during a vegetation season (Puliti et al. 2018). Airborne laser scanning data are obtained less frequently, but on the other hand such data provide important auxiliary information about both site and vegetation conditions (Lidberg et al. 2020). Auxiliary covariate data are also available in the form of raster maps of, for example, predicted forest state attributes (Nilsson et al. 2017; Wallerman et al. 2021) and soil moisture classes (Ågren et al. 2021). Consequently, the possibility to model species abundance based on auxiliary data has increased considerably over the last couple of decades.

The objective of this study was to use P/A and remote sensing data for estimating expected plant density for a selection of species and for large regions and subregions, together with valid estimates of variance. In order to do this, we have used the following components, while incorporating auxiliary information: i) inhomogeneous Poisson point process models for plant locations; ii) generalised linear models (GLMs) to link P/A data to expected plant density; and iii) estimation of expected plant density for species across large regions within the framework of model-based inference (Chambers & Clark 2012; Ståhl et al. 2016). In contrast to design-based model-assisted approaches, it is of importance to protect inference against model misspecification when performing model-based inference. When models are (approximately) correctly specified, model-based estimators can be very useful, but model misspecifications may result in severely biased estimators (Chambers et al. 2006; Ekström & Nilsson 2021). In principle, the binary regression model implied by the inhomogeneous Poisson point process assumes independent response variables given the covariates. Therefore, a particular objective of this study was to apply a test for assessing whether this holds true, and to evaluate the test as well as the suggested estimators of expected plant density using Monte Carlo simulations and empirical data from environmental monitoring. We evaluate different

standard correlation tests based on model residuals, keeping in mind that the choice of the type of residuals is a key factor to take into consideration.

2 | Estimation of expected number of plants and plant density

A basic assumption of model-based inference is that the population consists of (realisations of) random variables, following some specific model; for example, a model with covariate data derived from remote sensing (Ståhl et al. 2016). To construct an estimator of a certain attribute of the population, the area of interest may be tesselated into a finite number of population units. In our setting, a square tesselation is used, given by a cell grid with N cells, each of area a_P. For simplicity, we let the ith cell be represented by its label i. Thus, we denote the finite population of cells as $U = \{1, 2, \ldots, N\}$. Throughout this paper, it is assumed that the (remotely sensed) covariate vector $\boldsymbol{x}_i = (x_{i1}, x_{i2}, ..., x_{iq})'$ is constant within each cell i, where $x_{i1} = 1$ for all i (cf. Baddeley et al. 2010).

We want to estimate the expected number of points (plants) in an area of interest, or the corresponding expected number of points per area unit. For doing this, let us consider a Poisson point process with log intensity

$$\log \lambda_i = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i \tag{1}$$

in cell i, and let M_i denote the number of points falling in cell i. Thus, the point process is an inhomogeneous Poisson point process, assumed to be homogeneous within each cell (cf. Baddeley et al. 2010). Under the Poisson model, the M_i are independent Poisson random variables with means $E(M_i) = a_P \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i)$, $i \in U$. Thus, with M denoting the total number of points in all the N cell cells, its expected value can be written as

$$E(M) = a_P \sum_{i \in U} \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i). \tag{2}$$

Once the parameter vector is estimated by $\hat{\boldsymbol{\beta}}$, we can use

$$\widehat{E(M)} = a_P \sum_{i \in U} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i).$$

as an estimator of the expected number of points within the area covered by the N cells. For finding the variance of $\widehat{E(M)}$, we only need to take into account the uncertainty introduced through the model parameter estimation (cf. Ståhl et al. 2016). For large-area surveys, the relative difference between M and its expected value can be expected to be small if the Poisson model is valid (cf. Ståhl et al. 2016).

If a sample of realisations of the random variables M_i is available, the parameter vector $\boldsymbol{\beta}$ of the Poisson model may be estimated through a loglinear Poisson regression (Baddeley et al. 2010). In our case, we do not have such data available. Instead, for each cell P_i in a sample of n of the N cells, there are two disjoint vegetation plots,

$A_{i1} \subset P_i$ and $A_{i2} \subset P_i$, $i = 1, ..., n$, $n << N$ (Fig. 1). All such plots A_{ij} are of size a_A, and in each one of them the presence or absence of any data points is recorded. Let Y_{ij} be equal to 1 if a point of the pattern is present in A_{ij}, , $i = 1, ..., n$, $j = 1, 2$, and zero otherwise. Based on the sample of Y_{ij} data, together with the corresponding covariate data, an estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ in (1) may be obtained using binary regression. Assuming that the point pattern is generated by a Poisson point process with intensity (1), it follows that the binary variables Y_{ij} satisfy a binary GLM, with complementary log-log link and offset $\log(a_A)$,

$$\log(-\log(1 - p_{ij})) = \log(a_A) + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i, \tag{3}$$

where $p_{ij} = p_{ij}(\boldsymbol{\beta}) = P(Y_{ij} = 1) = 1 - \exp(-a_A \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i))$ denotes the probability that there are data points in plot A_{ij} (Baddeley et al. 2010). Thus, $\hat{\boldsymbol{\beta}}$ is obtained from model (3). The fact that a pair of vegetation plots is available for each cell P_i in the sample of cells will make a correlation test possible for the model given in (3). In theory, the latter model has conditionally independent response variables Y_{ij} given the covariates; thus, it is of interest to test whether this property is fulfilled when evaluating the model.

For related studies of P/A data in ecology using the complementary log-log link, we refer to Yee & Mitchell (1991), Royle & Dorazio (2008), Lindenmayer et al. (2009), and Fithian et al. (2015). The most commonly used GLM for P/A data is the logistic regression model (e.g. Wintle et al. 2005; Foody 2008). Contrary to these studies, which focus on cell-wise estimation or prediction for, e.g., producing maps, our study focuses on obtaining large-area estimates of expected plant density.

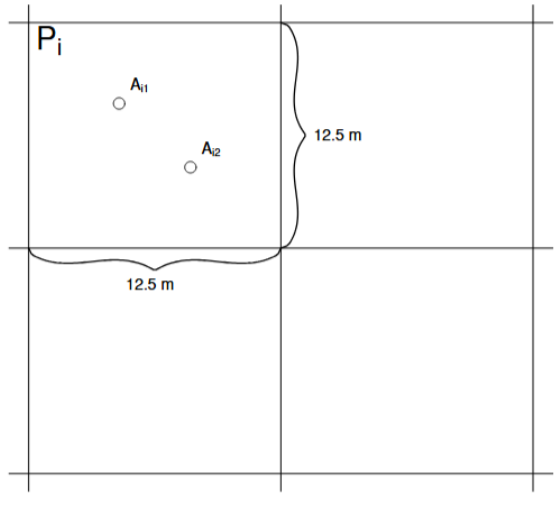


Fig. 1: Placement of the vegetation plots A_{i1} and A_{i2} within cell P_i. In our empirical study, each cell is of size 12.5×12.5 m^2 and the vegetation plots A_{i1} and A_{i2} are each of size 0.25 m^2, circular, and separated by 5 m.

If locations of plants are regarded as a realisation of an inhomogeneous Poisson point process, Baddeley et al. (2010) show that logistic regression models do not generally have a consistent meaning, independent of the choice of cell or plot size. The implication of this finding is that two researchers who apply logistic regression to the same point process data, but using different cell or plot sizes, may obtain results that cannot be reconciled. GLMs with a complementary log-log link do not suffer from this problem.

For estimating the variance of $\widehat{E(M)}$, we may use the fact that for large samples and under mild conditions, $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed,

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_q(\mathbf{0}, \boldsymbol{\Sigma}), \tag{4}$$

where

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{2}{[g'(p_{i1}(\boldsymbol{\beta}))]^2 p_{i1}(\boldsymbol{\beta})(1 - p_{i1}(\boldsymbol{\beta}))} \boldsymbol{x}_i \boldsymbol{x}_i'$$

is assumed to be finite and positive definite (Sen & Singer 1993, Theorem 7.4.1), and where $g(p) = \log(-\log(1-p))$. This implies that $\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i$ is approximately normally distributed with mean $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i$ and variance $n^{-1} \boldsymbol{x}_i' \boldsymbol{\Sigma} \boldsymbol{x}_i$, and the joint distribution of $\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i)$, $i = 1, ..., n$, is approximately a multivariate lognormal distribution (e.g. Kleiber & Kotz 2003). From the properties of the multivariate lognormal distribution, it follows that

$$\begin{aligned} \operatorname{var}(\widehat{E(M)}) &= a_P^2 \sum_{i\in U} \sum_{j\in U} \operatorname{cov}(\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i), \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_j)) \\ &\approx a_P^2 \sum_{i\in U} \sum_{j\in U} \left[\exp\left(\frac{\boldsymbol{x}_i' \boldsymbol{\Sigma} \boldsymbol{x}_j}{n} \right) - 1 \right] \exp\left(\boldsymbol{\beta}^{\mathrm{T}} (\boldsymbol{x}_i + \boldsymbol{x}_j) + \frac{\boldsymbol{x}_i' \boldsymbol{\Sigma} \boldsymbol{x}_i + \boldsymbol{x}_j' \boldsymbol{\Sigma} \boldsymbol{x}_j}{2n} \right), \end{aligned} \tag{5}$$

which in turn is estimated by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}$ by

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \frac{2}{[g'(p_{i1}(\hat{\boldsymbol{\beta}}))]^2 p_{i1}(\hat{\boldsymbol{\beta}})(1 - p_{i1}(\hat{\boldsymbol{\beta}}))} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

2.1 | Estimation of expected plant density in large regions

When analysing spatial point patterns of plants it is typically more relevant to study plant density than the total number of plants in a region. The expected plant density, defined as the mean number of plants per unit area, is given by $\mu_U = (N a_P)^{-1} E(M)$. It is estimated by

$$\hat{\mu}_U = \frac{\widehat{E(M)}}{N a_P} = \frac{1}{N} \sum_{i\in U} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i). \tag{6}$$

We denote the corresponding variance by $\sigma_U^2 = \operatorname{var}(\hat{\mu}_U) = (N a_P)^{-2} \operatorname{var}(\widehat{E(M)})$, and by arguing as in (5), its estimate is given by

$$\hat{\sigma}_U^2 = \frac{1}{N^2} \sum_{i\in U} \sum_{j\in U} \left[\exp\left(\frac{\boldsymbol{x}_i' \hat{\boldsymbol{\Sigma}} \boldsymbol{x}_j}{n} \right) - 1 \right] \exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}} (\boldsymbol{x}_i + \boldsymbol{x}_j) + \frac{\boldsymbol{x}_i' \hat{\boldsymbol{\Sigma}} \boldsymbol{x}_i + \boldsymbol{x}_j' \hat{\boldsymbol{\Sigma}} \boldsymbol{x}_j}{2n} \right). \tag{7}$$

Remark 1. If the number of cells N is very large, it may be computationally demanding to compute the sum of the N^2 terms that define the estimate of (7). In such cases, a large sample s may be selected from the population of N cells. Given the values of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$, an unbiased estimator of (7) is given by

$$\frac{1}{N^2}\sum_{i\in s}\sum_{j\in s}\frac{1}{\pi_{ij}}\left[\exp\left(\frac{\boldsymbol{x}_i'\hat{\boldsymbol{\Sigma}}\boldsymbol{x}_j}{n}\right)-1\right]\exp\left(\hat{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{x}_i+\boldsymbol{x}_j)+\frac{\boldsymbol{x}_i'\hat{\boldsymbol{\Sigma}}\boldsymbol{x}_i+\boldsymbol{x}_j'\hat{\boldsymbol{\Sigma}}\boldsymbol{x}_j}{2n}\right), \tag{8}$$

where π_{ij} denotes the second-order inclusion probabilities for the sampling design used for taking the sample s. If simple random sampling is used for selecting the sample s, then $\pi_{ij} = n_s/N$ if $i = j$ and $\pi_{ij} = n_s(n_s-1)/(N(N-1))$ if $i \neq j$, where n_s is the size of the sample s (Särndal et al. 1992).

2.2 | Estimation of expected plant density in subregions

The parameter vector $\boldsymbol{\beta}$ is estimated from sample data from a region tesselated into N cells. Once this is done (and the model has been validated), one may use this estimated parameter vector and the fitted model not only for computing an estimate of expected plant density in this region, but also for subregions. Let U_r be the subset of cells in U that tesselates the subregion of interest. For example, U_r may consist of the subset of cells representing a municipality within the region. Similarly to (6), the expected plant density in the subregion, denoted by μ_{U_r}, is then estimated by

$$\hat{\mu}_{U_r} = \frac{1}{N_r}\sum_{i\in U_r}\exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i), \tag{9}$$

where N_r is the number of cells in U_r. Likewise, the estimate $\hat{\sigma}^2_{U_r}$ of the variance of (9) is computed as in (7), but with N and U replaced by N_r and U_r, respectively. If N_r is large, one may use the ideas put forward in Remark 1 for computing an estimate of the of variance of (9).

3 | Correlation test for the binary regression model

The binary regression model (3) implied by the inhomogeneous Poisson model assumes independent response variables given the covariates. We propose testing this assumption using the residuals from the fitted binary regression model. Thus, the null hypothesis is that Y_{i1} and Y_{i2} are conditionally independent given the covariates. The alternative hypothesis says that the null hypothesis is incorrect. It should be noted that if the inhomogeneous Poisson point process model with intensity given by (1) is correct, then the null hypothesis is valid. The reverse implication is not necessarily true. However, if the null hypothesis is incorrect, then the corresponding model for the point pattern is also incorrect.

Each sampling plot has two subplots, A_{i1} and A_{i2}, which implies that we have paired observations at our disposal. Let the residuals corresponding to A_{i1} and A_{i2} be denoted

by r_{i1} and r_{i2}. To test the null hypothesis, a correlation coefficient computed from the paired residuals $(r_{i1}, r_{i2}), i = 1, ..., n$, is used. Different correlation coefficients, as well as different types of residuals, can be used for performing the test. We consider two correlation coefficients: the Pearson coefficient r and the Spearman coefficient ρ.

The test statistic for the test using r is (Rahman 1968)

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n - 2}.$$

If t is larger in absolute value than the 97.5th percentile from a t-distribution with $n - 2$ degrees of freedom, then the null hypothesis is rejected at the 5% significance level. Pearson correlation is most effective when the data (i.e. (r_{i1}, r_{i2}) in our case) are known to come from a bivariate normal distribution (Agresti 2006). In case ρ is used, then ρ itself becomes the test statistic and the p-values for the test are computed according to the AS 89 algorithm (Best & Roberts 1975).

Another choice to be made is the type of residuals to use. There exist several types of residuals that can be computed in case one uses a binary regression model. These include deviance residuals, response residuals, working residuals, Pearson residuals and randomised quantile residuals (Dunn & Smyth 2018). All types have been tested during the simulation study, but the latter three produced the best results. Thus, only these three types of residuals and their results will be presented in this article.

Remark 2. A model may be judged unsuitable not only if it does not pass the chosen test based on the correlation of the residuals, but also if the fitted model is unreasonable from an ecological context. This may happen, for example, if the fitted model produces unrealistic predicted values due to extrapolation. One may in an informal way compare the estimated expected plant density $\hat{\mu}_U$ in (6) with the corresponding estimate based only on data from the n cells that were used for fitting the model, i.e., from

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \exp(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i). \tag{10}$$

If these two estimates deviate "much" from each other, this may indicate that extrapolation is a problem. Both $\hat{\mu}_U$ and $\hat{\mu}_n$ may be used for estimating the expected plant density μ_U in the larger region. However, even if the number of cells used for fitting the model is large enough to support the use of $\hat{\mu}_n$ as an estimate for μ_U, it may well not be large enough to support corresponding estimates for subregions (cf. Rao & Molina 2015). In contrast, if the Poisson model is valid, the estimate $\hat{\mu}_{U_r}$ defined in (9) may be used for estimating expected plant density in subregions also in cases where they contain at most a few cells with P/A observations.

4 | Empirical study with P/A data from environmental monitoring

For this study, we used empirical P/A data from the Swedish National Forest Inventory (NFI; Fridman et al. 2014) that were collected between 2011 and 2013 for the

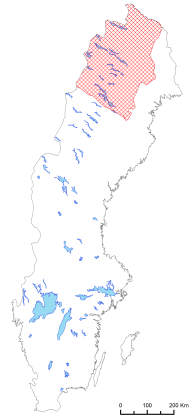


Fig. 2: Map of Sweden showing Norrbotten's Lappmark (crosshatched area).

Lappmark region in Norrbotten (Sweden, Fig. 2). The Swedish NFI is a field sample plot inventory that includes both temporary and permanent plots. At each permanent plot, the presence or absence of a predetermined set of species is recorded every 10th year on two circular 0.25 m^2 vegetation plots (A_{i1} and A_{i2}), separated by 5 m and located 2.5 m from the plot centre (Fig. 1). Covariates were obtained from several forest raster map products: the SLU Forest Map (Reese et al. 2003; Wallerman et al. 2021), the National Forest Attribute Map (NFAM, Nilsson et al. 2017), and a soil moisture map derived by Ågren et al. (2021). The SLU Forest Map comprises raster maps of the Swedish forest state, generated from satellite images using the Swedish NFI sample plots as reference data. A similar method was used in the NFAM to create forest raster maps from airborne laser scanning data collected in Sweden between 2009 and 2016. The soil moisture map by Ågren et al. (2021) was derived from terrain indices generated from a national LIDAR digital elevation model and environmental features. Since the NFAM provides data only for forested areas with a minimum tree canopy height of 3 m, estimates of expected plant density were computed for the corresponding area. Covariates from the forest map products for each sample plot were extracted using nearest neighbour interpolation at the plot centre coordinates.

To ensure that the covariates used in the models represent the conditions at the time the P/A data were acquired, cells that showed differences of more than 5 m in basal area-weighted mean tree height and 100 m^3/ha in tree stem volume, compared to corresponding field plot measurements, were excluded from the sample used for fitting binary regression models. This exclusion indicated clear-felling or other harvesting activities conducted after the P/A data collection. After this adjustment, the sample size was $n = 293$. For the defined region, comprising forested areas with a minimum tree canopy height of 3 m in Norrbotten's Lappmark, the cell size was 12.5 m by 12.5 m, and the total number of cells (N) exceeded 253 million. To address the potential for nonlinear relationships when fitting binary regression models, multivariable fractional polynomials (Sauerbrei & Royston 1999) were employed using the R library mfp

Table 1: Estimated expected plant densities ($\hat{\mu}_U$) in m^{-2} and corresponding estimates of variance ($\hat{\sigma}^2_U$) for two species in a region defined as the forested area with at least 3 m tree canopy height in Norrbotten's Lappmark. The estimated densities were based on binary regression models, for which covariates and estimated coefficients are presented (with standard errors in parentheses). The binary regression models were evaluated using the correlation tests presented in Section 3, and the resulting p-values are presented in the last three columns (first p_{pp} with Pearson residuals and a Pearson test and then quantile residuals with a Pearson test (p_{qp}) and a Spearman test (p_{qs})). The variances were estimated using formula (8), and were based on simple random samples of 50 000 cells.

Species	Estimated parameters ($\hat{\boldsymbol{\beta}}$)		$\hat{\mu}_U$	$\hat{\sigma}^2_U$	p_{pp}	p_{qp}	p_{qs}
L. europaea	Intercept	-1.680 (0.806)	0.111	0.00125	0.909	0.880	0.992
	Prop. tree stem volume of deciduous trees	3.520 (1.748)					
	Basal area	-0.143 (0.073)					
L. pilosa	Intercept	0.127 (1.029)	0.054	0.00052	0.821	0.867	0.853
	Basal area-weighted mean tree DBH	-0.138 (0.067)					
	Index of soil moisture	-0.025 (0.012)					

(Ambler & Benner 2015).

For the analysis, we chose two species, *Lysimachia europaea* (L.) U. Manns & Anderb. (arctic starflower) and *Luzula pilosa* (L.) Willd. (hairy wood-rush) as they are relatively common, easy to identify and with known habitat requirement. For the binary regression model for *L. europaea*, the covariates used were basal area (m^2/ha) from the NFAM and the proportion of tree stem volume of deciduous trees from the SLU Forest Map. For *L. pilosa*, the covariates used were basal area-weighted mean tree diameter at breast height (DBH; cm) from the NFAM and the 0–100% index (dry to wet) of soil moisture derived by Ågren et al. (2021). Additional covariates at our disposal were basal area-weighted mean tree height (m) and tree stem volume (m^3/ha) from the NFAM and the proportion of tree stem volume of spruce from the SLU Forest Map, but none of these covariates were included in our final models for *L. europaea* or *L. pilosa*.

As in Ståhl et al. (2017), the theory assumes that plant occurrences on a vegetation plot are recorded when a predetermined reference point of a plant is located on the vegetation plot in the field. However, presence is recorded if any part of a plant is located on a vegetation plot, and a correction was applied by adding a presumed average plant radius to each vegetation plot's radius in the calculations. The assumed plant radius was set to 3 cm for *L. europaea* and 10 cm for *L. pilosa*.

The empirical results based on monitoring data are presented in Table 1. The p-values for the correlation tests of the binary regression models are given for both *L. europaea* and *L. pilosa*. The null hypothesis was not rejected at level 5% with neither the test based on Pearson residuals nor the tests based on quantile residuals. The working residuals were not used to produce a p-value because of their lacklustre performance during the simulation study. The fitted model for *L. europaea* in Table 1 suggests that the larger proportion of tree stem volume of deciduous trees, the larger mean number of plants in a cell (and the higher probability of presence in a plot or cell). For basal area, the relation is the opposite. The fitted model for *L. pilosa*

Table 2: Estimated expected plant densities ($\hat{\mu}_{U_r}$) in m^{-2} and corresponding estimates of variance ($\hat{\sigma}^2_{U_r}$) for two species in the forested areas with at least 3 m tree canopy height in each of the five municipalities of Norrbotten's Lappmark.

Species	Municipality	$\hat{\mu}_{U_r}$	$\hat{\sigma}^2_{U_r}$
L. europaea	Arjeplog	0.123	0.00438
	Arvidsjaur	0.057	0.00018
	Gällivare	0.115	0.00117
	Jokkmokk	0.080	0.00046
	Kiruna	0.162	0.00269
L. pilosa	Arjeplog	0.044	0.00012
	Arvidsjaur	0.044	0.00013
	Gällivare	0.050	0.00019
	Jokkmokk	0.044	0.00013
	Kiruna	0.083	0.00085

implies that the larger basal area-weighted mean tree DBH and the larger index of soil moisture, the smaller probability of presence in a cell (and the smaller mean number of plants). The estimated expected plant density for *L. europaea* is about twice as high as for *L. pilosa*. Table 2 contains estimated expected plant densities and corresponding estimates of variance for *L. europaea* and *L. pilosa* in each of the municipalities of Norrbotten's Lappmark. For both species, the largest estimated expected plant density was obtained for Kiruna Municipality.

5 | A Monte Carlo study

For evaluating the performance of the suggested correlation tests, we created a population consisting of 100 000 cells, chosen as a simple random sample from the forested area of the Lappmark region in Norrbotten with at least 3 m tree canopy height. Thus, complete covariate information was available in the created population. A sample of n cells was then selected from the created population, with n equal to either 300 or 1 000 depending on the case. All sample units remained the same throughout the simulation process. The sample units were drawn according to a simple random sampling design and appeared to represent the population quite well (Table 3).

For investigating the power properties of the correlation tests and actual levels of the tests, corresponding point process data for all cells in the sample were generated by log-Gaussian Cox processes, Matérn cluster processes, or (generalised) Thomas cluster processes. Such processes are useful for modelling clustered point patterns (Penttinen et al. 1992; Tanaka et al. 2008; Renner et al. 2015). Point process data were generated using the `spatstat` package (Baddeley et al. 2016) in `R` (R Core Team 2025). For each point process model considered, 1 000 replicates of point data were generated in simulations to estimate the power of the tests, and 10 000 replicates were used to compute the actual significance levels.

Table 3: Summary statistics for the covariates in the population, the large sample consisting of 1 000 cells and the medium sample consisting of 300 cells used during the simulation study.

Covariate	Summary statistic	Population	Large sample	Medium sample
Basal area	Tenth percentile	4.00	4.00	4.00
	First quartile	6.00	6.00	6.00
	Median	9.00	9.00	9.00
	Mean	10.87	10.59	10.48
	Third quartile	14.00	14.00	14.00
	Ninetieth percentile	20.00	19.00	18.00
Proportion of deciduous trees	Tenth percentile	0	0	0.0054
	First quartile	0.0380	0.0380	0.0360
	Median	0.111	0.111	0.110
	Mean	0.178	0.176	0.172
	Third quartile	0.286	0.286	0.273
	Ninetieth percentile	0.444	0.444	0.408

In our main setting, the value of λ_i for cell i in the created population was set equal to the fitted intensity for *L. europaea* in Table 1, i.e., λ_i was computed using the covariates basal area and proportion of tree stem volume of deciduous trees. In an alternative setting, the intercept coefficient in λ_i was enlarged by two units, representing a situation with a higher overall density of the species.

For log-Gaussian Cox processes, points in cell i were generated with intensity $\lambda_i \exp(\eta(\boldsymbol{u}))$ at location $\boldsymbol{u}$ within cell i, where $\eta(\boldsymbol{u})$ is a mean zero Gaussian process with an isotropic exponential covariance function $C(\boldsymbol{u}, \boldsymbol{u}') = \sigma^2 \exp(r/\alpha)$, where r is the distance between the locations $\boldsymbol{u}$ and $\boldsymbol{u}'$, $\alpha > 0$ is a scale parameter, and $\sigma > 0$ is a standard deviation parameter. Conditional on $\eta(\boldsymbol{u})$, the point events are inhomogeneous Poisson. That is, any spatial dependence in the point location data is entirely captured by $\eta(\boldsymbol{u})$. In addition, the Poisson process with intensity λ_i in cell i may be considered as the limit of the log-Gaussian Cox process as σ tends to zero. Ideally, the rejection rate of the test should be close to the nominal level of significance, 5%, when the population of points is generated through an inhomogeneous Poisson point process, and then, as the point process gets further and further away from the inhomogeneous Poisson point process (i.e., as the value of σ increases), the power of the test should increase.

For Matérn and Thomas cluster processes, the intensity of the parent process was set equal to λ_i in cell i. In `spatstat`, generating Matérn cluster processes with the cluster radius parameter σ set to 0 or Thomas cluster processes with the standard deviation parameter σ set to 0 is not possible. Instead, the smallest value utilised for these parameters was 0.00001. For both these cluster processes, there is a third parameter, μ, which denotes the mean number of points per cluster.

Binary regression models of the form (3) were fitted using two covariates, basal area and proportion of tree stem volume of deciduous trees (Table 1). That is, the models were fitted as if the point data were generated from an inhomogeneous Poisson point process model. As in the Swedish NFI, the size of each of the vegetation plots A_{i1} and

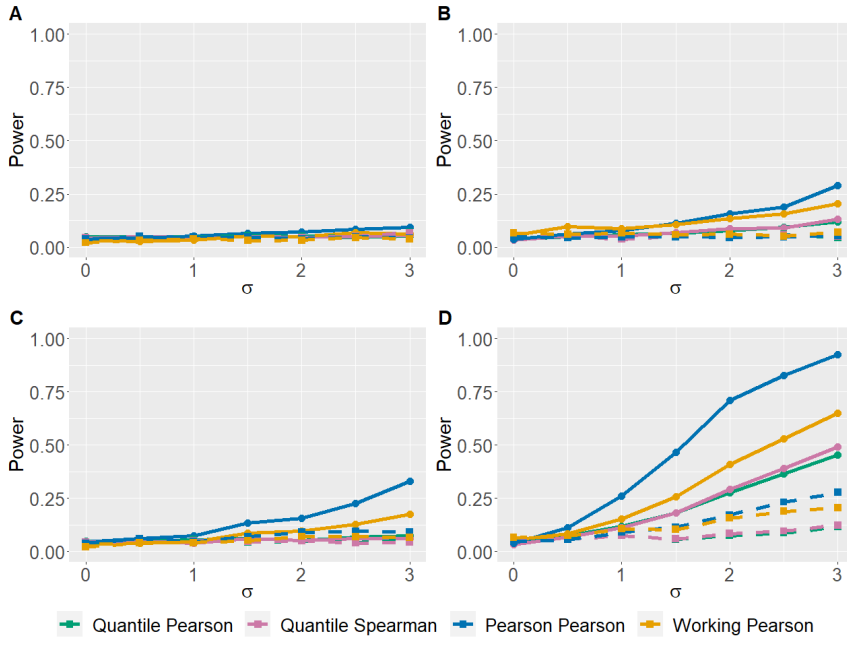


Fig. 3: Power curves ($n = 300$). Point data were generated according to a log-Gaussian Cox process with intensity $\lambda_i \exp(\eta(\boldsymbol{u}))$ at location $\boldsymbol{u}$ within cell i, where $\eta(\boldsymbol{u})$ is a mean zero Gaussian process with an isotropic exponential covariance function with scale parameter α and standard deviation parameter σ. For plots A and B, $\alpha = 0.5$, while for C and D, $\alpha = 5$. The special case with $\sigma = 0$ corresponds to data from a Poisson point process with intensity λ_i. The value of λ_i was set equal to the fitted intensity for *L. europaea* in Table 1 for plots A and C, whereas for B and D the intercept coefficient in λ_i was enlarged by two units. The solid lines and round points represent the cases where the distance between the plot centres is the minimum possible, while the dashed lines and square points represent the cases where the distance between the plot centres is 5 metres.

A_{i2} was set to 0.25 m^2 and the distance between their centres was 5 m (Fig. 1). In this simulation study, we considered the use of the distance 0.6242 m as well. With this latter distance and assuming that the P/A sampling is conducted as in the Swedish NFI, a *L. europaea* plant of presumed radius can only be registered as present in one of the two plots A_{i1} and A_{i2}. If this distance would be decreased even further, this would no longer be the case.

Four combinations of residual and correlation types were tested: quantile residuals with Pearson correlation, quantile residuals with Spearman correlation, Pearson residuals with Pearson correlation and working residuals with Pearson correlation. These are the variants that produced the most satisfying results out of all the possible combinations.

For log-Gaussian Cox processes and sample size $n = 300$, the results of the Monte Carlo simulations are shown in Fig. 3. The actual significance levels are close to the nominal level for both variants of the quantile residual tests (ranging from 0.049 to 0.053) when the scale parameter is 0.5, but are somewhat lower for the Pearson test

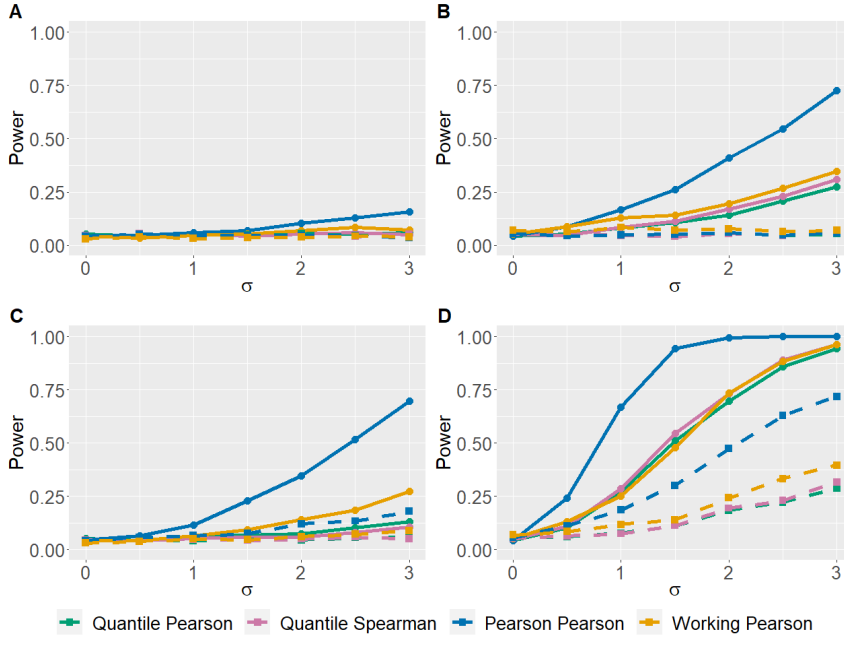


Fig. 4: Power curves ($n = 1\,000$). Point data were generated according to a log-Gaussian Cox process with intensity $\lambda_i \exp(\eta(\boldsymbol{u}))$ at location $\boldsymbol{u}$ within cell i, where $\eta(\boldsymbol{u})$ is a mean zero Gaussian process with an isotropic exponential covariance function with scale parameter α and standard deviation parameter σ. For plots A and B, $\alpha = 0.5$, while for C and D, $\alpha = 5$. The special case with $\sigma = 0$ corresponds to data from a Poisson point process with intensity λ_i. The value of λ_i was set equal to the fitted intensity for *L. europaea* in Table 1 for plots A and C, whereas for B and D the intercept coefficient in λ_i was enlarged by two units. The solid lines and round points represent the cases where the distance between the plot centres is the minimum possible, while the dashed lines and square points represent the cases where the distance between the plot centres is 5 metres.

based on Pearson residuals (ranging from 0.040 to 0.047). Figs. 3A and 3B show that the power may be very low if the scale parameter α is as low as 0.5 and the distance between the centres of the vegetation plots is as large as in the Swedish NFI (5 m). If, however, this distance is decreased, the power properties are slightly improved. When the value of α is increased to 5, the power properties are improved for both considered distances between the centres of the vegetation plots (Figs. 3C and 3D). Here, as expected, the power increases as the point process data gets further and further away from being generated from an inhomogeneous Poisson process, i.e., as the parameter σ of the log-Gaussian Cox process increases. When comparing Fig. 3B with 3A and Fig. 3D with 3C, we see that the power increases when the intensity of the log-Gaussian Cox process is enlarged. When the sample size is increased to $n = 1\,000$, the actual significance levels of the test stays close to the nominal level if the quantile residuals are used (now falling between 0.046 and 0.051), whereas the power gets larger (Fig. 4). In case Pearson residuals are used, the actual significance levels remain in the same region (between 0.041 and 0.047), and an improvement in power occurs the more distant from a

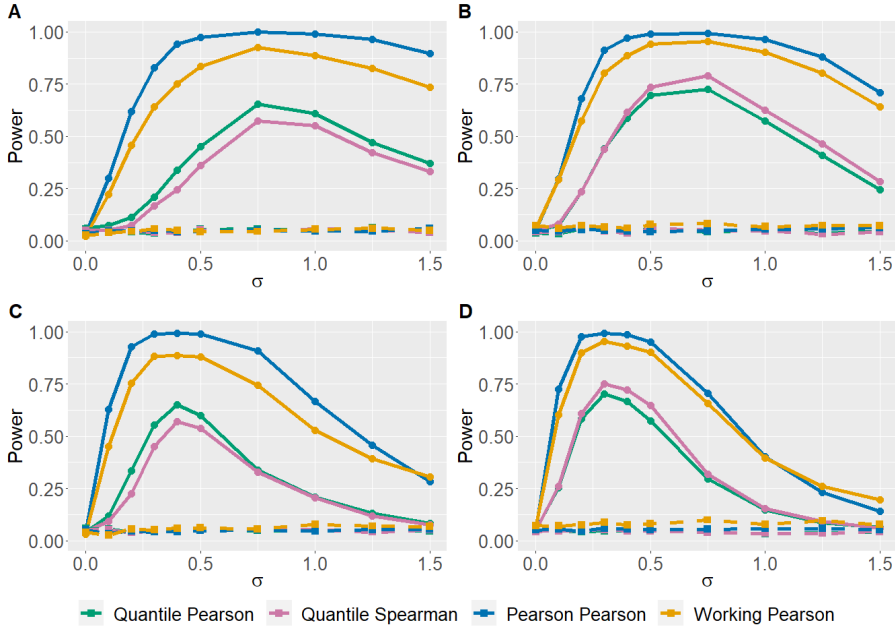


Fig. 5: Power curves ($n = 300$). Point data were generated according to Matérn cluster processes (plots A and B) and Thomas cluster processes (plots C and D). The intensity λ_i in cell i for the parent process was set equal to the fitted intensity for *L. europaea* in Table 1 for plots A and C, whereas for B and D the intercept coefficient in λ_i was enlarged by two units. The parameter μ, representing the mean number of points per cluster, was equal to 10 throughout. The solid lines and round points represent the cases where the distance between the plot centres is the minimum possible, while the dashed lines and square points represent the cases where the distance between the plot centres is 5 metres.

Poisson process the process becomes. For the working residuals, the actual significance levels are quite distant from 0.05 in general, showing strong variation depending on the cases (from 0.028 to 0.064).

For Thomas and Matérn cluster processes with $\mu = 10$ and a sample size of $n = 300$, the results of the Monte Carlo simulations are shown in Fig. 5. For both of these processes, all versions of the test are expected to exhibit negligible power when the parameter σ is approximately 0 since in this case, all the clusters will approximately be concentrated at single points. Then, intuitively, with only P/A data from the vegetation plots, it should be practically impossible to distinguish the underlying true point process from a (thinned) Poisson point process. The power values for the correlation test in such instances range between 0.047 and 0.053 when quantiles residuals are examined, and between 0.040 and 0.048 with Pearson residuals. The power increases for all variants and types of residuals when the cluster radius parameter gets away from 0, but then shows a dip starting from a certain value. When the distance between the centres of vegetation plots is as large as in the Swedish NFI (5 m), all versions of the test demonstrate negligible power, a pattern also observed when the sample size is increased to $n = 1\,000$

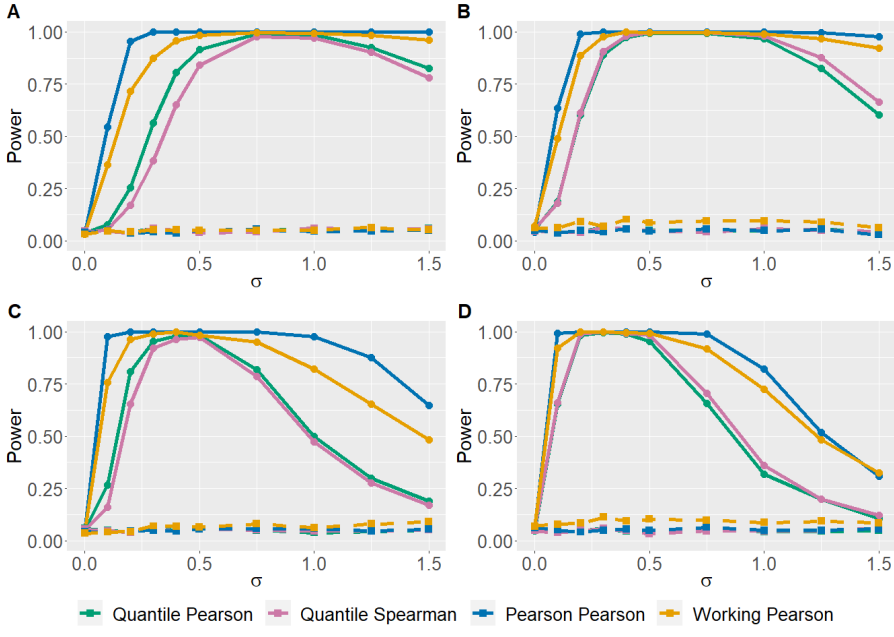


Fig. 6: Power curves ($n = 1\,000$). Point data were generated according to Matérn cluster processes (plots A and B) and Thomas cluster processes (plots C and D). The intensity λ_i in cell i for the parent process was set equal to the fitted intensity for *L. europaea* in Table 1 for plots A and C, whereas for B and D the intercept coefficient in λ_i was enlarged by two units. The parameter μ, representing the mean number of points per cluster, was equal to 10 throughout. The solid lines and round points represent the cases where the distance between the plot centres is the minimum possible, while the dashed lines and square points represent the cases where the distance between the plot centres is 5 metres.

(Fig. 6). In the latter case, the actual significance levels of the test lie between 0.046 and 0.052 (quantile) or 0.038 and 0.051 (Pearson). Across both sample sizes, power properties notably improve when reducing the distance between vegetation plots, with powers close to 1 for some values of σ. Enhancing the parameter μ to 100 results in improved power properties for both vegetation plot distance scenarios (Fig. 7). For the larger sample size, the actual significance levels of the test are between 0.046 and 0.054 (quantile) or 0.040 and 0.050 (Pearson), aligning reasonably well with the test's nominal level.

In every case, when the minimal distance is used, the test based on the Pearson residuals and the Pearson correlation coefficient is the one that produces the higher power, followed by the test based on the working residuals and Pearson correlation coefficient and finally the ones based on quantile residuals (both with the Pearson and Sperman coefficients). On the other hand, when the standard NFI distance (i.e. 5 meters) is in place, the test based on the working residuals and the Pearson coefficient is the one producing the highest power in most cases, although the power remains relatively low throughout. In some cases when the Matérn or Thomas processes are

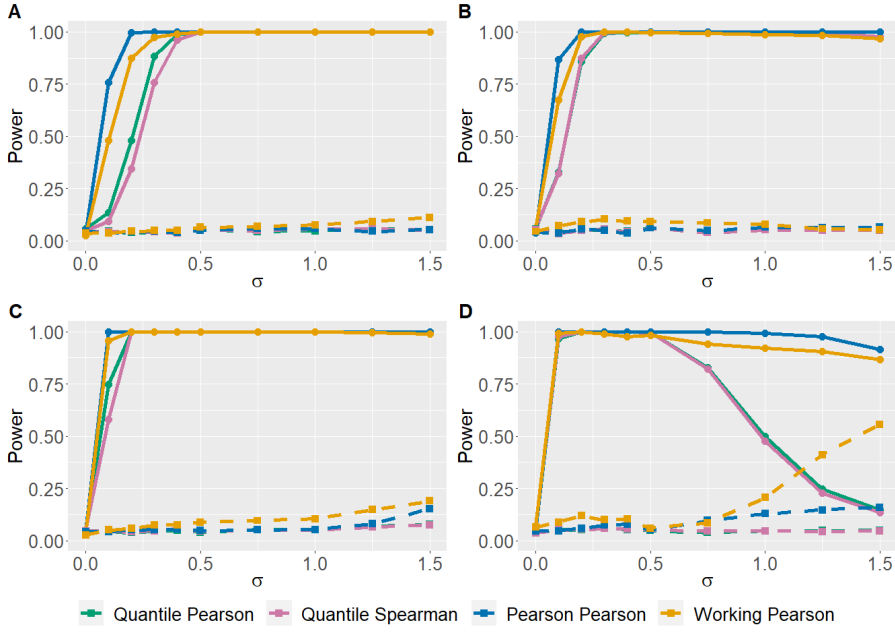


Fig. 7: Power curves ($n = 1\,000$). Point data were generated according to Matérn cluster processes (plots A and B) and Thomas cluster processes (plots C and D). The intensity λ_i in cell i for the parent process was set equal to the fitted intensity for *L. europaea* in Table 1 for plots A and C, whereas for B and D the intercept coefficient in λ_i was enlarged by two units. The parameter μ, representing the mean number of points per cluster, was equal to 100 throughout. The solid lines and round points represent the cases where the distance between the plot centres is the minimum possible, while the dashed lines and square points represent the cases where the distance between the plot centres is 5 metres.

investigated, it can be seen that the power experiences a drop after a certain threshold value of the cluster radius. At least in the Matérn cluster process case, this phenomenon could be explained by the fact that the larger the clusters, the higher the probabilities that they overlap. The superposition of several independent Poisson point processes is also a Poisson point process (Baddeley et al. 2016). It can be observed that the power seems to be higher whatever the case when the sample size is equal to 1 000 compared to when it is equal to 300. Another observation that can be made based on the results is that the test based on the Pearson residuals appears to be conservative in most cases. In other words, the actual significance levels rarely exceed the nominal level of 0.05.

6 | Discussion and conclusion

In this paper, it has been shown in a model-based setting how to use P/A data from environmental monitoring for estimating expected plant density for species of plants in regions and subregions of interest. Locations of plants were modelled through

inhomogeneous Poisson point processes, and GLMs and auxiliary covariate information were used for linking the P/A data to plant density. More specifically, for regions tesselated into cells, cell-wise estimates of the expected number of plants were obtained, and such estimates were thereafter aggregated to obtain regional estimates of expected plant density, together with corresponding estimates of variances. One way to see this is that the estimates of expected plant density are derived from created maps of cell-wise estimates of expected number of plants. The increasing quality of remote sensing data has greatly improved the possibility to create such maps. Although they may have a poor accuracy at cell-level, they can be useful for identifying parts of a region which are likely to have larger (or smaller) abundances of plants. For a specific species, maps and estimates of expected plant density for regions can all be obtained from the same fitted GLM.

An advantage with the presented approach is that it also enables estimation of expected plant density in subregions. If estimates for subregions were based only on the subregion-specific data, then, due to the costs involved, it would often not be possible to have a large enough overall sample size of P/A data to support reliable estimates for all subregions of interest (cf. Rao & Molina 2015). Therefore, we "borrow strength" by using P/A data collected from neighboring subregions (e.g., municipalities), thus increasing the effective sample size. When doing this, reliable estimates of regression model parameters can be obtained and estimates of the expected plant density can be computed for each cell in the subregion, based on which it is possible to estimate the expected plant density and the corresponding variance for the subregion. In comparison, a large enough sample of plots would also support such estimation in the larger region using only the cells containing the collected P/A data, but such a sample size may not be large enough to support corresponding estimates for subregions containing at most a few cells with P/A data.

A fitted binary regression model may be judged unsuitable if it produces unrealistic predicted values due to extrapolation, and an informal way to judge if such problems exist has been suggested. To minimise the problem of extrapolation, samples that are well spread in the covariate space are desired. Such samples may be obtained through local pivotal methods and other related methods (Grafström et al. 2012; Tillé 2020).

Inferences from our model-based estimators rely on the distribution implied by the assumed model (cf. Rao & Molina 2015). First of all, one should apply standard methods for model selection and validation of the binary regression model. In addition to such standard methods, we have derived and evaluated correlation tests for this model. If this binary regression model is invalid, then so is the corresponding inhomogeneous Poisson point process model for the point pattern (but the reverse implication does not hold). In a simulation study, it was shown that the proposed correlation tests have reasonably good power properties if the two vegetation plots in each pair of vegetation plots used for collecting P/A data are not located too far from each other, indicating that the power of the tests can be improved if the distance between the vegetation plots used by the Swedish NFI is decreased. The actual levels of significance were closer to the nominal when quantile residuals were used instead of Pearson, while the power was higher when the latter was used instead of the former.

In the simulation study, the test variant with the quantile residuals showed promising

results when it comes to actual levels of significance (that were very close to the nominal level of 0.05), although the power of the test was a little unsatisfactory in some cases compared to the variant with Pearson residuals. However, unlike Pearson residuals, quantile residuals are randomised, which means that the test statistic and p-value will vary each time the test is run, even on the same empirical data (i.e., for the exact same values of response variable and covariate data). In the simulations, we have fairly often observed that if a new set of randomised quantile residuals is computed based on the same fitted binary regression model, the test outcome might change. This randomisation has nothing to do with environmental factors or P/A data, which makes it problematic. For this reason, we suggest using Pearson residuals rather than randomised quantile residuals.

An extension to estimation of change of expected plant density between two time points, t_1 and t_2, serve as an important topic for further studies. Here, modelling of P/A data from two point processes will be needed, and the extension is straightforward if the time between t_1 and t_2 is long enough. On the other hand, if this time interval is relatively short, then the two point processes will be dependent, which would need to be taken into consideration in the modelling. Ideas in this direction were put forward in Ståhl et al. (2017), but only for homogeneous Poisson point processes.

To conclude, our study has suggested an approach to assessing state and trends (if repeated across time) of plant populations in regions and subregions, based on a combination of P/A and remote sensing data. Formal estimates of this kind could be useful for several purposes, e.g., for reporting to the EU's habitats directive.

Acknowledgements

We acknowledge the financial support from the Kempe Foundations (SMK-1955).

References

Agresti, A. (2006). An Introduction to Categorical Data Analysis. John Wiley & Sons. https://doi.org/10.1002/0470114754

Ambler, G., & Benner, A. (2015). mfp: Multivariable fractional polynomials. R package version 1.5.2. https://CRAN.R-project.org/package=mfp

Baddeley, A., Berman, M., Fisher, N. I., Hardegen, A., Milne, R. K., Schuhmacher, D., & Turner, R. (2010). Spatial logistic regression and change-of-support for Poisson point processes. Electronic Journal of Statistics, 4, 1151–1201. https://doi.org/10.1214/10-EJS581

Baddeley, A., Rubak, E., & Turner, R. (2016). Spatial point patterns: Methodology and applications with R. CRC Press, Boca Raton.

Best, D.J., & Roberts, D.E. (1975). Algorithm AS89: The Upper Tail Probabilities of Spearman's ρ. Applied Statistics, 24, 377-379. https://doi.org/10.2307/2347111

Bonham, C. D. (2013). Measurements for terrestrial vegetation (2nd ed.). John Wiley and Sons, New York.

Chambers, R., van den Brakel, J., Hedlin, D., Lehtonen, R., & Zhang, L.-C. (2006). Future challenges of small area estimation. Statistics in Transition, 7, 759–769.

Chambers, R., & Clark, R. (2012). An introduction to model-based survey sampling with applications. Oxford University Press, Oxford.

Dunn, P.K., & Smyth, G.K. (2018). Generalized Linear Models With Examples in R. Springer, New York.

Ekström, M., Sandring, S., Grafström, A., Esseen, P.-A., Jonsson, B. G., & Ståhl, G. (2020). Estimating density from presence-absence data in clustered populations. Methods in Ecology and Evolution, 11, 390–402. https://doi.org/10.1111/2041-210X.13347

Ekström, M., & Nilsson, M. (2021). A comparison of model-assisted estimators, with and without data-driven transformations of auxiliary variables, with application to forest inventory. Frontiers in Forests and Global Change, 4, 764495. https://doi.org/10.3389/ffgc.2021.764495

Elzinga, C. L., Salzer, D. W., & Willoughby, J. W. (1998). Measuring and monitoring plant populations. BLM Technical Reference 1730-1. BLM National Applied Resource Sciences Center. Denver, CO.

Fithian, W., Elith, J., Hastie, T., & Keith, D. A. (2015). Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution, 6, 424–438. https://doi.org/10.1111/2041-210X.12242

Foody, G. M. (2008). Refining predictions of climate change impacts on plant species distribution through the use of local statistics. Ecological Informatics, 3, 228–236. https://doi.org/10.1016/j.ecoinf.2008.02.002

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H., & Ståhl, G. (2014). Adapting National Forest Inventories to changing requirements – the case of the Swedish National Forest Inventory at the turn of the 20th century. Silva Fennica, 48 (3). https://doi.org/10.14214/sf.1095

Gallegos-Torell, Å., & Glimskär, A. (2009). Computer-aided calibration for visual estimation of vegetation cover. Journal of Vegetation Science, 20, 973–983. https://doi.org/10.1111/j.1654-1103.2009.01111.x

Gelfand, A. E., & Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. Ecological Monographs, 89, e01372. https://doi.org/10.1002/ecm.1372

Godínez-Alvarez, H., Herrick, J. E., Mattocks, M., Toledo, D., & Van Zee, J. (2009), Comparison of three vegetation monitoring methods: Their relative utility for ecological assessment and monitoring. Ecological Indicators, 9, 1001–1008. https://doi.org/10.1016/j.ecolind.2008.11.011

Grafström, A., Lundström, N. L., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. Biometrics, 68, 514–20. https://doi.org/10.1111/j.1541-0420.2011.01699.x

Hwang, W.-H., & He, F. (2011). Estimating abundance from presence/absence maps. Methods in Ecology and Evolution, 2, 550–559. https://doi.org/10.1111/j.2041-210X.2011.00105.x

Kleiber, C., & Kotz, S. (2003). Statistical size distributions in economics and actuarial science. Wiley, Hoboken.

Lidberg, W., Nilsson, M., & Ågren, A. (2020). Using machine learning to generate high-resolution wet area maps for planning forest management: A study in a boreal forest landscape. Ambio, 49, 475–486. https://doi.org/10.1007/s13280-019-01196-9

Lindenmayer, D. B., Welsh, A., Donnelly, C., Crane, M., Michael, D., Macgregor, C., McBurney, L., Montague-Drake, R., Gibbons, P. (2009), Are nestboxes a viable alternative source of cavities for hollow-dependent animals? Long-term monitoring of nest box occupancy, pest use and attrition. Biological Conservation, 142, 33–42. https://doi.org/10.1016/j.biocon.2008.09.026

Lindgren, N., Olsson, H., Nyström, K., Nyström, M., & Ståhl, G. (2021). Data assimilation of growing stock volume using a sequence of remote sensing data from different sensors. Canadian Journal of Remote Sensing, 48, 127–143. https://doi.org/10.1080/07038992.2021.1988542

Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J., & Olsson, H. (2017). A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. Remote Sensing of Environment, 194, 447–454. https://doi.org/10.1016/j.rse.2016.10.022

Penttinen, A., Stoyan, D., & Henttonen, H. M. (1992). Marked point processes in forest statistics. Forest Science, 38, 806–824. https://doi.org/10.1093/forestscience/38.4.806

Puliti, S., Saarela, S., Gobakken, T., Ståhl, G., & Næsset, E. (2018). Combining UAV and Sentinel-2 auxiliary data for forest growing stock volume estimation through hierarchical model-based inference. Remote Sensing of Environment, 204, 485–497. https://doi.org/10.1016/j.rse.2016.10.022

R Core Team (2025). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rahman, N. A. (1968). A Course in Theoretical Statistics. Charles Griffin and Company.

Rao, J. N. K., & Molina, I. (2015). Small area estimation. Wiley, Hoboken.

Reese, H., Nilsson, M., Pahlén, T. G., Hagner, O., Joyce, S., Tingelöf, U., Egberth, M., & Olsson, H. (2003). Countrywide estimates of forest variables using satellite data and field data from the National Forest Inventory. Ambio, 32, 542–548. https://doi.org/10.1579/0044-7447-32.8.542

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

Ringvall, A., Petersson, H., Ståhl, G., & Lämås, T. (2005). Surveyor consistency in P/A sampling for monitoring vegetation in a boreal forest. Forest Ecology and Management, 212, 109–117. https://doi.org/10.1016/j.foreco.2005.03.002

Royle, J. A., & Nichols, J. D. (2003). Estimating abundance from repeated presence-absence data or point counts. Ecology, 84, 777–790. https://doi.org/10.1890/0012-9658(2003)084[0777:EAFRPA]2.0.CO;2

Royle, J. A., & Dorazio, R. M. (2008). Hierarchical modelling and inference in ecology: The analysis of data from populations, metapopulations and communities. Academic Press, London.

Sauerbrei, W., & Royston, P. (1999). Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. Journal of the Royal Statistical Society, Series A, 162, 71–94. https://doi.org/10.1111/1467-985X.00122

Sen, P. K., & Singer. J. M. (1993). Large sample methods in statistics: An introduction with applications. Chapman and Hall, New York.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., Patterson, P. L., Magnussen, S., Næsset, E., & McRoberts, R. E. (2016). Use of models in large-area forest surveys: Comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems, 3, 5. https://doi.org/10.1186/s40663-016-0064-9

Ståhl, G., Ekström, M., Dahlgren, J., Esseen, P.-A., Grafström, A., & Jonsson, B.G. (2017). Informative plot sizes in presence-absence sampling of forest floor vegetation. Methods in Ecology and Evolution, 8, 1284–1291. https://doi.org/10.1111/2041-210X.12749

Ståhl, G., Ekström, M., Dahlgren, J., Esseen, P.-A., Grafström, A., & Jonsson, B.G. (2020). Presence-absence sampling for estimating plant density using survey data with variable plot size. Methods in Ecology and Evolution, 11, 580–590. https://doi.org/10.1111/2041-210X.13348

Särndal, C. E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling. Springer-Verlag, New York.

Tanaka, U., Ogata, Y., & Stoyan, D. (2008). Parameter estimation and model selection for Neyman Scott point processes. Biometrical Journal, 50, 43– 57. https://doi.org/10.1002/bimj.200610339

Tillé, Y. (2020). Sampling and estimation from finite populations. Wiley, Hoboken.

Wallerman, J., Axensten, P., Egberth, M., Jonzén, J., Sandström, E., Fransson, J. E. S., & Nilsson, M. (2021). SLU Forest Map – Mapping Swedish forests since year 2000. In: Proceedings of IGARSS 2021, Crossing Borders, Virtual Symposium, Brussels, Belgium, 11-16 July, 2021, pp. 6056–6059.

Wintle, B. A., Elith, J., & Potts, J. (2005). Fauna habitat modelling and mapping in an urbanising environment: A review and case study in the Lower Hunter Central Coast region of NSW. Austral Ecology, 30, 729–748. https://doi.org/10.1111/j.1442-9993.2005.01514.x

Yee, T. W., & Mitchell, N. D. (1991). Generalized additive models in plant ecology. Journal of Vegetation Science, 2, 587–602. https://doi.org/10.2307/3236170

Ågren, A. M., Larson, J., Paul, S. S., Laudon, H., & Lidberg, W. (2021). Use of multiple LIDAR-derived digital terrain indices and machine learning for high-resolution national-scale soil moisture mapping of the Swedish forest landscape. Geoderma, 404, 115280. https://doi.org/10.1016/j.geoderma.2021.115280

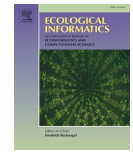
II

Ecological Informatics 80 (2024) 102377

Contents lists available at ScienceDirect

Ecological Informatics

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



Estimation of plant density based on presence/absence data using hybrid inference

Léna Gozé [a,*], Magnus Ekström [a,b], Saskia Sandring [a], Bengt-Gunnar Jonsson [c], Jörgen Wallerman [a], Göran Ståhl [a]

[a] Department of Forest Resource Management, Swedish University of Agricultural Sciences, Skogsmarksgränd, 901 83 Umeå, Sweden
[b] Department of Statistics, USBE, Umeå University, Stastistics, 901 87 Umeå, Sweden
[c] Department of Natural Sciences, Design and Sustainable Development, Mid Sweden University, 851 70 Sundsvall, Sweden

ARTICLE INFO

Keywords:
Binary regression
Forest inventory data
Inhomogeneous Poisson point processes
Plant monitoring
Vegetation survey

ABSTRACT

Monitoring of plant populations has become more and more important, especially in the current context of environmental change. In this paper, we propose methods to estimate plant density from presence/absence surveys, wherein the presence or absence of each species is recorded on sample plots. Presence/absence sampling is a useful and relatively simple method for monitoring state and change of plant communities. Moreover, it has advantages compared to traditional plant cover assessment, the latter being more prone to observer bias. We present a hybrid estimation framework, that combines model- and design-based inference features, in which a generalised linear model (for binary presence/absence data) and an inhomogeneous Poisson model (for plant locations) are used to estimate plant density in a region of interest. We look at two different cases, the first one with a known area and the second one where the area is unknown and must be estimated. Our methods are applied to real data on *Vaccinium vitis-idaea* from the Swedish National Forest Inventory as well as simulated data to assess the performance of our estimators of plant density and corresponding variance estimators. The results obtained are promising and indicate that this method has a potential to add considerable analytic strength to monitoring programmes that collect presence/absence data.

1. Introduction

Collecting data on ground vegetation in forests is an important part of environmental monitoring, e.g., as part of initiatives for assessing trends in biodiversity (e.g., Pain et al. 2020; CBD 2002) or reporting within international agreements, such as the EU's Habitats Directive (Commission of the European Communities 2003). The demands for such monitoring programmes are currently increasing (e.g. O'Connor et al. 2020). However, monitoring plant populations is far from trivial. The methods applied should preferably be cost-efficient, easy to apply, and use protocols that avoid assessment errors. Methods based on assessing plant cover fulfil the first two requirements, but they tend to be prone to observer bias and variability due to phenology (e.g., Gallegos Torell & Glimskär 2009; Futschik et al. 2020; Kennedy & Addison 1987; Kercher et al. 2003).

In some cases, especially if the sample plots are not too large, methods based on presence/absence (P/A) sampling are less prone to errors of the kinds mentioned above (e.g., Ringvall et al. 2005; Kercher et al. 2003), since only the presence or absence of target species within plots needs to be registered. Some studies also suggest that P/A-data could be more useful than cover data in characterizing plant communities (e.g., Bastow Wilson 2012). On the other hand, whereas state and change in terms of vegetation cover or plant density are straightforward to interpret, state and change in terms of presence or absence frequencies are vaguer measures, which depend on sample plot size (e.g., Ståhl et al. 2017). However, if plant spatial occurrences are modelled, large-area estimates in terms of state and change of plant density or vegetation cover can be derived from P/A data (Ekström et al. 2020; Ståhl 2003) through application of model-based inference (e.g., Cassel et al. 1977; Warton et al. 2015). In addition, if a model for the probability that at least one plant will occur on a given plot (or pixel) depends on one or more auxiliary variables, then the model-based inferential framework assumes the availability of wall-to-wall auxiliary variables (cf. Fortin et al. 2023).

Auxiliary information is becoming increasingly available through different remote sensing techniques (e.g., Olsson 2020; Baena et al.

* Corresponding author.
E-mail address: lena.goze@slu.se (L. Gozé).

https://doi.org/10.1016/j.ecoinf.2023.102377
Received 12 July 2023; Received in revised form 10 November 2023; Accepted 11 November 2023
Available online 21 November 2023

2018; Dubayah et al. 2022) and so are data about presence of species through citizen science data collection programs (e.g., the Species Observation System in Sweden (Artdatabanken 2022) or the Atlas of Living Australia and its citizen science data portal (Belbin 2011)), which can be combined with P/A data (Fithian et al. 2015). Thus, opportunities for modelling plant occurrence are much better today compared to some decades ago. This type of modelling, with the availability of wall-to-wall auxiliary information from, e.g., remote sensing, can offer information in terms of both estimates and maps. Estimates are needed, e.g., for trend analysis and reporting to agreements such as the Habitats Directive mentioned above. Maps are useful for implementing management plans related to preserving threatened species (Baena et al. 2018) or limiting the impact of invasive species.

As the degree of detail in the auxiliary data increases, it will be possible to develop better models for plant occurrences, thus facilitating model-based estimation of plant density with higher precision. Dense networks of field plots from National Forest Inventories (NFI, e.g., Fridman et al. 2014; Tomppo et al. 2010) could provide such auxiliary data, because very detailed descriptions of biotic and abiotic conditions, including soil variables, are made on such plots. However, with sample plot data alone, i.e. without wall-to-wall data, it is not possible to apply the standard theory of model-based inference. Instead, hybrid inference can be an alternative (e.g., Corona et al. 2014; Ståhl et al. 2016), where features of model-based and design-based inference are combined.

Examples of applications of hybrid inference include biomass surveys based on LiDAR sample data in Norway (Ståhl et al. 2011) and North America (Margolis et al. 2015), biomass prediction for temperate and pan-tropical regions in the context of the Global Ecosystem Dynamics Investigation project (Saarela et al. 2022), comparison of forest biomass estimates based on coarse and fine resolution data in the USA (McRoberts et al. 2019), and estimation of growing stock volume in Italy (Corona et al. 2014), Finland (Saarela et al. 2015), and Spain (Condés and McRoberts 2017). It has been applied to a broad variety of models, such as mixed-effect models (Fortin et al. 2016) and more complex models where variance estimation requires resampling methods such as the parametric bootstrap (Fortin et al. 2018).

Using conventional model-based inference, Ekström et al. (Unpublished results) investigated the use of P/A data for regional estimation of plant density for a selection of plant species occurring mainly in forests. The main components of the study were inhomogeneous Poisson point processes for modelling the spatial locations of plants and generalised linear models (GLMs) with a complementary log-log link function for associating P/A data with the intensity of the point process, taking auxiliary remotely sensed data into account. As will be described in detail later, a similar modelling approach is used in the present study, with the important difference that auxiliary data were obtained from a large probability sample rather than from wall-to-wall remote sensing. A GLM with a complementary log-log link function for modelling P/A data has also been used in other studies, such as Yee & Mitchell (1991), Royle & Dorazio (2008), Lindenmayer et al. (2009), Baddeley et al. (2010) or Fithian et al. (2015). However, contrary to these articles, which focus on pixel-wise estimation for, e.g., producing maps, our study focuses on obtaining large-area estimates of plant density based on data collected exclusively from sample plots. To our knowledge, no previous studies that make use of hybrid inference have been conducted based on GLMs.

A complementary log-log link function has also been used for modelling of presence-only data (e.g., Phillips et al. (2017); Wan et al. (2017); Sreekumar & Nameer (2022)), although none of them make use of hybrid inference. In addition, it should be mentioned that the standard logit link is frequently used in studies analysing P/A data of species occurrences (e.g., Foody 2008; Ekström et al. 2018; Esseen et al. 2022; Esseen & Ekström 2023). However, for the case where the locations of plants are regarded as a realisation of an inhomogeneous Poisson point process, Baddeley et al. (2010) provide an explanation of why the complementary log-log link function should be preferred for modelling P/A data.

The objective of this study is to assess the usefulness of hybrid inference for estimating plant density, where GLMs estimated from a small sample of P/A data (and auxiliary data) were applied to a large sample of auxiliary data from the Swedish NFI. An important part of the study is to develop formal plant density estimators, variances, and variance estimators for this approach, because no previous studies are available where hybrid inference has been applied in this modelling context. The performance of our estimators and corresponding variance estimators was examined through Monte Carlo simulations and the use of empirical NFI data on a common dwarf shrub, *Vaccinium vitis-idaea*.

We choose to focus our study on estimating the expected plant density (we refer to (13) for a precise definition) rather than on predicting the actual plant density (which is a random quantity in our study setting). The main reason is that this approach simplifies the analyses to some extent meanwhile, for large-area surveys, the relative difference between actual plant density and its expected value is very small, if the models used are approximately correct (cf. Ståhl et al. 2016). The motivation for studying plant density rather than the absolute number of plants is that density is a more relevant measure for plants with large populations (in contrast to many animals), and because the measure allows for comparison between regions of different size.

2. Methods

In this section, we first explain the necessary basis for our derivations, then propose estimators of the expected number of plants in a region of interest U, where U can be, e.g., a municipality, a province or a country. Furthermore, we develop variance formulas and corresponding variance estimators. The estimator of the expected density, defined as the expected number of plants per unit area, is thereafter obtained via the estimator of the expected number of plants and is presented for two cases: one with known area a_U of U and one with unknown area. We also look at the case where we want to estimate the expected density for a specific domain within U, for example the forested part of U. Two different sampling designs are considered. In the first design, plot centres are sampled according to some joint probability density function on U, or rather the union of U and a so-called "buffer" for handling edge effects (Subsections 2.2–2.4). In the second design, centres of clusters of plots are sampled rather than individual plot centres (Subsection 2.5).

2.1. Models

Assume that the plant population is generated by an inhomogeneous Poisson point process with intensity

$$\lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) = \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{u})\right), \boldsymbol{u} \in U \subset \mathbb{R}^2 \tag{1}$$

(Baddeley et al. 2010), where $\boldsymbol{\beta} \in \mathbb{R}^q$ denotes the vector of model parameters and $\boldsymbol{x}(\boldsymbol{u})$ denotes a covariate vector (of length q) at point $\boldsymbol{u}$. The expected number of plants in U is then given by

$$\Lambda(\boldsymbol{\beta}) = \int_U \lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) d\boldsymbol{u}. \tag{2}$$

We consider plots $C(\boldsymbol{u}_i)$, where index i designates plot i, and where the plot centres $\{\boldsymbol{u}_i\}$ are selected according to some specified sampling design. Let N_i denote the number of plants in $C(\boldsymbol{u}_i) \cap U$. Our assumptions imply that N_i is Poisson distributed, and then

$$\mathbb{E}(N_i) = \int_{C(\boldsymbol{u}_i)\cap U} \lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) d\boldsymbol{u} = \int_{C(\boldsymbol{u}_i)\cap U} \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{u})\right) d\boldsymbol{u}.$$

Unless stated otherwise, we assume, as an approximation, that $\boldsymbol{x}(\boldsymbol{u})$ is constant in a sample plot, and thus $\boldsymbol{x}(\boldsymbol{u}) = \boldsymbol{x}(\boldsymbol{u}_i) = \boldsymbol{x}_i$ for all $\boldsymbol{u} \in C(\boldsymbol{u}_i)$, and

$$\mathbb{E}(N_i) = a_i \exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i\right), \tag{3}$$

with a_i being the area of the intersection of plot $C(\boldsymbol{u}_i)$ and the region of interest U (cf. Baddeley et al. 2010). Since N_i is Poisson-distributed, the probability of presence can be expressed by

$$p_i = 1 - P(N_i = 0) = 1 - \exp\big(-a_i \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)\big) \tag{4}$$

so that the loglikelihood for the binary response variables (i.e. P/A data from $C(\boldsymbol{u}_i) \cap U$) becomes the loglikelihood of a complementary log-log regression with an offset equal to the log of the plot area, i.e. of the binary regression model given by

$$g(p_i) = \log(a_i) + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i, \text{ where } g(p) = \log(-\log(1-p)). \tag{5}$$

According to Baddeley et al. (2010), the corresponding likelihood may be regarded as an approximation of the likelihood that would have been obtained without the assumption of constant covariate data in a plot.

2.2. Estimation of the expected number of plants in U

Hybrid inference can be used when covariate information is not available everywhere in the region of interest but only at sample plot level, for example for budgetary reasons (Ståhl et al. 2016). As stated in the introduction, this hybrid method includes aspects of both design-based and model-based inference. As in, amongst others, the papers by Ståhl et al. (2011), Nelson et al. (2012), Corona et al. (2014), Saarela et al. (2015) or Saarela et al. (2022) on hybrid inference, we utilise two samples that are readily available, for instance in monitoring programme databases. Our first sample S_1 of size n_1 contains plot centre locations for plots with both binary response data and covariate data, while our second sample S_2 of size n_2 contains plot centre locations for plots with only covariate data. Typically, n_2 is much larger than n_1. Sample S_1 is used only to establish a model and estimate the vector of model coefficients in a GLM (as opposed to, e.g., Ståhl et al. (2011), where a standard linear model is used). Thereafter, the fitted GLM and covariate information from S_2 are used to predict expected numbers of plants on all plots with centres in S_2, and subsequently the expected plant density in the region of interest, using design-based estimation and Horvitz-Thompson-like estimators. Sample plots with centre locations in S_1 and S_2 do not necessarily need to have the same size, and the sampling designs used to obtain the data in S_1 and S_2 are allowed to differ.

When sampling from a finite population, the well-known Horvitz-Thompson estimator (Horvitz & Thompson 1952) is often used for obtaining estimates of population parameters. However, in our case the population is not finite but a continuous set of locations, and therefore we use Cordy's continuous analogue of the Horvitz-Thompson estimator (Cordy 1993), which we introduce next.

Let f be the joint probability density function (pdf) for sample $S_2 = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{n_2}\}$, and $f_i(\boldsymbol{u})$ the marginal pdf for point $\boldsymbol{u}_i$. The inclusion density function is

$$\pi(\boldsymbol{u}) = \sum_{i=1}^{n_2} f_i(\boldsymbol{u}), \tag{6}$$

and it can intuitively be considered as a local measure of the number of sample points to be selected per unit area (Cordy 1993). If, for example, the points in S_2 are independent and identically distributed (iid), this means that $\pi(\boldsymbol{u}) = n_2 f_1(\boldsymbol{u})$.

The inclusion zone for a point $\boldsymbol{u} \in U$ consists of all points in the frame that would result in the inclusion of $\boldsymbol{u}$ if they were selected to the sample. It may be formally written as $K(\boldsymbol{u}) = \{\boldsymbol{u}' \in U : \boldsymbol{u} \in C(\boldsymbol{u}')\}$, where $C(\boldsymbol{u}')$ is a plot centred around point $\boldsymbol{u}'$. For simplicity purposes, we assume from here on that all plots $C(\boldsymbol{u}_i)$, $i \in S_2$, are circular and have the same area a. The area of the inclusion zone of $\boldsymbol{u} \in U$ is $\widetilde{a_{\boldsymbol{u}}} = \int_U I(\boldsymbol{u} \in C(\boldsymbol{u}'))d\boldsymbol{u}'$. If point $\boldsymbol{u}$ is sufficiently into the interior of U, then its inclusion zone will have the same shape and size as each of the circular plots. On the other hand, if $\boldsymbol{u}$ is close enough to the boundary of U, then its inclusion zone will have a smaller size than a. The Horvitz–Thompson-type estimator presented below has the ability to take this into account, but would require the inclusion zone area to be determined for each point $\boldsymbol{u}_i \in S_2$ near the edge (cf. Gregoire & Valentine 2007). A less labour-intensive way to solve this problem is to use the so-called buffer method, which applies to both the single-plot and cluster-plot designs. Thus, we suppose that a buffer at least as large as the plot radius is used around U (Gregoire & Valentine 2007). This allows sample points $\boldsymbol{u}_i$ to fall outside U, i.e. in some larger region $U^{\bullet}$, defined as the union of U and the buffer. The use of a buffer impacts the definitions of $\widetilde{a_{\boldsymbol{u}}}$ and $K(\boldsymbol{u})$, in which U needs to be replaced by $U^{\bullet}$. The introduction of a buffer implies that all points in U have the same inclusion zone area, and thus $\widetilde{a_{\boldsymbol{u}}} = a_{\boldsymbol{u}} = a$ for all $\boldsymbol{u} \in U$, where $a_{\boldsymbol{u}}$ denotes the area of $C(\boldsymbol{u})$. In this setting, we set $\lambda_\beta(\boldsymbol{u}) = 0$ for all $\boldsymbol{u} \in U$ (cf. Gregoire & Valentine 2007).

The "generalised" Horvitz-Thompson estimator of the expected number of plants in U is then given by

$$\widehat{\Lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n_2} \frac{\lambda(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}, \tag{7}$$

where $\pi(\boldsymbol{u})$ is given by (6) and

$$\lambda(\boldsymbol{u}) = \int_{C(\boldsymbol{u})} \frac{\lambda_\beta(\boldsymbol{u}')}{a_{\boldsymbol{u}'}} d\boldsymbol{u}', \boldsymbol{u} \in U^{\bullet},$$

is the average intensity over $C(\boldsymbol{u}_i)$, where $a_{\boldsymbol{u}'} = a$ by our assumptions (Cordy 1993, Grafström et al. 2017). Note that

$$\begin{aligned} \int_{U^{\bullet}} \lambda(\boldsymbol{u}) d\boldsymbol{u} &= \int_{U^{\bullet}} \int_{C(\boldsymbol{u})} \frac{\lambda_\beta(\boldsymbol{u}')}{a_{\boldsymbol{u}'}} d\boldsymbol{u}' d\boldsymbol{u} = \int_{U^{\bullet}} \frac{\lambda_\beta(\boldsymbol{u}')}{a_{\boldsymbol{u}'}} \int_{U^{\bullet}} I(\boldsymbol{u}' \in C(\boldsymbol{u})) d\boldsymbol{u} d\boldsymbol{u}' \\ &= \int_U \frac{\lambda_\beta(\boldsymbol{u}')}{a_{\boldsymbol{u}'}} \int_{U^{\bullet}} I(\boldsymbol{u}' \in C(\boldsymbol{u})) d\boldsymbol{u} d\boldsymbol{u}' = \int_U \lambda_\beta(\boldsymbol{u}') d\boldsymbol{u}' = \Lambda(\boldsymbol{\beta}) \end{aligned} \tag{8}$$

and, according to Theorem 1 in Cordy (1993), this implies that the Horvitz-Thompson estimator of $\Lambda(\boldsymbol{\beta})$ is unbiased if $\pi(\boldsymbol{u}) > 0$ for all $\boldsymbol{u} \in U^{\bullet}$. Hence, with a buffer for handling edge effects, we obtain an unbiased estimator of $\Lambda(\boldsymbol{\beta})$. The price to be paid is that the buffer method tends to inflate the variance of the estimator (Gregoire & Valentine 2007). If the area of the buffer is small relative to the area of U, this increase in variance can be expected to be small. Using (3), $\lambda(\boldsymbol{u}_i)$ can be rewritten as

$$\lambda(\boldsymbol{u}_i) = \int_{C(\boldsymbol{u}_i) \cap U} \frac{\exp(\boldsymbol{\beta}^{\mathrm{T}} x(\boldsymbol{u}))}{a_{\boldsymbol{u}}} d\boldsymbol{u} = \frac{a_i}{a} \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) = r_i \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i) = \widetilde{\lambda}_\beta(\boldsymbol{u}_i),$$

where r_i is the ratio of the area a_i of $C(\boldsymbol{u}_i) \cap U$ and the area of $C(\boldsymbol{u}_i)$. With $\widetilde{\lambda}_\beta(\boldsymbol{u}_i)$ defined as above, note that if $C(\boldsymbol{u}_i) \subseteq U$, then $\widetilde{\lambda}_\beta(\boldsymbol{u}_i) = \lambda_\beta(\boldsymbol{u}_i)$. This implies that

$$\widehat{\Lambda}(\boldsymbol{\beta}) = \sum_{i=1}^{n_2} \frac{\widetilde{\lambda}_\beta(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} = \sum_{i=1}^{n_2} \frac{r_i \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)}{\pi(\boldsymbol{u}_i)}.$$

$\widehat{\Lambda}(\boldsymbol{\beta})$ can also be regarded as a natural predictor of the actual number of plants, given the available information and in the context of the inhomogeneous Poisson point process. As $\boldsymbol{\beta}$ is usually unknown, we will use $\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})$ as our estimator of the expected number of plants, where $\widehat{\boldsymbol{\beta}}$ is an

estimator of $\boldsymbol{\beta}$ obtained using model (5) based on data from S_1.

2.3. Variance estimation

To estimate the variance of the estimator $\widehat{\Lambda}(\boldsymbol{\beta})$ of $\Lambda(\boldsymbol{\beta})$, we use the Sen-Yates-Grundy variance formula defined in Cordy (1993),

$$\mathrm{Var}(\widehat{\Lambda}(\boldsymbol{\beta}))=\frac{1}{2}\int_{U^\bullet}\int_{U^\bullet}\Delta(\boldsymbol{u},\boldsymbol{u}')\left(\frac{\lambda(\boldsymbol{u})}{\pi(\boldsymbol{u})}-\frac{\lambda(\boldsymbol{u}')}{\pi(\boldsymbol{u}')}\right)^2 d\boldsymbol{u}d\boldsymbol{u}',$$

where

$$\Delta(\boldsymbol{u},\boldsymbol{u}')=\pi(\boldsymbol{u})\pi(\boldsymbol{u}')-\pi(\boldsymbol{u},\boldsymbol{u}') \quad \text{and} \quad \pi(\boldsymbol{u},\boldsymbol{u}')=\sum_{i\in I_n}\sum_{j\in J_{n,i}} f_{ij}(\boldsymbol{u},\boldsymbol{u}'), \tag{9}$$

the latter being the pairwise inclusion density function with $I_n=\{1,\dots,n_2\}$, $J_{n,i}=\{1,\dots,n_2\}\backslash\{i\}$, and f_{ij} the joint marginal pdf of $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$. As advised by, e.g., Tillé (2006), the Sen-Yates-Grundy formula should be used in case a fixed sample size is used. By Cordy (1993), if $\pi(\boldsymbol{u})$ and $\pi(\boldsymbol{u},\boldsymbol{u}')$ are strictly positive for all $(\boldsymbol{u},\boldsymbol{u}')\in U^\bullet$, an unbiased estimator of the Sen-Yates-Grundy variance is given by

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\Lambda}(\boldsymbol{\beta}))&=\frac{1}{2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{\lambda(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}-\frac{\lambda(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}\right)^2\\&=\frac{1}{2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{r_i\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i)}{\pi(\boldsymbol{u}_i)}-\frac{r_j\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_j)}{\pi(\boldsymbol{u}_j)}\right)^2,\end{aligned} \tag{10}$$

and that is in effect the part of the variance due to sampling of the plot centres in S_2, treating the model coefficients as known. With unknown $\boldsymbol{\beta}$, i.e. where $\boldsymbol{\beta}$ needs to be estimated by $\widehat{\boldsymbol{\beta}}$, an estimate of the variance of $\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})$ can be expressed as

$$\begin{aligned}\widehat{\mathrm{Var}}(\widehat{\Lambda}(\widehat{\boldsymbol{\beta}}))&=\frac{1}{2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i\boldsymbol{u}_j)}\left(\frac{r_i\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i)}{\pi(\boldsymbol{u}_i)}-\frac{r_j\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_j)}{\pi(\boldsymbol{u}_j)}\right)^2\\&\quad+\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k\widehat{\beta}_l)\widehat{v}_k\widehat{v}_l,\end{aligned} \tag{11}$$

with

$$\widehat{v}_k=\sum_{i=1}^{n_2}\frac{1}{\pi(\boldsymbol{u}_i)}\widetilde{\lambda}^{(k)}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i), \tag{12}$$

where $\widehat{\beta}_k$ denotes the kth component of the $\widehat{\boldsymbol{\beta}}$ vector, and

$$\widetilde{\lambda}^{(k)}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i)=\frac{\partial\widetilde{\lambda}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i)}{\partial\widehat{\beta}_k}=r_i x_{ik}\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i)$$

with x_{ik} denoting the kth component of $\boldsymbol{x}_i$. The different $\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)$ terms can be obtained from statistical software, for example using the *glm* function in R. The derivation of (11) can be found in Appendix A. Another case, where S_2 is a sample of centres of plot clusters, is considered in Subsection 2.5.

2.4. Estimation of the expected plant density

In this section, we utilise our estimator of the total number of plants for estimating the expected plant density. First, we assume that the area of the region of interest is known. In this case, the expected density $R(\boldsymbol{\beta})$ is defined as the expected number of plants in the region divided by the area a_U of U,

$$R(\boldsymbol{\beta})=\frac{\Lambda(\boldsymbol{\beta})}{a_U}, \tag{13}$$

where $\Lambda(\boldsymbol{\beta})$ is defined in (2). This quantity can be estimated by

$$\widehat{R}(\widehat{\boldsymbol{\beta}})=\frac{\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})}{a_U}, \tag{14}$$

where $\widehat{\Lambda}(\boldsymbol{\beta})$ is defined in (7). Its corresponding variance estimator is given by

$$\widehat{\mathrm{Var}}(\widehat{R}(\widehat{\boldsymbol{\beta}}))=\frac{\widehat{\mathrm{Var}}(\widehat{\Lambda}(\widehat{\boldsymbol{\beta}}))}{a_U}, \tag{15}$$

where $\widehat{\mathrm{Var}}(\widehat{\Lambda}(\widehat{\boldsymbol{\beta}}))$ is the same as in (11).

However, information about the area of the region of interest may not be available, or we may wish to estimate expected plant density in a subregion of unknown area, for example in the forested area of a region. In such cases, the area has to be estimated. Thus, $\Lambda(\boldsymbol{\beta})$ needs to be modified as

$$\Lambda^\star(\boldsymbol{\beta})=\int_U\lambda_{\boldsymbol{\beta}}(\boldsymbol{u})I_{\boldsymbol{u}}d\boldsymbol{u},$$

with $I_{\boldsymbol{u}}$ being an indicator function taking the value 1 if $\boldsymbol{u}$ is situated in the target part of the landscape and 0 otherwise; $I_{\boldsymbol{u}}$ is set to 0 outside of U. The area of the target part of the landscape in U can be written as

$$A=\int_U I_{\boldsymbol{u}}d\boldsymbol{u}$$

and the expected plant density in the area of interest is given by

$$R^\star(\boldsymbol{\beta})=\frac{\Lambda^\star(\boldsymbol{\beta})}{A}. \tag{16}$$

This quantity can be estimated by

$$\widehat{R}^\star(\boldsymbol{\beta})=\frac{\widehat{\Lambda}^\star(\boldsymbol{\beta})}{\widehat{A}}, \tag{17}$$

where $\widehat{\Lambda}^\star(\boldsymbol{\beta})$ is defined as

$$\widehat{\Lambda}^\star(\boldsymbol{\beta})=\sum_{i=1}^{n_2}\frac{\lambda^\star(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}, \tag{18}$$

where

$$\lambda^\star(\boldsymbol{u}_i)=\int_{C(\boldsymbol{u}_i)}\frac{\lambda_{\boldsymbol{\beta}}(\boldsymbol{u})I_{\boldsymbol{u}}}{a_{\boldsymbol{u}}}d\boldsymbol{u},$$

and

$$\widehat{A}=\sum_{i=1}^{n_2}\frac{z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} \tag{19}$$

is an estimator of the area A, with

$$z(\boldsymbol{u}_i)=\int_{C(\boldsymbol{u}_i)}\frac{I_{\boldsymbol{u}}}{a_{\boldsymbol{u}}}d\boldsymbol{u}.$$

Note that, if we adopt a reasoning similar to the one in (8), $\widehat{A}$ is an unbiased estimator of A if $\pi(\boldsymbol{u})>0$ for all $\boldsymbol{u}\in U^\bullet$ (Cordy 1993).

In Appendix A, the following estimator of the variance of $\widehat{R}^\star(\widehat{\boldsymbol{\beta}})$ is derived:

$$\widehat{\mathrm{Var}}(\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}}))=\frac{1}{2\widehat{A}^2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{\widehat{\lambda}^{\star}(\boldsymbol{u}_i)-\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}-\frac{\widehat{\lambda}^{\star}(\boldsymbol{u}_j)-\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}\right)^2$$
$$+\frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{\widehat{d}_{1,k}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)\widehat{d}_{2,k}/\widehat{A}}{\pi(\boldsymbol{u}_i)}\sum_{j\in I_n}\frac{\widehat{d}_{1,l}(\boldsymbol{u}_j)-z(\boldsymbol{u}_j)\widehat{d}_{2,l}/\widehat{A}}{\pi(\boldsymbol{u}_j)} \tag{20}$$
$$+\frac{2}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{d}_{2,l}\sum_{i\in I_n}\frac{\widehat{d}_{1,k}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}-\frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{d}_{2,k}\widehat{d}_{2,l},$$

where

$$\widehat{\lambda}^{\star}(\boldsymbol{u}_i)=\int_{C(\boldsymbol{u}_i)}\frac{\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u})I_{\boldsymbol{u}}}{a_{\boldsymbol{u}}}d\boldsymbol{u},$$

$$\widehat{d}_{1,k}(\boldsymbol{u}_i)=\int_{C(\boldsymbol{u}_i)}\frac{I_{\boldsymbol{u}}}{a_{\boldsymbol{u}}}\lambda_{\widehat{\boldsymbol{\beta}}}^{(k)}(\boldsymbol{u})d\boldsymbol{u},\quad \widehat{d}_{2,k}=\sum_{i=1}^{n_2}\frac{I_{\boldsymbol{u}_i}\lambda_{\widehat{\boldsymbol{\beta}}}^{(k)}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}, \tag{21}$$

$$\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})=\sum_{j\in I_n}\frac{\lambda(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}=\sum_{j\in I_n}\frac{1/k_j\sum_{i=1}^{k_j}r_i\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^j)}{\pi(\boldsymbol{u}_j)}. \tag{23}$$

Using the same reasoning that led us to (11), we obtain the following variance estimators for $\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})$;

$$\widehat{\mathrm{Var}}(\widehat{\Lambda}(\widehat{\boldsymbol{\beta}}))=\frac{1}{2}\sum_{j\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{1}{\pi(\boldsymbol{u}_j)k_j}\sum_{i=1}^{k_j}r_i\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^j)-\frac{1}{\pi(\boldsymbol{u}_j)k_j}\sum_{l=1}^{k_j}r_l\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_l^j)\right)^2+\sum_{k=1}^{p}\sum_{k'=1}^{p}\widehat{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_{k'})\widehat{v}_k\widehat{v}_{k'}, \tag{24}$$

and

$$\lambda_{\widehat{\boldsymbol{\beta}}}^{(k)}(\boldsymbol{u}_i)=\frac{\partial\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i)}{\partial\widehat{\beta}_k}=x_{ik}\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i).$$

It can happen that sample plots are divided into several parts, for example if one part of the plot is in forests and other parts are in other landscape categories. In such cases, some adjustments of the above estimators of the expected plant density and variance are needed. See Appendix B.

2.5. Cluster sampling case

It is also of interest to consider the case where S_2 is a sample of centres of clusters (sometimes called tracts) of plots rather than a sample of centres of individual plots. Indeed, this sampling procedure is used in, e.g., the Swedish NFI (Anon 2014). In this case, $C(\boldsymbol{u}_j)$ denotes a cluster j of k_j plots centred around $\boldsymbol{u}_j$, and we denote the area of the plots within the cluster by $a_{\boldsymbol{u}_j}=k_j s$, where s is the area of a single plot (all plots are assumed to have the same area). A buffer is also used in this case, although it will be larger (at least as large as the radius of the tract, see Grafström et al. 2017). We can still use the Horvitz-Thompson estimator (7) to get our estimator of the expected number of plants in U; the resulting expression will just be slightly different.

Using approximation (3) and if no plot is divided,

$$\lambda(\boldsymbol{u}_j)=\int_{C(\boldsymbol{u}_j)\cap U}\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{u}))}{a_{\boldsymbol{u}}}d\boldsymbol{u}=\frac{1}{k_j}\sum_{i=1}^{k_j}r_i\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i^j), \tag{22}$$

where $\boldsymbol{x}_i^j$ denotes the (constant) covariate information in plot i of cluster j, and r_i is the ratio of the area of the intersection of plot i n cluster j and U to the area of a single plot. Then, the Horvitz-Thompson estimator $\widehat{\Lambda}(\boldsymbol{\beta})$ may be written as

with

$$\widehat{v}_k=\sum_{j\in I_n}\frac{1}{\pi(\boldsymbol{u}_j)}\frac{1}{k_j}\sum_{i=1}^{k_j}r_i x_{ik}^j\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i^j),$$

where $\boldsymbol{x}_{ik}^j$ denotes the kth component of vector $\boldsymbol{x}_i^j$. Similar changes are made in case we want to estimate expected plant density in (sub)regions with unknown area.

2.6. Statistical testing

The estimates of the expected plant density and corresponding variance estimators rely on the condition that the binary regression model (5) is realistic. For this reason, it is of importance to assess whether said model, used to estimate $\boldsymbol{\beta}$, holds true. In order to do that, we use a parametric bootstrap test suggested by Ekström et al. (Unpublished results). It should be noted that if model (5) is incorrect, then so is the underlying Poisson model assumption. Details on how to perform the test are given in Appendix C.

3. Real data study

The Swedish NFI (Fridman et al. 2014) is a field sample plot inventory of Swedish forests that consists of both temporary and permanent tracts, each composed of several plots. The temporary plots (which have a radius of 7 m) are only inventoried once, while the permanent plots are inventoried once every 5 years. Moreover, the permanent tracts are separated into two subcategories, "C_1", where both terrain and vegetation inventories are conducted, and "C_2", which denotes all other tracts. At each permanent "C_1" plot, P/A data for a set of plant species are recorded on each of two small circular "vegetation plots"; those small vegetation plots have an area of 0.25 m^2 each and are separated by 5 m and located 2.5 m from the main plot centre, the main plot having a radius of 10 m. Those registrations are not made during each visit, but rather once every two visits (i.e. every tenth year). Vegetation

registrations are not made on temporary plots. The covariates are registered at main plot level for both temporary and permanent plots. Thus, values of the covariates are always the same in each pair of small vegetation plots. The registrations are performed by experienced field workers on plots for which the positions are defined in advance according to the given sampling design.

We chose to study Lingonberry (*Vaccinium vitis-idaea*) data in the Norrbotten Lappmarken region (in northern Sweden) during the years 2008–2012. According to the Swedish NFI, region Norrbotten Lappmarken has a known area of 7,785,748 ha. The particular landscape category we chose for the estimation of $R^\star$ and its corresponding variance is productive forestland (i.e. land that can produce on average at least 1 m^3 of wood per hectare and per year and that is not significantly used for other purposes, according to Anon (2014)), whose area is unknown.

Sample S_1 consists of the centres of the small vegetation plots included in permanent "C_1" plots, in Norrbotten Lappmarken during 2008–2012. Sample S_1 has size $n_1 = 724$, corresponding to 362 pairs of vegetation plots that were used for the parametric bootstrap test. Cluster sampling was used to obtain sample S_2. It originally consists of the centres of the tracts of temporary circular plots. This sample has a size of $n_2 = 111$ tract centres, which corresponds to 1132 sample plots in total. There are one to twelve plots with available data in each (quadratic) tract, and the plots are separated by at least 600 m (Anon 2014).

In Table 1, the fitted binary regression model for *Vaccinium vitis-idaea* is presented for productive forestland in Norrbotten Lappmarken for years 2008–2012. The model was not rejected by the parametric bootstrap test (p-value = 0.184). Its explanatory variables are a transformation of the number of tree stems per hectare, multiplied by 100, and an indicator variable stating whether the soil is humid/wet. It can be seen that *Vaccinium vitis-idaea* seem less likely to be found on humid/wet soil, compared to dry soils. On the other hand, the model suggests that the more tree stems per hectare, the higher the probability of presence of *Vaccinium vitis-idaea*.

Table 2 contains estimated expected densities in two different cases. The first case is cluster sampling, where centres of clusters of plots were assumed to be sampled independently and uniformly on $U^\bullet$. In the second case, the computations were made by (incorrectly) assuming that centres of individual plots were sampled rather than centres of clusters. The densities were estimated using two different estimators (expected density estimator with known area (14) and unknown area (17), and their cluster sampling case counterparts). The corresponding variance estimates, (15) and (20) respectively (as well as their cluster sampling case counterparts), are also given. In both cases, the variance estimate of the expected density estimator in productive forestland is almost twice as high as the variance estimate using the whole region. It can be explained by the relatively small amount of plots that are situated in productive forestland in Norrbotten Lappmarken in the Swedish NFI data (approximately 50% of the total).

Table 1
Estimated model coefficients $\hat{\boldsymbol{\beta}}$ for *Vaccinium vitis-idaea* in productive forestland in Norrbotten Lappmarken. The intercept was offset-adjusted. $\mathbf{1}_{wet}$ is an indicator variable stipulating whether a plot is humid/wet or not. $((\text{No.stems/ha} + 0.6)/1000)^{-0.5}$ is a non-linear transformation of the "number of tree stems per hectare" (in hundreds per hectare) covariate, found by using the mfp R package (Ambler & Benner 2015), which applies multivariable fractional polynomials (Sauerbrei & Royston 1999).

Species	Estimated parameters ($\hat{\boldsymbol{\beta}}$)	
Vaccinium vitis-idaea (Lingonberry)	Offset-adjusted Intercept	2.423
	$\mathbf{1}_{wet}$	−0.667
	$((\text{No.stems/ha} + 0.6)/1000)^{-0.5}$	−0.025

Table 2
Estimated expected plant densities in m^{-2} and corresponding estimates of variance for *Vaccinium vitis-idaea* in Norrbotten Lappmarken. Two cases were considered: one where the computations were made assuming cluster sampling and another where it was (incorrectly) assumed that single plots were sampled. $\widehat{R}(\hat{\boldsymbol{\beta}})$ and $\widehat{\text{Var}}(\widehat{R}(\hat{\boldsymbol{\beta}}))$ are computed for the whole Norrbotten Lappmarken region, while $\widehat{R}^\star(\hat{\boldsymbol{\beta}})$ and $\widehat{\text{Var}}(\widehat{R}^\star(\hat{\boldsymbol{\beta}}))$ are computed for the productive forestland area of Norrbotten Lappmarken only.

Case	$\widehat{R}(\hat{\boldsymbol{\beta}})$	$\widehat{R}^\star(\hat{\boldsymbol{\beta}})$	$\widehat{\text{Var}}(\widehat{R}(\hat{\boldsymbol{\beta}}))$	$\widehat{\text{Var}}(\widehat{R}^\star(\hat{\boldsymbol{\beta}}))$
Tracts	7.61	9.72	0.205	0.406
Single plots	7.49	9.73	0.209	0.411

4. Monte Carlo study

The aim of the Monte Carlo study was to evaluate our estimators of expected plant density and variance estimators and assess whether they performed well. The simulations, all performed in **R** (R Core Team 2022), were conducted as follows.

- We created a quadratic grid of 1024 cells that corresponds to our area frame U, as well as a buffer zone around U. Each grid cell had an area of 1 ha and artificial covariates.
- The created covariates were based on the ones included in the model for *Vaccinium vitis-idaea*. The indicator variable stipulating whether a plot is humid/wet or not was built on actual data in the Norrbotten Lappmarken region between 2008 and 2012, which had approximately 16.85% of plots being considered as humid/wet. This particular covariate was created as realisations of a Bernoulli distribution with parameter $p = 0.1685$ in each cell. As for the number of stems per hectare, we used fitted Weibull distributions as described below. Two cases were considered:
  1. In the first case, we assumed that the whole grid was productive forestland, and the area of the area frame (the cell grid) was assumed to be known. In that case, we supposed that the number of stems per hectare varied only depending on whether the soil was humid/wet or dry. Based on Swedish NFI data in productive forestland, Weibull distributions were fitted using the *fitdist* function from the fitdistrplus package (Delignette-Muller & Dutang 2015). On humid/wet grid cells, the fitted distribution was a Weibull distribution with shape parameter $k = 1.047$ and scale parameter $\lambda = 3898.3$. For the dry grid cells, a two-step procedure was used since 4% of the original data had values equal to 0. Therefore, a random number between 0 and 1 was generated for each grid cell; if this number was smaller than 0.04, the number of stems per hectare for that grid cell was set to 0; otherwise it was a realisation of a Weibull-distributed random variable with parameters $k = 0.903$ and $\lambda = 2076.5$.
  2. In the second case, we created an indicator variable which was assigned the value 1 if the cell was in productive forestland, and 0 otherwise. As 49.8% of the original sample plots are in productive forestland, each cell was assigned the value 1 with a probability of 0.498. The number of stems per hectare was supposed to vary according to both humidity of the soil and type of landscape (productive forestland or not), which means that four different subcases had to be considered. The area of productive forestland in the grid was estimated by (19). The covariates were generated exclusively for the cells that are situated in productive forestland (which means in two of the subscases), and in such case were generated exactly as in case 1.
- Each Monte Carlo simulation consisted of 2000 replicates; P/A data were generated from an inhomogeneous Poisson point process with the *rpoispp* function from the spatstat package (Baddeley et al. 2016) in each replicate; plot centres in S_2 were sampled independently according to a uniform distribution over $U^\bullet$, while a two-step generation procedure was used for S_1: first, plot centres for the

Table 3
Actual expected plant densities $R(\boldsymbol{\beta})$ (resp. $R^{\star}(\boldsymbol{\beta})$), estimated mean values of the estimated expected densities $\widehat{\mathbb{E}}(\widehat{R}(\widehat{\boldsymbol{\beta}}))$ (resp. $\widehat{\mathbb{E}}(\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}}))$), estimated mean value of the variance estimates $\widehat{\mathbb{E}}(\widehat{\mathrm{Var}}(\widehat{R}(\widehat{\boldsymbol{\beta}})))$ (resp. $\widehat{\mathbb{E}}(\widehat{\mathrm{Var}}(\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})))$) and s^2, the sample variance of the $\widehat{R}(\widehat{\boldsymbol{\beta}})$ (resp. $\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})$), for simulated *Vaccinium vitis-idaea* data in a grid of 1024 cells, each cell having an area of 1 ha. In the known area case, the area is a_U, the area of the grid. In the unknown area case, the area is estimated according to (19). The variances were estimated using formulas (15) and (20). "/" means that the formula does not apply to the specific case.

Case	$R(\boldsymbol{\beta})$	$R^{\star}(\boldsymbol{\beta})$	$\widehat{\mathbb{E}}(\widehat{R}(\widehat{\boldsymbol{\beta}}))$	$\widehat{\mathbb{E}}(\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}}))$	$\widehat{\mathbb{E}}(\widehat{\mathrm{Var}}(\widehat{R}(\widehat{\boldsymbol{\beta}})))$	$\widehat{\mathbb{E}}(\widehat{\mathrm{Var}}(\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})))$	s^2
Known area	9.740	/	9.606	/	0.191	/	0.196
Unknown area	/	9.715	/	9.657	/	0.187	0.189

permanent plots were sampled independently according to a uniform distribution over U, and then the small vegetation plots in S_1 were created for each permanent plot as described in Section 3. The value of the vector of coefficients $\boldsymbol{\beta}$ was set equal to the one from the fitted model for *Vaccinium vitis-idaea* in Norrbotten Lappmarken in years 2008–2012 (Table 1). Estimated model coefficients $\widehat{\boldsymbol{\beta}}$ were computed for every replicate using the S_1 data, while the estimated expected plant density and its corresponding variance estimate were computed for every replicate using the S_2 data. The sample sizes were $n_1 = 1500$ and $n_2 = 1500$. The same plot radii as in the Swedish NFI were used (see Section 3). The plots in S_2 were divided when they overlapped different grid cells (see details in Appendix B). In accordance with the Swedish NFI (Jonas Dahlgren, personal communication), the small vegetation plots within S_1 were not divided.

The results for the simulation study are presented in Table 3. The estimator $\widehat{R}(\widehat{\boldsymbol{\beta}})$ was used for Case 1 and $\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})$ was used for Case 2. In Case 1, the estimator $\widehat{R}(\widehat{\boldsymbol{\beta}})$ was on average close to but a little lower than the real expected plant density. In Case 2, the estimator $\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})$ was even closer to the true value, but even in that case a slight negative bias occurred. The two variance estimators seem to have a very small bias and have low values. Based on these observations, we can conclude that our estimators performed quite well.

5. Discussion

In this study, we show how P/A data can be used for modelling and monitoring plant population densities. We argue that this approach offers advantages over methods based on visual assessment of vegetation cover, since studies indicate that P/A sampling may not be as prone to observer bias as methods based on assessing vegetation cover, and since P/A sampling is a rapid and thus cheap method to apply (e.g., Ringvall et al. 2005).

Since the auxiliary modelling data are available for both considered samples, but the binary response data are available for only one sample, we apply methods from hybrid inference (e.g., Corona et al. 2014) for estimating the expected value of plant density and the corresponding variance. This concerns taking into account both modelling and sampling uncertainty, and to our knowledge, our study is the first one that involves GLMs in hybrid inference. This type of inference is important in this context since, in many cases, detailed descriptions of environmental conditions, needed for the modelling, may not be available wall-to-wall but only from sampling locations, e.g., from sample plots within environmental monitoring programmes. In this article, we extend the already existing theory on hybrid inference to GLMs with binary response data.

Our method is most suitable when n_2, the sample size of S_2, is much larger than n_1, the sample size of S_1. Indeed, the main purpose in applying this method is to gather a minimum of information to develop a reliable model on the smallest sample possible (principally due to budgetary reasons), to then apply this model in connection with covariates that come from a larger sample whose units do not contain the desired response data. However, with our available data, n_2 was only a little larger than n_1. This shows that our method works even in that particular case.

In regions with high perimeter-to-area ratios, a large or very large proportion of the sampling plots will extend beyond the region's boundary. In such cases, our suggested methodology, which uses a "buffer" to address edge effects, may be unsuitable and could result, for example, in estimators with larger variances than desired.

An important part of the study involves making the proposed hybrid inference framework available for practical application in monitoring programmes, in which case we need to take into account that sample plots are often allocated in clusters and that the area of the domain of study is unknown (e.g., Fridman et al. 2014). This introduces several additional details to the general framework, which are important for the usefulness of the framework in practice.

The Monte Carlo simulations we performed show that our framework for estimating the expected plant density provides accurate estimates when the modelling assumptions are valid. In the study based on empirical data from the Swedish NFI, we obtained estimates of expected Lingonberry (*Vaccinium vitis-idaea*) densities in Northern Sweden that appear to be realistic, although we cannot check them since no reference data are available.

For the sake of simplicity, we assumed that the sampling design of S_1 was non-informative (see Appendix A), i.e. the design was not taken into account during model parameter estimation. Ignoring an informative sampling design may yield biased estimates of regression coefficients. For handling informative designs, methods using probability weighting may be used (e.g., Heeringa et al. 2010; Ekström et al. 2018).

It is possible to generalise the considered hybrid inference framework to other types of GLMs. Instead of P/A data as a response variable, one could use a continuous variable (such as biomass) or a discrete variable such as a count variable (number of trees, birds etc.). The main requirement is to have two samples; one to estimate model coefficients, with both covariate and response data, and another one, with only covariate data, for estimation of, e.g., expected biomass per hectare or expected plant density based on the estimated model coefficients. As long as this requirement is met, then hybrid inference should work, in principle, with any kind of response variable. The statistical developments would, however, be different from the ones derived in the present paper; although with counts instead of P/A, the difference would not be that significant (in both cases, it would be possible to use an inhomogeneous Poisson model). With count data that are not subject to too many errors, it should be possible to obtain better estimators than the ones obtained from P/A data. However, the survey would be more expensive to conduct.

There is one key condition for the developed technique to be applicable; the underlying point process should be, at least approximately, an inhomogeneous Poisson point process. We estimate models that utilise a combination of P/A and auxiliary data to estimate expected plant density, assuming that the spatial distribution of plants follow an inhomogeneous Poisson process, i.e. the plant densities vary due to the environmental conditions. In the article, we check the suitability of the binary regression model implied by the underlying inhomogeneous Poisson point process through a statistical test specifically developed for the purpose (cf. Appendix C). Recognising that plants can occur in clustered spatial patterns, extensions from inhomogeneous Poisson point processes to inhomogeneous cluster point processes serve as an

important topic for further studies. However, if we would like to use a similar methodology as in Ekström et al. (2020), we would need to gather data on more than two subplots for each main plot.

In our paper, the intensity of the inhomogeneous Poisson point process is determined via a log-linear model that involves a number of covariates. This model cannot be fitted directly, since no observed point pattern or observed values of counts of points in plots are available. This problem is circumvented by making use of observable P/A variables. Given that the pattern is a realisation of an inhomogeneous Poisson point process (whose intensity on the ith cell is given by (1)), it follows that the P/A variables satisfy a binary GLM, with complementary log-log link and an offset, with the same parameter vector as that which appears in the intensity of the inhomogeneous Poisson point process. Thus, for extending the current approach to other inhomogeneous point processes than the Poisson, the parameters of their intensities must be estimable from P/A data and corresponding covariate data at plot level. In addition, estimates of covariance matrices of estimators of parameters are also needed. One possibility to achieve this is to extend the intensity estimator in Ekström et al. (2020) from homogeneous cluster point processes such as the Matérn and Thomas processes to corresponding heterogeneous processes, whose intensities are functions of on one or more covariates (Waagepetersen 2007).

When the point pattern is generated by an inhomogeneous Poisson point process, the binary GLM model in (5) will have independent binary (P/A) response variables conditional on the covariates. For other point processes, responses cannot be expected to fulfill this property of conditional independence. Then, instead of using a standard GLM, other estimation methods such as generalised estimating equations (Albert & McShane 1995; Gotway & Stroup 1997) and a composite likelihood approach for spatial binary data (Heagerty & Lele 1998) can be used. However, as mentioned, this is not enough for extending the current approach to more general point processes. Most importantly, the estimable unknown parameters in the regression model for the P/A data must also include all unknown parameters in the intensity function of the point process model.

For a Poisson point process with a homogeneous intensity λ, the species abundance N in a plot C of area a follows a Poisson distribution with mean $a\lambda$, and the probability of presence of at least one plant in the plot C equals $p = 1 - \exp(-a\lambda)$. Rearranging this equation, we can estimate the intensity (plant density) λ from the proportion $\widehat{p}$ of plots with plant occurrences, i.e., by $\widehat{\lambda} = -a^{-1}\log(1-\widehat{p})$ (e.g., Ståhl et al. 2017). A homogeneous spatial Poisson process is synonymous with complete spatial randomness. However, in nature, individuals of many species are typically aggregated (Pielou 1977; He & Gaston 2000). For plot abundance N, the model most commonly used to describe such aggregation is the negative binomial distribution (He et al. 2002), which implies the following relationship between the presence probability p and plant density λ, $p = 1 - \left(1 + k^{-1}\lambda\right)^{-k}$, where k is referred to as a "clumping" parameter, with small $k > 0$ representing strong aggregation (Wright 1991; He & Gaston 2000; He et al. 2002). Under this model, Conlisk et al. (2007) specify the likelihood function and conclude that the clumping parameter cannot be estimated from P/A data, i.e., that it must be specified from outside the model. The suitability of the negative binomial distribution has also been much debated (Holt et al. 2002; Gaston et al. 2011) and only two known homogeneous point processes give the negative binomial distribution for plot abundances, and both are extreme cases (Daley & Vere-Jones 2003). For some further developments of the negative binomial distribution model, we refer to Solow & Smith (2010), Hwang & Huggins (2016), Huggins et al. (2018), Hwang et al. (2022), and Stoklosa et al. (2022). For other suggested models than those based on the Poisson and the negative binomial distributions for describing the relationship between the presence probability p and plant density λ, see, e.g., Holt et al. (2002), He et al. (2002), and the references therein. Extensions of the negative binomial model and other related models to an inhomogeneous setting would be useful for extending the approach presented in the current article to more general settings.

Many monitoring and citizen science programmes already have large amounts of P/A data in their databases (e.g., the Norwegian Biodiversity Information Center in Norway (Hoem 2022); the Global Biodiversity Information Facility GBIF (GBIF 2022)). Therefore, the techniques and estimators developed in the present study can be applied to already available data, especially since new fine-scaled covariate data are becoming increasingly common in such databases. Promising results were obtained in this study, which means that the proposed framework for monitoring plant population density through P/A sampling and modelling holds promise for future practical application, e.g., in national reporting of trends in declining species.

Author contributions

LG wrote the main draft, performed the analyses and simulations and contributed to the theoretical developments; ME conceived the idea, contributed to the theoretical developments and contributed critically to the drafts; SS, BGJ, JW and GS contributed critically to the drafts.

Funding

This work was supported by Kempestiftelserna (SMK-1955).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data will be made available upon request.

Acknowledgements

We thank Jonas Dahlgren for having provided the data used in Section 3.

Appendix A. Theoretical developments in the case of single plots

A.1. Case with known area

For simplicity, we assume that the sampling design of S_1 is non-informative, i.e. the vector of model parameters is estimated without taking this sampling design into account. Under this assumption, for large samples and under mild conditions (see for example Sen & Singer 1993),

$$\sqrt{n_2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{N}\left(0, I^{-1}(\boldsymbol{\beta})\right), \tag{A.1}$$

where $I(\boldsymbol{\beta})$ denotes the Fisher information matrix and can be estimated by

$$\widehat{I}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n_2}\sum_{i\in I_n}\frac{1}{[g'(p_i(\widehat{\boldsymbol{\beta}}))]^2 v_i(\widehat{\boldsymbol{\beta}})}\ \boldsymbol{x}_i\boldsymbol{x}_i', \tag{A.2}$$

with $\widehat{\boldsymbol{\beta}}$ being the estimate of $\boldsymbol{\beta}$, g defined by (5), p_i defined by (4), and $v_i(\boldsymbol{\beta}) = \mathrm{Var}(Y_i) = p_i(1-p_i)$, where $Y_i = 1$ if there is presence of plants in plot i, and $Y_i = 0$ otherwise.

Using a similar reasoning as in [Ståhl et al. 2011], we start with the decomposition

$$\widehat{\Lambda}(\widehat{\boldsymbol{\beta}}) - \Lambda(\boldsymbol{\beta}) = \sum_{i\in I_n}\frac{\widetilde{\lambda}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} - \Lambda = D_1 + D_2, \tag{A.3}$$

where

$$D_1 = \sum_{i\in I_n}\frac{\widetilde{\lambda}_{\boldsymbol{\beta}}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} - \Lambda \quad \text{and} \quad D_2 = \sum_{i\in I_n}\frac{\widetilde{\lambda}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i) - \widetilde{\lambda}_{\boldsymbol{\beta}}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}.$$

Our objective is to compute the variance

$$\mathrm{Var}(D_1 + D_2) = \mathrm{Var}(D_1) + \mathrm{Var}(D_2) + 2\ \mathrm{Cov}(D_1, D_2).$$

Using the Sen-Yates-Grundy formula presented in Cordy (1993), an unbiased estimator of $\mathrm{Var}(D_1)$ is given by (10). If $\boldsymbol{\beta}$ is unknown, we estimate this variance with

$$\widehat{\mathrm{Var}}(D_1) = \frac{1}{2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{r_i\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i)}{\pi(\boldsymbol{u}_i)} - \frac{r_j\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_j)}{\pi(\boldsymbol{u}_j)}\right)^2. \tag{A.4}$$

The law of total variance is used in order to compute $\mathrm{Var}(D_2)$, i.e

$$\mathrm{Var}(D_2) = \mathrm{Var}_{S_2}[\mathbb{E}_{S_1}(D_2|S_2)] + \mathbb{E}_{S_2}[\mathrm{Var}_{S_1}(D_2|S_2)]. \tag{A.5}$$

For non-linear models, a Taylor approximation can be applied, i.e.

$$\widetilde{\lambda}_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}) \approx \widetilde{\lambda}_{\boldsymbol{\beta}}(\boldsymbol{u}) + \sum_{k=1}^{q}(\widehat{\beta}_k - \beta_k)\widetilde{\lambda}^{(k)}_{\boldsymbol{\beta}}(\boldsymbol{u}), \tag{A.6}$$

where

$$\widetilde{\lambda}^{(k)}_{\boldsymbol{\beta}}(\boldsymbol{u}_i) = r_i x_{ik}\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i).$$

Then,

$$D_2 \approx \sum_{i\in I_n}\sum_{k=1}^{q}\frac{(\widehat{\beta}_k - \beta_k)}{\pi(\boldsymbol{u}_i)}\ \widetilde{\lambda}^{(k)}_{\boldsymbol{\beta}}(\boldsymbol{u}_i) = \sum_{k=1}^{q}(\widehat{\beta}_k - \beta_k)v_k,$$

where

$$v_k = \sum_{i\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)}\widetilde{\lambda}^{(k)}_{\boldsymbol{\beta}}(\boldsymbol{u}_i)$$

and q being the number of model coefficients. Conditioned on S_2, v_k is a constant. Then, by (A.1), $\mathbb{E}_{S_1}(D_2|S_2) \approx \sum_{k=1}^{q}\mathbb{E}_{S_1}(\widehat{\beta}_k - \beta_k|S_2)v_k \approx 0$ for large samples, and thus $Var_{S_2}[\mathbb{E}_{S_1}(D_2|S_2)] \approx 0$. Furthermore,

$$
\begin{aligned}
\mathrm{Var}_{S_1}(D_2|S_2) &\approx \mathrm{Var}_{S_1}\left(\sum_{k=1}^{q}(\widehat{\beta}_k-\beta_k)v_k\,|S_2\right)\\
&\approx \sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)v_k v_l\\
&= \sum_{i\in I_n}\sum_{j\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)\pi(\boldsymbol{u}_j)}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)r_i\,r_j\,x_{ik}\,x_{jl}\exp\left(\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}_i+\boldsymbol{x}_j)\right)\\
&= \sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{r_i^2}{\pi(\boldsymbol{u}_i)^2}x_{ik}x_{il}\exp\left(2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}_i\right)\\
&+\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{r_i r_j}{\pi(\boldsymbol{u}_i)\pi(\boldsymbol{u}_j)}x_{ik}\,x_{jl}\,\exp\left(\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}_i+\boldsymbol{x}_j)\right).
\end{aligned}
$$

From the arguments in the proof of Theorem 2 in Cordy (1993), we get

$$
\begin{aligned}
\mathrm{Var}(D_2) &\approx \mathbb{E}_{S_2}[\mathrm{Var}_{S_1}(D_2|S_2)]\\
&= \sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\star}\frac{r_{\boldsymbol{u}}^2}{\pi(\boldsymbol{u})}\boldsymbol{x}^k(\boldsymbol{u})\boldsymbol{x}^l(\boldsymbol{u})\exp\left(2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{u})\right)d\boldsymbol{u}\\
&+\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\star}\int_{U^\star}\frac{\pi(\boldsymbol{u},\boldsymbol{u}')}{\pi(\boldsymbol{u})\pi(\boldsymbol{u}')}r_{\boldsymbol{u}}r_{\boldsymbol{u}'}\boldsymbol{x}^k(\boldsymbol{u})\boldsymbol{x}^l(\boldsymbol{u})\exp\left(\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x}(\boldsymbol{u})+\boldsymbol{x}(\boldsymbol{u}'))\right)d\boldsymbol{u}d\boldsymbol{u}',
\end{aligned}
$$

where $\boldsymbol{x}^k(\boldsymbol{u})$ denotes the kth component of the $\boldsymbol{x}$ vector and $r_{\boldsymbol{u}}$ is the ratio of the area of $C(\boldsymbol{u})\cap U$ and the area of $C(\boldsymbol{u})$. Thus, $Var(D_2)$ can be estimated by

$$
\begin{aligned}
\widehat{\mathrm{Var}}(D_2) &= \sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{r_i^2}{\pi(\boldsymbol{u}_i)^2}\,x_{ik}x_{il}\exp\left(2\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i\right)\\
&+\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{r_i r_j}{\pi(\boldsymbol{u}_i)\pi(\boldsymbol{u}_j)}\,x_{ik}x_{jl}\exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{x}_i+\boldsymbol{x}_j)\right)\\
&= \sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in I_n}\frac{r_i r_j}{\pi(\boldsymbol{u}_i)\pi(\boldsymbol{u}_j)}\,x_{ik}x_{jl}\exp\left(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{x}_i+\boldsymbol{x}_j)\right)\\
&= \sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{v}_k\widehat{v}_l,
\end{aligned}
\tag{A.7}
$$

where $\widehat{v}_k$ is defined in (12).

The next step is to compute the covariance between D_1 and D_2. According to the law of total covariance,

$$
\mathrm{Cov}(D_1,D_2)=\mathbb{E}_{S_2}[\mathrm{Cov}_{S_1}(D_1,D_2|S_2)]+\mathrm{Cov}_{S_2}[\mathbb{E}_{S_1}(D_1|S_2),\mathbb{E}_{S_1}(D_2|S_2)]. \tag{A.8}
$$

It can be deduced that $\mathrm{Cov}_{S_2}[\mathbb{E}_{S_1}(D_1|S_2),\mathbb{E}_{S_1}(D_2|S_2)]\approx 0$ because, as argued before, $\mathbb{E}_{S_1}(D_2|S_2)\approx 0$. Then, as the stochastic nature of D_1 is determined by sample S_2 and not by sample S_1, $\mathbb{E}_{S_1}(D_1D_2|S_2)=D_1\mathbb{E}_{S_1}(D_2|S_2)\approx 0$. Because of the latter, $\mathbb{E}_{S_2}[\mathrm{Cov}_{S_1}(D_1,D_2|S_2)]\approx 0$. Thus, $\mathrm{Cov}(D_1,D_2)\approx 0$ and we just need to add the variances of D_1 and D_2 to get an approximate variance of D_1+D_2. As a result, setting (A.4) and (A.7) together, the estimate becomes

$$
\begin{aligned}
\widehat{\mathrm{Var}}(\widehat{\Lambda}(\widehat{\boldsymbol{\beta}})) &= \widehat{\mathrm{Var}}(D_1)+\widehat{\mathrm{Var}}(D_2)+2\widehat{\mathrm{Cov}}(D_1,D_2)\\
&= \frac{1}{2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{r_i\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i)}{\pi(\boldsymbol{u}_i)}-\frac{r_j\exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_j)}{\pi(\boldsymbol{u}_j)}\right)^2+\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{v}_k\widehat{v}_l,
\end{aligned}
$$

with $\widehat{v}_k$ defined in (12).

A.2. Expected density estimator in a specific area of the landscape

Suppose we want to estimate the number of plants exclusively in a certain landscape category, for example forests. Then, the parameter vector $\boldsymbol{\beta}$ will be estimated only from the plots that are situated in this landscape category.

As in Result 5.6.2 in Särndal et al. (1992), for estimating the variance of $\widehat{R}^\star(\boldsymbol{\beta})$ we use a Taylor linearisation by introducing $\widehat{R}_0^\star(\boldsymbol{\beta})$, that is related to $\widehat{R}^\star(\boldsymbol{\beta})$ by the relation

$$
\widehat{R}^\star(\boldsymbol{\beta})\approx\widehat{R}_0^\star(\boldsymbol{\beta})=R^\star(\boldsymbol{\beta})+\frac{1}{A}\sum_{i\in I_n}\frac{\lambda^\star(\boldsymbol{u}_i)-R^\star(\boldsymbol{\beta})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}. \tag{A.9}
$$

Remember that the estimator of $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, is approximately normally distributed with mean $\boldsymbol{\beta}$ (see (A.1)). We estimate $R^{\star}(\boldsymbol{\beta})$ with $\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})$. The goal here is to derive an estimate of the variance of $\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}})$, or equivalently the variance of $\widehat{R}^{\star}(\widehat{\boldsymbol{\beta}}) - R^{\star}(\boldsymbol{\beta})$, which by the arguments in the proof of Result 5.6.2 in Särndal et al. (1992) is approximately the same as the one for

$$D(\widehat{\boldsymbol{\beta}}) = \widehat{R}_0^{\star}(\widehat{\boldsymbol{\beta}}) - R^{\star}(\widehat{\boldsymbol{\beta}}) = \frac{1}{A}\sum_{i\in I_u}\frac{\widehat{\lambda}^{\star}(\boldsymbol{u}_i) - R^{\star}(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)},$$

where

$$\widehat{\lambda}^{\star}(\boldsymbol{u}) = \int_{C(\mathbf{u})}\frac{\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}')I_{\boldsymbol{u}'}}{a_{\boldsymbol{u}'}}d\boldsymbol{u}', \boldsymbol{u}\in U^{\bullet}. \tag{A.10}$$

We can write

$$\widehat{R}_0^{\star}(\widehat{\boldsymbol{\beta}}) - R^{\star}(\boldsymbol{\beta}) = \left(\widehat{R}_0^{\star}(\widehat{\boldsymbol{\beta}}) - R^{\star}(\widehat{\boldsymbol{\beta}})\right) + \left(R^{\star}(\widehat{\boldsymbol{\beta}}) - R^{\star}(\boldsymbol{\beta})\right) = D(\widehat{\boldsymbol{\beta}}) + D_*(\widehat{\boldsymbol{\beta}}),$$

where $D_*(\widehat{\boldsymbol{\beta}}) = R^{\star}(\widehat{\boldsymbol{\beta}}) - R^{\star}(\boldsymbol{\beta})$. By the following Taylor approximation

$$\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}) \approx \lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) + \sum_{k=1}^{q}(\widehat{\beta}_k - \beta_k)\lambda_{\boldsymbol{\beta}}^{(k)}(\boldsymbol{u}), \tag{A.11}$$

where

$$\lambda_{\boldsymbol{\beta}}^{(k)}(\boldsymbol{u}) = \frac{\partial\lambda_{\boldsymbol{\beta}}(\boldsymbol{u})}{\partial\beta_k},$$

we obtain

$$\begin{aligned}\mathbb{E}[R^{\star}(\widehat{\boldsymbol{\beta}})] &= \mathbb{E}_{S_1}[R^{\star}(\widehat{\boldsymbol{\beta}})] = \frac{1}{A}\mathbb{E}_{S_1}[\Lambda^{\star}(\widehat{\boldsymbol{\beta}})] = \frac{1}{A}\int_U \mathbb{E}_{S_1}[\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u})]I_{\boldsymbol{u}}d\boldsymbol{u}\\ &\approx \frac{1}{A}\int_U \lambda_{\boldsymbol{\beta}}(\boldsymbol{u})I_{\boldsymbol{u}}d\boldsymbol{u} + \frac{1}{A}\sum_{k=1}^{q}\mathbb{E}_{S_1}[\widehat{\beta}_k - \beta_k]\int_U \lambda_{\boldsymbol{\beta}}^{(k)}(\boldsymbol{u})I_{\boldsymbol{u}}d\boldsymbol{u} \approx \frac{1}{A}\int_U \lambda_{\boldsymbol{\beta}}(\boldsymbol{u})I_{\boldsymbol{u}}d\boldsymbol{u} = R^{\star}(\boldsymbol{\beta})\end{aligned} \tag{A.12}$$

and

$$\begin{aligned}\mathbb{E}\left[(R^{\star}(\widehat{\boldsymbol{\beta}}))^2\right] &= \mathbb{E}_{S_1}\left[(R^{\star}(\widehat{\boldsymbol{\beta}}))^2\right] = \frac{1}{A^2}\mathbb{E}_{S_1}\left[(\Lambda^{\star}(\widehat{\boldsymbol{\beta}}))^2\right]\\ &= \frac{1}{A^2}\int_U\int_U \mathbb{E}_{S_1}[\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u})\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}')]I_{\boldsymbol{u}}I_{\boldsymbol{u}'}d\boldsymbol{u}d\boldsymbol{u}'\\ &\approx \frac{1}{A^2}\int_U\int_U \lambda_{\boldsymbol{\beta}}(\boldsymbol{u})\lambda_{\boldsymbol{\beta}}(\boldsymbol{u}')I_{\boldsymbol{u}}I_{\boldsymbol{u}'}d\boldsymbol{u}d\boldsymbol{u}' + \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,k}d_{2,l}\\ &= (R^{\star}(\boldsymbol{\beta}))^2 + \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,k}d_{2,l},\end{aligned}$$

where

$$d_{2,k} = \int_U I_{\boldsymbol{u}}\lambda_{\boldsymbol{\beta}}^{(k)}(\boldsymbol{u})d\boldsymbol{u} = \frac{\partial\Lambda^{\star}(\boldsymbol{\beta})}{\partial\beta_k}.$$

Thus,

$$\mathbb{E}[D_*(\widehat{\boldsymbol{\beta}})] = \mathbb{E}_{S_1}[D_*(\widehat{\boldsymbol{\beta}})] \approx 0 \tag{A.13}$$

and

$$\mathrm{Var}(D_*(\widehat{\boldsymbol{\beta}})) = \mathrm{Var}_{S_1}(D_*(\widehat{\boldsymbol{\beta}})) \approx \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,k}d_{2,l}. \tag{A.14}$$

Let us go further with $D(\widehat{\boldsymbol{\beta}})$. We have

$$\mathrm{Var}(D(\widehat{\boldsymbol{\beta}})) = \mathrm{Var}_{S_2}[\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})|S_2)] + \mathbb{E}_{S_2}[\mathrm{Var}_{S_1}(D(\widehat{\boldsymbol{\beta}})|S_2)]. \tag{A.15}$$

We see that

$$\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2) = \frac{1}{A}\sum_{i\in I_n}\frac{\mathbb{E}_{S_1}(\widehat{\lambda}^{\star}(\mathbf{u}_i)\,|S_2) - \mathbb{E}_{S_1}(R^{\star}(\widehat{\boldsymbol{\beta}})\,|S_2)z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}$$

and, by (A.11), we obtain

$$\begin{aligned}\mathbb{E}_{S_1}[\widehat{\lambda}^{\star}(\boldsymbol{u})\,|S_2] &= \int_{C(\boldsymbol{u})}\frac{1}{a_{\boldsymbol{u}}}E_{S_1}[\lambda_{\beta}(\boldsymbol{u}')]I_{\boldsymbol{u}}d\boldsymbol{u}'\\ &\approx \int_{C(\boldsymbol{u})}\frac{1}{a_{\boldsymbol{u}}}\lambda_{\beta}(\boldsymbol{u}')I_{\boldsymbol{u}}d\boldsymbol{u}' + \sum_{k=1}^{q}E_{S_1}[\widehat{\beta}_k-\beta_k]\int_{C(\boldsymbol{u})}\frac{1}{a_{\boldsymbol{u}}}I_{\boldsymbol{u}}\lambda_{\beta}^{(k)}(\boldsymbol{u}')d\boldsymbol{u}'\\ &\approx \int_{C(\boldsymbol{u})}\frac{1}{a_{\boldsymbol{u}}}\lambda_{\beta}(\boldsymbol{u}')I_{\boldsymbol{u}}d\boldsymbol{u}' = \lambda^{\star}(\boldsymbol{u}).\end{aligned}$$

Thus,

$$\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2) \approx \frac{1}{A}\sum_{i\in I_n}\frac{\lambda^{\star}(\boldsymbol{u}_i) - R^{\star}(\boldsymbol{\beta})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} = \widehat{R}_0^{\star}(\boldsymbol{\beta}) - R^{\star}(\boldsymbol{\beta}) \tag{A.16}$$

and, from the Sen-Yates-Grundy formula presented in Cordy (1993),

$$\begin{aligned}&\mathrm{Var}_{S_2}[\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)] \approx \mathrm{Var}_{S_2}\left(\widehat{R}_0^{\star}(\boldsymbol{\beta})\right)\\ &= \frac{1}{2A^2}\int_{U^{\star}}\int_{U^{\star}}\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)\left(\frac{\lambda^{\star}(\boldsymbol{u}) - R^{\star}(\boldsymbol{\beta})z(\boldsymbol{u})}{\pi(\boldsymbol{u})} - \frac{\lambda^{\star}(\boldsymbol{u}') - R^{\star}(\boldsymbol{\beta})z(\boldsymbol{u}')}{\pi(\boldsymbol{u}')}\right)^2.\end{aligned} \tag{A.17}$$

Then, we can look closer at

$$\mathrm{Var}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2) = \mathbb{E}_{S_1}\left(D^2(\widehat{\boldsymbol{\beta}})\,|S_2\right) - (\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2))^2,$$

which is a part of (A.15), where

$$\mathbb{E}_{S_1}\left(D^2(\widehat{\boldsymbol{\beta}})\,|S_2\right) = \frac{1}{A^2}\sum_{i\in I_n}\sum_{j\in I_n}\mathbb{E}_{S_1}\left[\left(\frac{\widehat{\lambda}^{\star}(\boldsymbol{u}_i) - R^{\star}(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}\right)\left(\frac{\widehat{\lambda}^{\star}(\boldsymbol{u}_j) - R^{\star}(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}\right)\middle|S_2\right]. \tag{A.18}$$

From (A.11), we see that

$$\begin{aligned}\mathbb{E}_{S_1}[\widehat{\lambda}^{\star}(\boldsymbol{u})\widehat{\lambda}^{\star}(\boldsymbol{u}')] &\approx \int_{C(\boldsymbol{u})}\int_{C(\boldsymbol{u}')}a_{\boldsymbol{v}}^{-1}a_{\boldsymbol{v}'}^{-1}\lambda_{\beta}(\boldsymbol{v})\lambda_{\beta}(\boldsymbol{v}')I_{\boldsymbol{v}}I_{\boldsymbol{v}'}d\boldsymbol{v}'d\boldsymbol{v}\\ &\quad+\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{1,k}(\boldsymbol{u})d_{1,l}(\boldsymbol{u}')\\ &= \lambda^{\star}(\boldsymbol{u})\lambda^{\star}(\boldsymbol{u}') + \sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{1,k}(\boldsymbol{u})d_{1,l}(\boldsymbol{u}'),\end{aligned} \tag{A.19}$$

where

$$d_{1,k}(\boldsymbol{u}) = \int_{C(\boldsymbol{u})}a_{\boldsymbol{u}}^{-1}I_{\boldsymbol{u}}\lambda_{\beta}^{(k)}(\boldsymbol{u}')d\boldsymbol{u}' = \int_{C(\boldsymbol{u})}a_{\boldsymbol{u}}^{-1}I_{\boldsymbol{u}}x(\boldsymbol{u}')_k\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{u}')\right)d\boldsymbol{u}',$$

and that

$$\begin{aligned}&\mathbb{E}_{S_1}[\widehat{\lambda}^{\star}(\boldsymbol{u})R^{\star}(\widehat{\boldsymbol{\beta}})] = \frac{1}{A}\int_{C(\boldsymbol{u})}a_{\boldsymbol{v}}^{-1}\mathbb{E}_{S_1}[\lambda_{\widehat{\beta}}(\boldsymbol{v})\Lambda^{\star}(\widehat{\boldsymbol{\beta}})]I_{\boldsymbol{v}}d\boldsymbol{v} = \frac{1}{A}\int_{U}\int_{C(\boldsymbol{u})}a_{\boldsymbol{v}}^{-1}\mathbb{E}_{S_1}[\lambda_{\widehat{\beta}}(\boldsymbol{v})\lambda_{\widehat{\beta}}(\boldsymbol{v}')]I_{\boldsymbol{v}}I_{\boldsymbol{v}'}d\boldsymbol{v}d\boldsymbol{v}'\\ &\approx\frac{1}{A}\int_{U}\int_{C(\boldsymbol{u})}a_{\boldsymbol{v}}^{-1}\lambda_{\beta}(\boldsymbol{v})\lambda_{\beta}(\boldsymbol{v}')I_{\boldsymbol{v}}I_{\boldsymbol{v}'}d\boldsymbol{v}d\boldsymbol{v}'\\ &\quad+\frac{1}{A}\sum_{k=1}^{q}\sum_{l=1}^{q}Cov_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U}\int_{C(\boldsymbol{u})}a_{\boldsymbol{v}}^{-1}x(\boldsymbol{v})_k\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{v})\right)x(\boldsymbol{v}')_l\exp\left(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}(\boldsymbol{v}')\right)I_{\boldsymbol{v}}I_{\boldsymbol{v}'}d\boldsymbol{v}d\boldsymbol{v}'\\ &= \lambda^{\star}(\boldsymbol{u})R^{\star}(\boldsymbol{\beta}) + \frac{1}{A}\sum_{k=1}^{q}\sum_{l=1}^{q}Cov_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{1,k}(\boldsymbol{u})d_{2,l}.\end{aligned} \tag{A.20}$$

From (A.18), (A.19) and (A.20), we obtain

$$\begin{aligned}\mathbb{E}_{S_1}\left[D^2(\widehat{\boldsymbol{\beta}})\,|S_2\right] &\approx \frac{1}{A^2}\sum_{i\in I_n}\sum_{j\in I_n}\left(\frac{\lambda^\star(\boldsymbol{u}_i)-R^\star(\boldsymbol{\beta})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}\right)\left(\frac{\lambda^\star(\boldsymbol{u}_j)-R^\star(\boldsymbol{\beta})z(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}\right)\\ &+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in I_n}\left(\frac{d_{1,k}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)d_{2,k}/A}{\pi(\boldsymbol{u}_i)}\right)\left(\frac{d_{1,l}(\boldsymbol{u}_j)-z(\boldsymbol{u}_j)d_{2,l}/A}{\pi(\boldsymbol{u}_j)}\right)\\ &=\left(\widehat{R}_0^\star(\boldsymbol{\beta})-R^\star(\boldsymbol{\beta})\right)^2\\ &+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in I_n}\left(\frac{d_{1,k}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)d_{2,k}/A}{\pi(\boldsymbol{u}_i)}\right)\left(\frac{d_{1,l}(\boldsymbol{u}_j)-z(\boldsymbol{u}_j)d_{2,l}/A}{\pi(\boldsymbol{u}_j)}\right).\end{aligned}$$

This, together with (A.16), gives

$$\begin{aligned}&\mathrm{Var}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\approx\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\left(\frac{d_{1,k}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)d_{2,k}/A}{\pi(\boldsymbol{u}_i)}\right)\sum_{j\in I_n}\left(\frac{d_{1,l}(\boldsymbol{u}_j)-z(\boldsymbol{u}_j)d_{2,l}/A}{\pi(\boldsymbol{u}_j)}\right)\\ &=\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)^2}\left(d_{1,k}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)d_{2,l}\frac{1}{A}\right)\\ &+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{1}{\pi(\boldsymbol{u}_i)\pi(\boldsymbol{u}_j)}\left(d_{1,k}(\boldsymbol{u}_i)-z(\boldsymbol{u}_i)d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u}_j)-z(\boldsymbol{u}_j)d_{2,l}\frac{1}{A}\right).\end{aligned}$$

It follows that

$$\begin{aligned}&\mathbb{E}_{S_2}[\mathrm{Var}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\,]\approx\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\star}\frac{1}{\pi(\boldsymbol{u})}\left(d_{1,k}(\boldsymbol{u})-z(\boldsymbol{u})d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u})-z(\boldsymbol{u})d_{2,l}\frac{1}{A}\right)d\boldsymbol{u}\\ &+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\star}\int_{U^\star}\frac{\pi(\boldsymbol{u},\boldsymbol{u}^{'})}{\pi(\boldsymbol{u})\pi(\boldsymbol{u}^{'})}\left(d_{1,k}(\boldsymbol{u})-z(\boldsymbol{u})d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u}^{'})-z(\boldsymbol{u}^{'})d_{2,l}\frac{1}{A}\right)d\boldsymbol{u}d\boldsymbol{u}^{'}.\end{aligned}\tag{A.21}$$

If we put (A.17) and (A.21) together, we obtain

$$\begin{aligned}&\mathrm{Var}(D(\widehat{\boldsymbol{\beta}})\,)=\mathrm{Var}_{S_2}[\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\,]+\mathbb{E}_{S_2}[\mathrm{Var}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\,]\\ &\approx\frac{1}{2A^2}\int_{U^\star}\int_{U^\star}\Delta(\boldsymbol{u}_i\boldsymbol{u}_j)\left(\frac{\lambda^\star(\boldsymbol{u})-R^\star(\boldsymbol{\beta})z(\boldsymbol{u})}{\pi(\boldsymbol{u})}-\frac{\lambda^\star(\boldsymbol{u}^{'})-R^\star(\boldsymbol{\beta})z(\boldsymbol{u}^{'})}{\pi(\boldsymbol{u}^{'})}\right)^2d\boldsymbol{u}d\boldsymbol{u}^{'}\\ &+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\star}\frac{1}{\pi(\boldsymbol{u})}\left(d_{1,k}(\boldsymbol{u})-z(\boldsymbol{u})d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u})-z(\boldsymbol{u})d_{2,l}\ \frac{1}{A}\right)d\boldsymbol{u}\\ &+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\star}\frac{\pi(\boldsymbol{u}\boldsymbol{u}^{'})}{\pi(\boldsymbol{u})\pi(\boldsymbol{u}^{'})}\left(d_{1,k}(\boldsymbol{u})-z(\boldsymbol{u})d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u}^{'})-z(\boldsymbol{u}^{'})d_{2,l}\ \frac{1}{A}\right)d\boldsymbol{u}d\boldsymbol{u}^{'}.\end{aligned}\tag{A.22}$$

Furthermore,

$$\mathrm{Cov}(D(\widehat{\boldsymbol{\beta}})\,,D^*(\widehat{\boldsymbol{\beta}})\,)=\mathrm{Cov}_{S_2}[\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\,,\mathbb{E}_{S_1}(D^*(\widehat{\boldsymbol{\beta}})\,|S_2)\,]+\mathbb{E}_{S_2}[\mathrm{Cov}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,,D^*(\widehat{\boldsymbol{\beta}})\,|S_2)\,].$$

From earlier calculations, we know that $\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\approx\widehat{R}_0^\star(\boldsymbol{\beta})-R^\star(\boldsymbol{\beta})$ and $\mathbb{E}_{S_1}(D^*(\widehat{\boldsymbol{\beta}})\,|S_2)\approx 0$, and thus $\mathrm{Cov}_{S_2}[\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\,,\mathbb{E}_{S_1}(D^*(\widehat{\boldsymbol{\beta}})\,|S_2)\,]\approx 0$. In addition, using (A.16),

$$\begin{aligned}\mathrm{Cov}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,,D^*(\widehat{\boldsymbol{\beta}})\,|S_2) &\approx\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})D^*(\widehat{\boldsymbol{\beta}})\,|S_2)=\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})R^\star(\widehat{\boldsymbol{\beta}})\,|S_2)-R^\star(\boldsymbol{\beta})\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})\,|S_2)\\ &\approx\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})R^\star(\widehat{\boldsymbol{\beta}})\,|S_2)-R^\star(\boldsymbol{\beta})\left(\widehat{R}_0^\star(\boldsymbol{\beta})-R^\star(\boldsymbol{\beta})\right)\end{aligned}$$

and, from (A.12) and (A.20),

$$\mathbb{E}_{S_1}(D(\widehat{\boldsymbol{\beta}})R^\star(\widehat{\boldsymbol{\beta}})\,|S_2) = \frac{1}{A}\mathbb{E}_{S_1}\left(\sum_{i\in I_n}\frac{\widehat{\lambda}^\star(\boldsymbol{u}_i)R^\star(\widehat{\boldsymbol{\beta}}) - (R^\star(\widehat{\boldsymbol{\beta}}))^2 z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}\,\middle|\,S_2\right)$$

$$\approx R^\star(\boldsymbol{\beta})\frac{1}{A}\sum_{i\in I_n}\frac{\lambda^\star(\boldsymbol{u}_i) - R^\star(\boldsymbol{\beta})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}$$

$$+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}Cov_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)}d_{1,k}(\boldsymbol{u}_i)d_{2,l}$$

$$-\frac{1}{A^3}\sum_{k=1}^{q}\sum_{l=1}^{q}Cov_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}d_{2,k}d_{2,l}.$$

As a consequence,

$$\begin{aligned}
\mathrm{Cov}(D(\widehat{\boldsymbol{\beta}})\,, D^*(\widehat{\boldsymbol{\beta}})\,) &\approx \mathbb{E}_{S_2}\left(R^\star(\boldsymbol{\beta})(\widehat{R}_0^\star(\boldsymbol{\beta}) - R^\star(\boldsymbol{\beta})\,) + \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}\left(\widehat{\beta}_k,\widehat{\beta}_l\right)\sum_{i\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)}d_{1,k}(\boldsymbol{u}_i)d_{2,l}\right.\\
&\quad\left.-\frac{1}{A^3}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}d_{2,k}d_{2,l} - R^\star(\boldsymbol{\beta})(\widehat{R}_0^\star(\boldsymbol{\beta}) - R^\star(\boldsymbol{\beta}))\right)\\
&= \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\mathbb{E}_{S_2}\left(\sum_{i\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)}d_{1,k}(\boldsymbol{u}_i)d_{2,l}\right) - \frac{1}{A^3}\sum_{k=1}^{p}\sum_{l=1}^{p}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\mathbb{E}_{S_2}\left(\sum_{i\in I_n}\frac{z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)}d_{2,k}d_{2,l}\right)\\
&= \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,l}\int_{U^\bullet}d_{1,k}(\boldsymbol{u})d\boldsymbol{u} - \frac{1}{A^3}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,k}d_{2,l}\int_{U}z(\boldsymbol{u})d\boldsymbol{u}\\
&= \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,l}\int_{U^\bullet}d_{1,k}(\boldsymbol{u})d\boldsymbol{u} - \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,k}d_{2,l}.
\end{aligned} \tag{A.23}$$

Finally, putting (A.22), (A.14) and (A.23) together,

$$\begin{aligned}
&\mathrm{Var}\left(\widehat{R}_0^\star(\widehat{\boldsymbol{\beta}}) - R^\star(\boldsymbol{\beta})\,\right) = \mathrm{Var}(D(\widehat{\boldsymbol{\beta}})\,) + \mathrm{Var}(D^*(\widehat{\boldsymbol{\beta}})\,) + 2\,\mathrm{Cov}(D(\widehat{\boldsymbol{\beta}})\,, D^*(\widehat{\boldsymbol{\beta}})\,)\\
&\approx \frac{1}{2A^2}\int_{U^\bullet}\int_{U^\bullet}\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)\left(\frac{\lambda^\star(\boldsymbol{u}) - R^\star(\boldsymbol{\beta})z(\boldsymbol{u})}{\pi(\boldsymbol{u})} - \frac{\lambda^\star(\boldsymbol{u}') - R^\star(\boldsymbol{\beta})z(\boldsymbol{u}')}{\pi(\boldsymbol{u}')}\right)^2\\
&\quad+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\bullet}\frac{1}{\pi(\boldsymbol{u})}\left(d_{1,k}(\boldsymbol{u}) - z(\boldsymbol{u})d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u}) - z(\boldsymbol{u})d_{2,l}\frac{1}{A}\right)d\boldsymbol{u}\\
&\quad+\frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\int_{U^\bullet}\int_{U^\bullet}\frac{\pi(\boldsymbol{u},\boldsymbol{u}')}{\pi(\boldsymbol{u})\pi(\boldsymbol{u}')}\left(d_{1,k}(\boldsymbol{u}) - z(\boldsymbol{u})d_{2,k}\frac{1}{A}\right)\left(d_{1,l}(\boldsymbol{u}') - z(\boldsymbol{u}')d_{2,l}\frac{1}{A}\right)d\boldsymbol{u}d\boldsymbol{u}'\\
&\quad+\frac{2}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,l}\int_{U^\bullet}d_{1,k}(\boldsymbol{u})d\boldsymbol{u} - \frac{1}{A^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\mathrm{Cov}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)d_{2,k}d_{2,l}.
\end{aligned}$$

By using Theorem 1 and the variance estimator based on the Sen-Yates-Grundy formula in Cordy (1993), this variance can be estimated by

$$\begin{aligned}
\widehat{\mathrm{Var}}\left(\widehat{R}_0^\star(\widehat{\boldsymbol{\beta}}) - R^\star(\boldsymbol{\beta})\,\right) &= \frac{1}{2\widehat{A}^2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{\widehat{\lambda}^\star(\boldsymbol{u}_i) - \widehat{R}^\star(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} - \frac{\widehat{\lambda}^\star(\boldsymbol{u}_j) - \widehat{R}^\star(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}\right)^2\\
&\quad+\frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{1}{\pi(\boldsymbol{u}_i)^2}\left(\widehat{d}_{1,k}(\boldsymbol{u}_i) - z(\boldsymbol{u}_i)\widehat{d}_{2,k}\frac{1}{\widehat{A}}\right)\left(\widehat{d}_{1,l}(\boldsymbol{u}_i) - z(\boldsymbol{u}_i)\widehat{d}_{2,l}\,\frac{1}{\widehat{A}}\right)\\
&\quad+\frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{1}{\pi(\boldsymbol{u}_i)\pi(\boldsymbol{u}_j)}\left(\widehat{d}_{1,k}(\boldsymbol{u}_i) - z(\boldsymbol{u}_i)\widehat{d}_{2,k}\frac{1}{\widehat{A}}\right)\left(\widehat{d}_{1,l}(\boldsymbol{u}_j) - z(\boldsymbol{u}_j)\widehat{d}_{2,l}\,\frac{1}{\widehat{A}}\right)\\
&\quad+\frac{2}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{d}_{2,l}\sum_{i\in I_n}\frac{\widehat{d}_{1,k}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} - \frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{d}_{2,k}\widehat{d}_{2,l}\\
&= \frac{1}{2\widehat{A}^2}\sum_{i\in I_n}\sum_{j\in J_{n,i}}\frac{\Delta(\boldsymbol{u}_i,\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_i,\boldsymbol{u}_j)}\left(\frac{\widehat{\lambda}^\star(\boldsymbol{u}_i) - \widehat{R}^\star(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} - \frac{\widehat{\lambda}^\star(\boldsymbol{u}_j) - \widehat{R}^\star(\widehat{\boldsymbol{\beta}})z(\boldsymbol{u}_j)}{\pi(\boldsymbol{u}_j)}\right)^2\\
&\quad+\frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\sum_{i\in I_n}\frac{\widehat{d}_{1,k}(\boldsymbol{u}_i) - z(\boldsymbol{u}_i)\widehat{d}_{2,k}/\widehat{A}}{\pi(\boldsymbol{u}_i)}\sum_{j\in I_n}\frac{\widehat{d}_{1,l}(\boldsymbol{u}_j) - z(\boldsymbol{u}_j)\widehat{d}_{2,l}/\widehat{A}}{\pi(\boldsymbol{u}_j)}\\
&\quad+\frac{2}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{d}_{2,l}\sum_{i\in I_n}\frac{\widehat{d}_{1,k}(\boldsymbol{u}_i)}{\pi(\boldsymbol{u}_i)} - \frac{1}{\widehat{A}^2}\sum_{k=1}^{q}\sum_{l=1}^{q}\widehat{\mathrm{Cov}}_{S_1}(\widehat{\beta}_k,\widehat{\beta}_l)\widehat{d}_{2,k}\widehat{d}_{2,l},
\end{aligned} \tag{A.24}$$

where $\widehat{d}_{1,k}$ and $\widehat{d}_{2,k}$ are as defined in (21).

Appendix B. Case with divided plots

It can happen that sample plots are divided into several parts, for example if one part of the plot is in forests and other parts are in other landscape categories, or if the plot overlaps borders between different regions, strata or forest stands (for example in the Swedish NFI, Anon. 2014). In such cases, the covariate information is not the same in different parts of the plot. Let us consider a case where we want to study expected plant densities in forests, and consider a particular plot $C(\boldsymbol{u}_i)$. Then let $I_{\mathbf{u}}$ be equal to 1 if $\boldsymbol{u}$ is in a forested area in U, and 0 otherwise. If the plot is divided and no part of the plot is in a forested area in U,

$$\lambda^\star(\boldsymbol{u}_i) = \int_{C(\boldsymbol{u}_i)} \frac{\lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) I_{\boldsymbol{u}}}{a_{\boldsymbol{u}}} d\boldsymbol{u} = 0 \quad \text{and} \quad \widehat{\lambda}^\star(\boldsymbol{u}_i) = 0. \tag{B.1}$$

If only one part of the plot is in a forested area in U, and if we denote the area of this part by $a_i^{(s)}$,

$$\lambda^\star(\boldsymbol{u}_i) = \int_{C(\boldsymbol{u}_i)} \frac{\lambda_{\boldsymbol{\beta}}(\boldsymbol{u}) I_{\boldsymbol{u}}}{a_{\boldsymbol{u}}} d\boldsymbol{u} = \lambda_{\boldsymbol{\beta}}(\boldsymbol{u}'_i) \frac{a_i^{(s)}}{a} \quad \text{and} \quad \widehat{\lambda}^\star(\boldsymbol{u}_i) = \lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}'_i) \frac{a_i^{(s)}}{a}, \tag{B.2}$$

where $\lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}_i) = \exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}(\boldsymbol{u}_i)) = \exp(\widehat{\boldsymbol{\beta}}^{\mathrm{T}} \boldsymbol{x}_i)$ and $\boldsymbol{u}'_i$ is an arbitrary point in the forested part of $C(\boldsymbol{u}_i) \cap U$. If $C(\boldsymbol{u}_i)$ has two parts that are in forests within U (with areas $a_i^{(s_1)}$ and $a_i^{(s_2)}$ respectively), then

$$\lambda^\star(\boldsymbol{u}_i) = \lambda_{\boldsymbol{\beta}}(\boldsymbol{u}'_i) \frac{a_i^{(s_1)}}{a} + \lambda_{\boldsymbol{\beta}}(\boldsymbol{u}''_i) \frac{a_i^{(s_2)}}{a} \quad \text{and} \quad \widehat{\lambda}^\star(\boldsymbol{u}_i) = \lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}'_i) \frac{a_i^{(s_1)}}{a} + \lambda_{\widehat{\boldsymbol{\beta}}}(\boldsymbol{u}''_i) \frac{a_i^{(s_2)}}{a} \tag{B.3}$$

where $\boldsymbol{u}'_i$ is an arbitrary point in the first forest part of $C(\boldsymbol{u}_i) \cap U$ and $\boldsymbol{u}''_i$ is an arbitrary point in the second forest part of $C(\boldsymbol{u}_i) \cap U$. And so on with three or more forest parts. Thus, the change of expression of $\widehat{\lambda}^\star(\boldsymbol{u}_i)$ will imply changes when applying formulas (16) and (A.24) for estimating the expected density and its variance estimator. Similar changes need to be done in the cluster sampling case presented in Section 2.5.

Appendix C. Details of the proposed goodness-of-fit test

Assume that there are two disjoint vegetation plots, A_{i1} and A_{i2}, contained in each (main) plot i, where all A_{ij} are of size a_A, $i = 1, \ldots, n$. Each vegetation plot A_{i1} and A_{i2} in a pair is separated by the same distance d. In each A_{ij}, the presence or absence of the plant species of interest is registered. Let M_i be the number of plants in plot A_i, $i = 1, \ldots, n$. Let Y_{ij} be 1 if presence in A_{ij}, and 0 otherwise, $i = 1, \ldots, n$, $j = 1, 2$. In our case, the M_i are not observed, contrary to the Y_{ij}, hence the necessity to develop a test based on the latter. Based on the sample of Y_{ij} data and corresponding covariate data x_i (assumed to be fixed in plot i), an estimator $\widehat{\boldsymbol{\beta}}$ of the parameter vector $\boldsymbol{\beta}$ is obtained using a binary regression with a complementary log-log link function (5). Let Y_i be 1 if there is at least one point in the union of A_{i1} and A_{i2}, and 0 otherwise. Based on a binary regression with a complementary log-log link function, offset $\log(2a_A)$, and the data $\{Y_i, x_i\}$, $i = 1, \ldots, n$, another estimator of $\boldsymbol{\beta}$ is constructed, denoted by $\widetilde{\boldsymbol{\beta}}$.

If the inhomogeneous Poisson point process model assumption is correct, then so is the model for the Y_{ij}. The reverse is not necessarily true. However, if the model for the Y_{ij} is incorrect, then so is the Poisson model for the M_i.

If the inhomogeneous Poisson point process model is correct, Y_{i1} and Y_{i2} will be independent conditional on the covariates, and binary regression model (5) implies the binary regression model based on the data $\{Y_i, x_i\}$. In this case, $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ will be close for large n. On the other hand, if Y_{i1} and Y_{i2} are not independent conditional on the covariates, then this implication will not hold and $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ will likely differ even if n is large. Based on this idea, Ekström et al. (Unpublished results) suggested the test statistic

$$S = (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^{\mathrm{T}} \ \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}), \tag{C.1}$$

where $\widehat{\boldsymbol{\Sigma}}$ is an estimate of the covariance matrix of $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$ given by

$$\widehat{\boldsymbol{\Sigma}} = n\big(\widehat{\boldsymbol{I}}_1^{-1}(\widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{I}}_2^{-1}(\widehat{\boldsymbol{\beta}}) - 2\widehat{\boldsymbol{I}}_1^{-1}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{C}}(\widehat{\boldsymbol{\beta}}) \widehat{\boldsymbol{I}}_2^{-1}(\widehat{\boldsymbol{\beta}})\big),$$

where

$$\widehat{\boldsymbol{I}}_1(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{2}{[g'(q_{i1}(\boldsymbol{\beta}))]^2 t_{i1}(\boldsymbol{\beta})} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}},$$

$$\widehat{\boldsymbol{I}}_2(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{[g'(q_i(\boldsymbol{\beta}))]^2 t_i(\boldsymbol{\beta})} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}},$$

$$\widehat{\boldsymbol{C}}(\boldsymbol{\beta}) = \frac{2}{n} \sum_{i=1}^{n} \frac{1}{g'(q_i(\boldsymbol{\beta})) t_i(\boldsymbol{\beta})} \frac{1}{g'(q_{i1}(\boldsymbol{\beta})) t_{i1}(\boldsymbol{\beta})} q_{i1}(\boldsymbol{\beta})(1 - q_i(\boldsymbol{\beta})) \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}},$$

$q_{ij}(\boldsymbol{\beta}) = 1 - \exp(-a_A \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i))$, $t_{ij}(\boldsymbol{\beta}) = q_{ij}(1 - q_{ij})$, $q_i(\boldsymbol{\beta}) = 1 - \exp(-2a_A \exp(\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i))$, $t_i(\boldsymbol{\beta}) = q_i(1 - q_i)$, and $g(p) = \log(-\log(1 - p))$.

If the Poisson model is valid, S is asymptotically distributed according to a chi-squared distribution with q degrees of freedom, where q is the length of $\boldsymbol{\beta}$. The binary model (5),and hence the Poisson model, is rejected if S is improbably large according to this chi-squared distribution. For small or

moderately large sample sizes, a better option might be to use parametric bootstrap (Davison and Hinkley, 1997). The bootstrap algorithm for computing the *p*-value of the test is given below.

For $b = 1,\ldots,B$, where B is a large integer:

i) For A_{ij}, generate points according to a Poisson point process with log intensity $\log\widehat{\lambda}_i = \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{x}_i$, $i = 1,\ldots,n$, $j = 1,2$.

ii) Based on the point data obtained in i), let Y^*_{ijb} be 1 if presence in A_{ij} and 0 otherwise, and let $Y^*_{ib} = max\left\{Y^*_{i1b}, Y^*_{i2b}\right\}$, $i = 1,\ldots,n$.

iii) Let S^* be defined as in (C.1), but based on $\left\{Y^*_{ijb}\right\}$ and $\left\{Y^*_{ib}\right\}$ rather than $\left\{Y_{ij}\right\}$ and $\{Y_i\}$.

The p-value of the test is given by the proportion of times S^* is larger than or equal to S.

References

Albert, P.S., McShane, L.M., 1995. A Generalized Estimating Equations Approach for Spatially Correlated Binary Data: Applications to the Analysis of Neuroimaging Data. Biometrics 51 (2), 627–638. https://doi.org/10.2307/2532950.

Ambler, G., Benner, A., 2015. mfp: Multivariable Fractional Polynomials. R Pack. Vers. 1 (5), 2. https://CRAN.R-project.org/package=mfp.

Anon, 2014. Fältinstruktion 2014, RIS, Riksinventeringen av skog [Field Instructions for the Swedish National Forest Inventory and the Swedish Forest Soil Inventory]. The Swedish University of Agricultural Sciences, Umeå, Sweden (In Swedish).

Artdatabanken, 2022. Artportalen. https://www.artdatabanken.se/sok-art-och-miljodata/artportalen/. Retrieved 17 August 2022.

Baddeley, A., Berman, M., Fisher, N.I., Hardegen, A., Milne, R.K., Schuhmacher, D., Turner, R., 2010. Spatial logistic regression and change-of-support for Poisson point processes. Electron. J. Stat. 4, 1151–1201. https://doi.org/10.1214/10-EJS581.

Baddeley, A., Rubak, E., Turner, R., 2016. Spatial Point Patterns: Methodology and Applications with R. CRC Press, Boca Raton. https://doi.org/10.1201/b19708.

Baena, S., Boyd, D.S., Moat, J., 2018. UAVs in pursuit of plant conservation-real world experiences. Eco. Inform. 47, 2–9. https://doi.org/10.1016/j.ecoinf.2017.11.001.

Bastow Wilson, J., 2012. Species presence/absence sometimes represents a plant community as well as species abundances do, or better. J. Veg. Sci. 23 (6), 1013–1023. https://doi.org/10.1111/j.1654-1103.2012.01430.x.

Belbin, L., 2011. The Atlas of Living Australia's Spatial Portal. In: Jones, M.B., Gries, C. (Eds.), Proceedings of the Environmental Information Management Conference 2011 (EIM 2011), pp. 39–43 (Santa Barbara).

Cassel, C., Särndal, C.E., Wretman, J.H., 1977. Foundations of inference in survey sampling. Wiley, New York. https://doi.org/10.2307/3314835.

CBD, 2002. Global Strategy for Plant Conservation. The Secretariat of the Convention on Biological Diversity, Montreal, Canada.

Commission of the European Communities, 2003. Council directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. Official Journal of the European Union 1. 236 99 23.9.2003, Brussels. European Commission 1992/95/2003.

Condés, S., McRoberts, R.E., 2017. Updating national forest inventory estimates of growing stock volume using hybrid inference. For. Ecol. Manag. 400, 48–57. https://doi.org/10.1016/j.foreco.2017.04.046.

Conlisk, E., Conlisk, J., Harte, J., 2007. The impossibility of estimating a negative binomial clustering parameter from presence-absence data: a comment on He and Gaston. Am. Nat. 170 (4), 651–659. https://doi.org/10.1086/521339.

Cordy, C.B., 1993. An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. Stat. Probab. Lett. 18 (5), 353–362. https://doi.org/10.1016/0167-7152(93)90028-H.

Corona, P., Fattorini, L., Franseschi, S., Scrinzi, G., Torresan, C., 2014. Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based and hybrid perspectives. Can. J. For. Res. 44 (11), 1303–1311. https://doi.org/10.1139/cjfr-2014-0203.

Daley, D.J., Vere-Jones, D., 2003. An introduction to the theory of point processes: volume I: elementary theory and methods. Springer. https://doi.org/10.1007/b97277.

Davison, A., Hinkley, D., 1997. Bootstrap Methods and their Application (Cambridge Series in Statistical and Probabilistic Mathematics). Cambridge University Press. https://doi.org/10.1017/CBO9780511802843.

Delignette-Muller, M.-L., Dutang, C., 2015. fitdistrplus: An R Package for Fitting Distributions. J. Stat. Softw. 64 (4), 1–34. https://doi.org/10.18637/jss.v064.i04.

Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Bryan Blair, J., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., Luthcke, S., 2022. GEDI launches a new era of biomass inference from space. Environ. Res. Lett. 17 (9), 095001 https://doi.org/10.1088/1748-9326/ac8694.

Ekström, M., Esseen, P.-A., Westerlund, B., Grafström, A., Jonsson, B.G., Ståhl, G., 2018. Logistic regression for clustered data from environmental monitoring programs. Eco. Inform. 43, 165–173. https://doi.org/10.1016/j.ecoinf.2017.10.006.

Ekström, M., Sandring, S., Grafström, A., Esseen, P.-A., Jonsson, B.G., Ståhl, G., 2020. Estimating density from presence-absence data in clustered populations. Methods Ecol. Evol. 11 (3), 390–402. https://doi.org/10.1111/2041-210X.13347.

Ekström, M., Gozé, L., Wallerman, J., Dahlgren, J., Jonsson, B.-G., Sandring, S., Ståhl, G., 2023. Model-based estimation and mapping of plant density based on remote sensing and presence/absence data.

Esseen, P.-A., Ekström, M., 2023. Influence of canopy structure and microclimate on three-dimensional distribution of the iconic lichen Usnea longissimi. For. Ecol. Manag. 529, 120667 https://doi.org/10.1016/j.foreco.2022.120667.

Esseen, P.-A., Ekström, M., Grafström, A., Jonsson, B.G., Palmqvist, K., Westerlund, B., Ståhl, G., 2022. Multiple drivers of large-scale lichen decline in boreal forest canopies. Glob. Chang. Biol. 28 (10), 3293–3309. https://doi.org/10.1111/gcb.16128.

Fithian, W., Elith, J., Hastie, T., Keith, D.A., 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods Ecol. Evol. 6 (4), 424–438. https://doi.org/10.1111/2041-210X.12242.

Foody, G.M., 2008. Refining predictions of climate change impacts on plant species distribution through the use of local statistics. Eco. Inform. 3 (3), 228–236. https://doi.org/10.1016/j.ecoinf.2008.02.002.

Fortin, M., Manso, R., Calama, R., 2016. Hybrid estimation based on mixed-effects models in forest inventories. Can. J. For. Res. 46 (11), 1310–1319. https://doi.org/10.1139/cjfr-2016-0298.

Fortin, M., Manso, R., Schneider, R., 2018. Parametric bootstrap estimators for hybrid inference in forest inventories. Forestry 91 (3), 354–365. https://doi.org/10.1093/forestry/cpx048.

Fortin, M., Lier, O.V., Côté, J.-F., 2023. Combining forest growth models and remotely sensed data through a hierarchical model-based inferential framework. Can. J. For. Res. 53, 1–13. https://doi.org/10.1139/cjfr-2022-0168.

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A.H., Ståhl, G., 2014. Adapting National Forest Inventories to changing requirements – the case of the Swedish National Forest Inventory at the turn of the 20th century. Silva Fennica 48 (3). https://doi.org/10.14214/sf.1095.

Futschik, A., Winkler, M., Steinbauer, K., Lamprecht, A., Rumpf, S.B., Barančok, P., Palaj, A., Gottfried, M., Pauli, H., 2020. Disentangling observer error and climate change effects in long-term monitoring of alpine plant species composition and cover. J. Veg. Sci. 31 (1), 14–25. https://doi.org/10.1111/jvs.12822.

Gallegos Torell, Å., Glimskär, A., 2009. Computer-aided calibration for visual estimation of vegetation cover. J. Veg. Sci. 20 (6), 973–983. https://doi.org/10.1111/j.1654-1103.2009.01111.x.

Gaston, K.J., He, F., Maguran, A., McGill, B., 2011. Species occurrence and occupancy. Biol. Divers. 141–151.

GBIF, 2022. What Is GBIF? Available from. https://www.gbif.org/what-is-gbif.

Gotway, C.A., Stroup, W.W., 1997. A Generalized Linear Model Approach to Spatial Data Analysis and Prediction. J. Agric. Biol. Environ. Stat. 2 (2), 157–178. https://doi.org/10.2307/1400401.

Grafström, A., Schnell, S., Saarela, S., Hubbell, S.P., Condit, R., 2017. The continuous population approach to forest inventories and use of information in the design. Environmetrics 28 (8). https://doi.org/10.1002/env.2480.

Gregoire, T.G., Valentine, H.T., 2007. Sampling Strategies for Natural Resources and the Environment. Chapman & Hall/CRC. https://doi.org/10.1201/9780203498880.

He, F., & Gaston, K.J. (2000). Estimating species abundance from occurrence. Am. Nat., 156 (5). 553–559. ISSN 0003–0147.

He, F., Gaston, K., Wu, J., 2002. On species occupancy-abundance models. Écoscience 9 (1), 119–126. https://doi.org/10.1080/11956860.2002.11682698.

Heagerty, P.J., Lele, S.R., 1998. A Composite Likelihood Approach to Binary Spatial Data. J. Am. Stat. Assoc. 93 (443), 1099–1111. https://doi.org/10.2307/2669853.

Heeringa, S.G., West, B.T., Berglund, P.A., 2010. Applied Survey Data Analysis. Chapman and Hall/CRC, Boca Raton. https://doi.org/10.1201/9781420080674.

Hoem, S., 2022. Norwegian Biodiversity Information Centre - Other Datasets. Version 13.236. The Norwegian Biodiversity Information Centre (NBIC). https://doi.org/10.15468/tm56sc. Occurrence dataset.

Holt, A.R., Gaston, K.J., He, F., 2002. Occupancy-abundance relationships and spatial distribution: a review. Basic Appl. Ecol. 3 (1), 1–13. https://doi.org/10.1078/1439-1791-00083.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Am. Stat. Assoc. 47 (260), 663–685. https://doi.org/10.1080/01621459.1952.10483446.

Huggins, R., Hwang, W.-H., Stoklosa, J., 2018. Estimation of abundance from presence–absence maps using cluster models. Environ. Ecol. Stat. 25, 495–522. https://doi.org/10.1007/s10651-018-0415-5.

Hwang, W.-H., Huggins, R., 2016. Estimating abundance from presence–absence maps via a paired Negative-Binomial Model. Scand. J. Stat. 43 (2), 573–586. https://doi.org/10.1111/sjos.12192.

Hwang, W.-H., Huggins, R., Stoklosa, J., 2022. A model for analyzing clustered occurrence data. Biometrics 78 (2), 598–611. https://doi.org/10.1111/biom.13435.

Kennedy, K.A., Addison, P.A., 1987. Some considerations for the use of visual estimates of plant cover in biomonitoring. J. Ecol. 151-157 https://doi.org/10.2307/2260541.
Kercher, S.M., Frieswyk, C.B., Zedler, J.B., 2003. Effects of sampling teams and estimation methods on the assessment of plant cover. J. Veg. Sci. 14 (6), 899–906. https://doi.org/10.1111/j.1654-1103.2003.tb02223.x.
Lindenmayer, D.B., Welsh, A., Donnelly, C., Crane, M., Michael, D., Macgregor, C., McBurney, L., Montague-Drake, R., Gibbons, P., 2009. Are nestboxes a viable alternative source of cavities for hollow-dependent animals? Long-term monitoring of nest box occupancy, pest use and attrition. Biol. Conserv. 142, 33–42. https://doi.org/10.1016/j.biocon.2008.09.026.
Margolis, H., Nelson, R., Montesano, P., Beaudoin, A., Sun, G., Andersen, H.-E., Wulder, M., 2015. Combining Satellite lidar, Airborne lidar and Ground Plots to Estimate the Amount and Distribution of Aboveground Biomass in the Boreal Forest of North America. Can. J. For. Res. 45 (7), 838–855. https://doi.org/10.1139/cjfr-2015-0006.
McRoberts, R.E., Næsset, E., Liknes, G.C., Chen, Q., Walters, B.F., Saatchi, S., Herold, M., 2019. Using a Finer Resolution Biomass Map to Assess the Accuracy of a Regional, Map-Based Estimate of forest Biomass. Surv. Geophys. 40 (4), 1001–1015. https://doi.org/10.1007/s10712-019-09507-1.
Nelson, R., Gobakken, T., Næsset, E., Gregoire, T.G., Ståhl, G., Holm, S., Flewelling, J., 2012. Lidar sampling - using an airborne profiler to estimate forest biomass in Hedmark County, Norway. Remote Sens. Environ. 123, 563–578. https://doi.org/10.1016/j.rse.2011.10.036.
O'Connor, B., Bojinski, S., Röösli, C., Schaepman, M.E., 2020. Monitoring global changes in biodiversity and climate essential as ecological crisis intensifies. Eco. Inform. 55, 101033 https://doi.org/10.1016/j.ecoinf.2019.101033.
Olsson, B., 2020. National Land Cover Database. Swedish Environmental Protection Agency, Stockholm. https://www.naturvardsverket.se/en/services-and-permits/maps-and-map-services/national-land-cover-database/.
Pain, D.J., Bardin, P., Hutchinson, N., Pénzesné Kónya, E., Krause, M., 2020. A review of European progress towards the Global Strategy for Plant Conservation 2011-2020. Planta Eur. Plantlife Int. XXXpp.
Phillips, S.J., Anderson, R.P., Dudík, M., Schapire, R.E., Blair, M.E., 2017. Opening the black box: an open-source release of Maxent. Ecography 40, 887–893. https://doi.org/10.1111/ecog.03049.
Pielou, E.C., 1977. Mathematical Ecology. Wiley.
R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.
Ringvall, A., Petersson, H., Ståhl, G., Lämås, T., 2005. Surveyor consistency in presence/absence sampling for monitoring vegetation in a boreal forest. For. Ecol. Manag. 212 (1–3), 109–117. https://doi.org/10.1016/j.foreco.2005.03.002.
Royle, J.A., Dorazio, R.M., 2008. Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities. Academic Press, London. https://doi.org/10.1016/B978-0-12-374097-7.50001-5.
Saarela, S., Schnell, S., Grafström, A., Tuominen, S., Nordkvist, K., Hyyppä, J., Kangas, A., Ståhl, G., 2015. Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume. Can. J. For. Res. 45 (11), 1524–1534. https://doi.org/10.1139/cjfr-2015-0077.
Saarela, S., Holm, S., Healey, S.P., Patterson, P.L., Yang, Z., Andersen, H.-E., Dubayah, R. O., Qi, W., Duncanson, L.I., Armston, J.D., Gobakken, T., Næsset, E., Ekström, M., Ståhl, G., 2022. Comparing frameworks for biomass prediction for the global ecosystem dynamics investigation. Remote Sens. Environ. 278 https://doi.org/10.1016/j.rse.2022.113074.
Särndal, C.-E., Swensson, B., Wretman, J., 1992. Model Assisted Survey Sampling. Springer. https://doi.org/10.1007/978-1-4612-4378-6.
Sauerbrei, W., Royston, P., 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. J. R. Stat. Soc. Ser. A 162 (1), 71–94. https://doi.org/10.1111/1467-985X.00122.
Sen, P.K., Singer, J.M., 1993. Large Sample Methods in Statistics: An Introduction with Applications. Chapman & Hall, New York.
Solow, A.R., Smith, W.K., 2010. On predicting abundance from occupancy. Am. Nat. 176 (1), 96–98. https://doi.org/10.1086/653077.
Sreekumar, E.R., Nameer, P.O., 2022. A MaxEnt modelling approach to understand the climate change effects on the distributional range of white-bellied Sholakili Sholicola albiventris (Blanford, 1868) in the Western Ghats, India. Ecol. Inform. 70 https://doi.org/10.1016/j.ecoinf.2022.101702.
Ståhl, G., 2003. Presence/absence sampling as a substitute for cover assessment in vegetation monitoring. In: Corona, P., Köhl, M., Marchetti, M. (Eds.), Advances in Forest Inventory for Sustainable Forest Management and Biodiversity Monitoring. Kluwer Academic Publishers, pp. 137–142. https://doi.org/10.1007/978-94-017-0649-0_11.
Ståhl, G., Holm, S., Gregoire, T.G., Gobakken, T., Næsset, E., Nelson, R., 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Can. J. For. Res. 41 (1), 96–107. https://doi.org/10.1139/X10-161.
Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., Gregoire, T.G., 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosyst. 3 (5) https://doi.org/10.1186/s40663-016-0064-9.
Ståhl, G., Ekström, M., Dahlgren, J., Esseen, P.-A., Grafström, A., Jonsson, B.G., 2017. Informative plot sizes in presence-absence sampling of forest floor vegetation. Methods Ecol. Evol. 8 (10), 1284–1291. https://doi.org/10.1111/2041-210X.12749.
Stoklosa, J., Blakey, R.V., Hui, F.K.C., 2022. An Overview of Modern Applications of Negative Binomial Modelling in Ecology and Biodiversity. Diversity 14 (5), 320. https://doi.org/10.3390/d14050320.
Tillé, Y., 2006. Sampling Algorithms. Springer, New York. ISBN 0-387-30814-8.
Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E. (Eds.), 2010. National Forest Inventories. Pathways for Common Reporting, 1. European Science Foundation, pp. 541–553. https://doi.org/10.1007/978-90-481-3233-1.
Waagepetersen, P., 2007. An estimating function approach to inference for inhomogeneous Neyman–Scott processes. Biometrics 63, 252–258. https://doi.org/10.1111/j.1541-0420.2006.00667.x.
Wan, J.-Z., Wang, C.-J., Yu, F.-H., 2017. Wind effects on habitat distributions of wind-dispersed invasive plants across different biomes on a global scale: assessment using six species. Eco. Inform. 42, 38–45. https://doi.org/10.1016/j.ecoinf.2017.09.002.
Warton, D.I., Foster, S.D., De'ath, G., Stoklosa, J., Dunstan, P.K., 2015. Model-based thinking for community ecology. Plant Ecol. 216, 669–682. https://doi.org/10.1007/s11258-014-0366-3.
Wright, D.H., 1991. Correlations Between Incidence and Abundance are Expected by Chance. J. Biogeogr. 18 (4), 463–466. https://doi.org/10.2307/2845487.
Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2, 587–602. https://doi.org/10.2307/3236170.

III

Estimation of parameters in inhomogeneous Neyman-Scott processes using presence/absence data

Magnus Ekström[a,b,1], Léna Gozé[a], Saskia Sandring[a], Bengt Gunnar Jonsson[c,d], Jörgen Wallerman[a], Göran Ståhl[a]

[a] Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

[b] Department of Statistics, USBE, Umeå University, SE-901 87 Umeå, Sweden

[c] Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden

[d] Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

Abstract. Environmental monitoring is of particular importance for studying biodiversity and ecosystems. Many environmental monitoring programs emphasize plant registrations as part of their inventory responsibilities. Among the various methods available for surveying plant communities, we focus on presence/absence (P/A) sampling due to its underutilized potential. P/A sampling offers several advantages over other methods, particularly its efficiency in terms of time and cost. However, interpreting direct information from this type of data can be challenging, as the results are heavily dependent on plot size and species distribution patterns. To overcome these difficulties, model-based assumptions are necessary. In this article, we propose a method for estimating parameters of an inhomogeneous Neyman-Scott point process, specifically a Matérn cluster process, using P/A data. The inhomogeneity is modeled by allowing the offspring process intensity to vary with environmental covariates. The proposed estimators and their corresponding confidence intervals are evaluated through Monte Carlo simulations and empirical data (P/A registrations for three plant species) collected by surveyors in Northern Sweden. The results indicate that the method generally produces nearly unbiased estimators, particularly when the sample size is sufficiently large. These parameter estimates from the underlying inhomogeneous Neyman-Scott point process can subsequently be used to compute local estimates of expected plant density.

KEYWORDS: Matérn cluster point process, multinomial regression, plant monitoring, remote sensing, vegetation survey

1 | Introduction

Monitoring plant communities has become increasingly important in the current context of global change. There is a growing demand for accurate inventories, including regular reporting for the EU Habitats Directive (EU, 1992). However, conducting inventories of plant communities presents significant challenges (e.g., Godínez-Alvarez et al., 2009). Counting individual plants can be challenging due to their high abundance or clonal structure (Ståhl et al., 2017). Consequently, common vegetation inventory

[1]Corresponding author: Magnus Ekström (Magnus.Ekstrom@slu.se)

methods often rely on visual cover assessments or point intercept sampling. However, visual cover assessments are susceptible to surveyor bias (e.g., Gallegos Torell and Glimskär, 2009), and point intercept sampling is generally both time-consuming and costly (Ståhl et al., 2017).

Presence/absence (P/A) sampling is a relatively simple and cost-effective alternative (Elzinga et al., 1998). In this method, the surveyor determines only whether a particular plant species is present in a sample plot. As a result, P/A sampling is quick to apply, and studies indicate that it is less prone to assessment errors than visual cover assessments (e.g., Ringvall et al., 2005). However, it also has drawbacks. P/A data can be challenging to interpret because occurrence proportions largely depend on plot dimensions and species distribution patterns. In this study, we propose using P/A data to estimate the parameters of inhomogeneous Neyman-Scott processes—and thereby intensity (local expected plant density)—as a promising and potentially useful approach for vegetation studies.

In nature, most plant species tend to have offspring that aggregate around parent plants due to seed dispersal and clonal growth (Schulze et al., 2019). Therefore, in this paper, we focus on plants that grow in clusters, i.e., groups of individuals located close to each other. Consequently, the commonly applied homogeneous Poisson point process is inadequate for modeling plant occurrences for such species, as it does not account for spatial dependence (Bonham, 2013). Instead, we employ Neyman–Scott (cluster) processes, which provide a simplified yet effective approach for modeling plants that grow in clusters (Baddeley et al., 2016). Biological justifications for Neyman–Scott processes can be found in Batista and Maguire (1998).

Ekström et al. (2020) developed a method for estimating the parameters of a homogeneous Neyman–Scott process (specifically, a Matérn cluster process) and, consequently, expected plant density from P/A data in clustered populations. However, this model assumes a constant intensity of the process across the entire study area, ignoring environmental variation. In reality, environmental conditions often vary considerably within a region. While homogeneous Neyman–Scott processes are effective for modeling plants with clustered growth patterns, they fail to capture local variations in environmental conditions that affect plant occurrences.

In this paper, we extend the method proposed by Ekström et al. (2020) to inhomogeneous Neyman–Scott processes, allowing the intensity of the process to vary with environmental covariates and thus become location-dependent. Inspired by Waagepetersen (2007), we chose to vary the intensity of the offspring process. However, other parameters could also be varied, as suggested by Mrkvička (2014).

Inhomogeneous Poisson point processes have commonly been used to model plant populations, as in studies by Uria-Diez et al. (2013), Pauchard et al. (2016), and Gozé et al. (2024). Notably, the latter study also developed estimators of expected plant density based on P/A data. Using P/A data, Gelfand and Shirota (2019) applied a specific type of cluster process, the log-Gaussian Cox process, to model plant populations under the assumption of preferential sampling. To our knowledge, no study has yet explored the use of P/A data of plants to estimate the parameters of an inhomogeneous Neyman–Scott process.

The objective of this study is to assess the suitability of P/A data sampling for

estimating the parameters of an inhomogeneous Neyman–Scott process and, consequently, local expected plant density, using auxiliary covariates derived from remotely sensed data. Estimates of all parameters characterizing the underlying inhomogeneous Neyman–Scott process (specifically, a Matérn cluster process) are derived through a multinomial regression model, which is implied by the assumed point process model and the sample plot design used for collecting P/A data. A key component of the study is to evaluate the performance of the proposed parameter estimation procedure through Monte Carlo simulations and an empirical analysis based on field data for three plant species: *Luzula pilosa* (L.) Willd., *Maianthemum bifolium* (L.) F.W. Schmidt, and *Lysimachia europaea* (L.) U. Manns & Anderb.

2 | Neyman-Scott processes

Neyman-Scott processes are examples of Poisson cluster processes, where the unobserved parent points follow a Poisson process with intensity τ (Baddeley et al., 2016). Each parent point produces a random number of offspring points according to a Poisson distribution with mean λ offspring per parent. Finally, for each parent point $x_i \in \mathbb{R}^2$, the offspring points $y_{ij} \in \mathbb{R}^2$ are independent and identically distributed, with a spatial offspring probability density $f_\gamma(y - x_i)$ depending on a parameter $\gamma > 0$. The set consisting of all the offspring points forms the Neyman-Scott process (e.g., Lawson and Denison, 2002). The intensity of the resulting homogeneous cluster process is $\tau\lambda$, representing the mean number of points per unit area.

It is possible to introduce spatial inhomogeneity into the above Neyman-Scott model by using inhomogeneous offspring processes (Waagepetersen, 2007). As before, parent points follow a homogeneous Poisson process with intensity τ. For a parent point at location x_i, the offspring follow a Poisson process with intensity $\lambda(u)f_\gamma(u-x_i)$, $u \in \mathbb{R}^2$, where $f_\gamma(u)$ is the offspring probability density. If $\lambda(u)$ is constant and equal to λ, the model reduces to the previously described homogeneous Neyman-Scott process.

We set

$$\lambda(u) = \lambda_{\boldsymbol{\beta}}(u) = \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i z_i(u)\right),$$

where each $z_i(u)$ denotes an environmental spatial covariate, $i = 1, ..., p$, and where $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)'$ is a vector of regression parameters. This is similar to Waagepetersen (2007), who used $\lambda(u) = \alpha \exp\left(\sum_{i=1}^{p} \beta_i z_i(u)\right)$. The set of offspring points forms the inhomogeneous Neyman-Scott process X and its intensity is $\tau\lambda_{\boldsymbol{\beta}}(u)$.

If $f_\gamma(u)$ is a uniform density function in a disc of radius γ, then the point process can be viewed as an inhomogeneous version of the Matérn cluster process (Matérn, 1960, 1986). If, instead, $f_\gamma(u)$ is the density function of an isotropic Gaussian distribution $N(0, \gamma^2 I)$, where I is the identity matrix, then the point process can be seen as an inhomogeneous version of the Thomas cluster process (Thomas, 1949).

For a compact set $S \subseteq \mathbb{R}^2$, let N_B denote the number of points that fall within a circular plot $B \subseteq S$. Then, $\{N_B > 0\}$ is the event indicating that at least one point is

present in B, and $\{N_B = 0\}$ denotes absence of points in B. Let

$$H_{\boldsymbol{\theta}}(B) = \exp\left(-\tau \int_{\mathbb{R}^2} \left(1 - \exp\left(-\int_B \lambda_{\boldsymbol{\beta}}(y) f_\gamma(y-x)dy\right)\right) dx\right),$$

where $\boldsymbol{\theta} = (\beta_0, ..., \beta_p, \tau, \gamma)'$.

Theorem 1. *Let $B_i \subset \mathbb{R}^2$, $i \in M = \{1, ..., m\}$, be disjoint circular plots in S, $M_s \subseteq M$, and $M_s^c = M \setminus M_s$. Then, for the defined inhomogeneous Neyman-Scott process,*

$$\begin{aligned} & P\{N_{B_i} > 0, i \in M_s, \text{ and } N_{B_i} = 0, i \in M_s^c\} \\ &= H_{\boldsymbol{\theta}}\left(\bigcup_{i \in M_s^c} B_i\right) - \sum_{i \in M_s} H_{\boldsymbol{\theta}}\left(B_i \cup \left[\bigcup_{j \in M_s^c} B_j\right]\right) \\ &\quad + \sum_{i_1, i_2 \in M_s, i_1 < i_2} H_{\boldsymbol{\theta}}\left(B_{i_1} \cup B_{i_2} \cup \left[\bigcup_{j \in M_s^c} B_j\right]\right) - ... + (-1)^{m_s} H_{\boldsymbol{\theta}}\left(\bigcup_{i \in M} B_i\right), \end{aligned}$$

where m_s is the number of elements in the set M_s.

The proof of Theorem 1 is given in the Appendix. Usage of Theorem 1 is illustrated in the next two examples.

Example 1. Assume that the point pattern consists of locations of plants and consider a concentric plot design, in which the jth innermost circle C_j has a radius r_j, $j = 1, ..., k$ (Fig. 1). Let $B_1 = C_1$ and $B_j = C_j \setminus C_{j-1}$, $j = 2, ..., k$. We assume that a surveyor starts with the innermost circle and moves outwards, until the first plant is observed. If no plants are present in $B_1, ..., B_{j-1}$, and at least one plant is present in B_j, where $j \leq k$, or if no plants are present in $C_k = \cup_{j=1}^k B_j$, then the surveyor is done. Hence, what is observed is whether the following mutually exclusive events occur or not,

$$\begin{aligned} A_0 &= \{\text{absence in } C_k\} = \{N_{C_k} = 0\}, \\ A_1 &= \{\text{presence in } C_1\} = \{N_{C_1} > 0\}, \\ A_j &= \{\text{presence in } B_j \text{ but not in } C_{j-1}\} = \{N_{C_{j-1}} = 0 \text{ and } N_{B_j} > 0\}, \quad j = 2, ..., k. \end{aligned}$$

It follows from Theorem 1 that the corresponding probabilities are given by

$$\begin{aligned} \pi_0(\boldsymbol{\theta}) &= P\{A_0\} = H_{\boldsymbol{\theta}}(C_k), \\ \pi_1(\boldsymbol{\theta}) &= P\{A_1\} = 1 - H_{\boldsymbol{\theta}}(C_1), \\ \pi_j(\boldsymbol{\theta}) &= P\{A_j\} = H_{\boldsymbol{\theta}}(C_{j-1}) - H_{\boldsymbol{\theta}}(C_j), \quad j = 2, ..., k. \end{aligned}$$

Note that $\sum_{j=0}^k \pi_j(\boldsymbol{\theta}) = 1$, which is to be expected since one and only one of the events A_j, $j = 0, ..., k$, will occur.

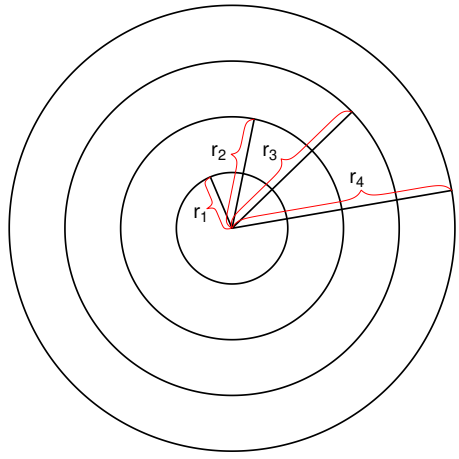


Fig. 1: Plot design with concentric circular sample plots with radii $r_1, ..., r_4$.

Example 2. In this example we consider a sample plot design used by the National Inventories of Landscapes in Sweden (NILS). For a list of plant species, P/A is recorded in three differently sized circular subplots, C_i, $i = 1, 2, 3$, per plot (Fig. 2). Thus, the mutually exclusive events that we consider are

$$\begin{aligned}
A_0 &= \{\text{absence in } C_1 \cup C_2 \cup C_3\},\\
A_i &= \{\text{presence only in } C_i\}, i = 1, 2, 3,\\
A_j &= \{\text{absence only in } C_{j-3}\}, j = 4, 5, 6,\\
A_7 &= \{\text{presence in each of } C_1, C_2, \text{ and } C_3\}.
\end{aligned}$$

By Theorem 1, the corresponding probabilities are given by

$$\begin{aligned}
\pi_0(\boldsymbol{\theta}) &= P\{A_0\} = H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_1(\boldsymbol{\theta}) &= P\{A_1\} = H_{\boldsymbol{\theta}}(C_2 \cup C_3) - H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_2(\boldsymbol{\theta}) &= P\{A_2\} = H_{\boldsymbol{\theta}}(C_1 \cup C_3) - H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_3(\boldsymbol{\theta}) &= P\{A_3\} = H_{\boldsymbol{\theta}}(C_1 \cup C_2) - H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_4(\boldsymbol{\theta}) &= P\{A_4\} = H_{\boldsymbol{\theta}}(C_1) - H_{\boldsymbol{\theta}}(C_1 \cup C_2) - H_{\boldsymbol{\theta}}(C_1 \cup C_3) + H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_5(\boldsymbol{\theta}) &= P\{A_5\} = H_{\boldsymbol{\theta}}(C_2) - H_{\boldsymbol{\theta}}(C_1 \cup C_2) - H_{\boldsymbol{\theta}}(C_2 \cup C_3) + H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_6(\boldsymbol{\theta}) &= P\{A_6\} = H_{\boldsymbol{\theta}}(C_3) - H_{\boldsymbol{\theta}}(C_1 \cup C_3) - H_{\boldsymbol{\theta}}(C_2 \cup C_3) + H_{\boldsymbol{\theta}}(C_1 \cup C_2 \cup C_3),\\
\pi_7(\boldsymbol{\theta}) &= P\{A_7\} = 1 - \sum_{i=0}^{6} \pi_i(\boldsymbol{\theta}).
\end{aligned}$$

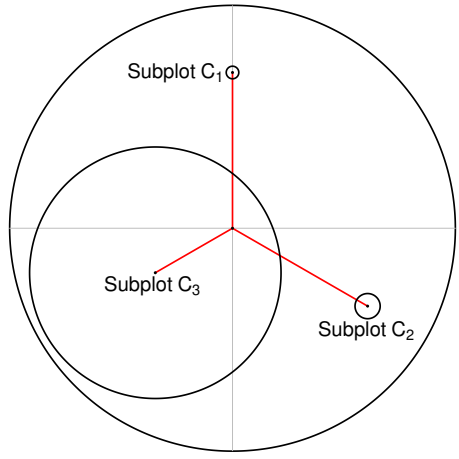


Fig. 2: Field subplot layout within a circular plot of radius 10 m in Example 2. The circular subplots C_i, $i = 1, 2, 3$, have areas 0.25, 1, and 100 m^2, respectively. Their respective center points are located 7, 7, and 4 m from the center of the plot. The angle between each pair of red lines in the figure is 120 degrees.

3 | Estimation of parameters of the inhomogeneous Neyman-Scott process

In Example 1, assume that there are n sets of concentric circular plots, C_{ij}, $i = 1, ..., n$, $j = 1, ..., k$, or, in Example 2, assume that there are n sets of circular subplots, C_{ij}, $i = 1, ..., n$, $j = 1, ..., k$, where $k = 3$. Suppose that the $C_{i\bullet} = \cup_{j=1}^{k} C_{ij}$, $i = 1, ..., n$, are so far apart that it is not unreasonable to assume that the point patterns $Z_{i'} = X \cap C_{i'\bullet}$ and $Z_{i''} = X \cap C_{i''\bullet}$ are independent for all $i' \neq i''$. For the ith set of plots, C_{ij}, $j = 1, ..., k$, define the events A_{ij},$j = 0, ..., m$, as in Example 1 or 2, with $m = k$ in Example 1 and $m = 7$ in Example 2. Let I_{ij} be the indicator of the event A_{ij}, $\pi_{ij}(\boldsymbol{\theta}) = P\{A_{ij}\}$, and set Y_i equal to j if the event A_{ij} occurs, $i = 1, ..., n$. The variable Y_i is then the dependent variable in a multinomial regression model (cf. Amemiya, 1985), for which

$$P\{Y_i = j\} = P\{A_{ij}\} = \pi_{ij}(\boldsymbol{\theta}).$$

Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. As in Rao (1973) and Amemiya (1985), we will use maximum likelihood estimation for estimating $\boldsymbol{\theta}_0$. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}_0$ is any parameter vector in $\Theta = \{\boldsymbol{\theta} = (\beta_0, ..., \beta_p, \tau, \gamma)' : \tau, \gamma > 0\}$ that maximizes the log-likelihood function,

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} \sum_{j=0}^{m} I_{ij} \log \pi_{ij}(\boldsymbol{\theta}).$$

By standard arguments (Rao 1973; Amemiya 1985),

$$i_{rsn}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_r}\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_s}\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_r \partial \theta_s}\right] = \sum_{i=1}^{n}\sum_{j=0}^{m}\frac{1}{\pi_{ij}(\boldsymbol{\theta})}\frac{\partial \pi_{ij}(\boldsymbol{\theta})}{\partial \theta_r}\frac{\partial \pi_{ij}(\boldsymbol{\theta})}{\partial \theta_s},$$

where $\theta_r = \beta_{r-1}$, $r = 1, ..., p+1$, $\theta_{p+2} = \tau$, and $\theta_{p+3} = \gamma$. Let $I_n(\boldsymbol{\theta}) = (i_{rsn}(\boldsymbol{\theta}))$. Similarly as in Sen and Singer (1993; Theorem 7.4.1), we assume that the limiting matrix $\lim_{n\to\infty} n^{-1} I_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ exists, is finite, and is positive definite. By a similar reasoning as for the general multinomial regression models considered in Rao (1973) and Amemiya (1985), it can be argued that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is consistent and asymptotically normally distributed,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(0, [I(\boldsymbol{\theta}_0)]^{-1}). \tag{1}$$

In this paper, we abstain from stating sufficient conditions under which the consistency and asymptotic normality results are valid. Similarly as in Sen and Singer (1993; Theorem 7.4.1), one of these conditions would be that the limiting matrix $\lim_{n\to\infty} n^{-1} I_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ exists, is finite, and is positive definite. This condition is difficult to verify in practice, and because of this a set of sufficient conditions would be of limited value in practical applications. Instead, we illustrate large sample properties of the maximum likelihood estimator through a Monte Carlo simulation study.

Let $i^{rsn}(\boldsymbol{\theta})$, $r, s = 1, 2, ..., p+3$, denote the elements of the inverse of the matrix $I_n(\boldsymbol{\theta})$. An approximate 95% confidence interval for an individual parameter θ_r is given by

$$\hat{\theta}_r \pm 1.96\sqrt{i^{rrn}(\hat{\boldsymbol{\theta}})}, \quad r = 1, 2, ..., p+3. \tag{2}$$

An approximate confidence interval for the intensity $\mu_{\boldsymbol{\theta}}(u) = \tau\lambda_{\boldsymbol{\beta}}(u) = \tau \exp(\beta_0 + \sum_{i=1}^{p}\beta_i z_i(u))$ of the point process at a spatial location u can also be constructed. By (1) and the delta method (e.g., Lehmann, 1999), we have

$$\sqrt{n}(\mu_{\hat{\boldsymbol{\theta}}}(u) - \mu_{\boldsymbol{\theta}_0}(u)) \xrightarrow{D} N(0, \nabla\mu_{\boldsymbol{\theta}_0}(u)'[I(\boldsymbol{\theta}_0)]^{-1}\nabla\mu_{\boldsymbol{\theta}_0}(u)),$$

where $\nabla\mu_{\boldsymbol{\theta}}(u)$ denotes the gradient, i.e., the vector formed by the partial derivatives $\partial\mu_{\boldsymbol{\theta}}(u)/\partial\theta_r$, $r = 1, ..., p+3$. That is, the vector containing the elements $\tau\lambda_{\boldsymbol{\beta}}(u)$, $\tau z_1(u)\lambda_{\boldsymbol{\beta}}(u)$, ..., $\tau z_p(u)\lambda_{\boldsymbol{\beta}}(u)$, $\lambda_{\boldsymbol{\beta}}(u)$, and 0. Thus, an approximate 95% confidence interval for the intensity at a given spatial location u is given by

$$\mu_{\hat{\boldsymbol{\theta}}}(u) \pm 1.96\sqrt{\sum_{r=1}^{p+2}\sum_{s=1}^{p+2} i^{rsn}(\hat{\boldsymbol{\theta}})\left.\frac{\partial\mu_{\boldsymbol{\theta}}(u)}{\partial\theta_r}\right|_{\hat{\boldsymbol{\theta}}}\left.\frac{\partial\mu_{\boldsymbol{\theta}}(u)}{\partial\theta_s}\right|_{\hat{\boldsymbol{\theta}}}}. \tag{3}$$

It should be noted that estimated standard errors for the estimators $\hat{\theta}_r$ and $\mu_{\hat{\boldsymbol{\theta}}}(u)$ are provided by the square root expressions in (2) and (3), respectively.

4 | Residuals for the multinomial regression model

It is well-known that for categorical binomial response variables, Pearson and deviance residuals can be clearly non-normal, leading to difficulties in graphical visualization and interpretation (Dunn and Smyth, 2018). The same problem appears for multinomial response variables (Araripe et al., 2023; Trijoulet et al., 2023; Gerber and Craig, 2024). For binomial models, randomized quantile residuals are often recommended, as they have the advantage of being exactly normally distributed apart from the sampling variability in estimating the model parameters (Dunn and Smyth, 2018). Trijoulet et al. (2023) extend these residuals to a multinomial setting using a conditional binomial method, and we have used their R package `compResidual` to compute these residuals.

Henceforth, we use the simplifying assumption that the environmental spatial covariate vector $\boldsymbol{z} = (z_1, ..., z_p)'$ is constant within each set of plots $C_{i\bullet}$, $i = 1, ..., n$. Let L denote the number of unique values of $\boldsymbol{z}$ in these sets of plots, and let r_l denote the number of sets of plots for which the covariate vector equals $\boldsymbol{z}_l$, $l = 1, ..., L$. Let the random variable Y_{lj} represent the number of times event A_j was observed in the r_l sets of plots, $l = 1, ..., L$. Then the random vector $(Y_{l0}, \ldots, Y_{lm})$ follows a multinomial distribution with parameters r_l and $(\pi_{l0}, \ldots, \pi_{lm})$.

For generating randomized quantile residuals for multinomial models, Trijoulet et al. (2023) use the fact that

$$\begin{aligned} P((Y_{l0}, \ldots, Y_{lm}) = (y_{l0}, \ldots, y_{lm})) = {} & P(Y_{l0} = y_{l0}) P(Y_{l1} = y_{l1} \,|\, Y_{l0} = y_{l0}) \\ & \cdots P(Y_{l,m-1} = Y_{l,m-1} \,|\, Y_{l0} = y_{l0}, \ldots, Y_{l,m-2} = y_{l,m-2}) \end{aligned}$$

and that

$$\begin{aligned} Y_{l0} &\sim \text{Bin}(r_l, \pi_{l0}), \\ Y_{l2} \mid Y_{l1} = y_{l1} &\sim \text{Bin}\left(r_l - y_{l1}, \frac{\pi_{l1}}{1 - \pi_{l0}}\right), \\ &\vdots \\ Y_{l,m-1} \mid Y_{l0} = y_{l0}, \ldots, Y_{l,m-2} = y_{l,m-2} &\sim \text{Bin}\left(r_l - y_{l1} \cdots - y_{l,m-2}, \frac{\pi_{l,m-1}}{1 - \sum_{j=0}^{m-2} \pi_{lj}}\right). \end{aligned}$$

Following Trijoulet et al. (2023), for each $l = 1, ..., L$, a randomized quantile residual is generated for each of the above m successive binomial distributions, in which the true unknown parameters $\{\pi_{lj}\}$ are first replaced by their estimated counterparts, $\{\pi_{lj}(\hat{\boldsymbol{\theta}})\}$. We refer to Trijoulet et al. (2023) for further details.

5 | Computational issues

Analytic expressions for maximum likelihood estimators in complex models are usually not easily available, necessitating the use of numerical methods to maximize

likelihood functions. The R software (R Core Team, 2025) provides several numerical procedures for maximizing likelihood functions. We used the optimization routine `constrOptim`, employing the BFGS algorithm, which is a quasi-Newton method for optimization (Lange, 1999). The log-likelihood function we aim to maximize is not necessarily well-behaved, meaning it may not be concave with a single maximum point. Therefore, we employed the standard method of identifying local maxima by using different starting values for the parameter vector $\boldsymbol{\theta}$. We then selected the highest of these local maxima. The numerical algorithm used is essentially the same as in Ekström et al. (2020), except that we used twenty starting values instead of five.

Numerical methods were also required to compute $H_{\boldsymbol{\theta}}(\cdot)$, on which the probabilities $\pi_{ij}(\boldsymbol{\theta})$ and the likelihood function are based. Under the simplifying assumption that the environmental spatial covariate vector $\boldsymbol{z} = (z_1, \ldots, z_p)'$ is constant within each set of plots $C_{i\bullet}$, $i = 1, \ldots, n$, the inner integral of $H_{\boldsymbol{\theta}}(\cdot)$ can be computed analytically for the Matérn cluster process to obtain the probabilities $\pi_{ij}(\boldsymbol{\theta})$ (cf. Ekström et al., 2020). For this reason, we considered this particular process in both our case example and Monte Carlo study. To compute the outer integral in $H_{\boldsymbol{\theta}}(\cdot)$, we used the `polyCub.SV` function in the R package `polyCub` (Meyer and Held, 2014, Supplement B).

6 | Case example

The P/A data for the case example were collected in the field during September 2022 at Kulbäcksliden Research Park, situated in Västerbotten County, Sweden (Fig. 3). In total, we obtained P/A data for 559 sets of concentric circular plots with radii of 0.94, 1.88, 2.82, 3.76, 4.70, and 5.64 meters. The sets of plots were systematically placed in the forested landscape of the research park, with a planned minimum distance of about 35 meters between them. In reality, some of these distances were shorter than 35 meters, occasionally less than 25 meters, but never shorter than 17 meters (Fig. 4). P/A data for the following plant species were recorded: *Luzula pilosa* (L.) Willd., *Maianthemum bifolium* (L.) F.W. Schmidt, and *Lysimachia europaea* (L.) U. Manns & Anderb. Covariate data were obtained from the forest raster map product Swedish open National Forest Attribute Maps (Skogsstyrelsen, 2025; Nilsson et al., 2017) and the SLU Soil Moisture Map (Ågren et al., 2021). These maps are all derived from the national airborne laser scanning campaigns in Sweden. The covariates were extracted using nearest neighbor interpolation at the coordinates of each concentric plot center. The available covariates were basal area of trees (m^2/ha), basal area-weighted mean stem diameter at breast height (cm), basal area-weighted mean tree height (dm), stem volume (m^3/ha), above-ground biomass (ton/ha), and soil moisture in three classes: wet - moist, moist - mesic, and mesic - dry.

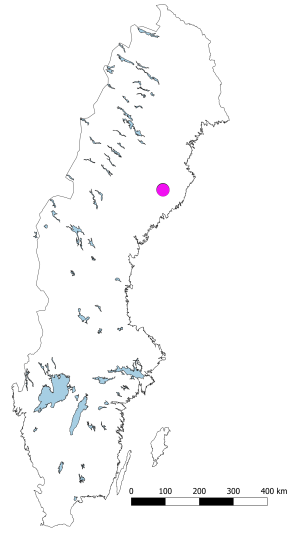


Fig. 3: The location of the Kulbäcksliden Research Park in Sweden.

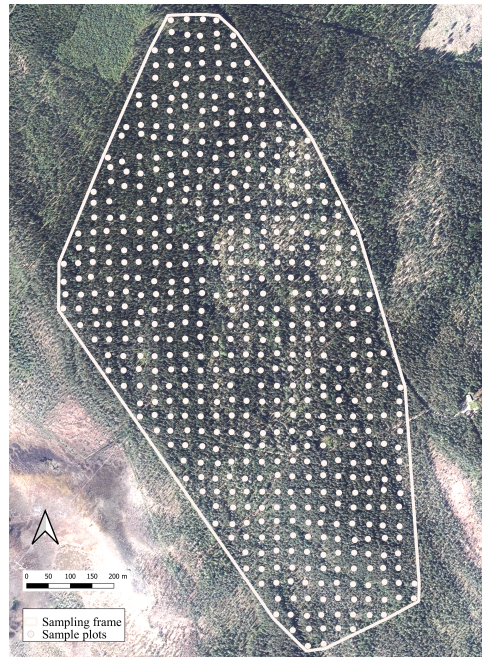


Fig. 4: Actual positions of the sample plots in Kulbäcksliden Research Park.

For each species, we fitted a model for each available covariate. We also explored models that included transformations of covariates, such as square roots or logarithms,

as well as models that combined two continuous covariates or one continuous covariate with the categorical soil moisture variable. To maintain simplicity, we avoided models with a large number of covariates, as the primary aim was to explore the methodological approach rather than to identify the "best" model for the species in question. Among the models with satisfactory residual plots and with $\beta_1, ..., \beta_p$ significantly different from zero, we chose the ones with the smallest Akaike Information Criterion (AIC) value. The fitted models are presented in Table 1 and the corresponding residual plots are shown in Fig. 5. Despite the presence of potential outliers in the residual plots for *M. bifolium* and *L. europaea*, the Shapiro-Wilk tests did not reject the hypothesis that the residuals are normally distributed (p-values 0.318 and 0.159, respectively).

In each model in Table 1, the parameter β_1 is significantly different from zero and has a positive estimate. Thus, as the logarithm and the square root of the basal area-weighted mean tree height increase, so does the intensity of the fitted Matérn cluster process for *L. pilosa* and *L. europaea*, respectively. When these covariates equal their sample median values, the estimated process intensities are 0.11 and 0.026 plants per m^2, with associated 95% confidence intervals of $(0.063, 0.17)$ and $(0.018, 0.035)$, respectively. For *M. bifolium*, the intensity of the process increases with the square root of the basal area. When this covariate equals its sample median value, the estimated intensity of the process is 0.052 plants per m^2 and its corresponding 95% confidence interval is $(0.034, 0.071)$. The estimated cluster radius for *M. bifolium* is larger than those for *L. pilosa* and *L. europaea*, at 13.7 m, 7.1 m, and 7.2 m, respectively.

Table 1: Estimated parameters of the Matérn cluster process are presented along with the corresponding 95% confidence intervals (given in parentheses). The covariates used in the models were the square root of the basal area for *M. bifolium* and the logarithm and the square root of the basal area-weighted mean tree height for *L. pilosa* and *L. europaea*, respectively. The product $\tau\lambda_i^*$ represents the intensity of the process at a location where the covariate equals its ith sample quartile, for $i = 1, 2, 3$.

Parameter	*M. bifolium*	*L. pilosa*	*L. europaea*
$\hat{\tau}$	0.00042 (0.00017, 0.00066)	0.0014 (0.00089, 0.0019)	0.0017 (0.00093, 0.0025)
$\hat{\beta}_0$	-2.9 (-4.9, -0.9)	-16.3 (-26.7, -5.8)	-18.6 (-22.8, -14.4)
$\hat{\beta}_1$	1.6 (1.2, 2.0)	4.0 (2.0, 6.1)	1.6 (1.3, 1.9)
$\hat{\gamma}$	13.7 (8.5, 18.9)	7.1 (5.1, 9.1)	7.2 (4.5, 9.8)
$\hat{\tau}\hat{\lambda}_1^*$	0.015 (0.0094, 0.021)	0.062 (0.036, 0.088)	0.0055 (0.0035, 0.0075)
$\hat{\tau}\hat{\lambda}_2^*$	0.052 (0.034, 0.071)	0.11 (0.063, 0.17)	0.026 (0.018, 0.035)
$\hat{\tau}\hat{\lambda}_3^*$	0.13 (0.068, 0.19)	0.18 (0.074, 0.29)	0.096 (0.050, 0.14)

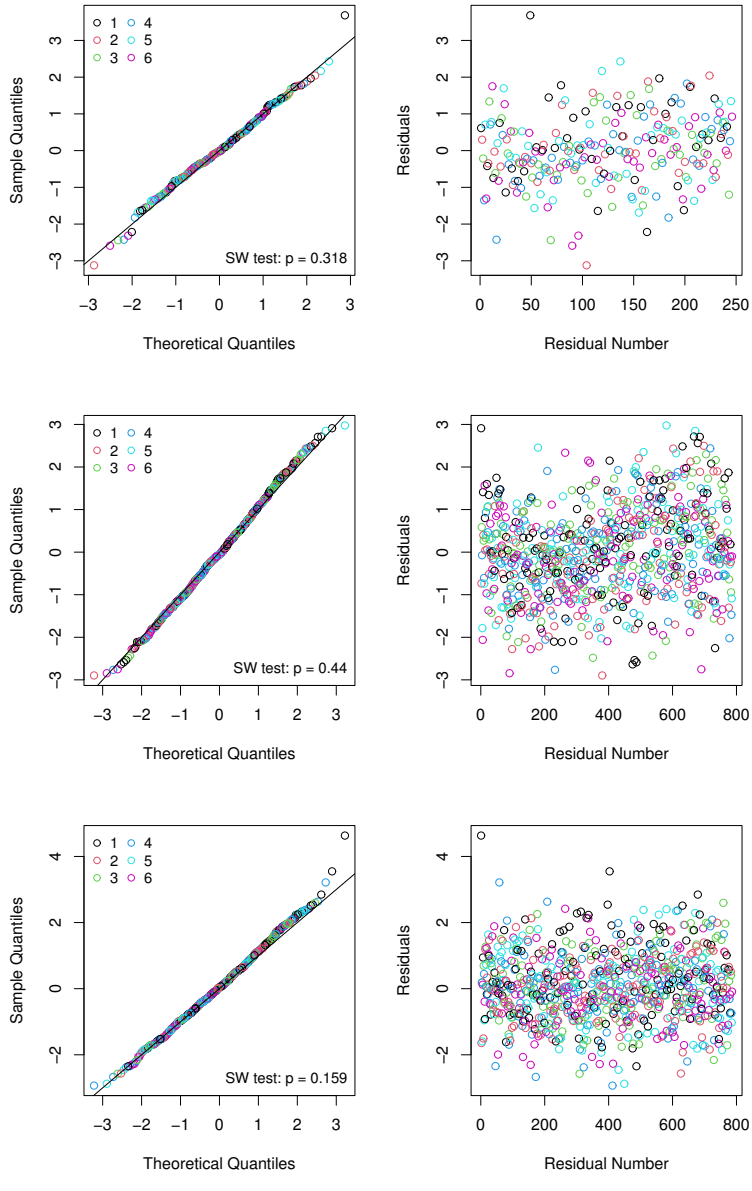


Fig. 5: Residual plots for the two fitted models in Table 1 are shown, along with p-values for the Shapiro-Wilk test of normality. The top two plots correspond to *M. bifolium*, the middle two to *L. pilosa*, and the bottom two to *L. europaea*. In each plot on the right-hand side, residuals for small and large values of the covariate are located on the left and right sides of the plot, respectively. The colors represent the six successive binomial distributions that were used to generate the residuals.

Realizations of the Matérn cluster process were generated using the `rMatClust` algorithm in the `spatstat` package (Baddeley et al., 2016). Maximum likelihood estimates of $\boldsymbol{\theta}_0$ were obtained from P/A data using a concentric plot design with radii of 0.94, 1.88, 2.82, 3.76, 4.70, and 5.64 meters (see Example 1).

For each parameter setup and sample size considered, we generated 1000 replicates of the process. For each replicate, we computed the maximum likelihood estimate of $\boldsymbol{\theta}_0$ and the confidence intervals for the individual parameters, as well as for $\tau\lambda_i^*$, which represents the intensity of the Matérn cluster process at a location where the covariate equals its ith sample quartile, for $i = 1, 2, 3$. Based on the replicate estimates of $\boldsymbol{\theta}_0$, we computed the median and the mean of the estimates for the individual parameters (τ, β_0, β_1, and γ) and the intensity $\tau\lambda_i^*$, $i = 1, 2, 3$, in each case considered. Using the same replicate estimates, we computed the actual confidence levels (ACLs) of the confidence intervals. The nominal confidence level was set to 95%.

In Table 2, we present the Monte Carlo results for a Matérn cluster process, where the true parameter values were set equal to those of the fitted model for *M. bifolium* in Table 1. In the data used to fit this model, the covariate—the square root of the basal area—had 41 unique values ranging from 2.00 to 6.64. To mimic the real data, we aimed to use the same covariate values in the Monte Carlo simulations. However, to reduce the number of numerical integrations required and to decrease computational time, we decided to round the covariate values to the nearest increments of 0.5, resulting in values of 2.0, 2.5, ..., 6.0, and 6.5. The sample size in the case example was $n = 559$. In the simulations, we also considered larger sample sizes, specifically $n = 4 \times 559 = 2236$ and $n = 10 \times 559 = 5590$. For the sample size of 4×559, we used four copies of each rounded covariate value from the original sample, and for a sample size of 10×559, we used ten copies. In Table 2, there is a trend indicating that the (mean and median) biases and standard deviations of the estimators decrease as the sample size increases. For the largest sample size considered, the biases are small or very small. For the smallest sample size, $n = 559$, the ACL is notably lower than the nominal level 95% for τ and $\tau\lambda_1^*$, but for the other parameters, including $\tau\lambda_2^*$ and $\tau\lambda_3^*$, the ACLs are quite close to 95%. For the largest sample size, all ACLs are close or quite close to 95%. The standard error estimators exhibit some biases, but they are not severe for any of the sample sizes considered.

Tables 3 and 4 are similar to Table 2, but here the true parameter values were set equal to those from the fitted models in Table 1 for *L. pilosa* and *L. europaea*, respectively. For *L. pilosa*, the values of the corresponding covariate—the logarithm of the basal area-weighted mean tree height—were rounded to the nearest multiple of 0.1 in the simulations, resulting in values ranging from 3.9 to 5.5. For *L. europaea*, the values of the covariate—the square root of the basal area-weighted mean tree height—were rounded to the nearest integer, resulting in values ranging from 7 to 16. The main trends observed in Table 2 are also evident in Tables 3 and 4. For the smallest sample size, some ACLs are noticeably too low; for example, the ACLs for τ and $\tau\lambda_1^*$ in Table 3 are as low as 91.6% and 91.7%, respectively, while the ACL for $\tau\lambda_1^*$ in Table 4 is

Table 2: Monte Carlo results for *M. bifolium*: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s).

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
559	τ	0.000416	0.000413	0.000423	0.000127	0.000128	92.2
	β_0	-2.89	-2.94	-3.00	1.05	1.04	95.4
	β_1	1.58	1.59	1.61	0.214	0.211	95.8
	γ	13.7	13.8	14.2	2.85	2.92	94.8
	$\tau\lambda_1^*$	0.0127	0.0126	0.0128	0.00274	0.00277	92.7
	$\tau\lambda_2^*$	0.0614	0.0621	0.0637	0.0130	0.0131	95.6
	$\tau\lambda_3^*$	0.135	0.138	0.145	0.0406	0.0421	94.1
2236	τ	0.000416	0.000415	0.000416	0.0000629	0.0000626	94.2
	β_0	-2.89	-2.94	-2.93	0.515	0.510	94.3
	β_1	1.58	1.59	1.59	0.104	0.103	95.3
	γ	13.7	13.7	13.9	1.38	1.36	95.1
	$\tau\lambda_1^*$	0.0127	0.0126	0.0127	0.00135	0.00135	93.8
	$\tau\lambda_2^*$	0.0615	0.0617	0.0621	0.00594	0.00589	94.6
	$\tau\lambda_3^*$	0.135	0.137	0.138	0.0176	0.0176	95.1
5590	τ	0.000416	0.000415	0.000416	0.0000404	0.0000396	94.2
	β_0	-2.89	-2.89	-2.90	0.325	0.321	94.8
	β_1	1.58	1.58	1.58	0.0655	0.0649	95.5
	γ	13.7	13.7	13.8	0.875	0.847	94.4
	$\tau\lambda_1^*$	0.0127	0.0127	0.0127	0.000823	0.000849	95.6
	$\tau\lambda_2^*$	0.0615	0.0614	0.0617	0.00368	0.00366	94.5
	$\tau\lambda_3^*$	0.135	0.136	0.136	0.0110	0.0108	94.9

92.9%. A possible reason for these low ACLs could be that the corresponding parameter estimators and/or standard error estimators have relatively large biases. For the two larger sample sizes, all ACLs fall within the range of 93.7% to 95.8%. For the largest sample size, the biases of the parameter estimators are small or negligible. In Table 3, note that the three replications with the most extreme outcomes were removed for the smallest sample size $n = 559$. This action had a significant impact on the values of $\overline{\text{SE}}$ for β_0, β_1, and $\tau\lambda_i^*$, as well as on the mean and standard deviation of $\tau\lambda_i^*$, for $i = 1, 2, 3$. Without this removal, these values would have been highly inflated.

In the Supporting Information, additional results from Monte Carlo simulations are presented, using a concentric plot design with radii of $1, 2, \ldots, 10$ meters (see Example 1). A single covariate was used in each considered model, generated according to a discrete uniform distribution taking values $0.1, 0.2, \ldots, 0.5$. Simulations were conducted for each of the 16 possible combinations of the following true parameters: τ was set to 0.001 or 0.002, β_0 to $\log 1.05$ or $\log 2$, β_1 to $\log 16$ or $\log 256$, and γ to 5 or 10. In other words, simulations were performed for 16 different models. We considered two sample sizes, $n = 1000$ and $n = 5000$. If, for some model with $n = 5000$, any of the ACLs deviated from the nominal level of 95% by more than 1.5 percentage points, an additional simulation was performed with $n = 10000$. Furthermore, due to the large biases observed in the estimators of β_0 and γ and their standard errors in Table S10, an additional simulation with $n = 10000$ was performed for this case. The simulations

Table 3: Monte Carlo results for *L. pilosa*: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s).

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
559[†]	τ	0.00138	0.00136	0.00138	0.000270	0.000251	91.6
	β_0	-16.3	-17.3	-17.6	5.74	5.84	95.3
	β_1	4.00	4.22	4.27	1.13	1.16	95.7
	γ	7.11	7.20	7.29	1.19	1.08	93.3
	$\tau\lambda_1^*$	0.0580	0.0590	0.0609	0.0174	0.0187	91.7
	$\tau\lambda_2^*$	0.129	0.134	0.146	0.0560	0.0730	92.0
	$\tau\lambda_3^*$	0.193	0.201	0.231	0.114	0.159	92.4
2236	τ	0.00138	0.00136	0.00137	0.000127	0.000125	93.9
	β_0	-16.3	-16.3	-16.5	2.76	2.71	94.0
	β_1	4.00	4.02	4.05	0.542	0.532	94.2
	γ	7.11	7.17	7.18	0.527	0.520	94.9
	$\tau\lambda_1^*$	0.0580	0.0581	0.0584	0.00644	0.00648	95.3
	$\tau\lambda_2^*$	0.129	0.123	0.132	0.0177	0.0176	94.3
	$\tau\lambda_3^*$	0.193	0.194	0.199	0.0350	0.0349	93.8
5590	τ	0.00138	0.00137	0.00138	0.0000833	0.0000793	94.3
	β_0	-16.3	-16.4	-16.4	1.77	1.70	94.2
	β_1	4.00	4.03	4.04	0.349	0.335	94.2
	γ	7.11	7.12	7.12	0.336	0.326	94.1
	$\tau\lambda_1^*$	0.0580	0.0582	0.0583	0.00422	0.00405	93.7
	$\tau\lambda_2^*$	0.129	0.130	0.131	0.0108	0.0105	95.7
	$\tau\lambda_3^*$	0.193	0.195	0.196	0.0210	0.0204	94.8

[†]For these results, the three replications with the most extreme outcomes were removed.

were otherwise conducted in the same manner as those presented in Tables 2-4.

Simulations with $n = 10000$ were conducted for five models: those with $\boldsymbol{\theta}_0 = (0.001, \log 1.05, \log 16, 5)$ (Table S1), $\boldsymbol{\theta}_0 = (0.001, \log 1.05, \log 16, 10)$ (Table S2), $\boldsymbol{\theta}_0 = (0.001, \log 2, \log 16, 5)$ (Table S5), $\boldsymbol{\theta}_0 = (0.001, \log 2, 2\log 16, 5)$ (Table S7), $\boldsymbol{\theta}_0 = (0.01, \log 1.05, \log 16, 10)$ (Table S10), and $\boldsymbol{\theta}_0 = (0.01, \log 2, \log 16, 5)$ (Table S13). For this increased sample size, all ACLs were within 1.5 percentage points of the nominal level, except for $\tau\lambda_3^*$ in Table 1, for which the ACL was 1.6 percentage points from the nominal level. With $n = 1000$, all ACLs were within 5.1 percentage points of the nominal level, and for $n = 5000$, this difference was 2.5 percentage points.

For models with $\beta_0 = \log 1.05$, the estimator for this parameter often exhibited notable mean and median bias, even for larger sample sizes. However, these biases were small relative to the standard errors, particularly for larger sample sizes. Apart from this case, the parameter estimators generally showed small or negligible mean biases when the sample size was $n = 5000$. Nevertheless, the relative mean bias was occasionally as high as about 5%, as observed for the estimators of $\gamma = 10.0$ in Table S10 and $\beta_0 = \log 2$ in Table S14.

For the smallest sample size, the standard error estimators showed considerable bias for several models. In contrast, these biases were fairly small when $n = 5000$, except for the models with $\boldsymbol{\theta}_0 = (0.001, \log 1.05, \log 16, 10)$ (Table S2) and $\boldsymbol{\theta}_0 =$

Table 4: Monte Carlo results for *L. europaea*: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s).

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
559	τ	0.00174	0.00170	0.00174	0.000437	0.000415	93.7
	β_0	-18.6	-18.7	-18.8	2.25	2.21	93.5
	β_1	1.61	1.62	1.62	0.178	0.168	93.8
	γ	7.16	7.21	7.43	1.61	1.50	93.2
	$\tau\lambda_1^*$	0.00344	0.00338	0.00345	0.000779	0.000746	92.9
	$\tau\lambda_2^*$	0.0171	0.0170	0.0173	0.00308	0.00300	94.1
	$\tau\lambda_3^*$	0.0854	0.0847	0.0902	0.0313	0.0309	93.1
2236	τ	0.00174	0.00175	0.00175	0.000206	0.000206	94.9
	β_0	-18.6	-18.6	-18.6	1.11	1.09	94.6
	β_1	1.61	1.61	1.61	0.0837	0.0829	94.9
	γ	7.16	7.12	7.19	0.706	0.684	93.9
	$\tau\lambda_1^*$	0.00344	0.00342	0.00344	0.000380	0.000366	93.9
	$\tau\lambda_2^*$	0.0171	0.0171	0.0171	0.00138	0.00139	94.7
	$\tau\lambda_3^*$	0.0854	0.0856	0.0859	0.0104	0.0108	95.3
5590	τ	0.00174	0.00173	0.00174	0.000129	0.000129	94.8
	β_0	-18.6	-18.6	-18.6	0.684	0.691	94.9
	β_1	1.61	1.61	1.61	0.0519	0.0526	95.0
	γ	7.16	7.16	7.17	0.429	0.427	94.6
	$\tau\lambda_1^*$	0.00344	0.00343	0.00344	0.000234	0.000232	94.9
	$\tau\lambda_2^*$	0.0171	0.0172	0.0172	0.000873	0.000879	95.5
	$\tau\lambda_3^*$	0.0854	0.0860	0.0861	0.00664	0.00682	95.8

$(0.01, \log 1.05, \log 16, 10)$ (Table S10), where the standard error estimators for the estimators of β_0 and γ showed the largest biases. The additional simulations with $n = 10000$ for these models substantially reduced these biases.

8 | Discussion

Based on P/A data and multinomial regression, we derived the sampling distributions of parameter estimators for a Neyman–Scott process and the corresponding confidence intervals using heuristic arguments. As the sample size increased, our Monte Carlo simulations generally showed that the biases of the parameter estimators became small or negligible, except possibly for cases where the intercept parameter β_0 is close to zero. However, even in these cases, the biases were small relative to the standard errors. The actual confidence levels were also close to the nominal 95% level for sufficiently large sample sizes. These results suggest that it is possible to accurately estimate the parameters of an inhomogeneous Neyman–Scott process and, consequently, produce local estimates of expected plant density when P/A data are used together with environmental covariate data from remote sensing.

In the empirical study based on data collected at the Kulbäcksliden Research Park in Northern Sweden, we estimated the parameters for the underlying cluster point pro-

cesses used to model plant positions for three different species. Using these parameters, we also computed estimates of expected plant density at locations where the covariate equaled its first, second, and third sample quartiles. Estimates of expected plant density can also be computed for other values of the covariates within a region of interest.

We used Monte Carlo simulations to illustrate the large-sample properties of our estimators. Although it would be possible to derive asymptotic theory using approaches similar to those outlined in Amemiya (1985) and Sen and Singer (1993), this was deemed beyond the scope of this paper. A method based on randomized quantile residuals, developed by Trijoulet et al. (2023), was used to assess the adequacy of the fitted multinomial regression models. Alternative approaches to using randomized quantile residuals for this purpose have been discussed by Araripe et al. (2024) and Gerber and Craig (2024).

Another important consideration is the selection of a specific Neyman-Scott process for modeling plant positions. In this study, we used a Matérn cluster process, primarily because it allows for more straightforward calculations, as some integrals can be computed analytically, significantly reducing the extensive computational effort required. However, other Neyman–Scott processes, such as the Thomas or Cauchy cluster process models, could also be considered, although these would increase the complexity of the numerical computations. As a general guideline, we recommend starting with a Matérn cluster model unless the data strongly suggest a different choice.

As in Waagepetersen (2007), we introduced inhomogeneity by allowing λ, the intensity of the offspring process, to vary with environmental covariates. However, other methods for introducing inhomogeneity into the point process exist, such as varying τ, the intensity of the parent process (e.g., Baddeley et al., 2016). For additional approaches to incorporating inhomogeneities into the Neyman–Scott point process, including allowing the cluster size (γ) to be inhomogeneous, we refer to Mrkvička (2014) and references therein. Extensions of our current method to these other types of inhomogeneities serve as an important topic for further studies.

We applied our suggested estimation method to data collected using a concentric plot design. In Example 2, we mentioned the possibility of applying our method to data from the sampling design currently used by the National 148 Inventories of Landscapes in Sweden (NILS). An important direction for future research is to explore various P/A sampling designs to identify those that yield estimators of expected plant density with the highest possible precision while remaining cost-efficient, practical, and reliable.

Covariate data are often available as wall-to-wall maps derived from remote sensing sources. This makes it possible to estimate intensity (local expected plant density) for each pixel of the map, based on the parameter estimates of the point process model and the values of the covariate(s) in those pixels. Using these pixel-wise estimates and reasoning similar to that of Ståhl et al. (2016), it would be possible to derive an estimate of expected plant density for a region of interest along with a corresponding variance estimate. One should be aware that such an approach may face issues with extrapolation if the map contains covariate values more extreme than those in the sample used to fit the multinomial regression model.

To conclude, our study suggests that model-based inference of plant population characteristics can be achieved using spatial models that mimic natural patterns of

plant occurrence, using P/A data together with auxiliary data readily available from remote sensing. The quantity and quality of such auxiliary information from various remote sensing techniques are currently increasing (e.g., Francini et al., 2020; Phiri et al., 2020; Dainelli et al., 2021; Dubayah et al., 2022). This development opens up opportunities to base plant population surveys on P/A data, also in cases where information about plant densities is required.

Appendix

Lemma 1 (Ekström et al., 2020). *Let $A_0, A_1, ..., A_m$ be a sequence of events. Then*

$$P\left\{A_0 \cap \left[\bigcap_{i=1}^{m} A_i^c\right]\right\} = P\{A_0\} - \sum_{1\le i\le m} P\{A_0 \cap A_i\} + \sum_{1\le i_1<i_2\le m} P\{A_0 \cap A_{i_1} \cap A_{i_2}\} - \sum_{1\le i_1<i_2<i_3\le m} P\{A_0 \cap A_{i_1} \cap A_{i_2} \cap A_{i_3}\} + ... + (-1)^m P\left\{\bigcap_{i=0}^{m} A_i\right\}.$$

Proof of Theorem 1. Let $m_s = |M_s|$. Assume, without loss of generality, that $M_s = \{1, ..., m_s\}$ and that the origin is in S. Let $A_i = \{N_{B_i} = 0\}$, $i = 1, ..., m_s$, and $A_0 = \cap_{i=m_s+1}^{m}\{N_{B_i} = 0\} = \{N_{\cup_{i=m_s+1}^{m} B_i} = 0\}$.

For one parent point at the origin, let D denote the number of offspring points generated from the Poisson point process with intensity $\lambda_{\boldsymbol{\beta}}(t)f_\gamma(t)$. Conditional on $D = d$, this offspring point pattern follows a binomial point process with density $\lambda_{\boldsymbol{\beta}}(t)f_\gamma(t)/G_{\boldsymbol{\beta},\gamma,0}(S)$ (e.g., Møller and Waagepetersen, 2003), where $G_{\boldsymbol{\beta},\gamma,x}(B) = \int_B \lambda_{\boldsymbol{\beta}}(t)f_\gamma(t-x)dt$, $B \subseteq S$, implying that

$$P\{A_0|D = d\} = \left(1 - \frac{G_{\boldsymbol{\beta},\gamma,0}(\cup_{i=m_s+1}^{m} B_i)}{G_{\boldsymbol{\beta},\gamma,0}(S)}\right)^d, \tag{4}$$

and, analogously, by Lemma 1,

$$\begin{aligned} P\left\{A_0 \cap \left[\bigcap_{i=1}^{m_s} A_i^c\right] \middle| D = d\right\} &= \left(1 - \frac{G_{\boldsymbol{\beta},\gamma,0}(\cup_{i=m_s+1}^{m} B_i)}{G_{\boldsymbol{\beta},\gamma,0}(S)}\right)^d \\ &\quad - \sum_{1\le i\le m_s} \left(1 - \frac{G_{\boldsymbol{\beta},\gamma,0}\left(B_i \cup \left[\cup_{j=m_s+1}^{m} B_j\right]\right)}{G_{\boldsymbol{\beta},\gamma,0}(S)}\right)^d \\ &\quad + \sum_{1\le i_1<i_2\le m_s} \left(1 - \frac{G_{\boldsymbol{\beta},\gamma,0}\left(B_{i_1} \cup B_{i_2} \cup \left[\cup_{j=m_s+1}^{m} B_j\right]\right)}{G_{\boldsymbol{\beta},\gamma,0}(S)}\right)^d \\ &\quad - ... + (-1)^{m_s} \left(1 - \frac{G_{\boldsymbol{\beta},\gamma,0}\left(\cup_{i=1}^{m} B_i\right)}{G_{\boldsymbol{\beta},\gamma,0}(S)}\right)^d. \end{aligned} \tag{5}$$

By noting that D is Poisson distributed with mean $G_{\boldsymbol{\beta},\gamma,0}(S)$ and by using (4), we see that

$$\begin{aligned} P\{A_0\} &= \sum_{d=0}^{\infty} \frac{e^{-G_{\boldsymbol{\beta},\gamma,0}(S)}(G_{\boldsymbol{\beta},\gamma,0}(S))^d}{d!} P\{A_0|D=d\} \\ &= \sum_{d=0}^{\infty} \frac{e^{-G_{\boldsymbol{\beta},\gamma,0}(S)}(G_{\boldsymbol{\beta},\gamma,0}(S))^d}{d!} \left(1 - \frac{G_{\boldsymbol{\beta},\gamma,0}(\cup_{i=m_s+1}^{m} B_i)}{G_{\boldsymbol{\beta},\gamma,0}(S)}\right)^d \\ &= \exp\left(-G_{\boldsymbol{\beta},\gamma,0}(\cup_{i=m_s+1}^{m} B_i)\right) \end{aligned}$$

and analogously, by using (5), that

$$\begin{aligned} P\left\{A_0 \cap \left[\bigcap_{i=1}^{m_s} A_i^c\right]\right\} &= \exp\left(-G_{\boldsymbol{\beta},\gamma,0}(\cup_{i=m_s+1}^{m} B_i)\right) \\ &\quad - \sum_{1\le i\le m_s} \exp\left(-G_{\boldsymbol{\beta},\gamma,0}\left(B_i \cup \left[\cup_{j=m_s+1}^{m} B_j\right]\right)\right) \\ &\quad + \sum_{1\le i_1<i_2\le m_s} \exp\left(-G_{\boldsymbol{\beta},\gamma,0}\left(B_{i_1} \cup B_{i_2} \cup \left[\cup_{j=m_s+1}^{m} B_j\right]\right)\right) \\ &\quad -... + (-1)^{m_s} \exp\left(-G_{\boldsymbol{\beta},\gamma,0}\left(\cup_{i=1}^{m} B_i\right)\right). \end{aligned}$$

Finally, if parent points are generated by a Poisson point process with intensity τ,

$$P\{A_0\} = \exp\left(-\tau \int \left(1 - \exp\left(-G_{\boldsymbol{\beta},\gamma,x}(\cup_{i=m_s+1}^{m} B_i)\right)\right) dx\right)$$

and

$$\begin{aligned} P\left\{A_0 \cap \left[\bigcap_{i=1}^{m_s} A_i^c\right]\right\} &= \exp\left(-\tau \int \left(1 - \exp\left(-G_{\boldsymbol{\beta},\gamma,x}(\cup_{i=m_s+1}^{m} B_i)\right)\right) dx\right) \\ &\quad - \sum_{1\le i\le m_s} \exp\left(-\tau \int \left(1 - \exp\left(-G_{\boldsymbol{\beta},\gamma,x}\left(B_i \cup \left[\cup_{j=m_s+1}^{m} B_j\right]\right)\right)\right) dx\right) \\ &\quad + \sum_{1\le i_1<i_2\le m_s} \exp\left(-\tau \int \left(1 - \exp\left(-G_{\boldsymbol{\beta},\gamma,x}\left(B_{i_1} \cup B_{i_2} \cup \left[\cup_{j=m_s+1}^{m} B_j\right]\right)\right)\right) dx\right) \\ &\quad -... + (-1)^{m_s} \exp\left(-\tau \int \left(1 - \exp\left(-G_{\boldsymbol{\beta},\gamma,x}\left(\cup_{i=1}^{m} B_i\right)\right)\right) dx\right). \end{aligned}$$

This completes the proof of the theorem. □

Acknowledgements We acknowledge financial support from the Kempe Foundations (SMK-1955) and the Carl Trygger Foundation (CTS 20: 110).

References

Ågren, A.M., Larson, J., Paul, S.S., Laudon, H., and Lidberg, W. (2021). Use of multiple LIDAR-derived digital terrain indices and machine learning for high-resolution national-

scale soil moisture mapping of the Swedish forest landscape. *Geoderma*, 404, 115280. doi: 10.1016/j.geoderma.2021.115280

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Araripe, P.P., Rodrigues de Lara, I.A., Rodrigues Palma, G., Cahill, N., and de Andrade Moral, R. (2024). Diagnostics for categorical response models based on quantile residuals and distance measures. *Journal of Applied Statistics*, 1–23. doi: 10.1080/02664763.2024.2367150

Baddeley, A., Rubak, E., and Turner, R. (2016). *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton, FL: CRC Press. doi: 10.1201/b19708

Batista, J.L.F., and Maguire, D.A. (1998). Modeling the spatial structure of tropical forests. *Forest Ecology and Management*, 110, 293–314. doi: 10.1016/S0378-1127(98)00296-5

Bonham, C.D. (2013). *Measurements for Terrestrial Vegetation* (2nd ed.). New York, NY: Wiley. doi: 10.1002/9781118534540

Cordy, C.B. (1993). An extension of the Horvitz-Thompson theorem to point sampling from a continuous universe. *Statistics & Probability Letters*, 18, 353–362. doi: 10.1016/0167-7152(93)90028-H

Dainelli, R., Toscano, P., Di Gennaro, S. F., and Matese, A. (2021). Recent advances in unmanned aerial vehicles forest remote sensing — A systematic review. Part II: Research applications. *Forests*, 12, 397. doi: 10.3390/f12040397

Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Bryan Blair, J., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., and Luthcke, S. (2022). GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, 17, 095001 doi: 10.1088/1748-9326/ac8694

Dunn, P.K., and Smyth, G.K. (2018). *Generalized Linear Models With Examples in R*. New York, NY: Springer. doi: 10.1007/978-1-4419-0118-7

Ekström, M., Sandring, S., Grafström, A., Esseen, P.-A., Jonsson, B.G., and Ståhl, G. (2020). Estimating density from presence-absence data in clustered populations. *Methods in Ecology and Evolution*, 11, 390–402. doi: 10.1111/2041-210X.13347

Elzinga, C.L., Salzer, D.W., and Willoughby, J.W. (1998). Measuring and Monitoring Plant Populations. BLM Technical Reference 1730-1. BLM National Applied Resource Sciences Center. Denver, CO.

European Union (1992). Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. *Official Journal of the European Communities*, L 206.

Francini, S., McRoberts, R. E., Giannetti, F., Mencucci, M., Marchetti, M., Scarascia Mugnozza, G., and Chirici, G. (2020). Near-real time forest change detection using PlanetScope imagery. *European Journal of Remote Sensing*, 53, 233–244. doi: 10.1080/22797254.2020.1806734

Gallegos Torell, Å., and Glimskär, A. (2009). Computer-aided calibration for visual estimation of vegetation cover. *Journal of Vegetation Science*, 20, 973–983. doi: 10.1111/j.1654-1103.2009.01111.x

Gelfand, A.E., and Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence-only data. *Ecological Monographs*, 89, e01372. doi: 10.1002/ecm.1372

Gerber, W.A.E., and Craig, B.A. (2024). Residuals and diagnostics for multinomial regression models. *Statistical Analysis and Data Mining*, 17, e11645. doi: 10.1002/sam.11645

Godínez-Alvarez, H., Herrick, J.E., Mattocks, M., Toledo, D., and Van Zee, J. (2009). Comparison of three vegetation monitoring methods: Their relative utility for ecological assessment and monitoring. *Ecological Indicators*, 9, 1001–1008. doi: 10.1016/j.ecolind.2008.11.011

Gozé, L., Ekström, M., Sandring, S, Jonsson, B.G., Wallerman, J., and Ståhl, G. (2024). Estimation of plant density based on presence/absence data using hybrid inference. *Ecological Informatics*, 80, 102377. doi: 10.1016/j.ecoinf.2023.102377

Lange, K. (1999). *Numerical Analysis for Statisticians*. New York, NY: Springer. doi: 10.1007/b98850

Lawson, A.B., and Denison, D.G.T., editors (2002). *Spatial Cluster Modelling*. Boca Raton, FL: Chapman & Hall/CRC Press.

Lehmann, E.L. (1999). *Elements of Large-Sample Theory*. New York, NY: Springer. doi: 10.1007/b98855

Matérn, B. (1960). Spatial variation: stochastic models and their application to some problems in forest surveys and other sampling investigations. *Meddelanden från Statens Skogsforskningsinstitut*, 49(5), 1–144.

Matérn, B. (1986). *Spatial Variation*. Lecture Notes in Statistics 36. New York, NY: Springer. doi: 10.1007/978-1-4615-7892-5

Meyer, S., and Held, L. (2014). Power-law models for infectious disease spread. *Annals of Applied Statistics*, 8, 1612–1639. doi: 10.1214/14-AOAS743

Møller, J., and Waagepetersen, R.P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton, FL: Chapman & Hall/CRC. doi: 10.1201/9780203496930

Mrkvička, T. (2014). Distinguishing different types of inhomogeneity in Neyman-Scott processes. *Methodology of Computing and Applied Probability*, 16, 385–395. doi: 10.1007/s11009-013-9365-4

Nilsson, M., Nordkvist, K., Jonzén, J., Lindgren, N., Axensten, P., Wallerman, J., Egberth, M., Larsson, S., Nilsson, L., Eriksson, J., and Olsson, H. (2017). A nationwide forest attribute map of Sweden predicted using airborne laser scanning data and field data from the National Forest Inventory. *Remote Sensing of Environment*, 194, 447–454. doi: 10.1016/j.rse.2016.10.022

Pauchard, A., Escudero, A., García, R.A., de la Cruz, M., Langdon, B., Cavieres, L.A., and Esquivel, J. (2016). Pine invasions in treeless environments: dispersal overruns microsite heterogeneity. *Ecology and Evolution*, 6, 447–459. doi: 10.1002/ece3.1877

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., and Ranagalage, M. (2020). Sentinel-2 data for land cover/use mapping: A review. *Remote Sensing*, 12, 2291. doi: 10.3390/rs12142291

R Core Team (2025). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rao, C.R. (1973). *Linear Statistical Inference and its Applications.* New York, NY: Wiley. doi: 10.1002/9780470316436

Ringvall, A., Petersson, H., Ståhl, G., and Lämås, T. (2005). Surveyor consistency in presence/absence sampling for monitoring vegetation in a boreal forest. *Forest Ecology and Management*, 212, 109–117. doi: 10.1016/j.foreco.2005.03.002

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling.* New York, NY: Springer. doi: 10.1007/978-1-4612-4378-6

Schulze, E.-D., Beck, E., Buchmann, N., Clemens, S., Müller-Hohenstein, K., and Scherer-Lorenzen, M. (2019). *Plant Ecology* (2nd ed.). Berlin: Springer. doi: 10.1007/978-3-662-56233-8

Sen, P.K., and Singer. J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications.* New York, NY: Chapman & Hall.

Skogsstyrelsen (2025). Digital Open Forest Data. Accessed January 1, 2025. https://www.skogsstyrelsen.se/en/digital-open-forest-data/

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., and Gregoire, T.G. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems*, 3. doi: 10.1186/s40663-016-0064-9

Ståhl, G., Ekström, M., Dahlgren, J., Esseen, P.-A., Grafström, A., and Jonsson, B.G. (2017). Informative plot sizes in presence-absence sampling of forest floor vegetation. *Methods in Ecology and Evolution*, 8, 1284–1291. doi: 10.1111/2041-210X.12749

Thomas, M. (1949). A generalisation of Poisson's binomial limit for use in ecology. *Biometrika*, 36, 18–25. doi: 10.2307/2332526

Trijoulet, V., Albertsen, C.M., Kristensen, K., Legault, C.M., Miller, T.J., and Nielsen, A. (2023). Model validation for compositional data in stock assessment models: calculating residuals with correct properties. *Fisheries Research*, 257, 106487. doi: 10.1016/j.fishres.2022.106487

Uria-Diez, J., Ibáñez, R. and Mateu, J. (2013). Importance of habitat heterogeneity and biotic processes in the spatial distribution of a riparian herb (*Carex remota* L.): a point process approach. *Stochastic Environmental Research and Risk Assessment*, 27, 59–76. doi: 10.1007/s00477-012-0569-x

Waagepetersen, R. (2007). An estimating function approach to inference for inhomogeneous Neyman-Scott processes. *Biometrics*, 63, 252–258. doi: 10.1111/j.1541-0420.2006.00667.x

Supplementary document for the paper entitled "Estimation of parameters in inhomogeneous Neyman-Scott processes using presence/absence data"

Magnus Ekström[a,b,1], Léna Gozé[a], Saskia Sandring[a], Bengt Gunnar Jonsson[c,d], Jörgen Wallerman[a], Göran Ståhl[a]

[a] Department of Forest Resource Management, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

[b] Department of Statistics, USBE, Umeå University, SE-901 87 Umeå, Sweden

[c] Department of Natural Sciences, Mid Sweden University, SE-851 70 Sundsvall, Sweden

[d] Department of Wildlife, Fish, and Environmental Studies, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

This supplementary material provides results from additional Monte Carlo simulations. Realizations of the Matérn cluster process were generated using the `rMatClust` function in the `spatstat` package (Baddeley et al., 2016). Maximum likelihood estimates of $\boldsymbol{\theta}_0$ were derived from P/A data using a concentric plot design with radii of 1, 2, ..., 10 meters.

For each parameter setting and sample size, we generated 1000 replications of the process. In each replication, we calculated the maximum likelihood estimate of $\boldsymbol{\theta}_0$ and the confidence intervals for individual parameters, as well as for $\tau\lambda_i^*$, which represents the intensity of the Matérn cluster process at a location where the covariate equals the ith quartile of its distribution, for $i = 1, 2, 3$. The covariate values in each replication were generated from a discrete uniform distribution ranging from 0.1 to 0.5. Based on the replicate estimates of $\boldsymbol{\theta}_0$, we estimated the median and mean of the estimators for individual parameters (τ, β_0, β_1, and γ) and for the intensities $\tau\lambda_i^*$, $i = 1, 2, 3$, in each case. The replicate estimates were also used to compute the actual confidence levels (ACLs) of the confidence intervals, with the nominal level set at 95%. The results are presented in Tables S1–S16.

[1]Corresponding author: Magnus Ekström (Magnus.Ekstrom@slu.se)

Table S 1: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 1.05, \log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.000956	0.000979	0.0002102	0.000292	95.1
	β_0	0.0488	0.0842	0.0784	0.382	0.433	96.6
	β_1	2.77	2.92	3.05	1.16	1.47	95.4
	γ	5.00	5.10	5.54	2.44	3.43	92.5
	$\tau\lambda_1^*$	0.00183	0.00186	0.00196	0.000503	0.000648	94.7
	$\tau\lambda_2^*$	0.00241	0.00254	0.00270	0.000881	0.00137	94.7
	$\tau\lambda_3^*$	0.00318	0.00338	0.00378	0.00170	0.00311	93.4
5000	τ	0.00100	0.000995	0.000998	0.0000867	0.0000884	94.8
	β_0	0.0488	0.0468	0.0456	0.154	0.157	95.5
	β_1	2.77	2.81	2.84	0.510	0.500	93.9
	γ	5.00	5.01	5.12	1.14	1.07	92.5
	$\tau\lambda_1^*$	0.00183	0.00183	0.00185	0.000183	0.000180	94.3
	$\tau\lambda_2^*$	0.00241	0.00244	0.00246	0.000318	0.000308	94.1
	$\tau\lambda_3^*$	0.00318	0.00323	0.00329	0.000571	0.000558	93.5
10000	τ	0.00100	0.000996	0.000999	0.0000603	0.0000607	95.3
	β_0	0.0488	0.0466	0.0478	0.111	0.110	95.1
	β_1	2.77	2.77	2.80	0.359	0.343	93.7
	γ	5.00	5.03	5.07	0.761	0.729	93.9
	$\tau\lambda_1^*$	0.00183	0.00183	0.00183	0.000126	0.000125	95.3
	$\tau\lambda_2^*$	0.00241	0.00241	0.00243	0.000215	0.000207	93.9
	$\tau\lambda_3^*$	0.00318	0.00318	0.00322	0.000378	0.000360	93.4

Table S 2: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 1.05, \log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.000904	0.000895	0.000341	0.000737	98.1
	β_0	0.0488	0.205	0.233	0.448	1.10	98.2
	β_1	2.77	2.83	2.88	0.694	0.700	95.2
	γ	10.0	10.7	12.1	5.21	14.8	92.3
	$\tau\lambda_1^*$	0.00183	0.00184	0.00188	0.000286	0.000277	95.0
	$\tau\lambda_2^*$	0.00241	0.00244	0.00251	0.000437	0.000429	96.1
	$\tau\lambda_3^*$	0.00318	0.00322	0.00338	0.000763	0.000772	96.2
5000	τ	0.00100	0.000982	0.000960	0.000206	0.000212	94.4
	β_0	0.0488	0.0819	0.120	0.254	0.307	95.9
	β_1	2.77	2.77	2.77	0.296	0.291	94.6
	γ	10.0	10.1	10.8	3.15	3.89	93.3
	$\tau\lambda_1^*$	0.00183	0.00184	0.00184	0.000113	0.000113	94.3
	$\tau\lambda_2^*$	0.00241	0.00242	0.00243	0.000169	0.000165	93.9
	$\tau\lambda_3^*$	0.00318	0.00318	0.00321	0.000281	0.000272	93.0
10000	τ	0.00100	0.000986	0.000979	0.000129	0.000132	95.7
	β_0	0.0488	0.0665	0.0773	0.150	0.157	96.2
	β_1	2.77	2.78	2.79	0.212	0.204	94.0
	γ	10.0	10.0	10.3	1.68	1.75	95.2
	$\tau\lambda_1^*$	0.00183	0.00183	0.00183	0.0000783	0.0000788	94.9
	$\tau\lambda_2^*$	0.00241	0.00242	0.00242	0.000115	0.000114	94.0
	$\tau\lambda_3^*$	0.00318	0.00320	0.00320	0.000193	0.000188	95.0

Table S 3: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 1.05, 2\log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.000995	0.000996	0.000131	0.000133	94.2
	β_0	0.0488	0.0642	0.0495	0.348	0.349	95.3
	β_1	5.55	5.58	5.61	1.26	1.26	93.0
	γ	5.00	5.08	5.21	1.19	1.15	95.6
	$\tau\lambda_1^*$	0.00318	0.00317	0.00324	0.000654	0.000678	93.0
	$\tau\lambda_2^*$	0.00554	0.00559	0.00575	0.00157	0.00196	91.8
	$\tau\lambda_3^*$	0.00965	0.00969	0.0104	0.00430	0.00665	89.9
5000	τ	0.00100	0.000998	0.00100	0.0000570	0.0000573	94.6
	β_0	0.0488	0.0509	0.0540	0.156	0.152	94.3
	β_1	5.55	5.54	5.53	0.511	0.510	94.0
	γ	5.00	5.00	5.03	0.469	0.460	94.7
	$\tau\lambda_1^*$	0.00318	0.00317	0.00319	0.000266	0.000254	93.5
	$\tau\lambda_2^*$	0.00554	0.00554	0.00556	0.000532	0.000521	94.4
	$\tau\lambda_3^*$	0.00965	0.00965	0.00970	0.00125	0.00125	94.4

Table S 4: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 1.05, 2\log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.00100	0.000997	0.000201	0.000199	94.4
	β_0	0.0488	0.0490	0.0659	0.298	0.299	96.0
	β_1	5.55	5.60	5.58	0.648	0.656	95.2
	γ	10.0	10.0	10.4	2.23	2.20	94.7
	$\tau\lambda_1^*$	0.00318	0.00316	0.00320	0.000349	0.000361	95.9
	$\tau\lambda_2^*$	0.00554	0.00553	0.00559	0.000628	0.000631	94.7
	$\tau\lambda_3^*$	0.00965	0.00968	0.00982	0.00146	0.00145	93.6
5000	τ	0.00100	0.000995	0.000996	0.0000892	0.0000874	93.8
	β_0	0.0488	0.0574	0.0550	0.130	0.129	95.9
	β_1	5.55	5.55	5.55	0.299	0.291	94.5
	γ	10.0	10.1	10.1	0.874	0.868	95.1
	$\tau\lambda_1^*$	0.00318	0.00318	0.00318	0.000163	0.000159	93.8
	$\tau\lambda_2^*$	0.00554	0.00554	0.00555	0.000282	0.000275	93.6
	$\tau\lambda_3^*$	0.00965	0.00964	0.00967	0.000637	0.000623	94.1

Table S 5: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 2, \log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.000989	0.000990	0.000146	0.000144	94.5
	β_0	0.693	0.714	0.730	0.367	0.384	95.8
	β_1	2.77	2.75	2.84	1.10	1.24	94.5
	γ	5.00	5.06	5.30	1.65	1.58	93.4
	$\tau\lambda_1^*$	0.00348	0.00354	0.00369	0.000939	0.00107	93.7
	$\tau\lambda_2^*$	0.00459	0.00471	0.00492	0.00138	0.00170	93.0
	$\tau\lambda_3^*$	0.00606	0.00628	0.00667	0.00239	0.00334	92.3
5000	τ	0.00100	0.000998	0.000999	0.0000648	0.0000609	92.5
	β_0	0.693	0.701	0.698	0.163	0.161	94.2
	β_1	2.77	2.76	2.79	0.502	0.495	93.3
	γ	5.00	5.02	5.06	0.641	0.607	93.4
	$\tau\lambda_1^*$	0.00348	0.00350	0.00351	0.000339	0.000351	95.4
	$\tau\lambda_2^*$	0.00459	0.00463	0.00465	0.000505	0.000516	94.5
	$\tau\lambda_3^*$	0.00606	0.00613	0.00617	0.000859	0.000872	93.7
10000	τ	0.00100	0.00100	0.00100	0.0000437	0.0000428	94.2
	β_0	0.693	0.701	0.698	0.118	0.114	94.5
	β_1	2.77	2.78	2.78	0.357	0.349	94.1
	γ	5.00	5.00	5.01	0.439	0.423	94.1
	$\tau\lambda_1^*$	0.00348	0.00349	0.00351	0.000250	0.000246	94.6
	$\tau\lambda_2^*$	0.00459	0.00462	0.00463	0.000362	0.000360	94.3
	$\tau\lambda_3^*$	0.00606	0.00609	0.00612	0.000603	0.000604	94.8

Table S6: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 2, \log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.000999	0.000982	0.000247	0.000265	95.5
	β_0	0.693	0.728	0.744	0.332	0.382	97.2
	β_1	2.77	2.82	2.83	0.646	0.648	94.2
	γ	10.0	9.98	10.7	3.25	4.03	93.3
	$\tau\lambda_1^*$	0.00348	0.00349	0.00354	0.000458	0.000473	95.0
	$\tau\lambda_2^*$	0.00459	0.00464	0.00470	0.000657	0.000646	94.3
	$\tau\lambda_3^*$	0.00606	0.00612	0.00626	0.00113	0.00109	93.8
5000	τ	0.00100	0.00100	0.00100	0.000111	0.000109	94.8
	β_0	0.693	0.695	0.696	0.142	0.143	94.9
	β_1	2.77	2.79	2.79	0.287	0.283	94.3
	γ	10.0	9.96	10.1	1.26	1.23	95.1
	$\tau\lambda_1^*$	0.00348	0.00348	0.00349	0.000201	0.000202	95.3
	$\tau\lambda_2^*$	0.00459	0.00461	0.00461	0.000273	0.000270	94.6
	$\tau\lambda_3^*$	0.00606	0.00609	0.00610	0.000446	0.000438	94.3

Table S7: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 2, 2\log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.00100	0.00100	0.000111	0.000111	94.7
	β_0	0.693	0.653	0.665	0.395	0.398	94.5
	β_1	5.55	5.69	5.65	1.30	1.41	96.3
	γ	5.00	5.07	5.10	0.820	0.820	94.7
	$\tau\lambda_1^*$	0.00606	0.00589	0.00615	0.00145	0.00153	92.9
	$\tau\lambda_2^*$	0.0106	0.0104	0.0109	0.00310	0.00442	92.8
	$\tau\lambda_3^*$	0.0184	0.0183	0.0196	0.00844	0.0164	92.5
5000	τ	0.00100	0.00100	0.00100	0.0000510	0.0000487	94.0
	β_0	0.693	0.688	0.690	0.176	0.175	93.4
	β_1	5.55	5.58	5.57	0.570	0.562	94.8
	γ	5.00	4.99	5.00	0.361	0.350	94.5
	$\tau\lambda_1^*$	0.00606	0.00603	0.00611	0.000573	0.000576	95.3
	$\tau\lambda_2^*$	0.0106	0.0106	0.0107	0.00104	0.00105	96.2
	$\tau\lambda_3^*$	0.0184	0.0185	0.0187	0.00243	0.00246	95.4
10000	τ	0.00100	0.00100	0.00100	0.0000335	0.0000344	95.5
	β_0	0.693	0.688	0.692	0.120	0.123	95.4
	β_1	5.55	5.54	5.55	0.391	0.393	95.2
	γ	5.00	5.01	5.01	0.242	0.247	96.1
	$\tau\lambda_1^*$	0.00606	0.00605	0.00608	0.000387	0.000400	96.3
	$\tau\lambda_2^*$	0.0106	0.0106	0.0106	0.000708	0.000722	95.0
	$\tau\lambda_3^*$	0.0184	0.0184	0.0185	0.00165	0.00167	95.1

Table S 8: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.001, \log 2, 2\log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.00100	0.00100	0.00100	0.000150	0.000152	95.1
	β_0	0.693	0.698	0.694	0.270	0.265	94.2
	β_1	5.55	5.52	5.55	0.656	0.670	94.4
	γ	10.0	10.0	10.2	1.45	1.44	95.3
	$\tau\lambda_1^*$	0.00606	0.00602	0.00608	0.000748	0.000725	93.4
	$\tau\lambda_2^*$	0.0106	0.0105	0.0106	0.00119	0.00116	93.5
	$\tau\lambda_3^*$	0.0184	0.0182	0.0185	0.00256	0.00257	94.4
5000	τ	0.00100	0.000997	0.000997	0.0000657	0.0000670	95.1
	β_0	0.693	0.702	0.699	0.116	0.118	96.4
	β_1	5.55	5.54	5.55	0.301	0.299	94.6
	γ	10.0	10.0	10.1	0.607	0.617	95.1
	$\tau\lambda_1^*$	0.00606	0.00606	0.00608	0.000316	0.000319	95.4
	$\tau\lambda_2^*$	0.0106	0.0106	0.0106	0.000510	0.000512	96.0
	$\tau\lambda_3^*$	0.0184	0.0184	0.0185	0.00113	0.00113	95.6

Table S9: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 1.05, \log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.00979	0.00988	0.00157	0.00158	95.1
	β_0	0.0488	0.0737	0.0740	0.222	0.221	95.1
	β_1	2.77	2.77	2.80	0.429	0.429	95.0
	γ	5.00	5.05	5.23	1.35	1.34	94.4
	$\tau\lambda_1^*$	0.0183	0.0183	0.0184	0.00163	0.00164	94.9
	$\tau\lambda_2^*$	0.0241	0.0242	0.0244	0.00244	0.00248	94.9
	$\tau\lambda_3^*$	0.0318	0.0318	0.0324	0.00415	0.00422	94.6
5000	τ	0.0100	0.00997	0.00996	0.000672	0.000663	94.3
	β_0	0.0488	0.0528	0.0556	0.0959	0.0929	94.0
	β_1	2.77	2.77	2.78	0.188	0.187	95.3
	γ	5.00	5.01	5.05	0.504	0.511	95.8
	$\tau\lambda_1^*$	0.0183	0.0183	0.0183	0.000702	0.000711	95.5
	$\tau\lambda_2^*$	0.0241	0.0241	0.0242	0.00103	0.00105	95.0
	$\tau\lambda_3^*$	0.0318	0.0319	0.0320	0.00170	0.00174	95.3

Table S 10: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 1.05, \log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.00947	0.00976	0.00409	0.00759	98.4
	β_0	0.0488	0.114	0.168	0.468	1.038	97.6
	β_1	2.77	2.79	2.80	0.290	0.301	95.7
	γ	10.0	10.3	11.4	4.39	11.6	94.3
	$\tau\lambda_1^*$	0.0183	0.0183	0.0184	0.00116	0.00120	96.3
	$\tau\lambda_2^*$	0.0241	0.0242	0.0243	0.00156	0.00163	96.8
	$\tau\lambda_3^*$	0.0318	0.0320	0.0322	0.00249	0.00262	97.0
5000	τ	0.0100	0.00972	0.00977	0.00236	0.00263	96.1
	β_0	0.0488	0.0777	0.104	0.265	0.318	96.3
	β_1	2.77	2.77	2.78	0.137	0.133	94.7
	γ	10.0	10.1	10.5	2.45	3.05	93.7
	$\tau\lambda_1^*$	0.0183	0.0183	0.0183	0.000546	0.000526	94.0
	$\tau\lambda_2^*$	0.0241	0.0242	0.0242	0.000732	0.000705	93.9
	$\tau\lambda_3^*$	0.0318	0.0319	0.0319	0.00116	0.00112	94.8
10000	τ	0.0100	0.0101	0.00997	0.00189	0.00182	93.8
	β_0	0.0488	0.0457	0.0726	0.210	0.204	95.9
	β_1	2.77	2.77	2.77	0.0964	0.0935	93.8
	γ	10.0	9.93	10.3	1.83	1.82	94.6
	$\tau\lambda_1^*$	0.0183	0.0183	0.0183	0.000377	0.000371	93.7
	$\tau\lambda_2^*$	0.0241	0.0241	0.0241	0.000512	0.000497	93.7
	$\tau\lambda_3^*$	0.0318	0.0318	0.0319	0.000819	0.000788	93.6

Table S 11: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 1.05, 2\log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.0100	0.0101	0.00125	0.00127	95.2
	β_0	0.0488	0.0529	0.0475	0.198	0.200	95.8
	β_1	5.55	5.57	5.55	0.448	0.457	95.7
	γ	5.00	4.97	5.01	0.640	0.650	95.3
	$\tau\lambda_1^*$	0.0318	0.0318	0.0320	0.00243	0.00244	94.8
	$\tau\lambda_2^*$	0.0554	0.0556	0.0558	0.00438	0.00448	95.1
	$\tau\lambda_3^*$	0.0965	0.0968	0.0975	0.0100	0.0104	95.4
5000	τ	0.0100	0.0101	0.0101	0.000563	0.000557	94.5
	β_0	0.0488	0.0433	0.0452	0.0883	0.0891	95.3
	β_1	5.55	5.55	5.55	0.206	0.203	94.6
	γ	5.00	4.97	4.98	0.288	0.282	94.2
	$\tau\lambda_1^*$	0.0318	0.0319	0.0319	0.00108	0.00108	94.1
	$\tau\lambda_2^*$	0.0554	0.0556	0.0557	0.00197	0.00197	95.5
	$\tau\lambda_3^*$	0.0965	0.0969	0.0970	0.00457	0.00455	95.2

Table S 12: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 1.05, 2\log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.0104	0.0105	0.00445	0.00560	96.2
	β_0	0.0488	0.0135	0.101	0.491	0.652	97.3
	β_1	5.55	5.53	5.55	0.297	0.313	95.5
	γ	10.0	9.75	10.9	3.83	5.90	94.3
	$\tau\lambda_1^*$	0.0318	0.0318	0.0319	0.00184	0.00181	94.5
	$\tau\lambda_2^*$	0.0554	0.0554	0.0556	0.00311	0.00308	94.2
	$\tau\lambda_3^*$	0.0965	0.0966	0.0969	0.00664	0.00680	95.8
5000	τ	0.0100	0.0100	0.0100	0.00228	0.00225	94.2
	β_0	0.0488	0.0440	0.0707	0.251	0.245	94.3
	β_1	5.55	5.56	5.55	0.139	0.140	94.8
	γ	10.0	10.0	10.3	1.77	1.75	94.2
	$\tau\lambda_1^*$	0.0318	0.0318	0.0318	0.000776	0.000806	94.9
	$\tau\lambda_2^*$	0.0554	0.0554	0.0555	0.00135	0.00137	95.4
	$\tau\lambda_3^*$	0.0965	0.0966	0.0967	0.00302	0.00300	94.0

Table S 13: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 2, \log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.00992	0.00996	0.00128	0.00129	95.1
	β_0	0.693	0.699	0.704	0.191	0.198	95.7
	β_1	2.77	2.78	2.80	0.444	0.439	94.5
	γ	5.00	5.03	5.11	0.831	0.816	94.9
	$\tau\lambda_1^*$	0.0348	0.0349	0.0351	0.00301	0.00315	96.2
	$\tau\lambda_2^*$	0.0459	0.0460	0.0464	0.00413	0.00427	95.5
	$\tau\lambda_3^*$	0.0606	0.0607	0.0615	0.00693	0.00703	95.8
5000	τ	0.0100	0.0100	0.0100	0.000578	0.000570	94.5
	β_0	0.693	0.696	0.696	0.0920	0.0872	93.8
	β_1	2.77	2.77	2.77	0.202	0.195	94.4
	γ	5.00	4.99	5.01	0.359	0.347	94.0
	$\tau\lambda_1^*$	0.0348	0.0349	0.0349	0.00142	0.00138	95.8
	$\tau\lambda_2^*$	0.0459	0.0461	0.0461	0.00192	0.00185	93.5
	$\tau\lambda_3^*$	0.0606	0.0608	0.0609	0.00312	0.00300	93.0
10000	τ	0.0100	0.0100	0.0100	0.000401	0.000402	96.0
	β_0	0.693	0.695	0.693	0.0618	0.0616	95.1
	β_1	2.77	2.77	2.77	0.142	0.137	94.8
	γ	5.00	5.00	5.01	0.246	0.244	95.3
	$\tau\lambda_1^*$	0.0348	0.0348	0.0348	0.000966	0.000972	95.3
	$\tau\lambda_2^*$	0.0459	0.0460	0.0460	0.00131	0.00130	94.0
	$\tau\lambda_3^*$	0.0606	0.0607	0.0607	0.00215	0.00211	94.7

Table S 14: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 2, \log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.00995	0.00994	0.00430	0.00581	95.9
	β_0	0.693	0.723	0.807	0.498	0.771	97.5
	β_1	2.77	2.80	2.80	0.304	0.299	94.6
	γ	10.0	10.1	11.4	4.41	8.10	93.7
	$\tau\lambda_1^*$	0.0348	0.0349	0.0350	0.00228	0.00222	93.9
	$\tau\lambda_2^*$	0.0459	0.0462	0.0463	0.00293	0.00285	95.0
	$\tau\lambda_3^*$	0.0606	0.0610	0.0614	0.00462	0.00450	95.4
5000	τ	0.0100	0.00983	0.00989	0.00225	0.00238	94.1
	β_0	0.693	0.721	0.732	0.245	0.260	96.0
	β_1	2.77	2.77	2.77	0.139	0.133	94.4
	γ	10.0	10.1	10.4	1.98	2.12	95.9
	$\tau\lambda_1^*$	0.0348	0.0348	0.0348	0.000949	0.000976	95.7
	$\tau\lambda_2^*$	0.0459	0.0459	0.0459	0.00121	0.00124	95.8
	$\tau\lambda_3^*$	0.0606	0.0606	0.0606	0.00196	0.00195	94.4

Table S 15: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 2, 2\log 16, 5)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.0101	0.0101	0.00113	0.00116	95.4
	β_0	0.693	0.701	0.694	0.193	0.200	95.4
	β_1	5.55	5.57	5.57	0.504	0.511	95.2
	γ	5.00	4.95	4.98	0.532	0.537	95.4
	$\tau\lambda_1^*$	0.0606	0.0616	0.0617	0.00521	0.00526	95.4
	$\tau\lambda_2^*$	0.106	0.107	0.108	0.00943	0.00940	95.0
	$\tau\lambda_3^*$	0.184	0.186	0.189	0.0220	0.0220	96.1
5000	τ	0.0100	0.00996	0.00999	0.000507	0.000509	94.6
	β_0	0.693	0.696	0.696	0.0862	0.0886	96.4
	β_1	5.55	5.54	5.54	0.216	0.222	94.8
	γ	5.00	5.01	5.01	0.232	0.239	95.7
	$\tau\lambda_1^*$	0.0606	0.0606	0.0607	0.00227	0.00227	94.4
	$\tau\lambda_2^*$	0.106	0.106	0.106	0.00410	0.00395	93.8
	$\tau\lambda_3^*$	0.184	0.184	0.184	0.00931	0.00899	93.5

Table S 16: Monte Carlo results for $\boldsymbol{\theta}_0 = (0.01, \log 2, 2\log 16, 10)$: medians, means and standard deviations (SDs) of estimates, actual confidence levels (ACLs) of the associated confidence intervals, and means of estimated standard errors ($\overline{\text{SE}}$s), where $\boldsymbol{\theta}_0$ denotes the true parameter vector.

Sample size (n)	Parameter	True value	Median	Mean	SD	$\overline{\text{SE}}$	ACL (%)
1000	τ	0.0100	0.0101	0.0106	0.00450	0.00520	93.4
	β_0	0.693	0.683	0.733	0.475	0.552	96.7
	β_1	5.55	5.55	5.57	0.311	0.318	95.1
	γ	10.0	9.89	10.7	3.46	4.41	94.8
	$\tau\lambda_1^*$	0.0606	0.0606	0.0608	0.00362	0.00354	94.3
	$\tau\lambda_2^*$	0.106	0.106	0.106	0.00582	0.00585	95.6
	$\tau\lambda_3^*$	0.184	0.184	0.185	0.0123	0.0128	95.1
5000	τ	0.0100	0.0100	0.0102	0.00222	0.00218	95.0
	β_0	0.693	0.696	0.698	0.224	0.226	94.4
	β_1	5.55	5.55	5.55	0.148	0.141	94.5
	γ	10.0	9.99	10.1	1.45	1.48	93.9
	$\tau\lambda_1^*$	0.0606	0.0607	0.0607	0.00165	0.00158	93.5
	$\tau\lambda_2^*$	0.106	0.106	0.106	0.00268	0.00259	93.8
	$\tau\lambda_3^*$	0.184	0.184	0.184	0.00578	0.00562	94.5

Reference

Baddeley, A., Rubak, E., & Turner, R. (2016). *Spatial Point Patterns: Methodology and Applications with R.* Boca Raton, FL: CRC Press. doi: 10.1201/b19708

IV

A closer look at uncertainties in forest ecosystem surveys using remotely sensed data and model-based inference

Göran Ståhl[1*], Léna Gozé[1], Emanuele Papucci[1], Terje Gobakken[2], Svetlana Saarela[2], Magnus Ekström[1], Sean P. Healey[3], Zhiqiang Yang[3], James R. Kellner[4], Zhengyang Hou[5], Qing Xu[6], Hans Ole Ørka[2], Erik Næsset[2], Ronald E. McRoberts[7]

1) Swedish University of Agricultural Sciences, Department of Forest Resource Management, Umeå, Sweden.
2) Norwegian University of Life Sciences, Faculty of Environmental Sciences and Natural Resource Management, Ås, Norway.
3) USDA Forest Service, Rocky Mountain Research Station, USA.
4) Brown University, Department of Ecology and Evolutionary Biology, USA.
5) Beijing Forestry University, College of Forestry, China.
6) International Center for Bamboo and Rattan, Key Laboratory of National Forestry and Grassland Administration on Bamboo & Rattan Science and Technology, China.
7) University of Minnesota Twin Cities, Department of Forest Resources, USA.

*) Corresponding author (goran.stahl@slu.se)

Abstract

Many forest ecosystem surveys use remotely sensed data and model-based inference. Assessing uncertainties in connection with such surveys is far from trivial. Focusing on continuous study variables, such as biomass, we first motivate why the mean square error (MSE) of predictors of population quantities of interest is a relevant measure of overall uncertainty. Secondly, we separate the MSE into four components involving (i) the variance of the predictor, (ii) the model-bias of the predictor, (iii) the variance of the true value of the quantity predicted, and (iv) the covariance between the predictor and the true value, and show how the different components can be estimated. Thirdly, based on simulations mimicking conditions in Eastern Africa and Western USA, we assess which components are influential in different survey contexts. Our findings show that using the variance of a predictor, alone, as a measure of uncertainty is a fair approximation if the study area is large and if the model linking remotely sensed data with reference data is adequately specified and estimated. However, we identify substantial risks of underestimating the MSE if models are extrapolated beyond conditions for which they were calibrated, in which case predictors were found to be severely model-biased. The remaining terms contributed differently in large-area surveys compared to small-area surveys. In small-area surveys, the MSE component (iii) is the dominating part. An interesting observation was that the MSE of a predictor is sometimes smaller than the variance of the predictor, due to the negative contribution from term (iv).

Keywords: Model-based inference, superpopulation inference, remote sensing, uncertainty

1. Introduction

The demands for information about the state and change of the world's forest ecosystems are increasing due to ongoing global change (e.g., Trumbore et al. 2015). For example, information about greenhouse gas emissions is required for mitigating climate change (e.g., Schlamadinger et al. 2007) and information about biodiversity is required for efforts to halt biodiversity loss (e.g., Angelstam et al. 2004). Following guidance from the Intergovernmental Panel on Climate Change (Penman et al. 2003), methods to monitor and report greenhouse gas emissions from land use, land-use change, and forestry involve monitoring carbon stock changes in different pools, such as above- and belowground biomass. Monitoring biodiversity may involve individual species, but due to methodological challenges indicators of forest ecosystem structure, such as amounts of deadwood (e.g., Fridman and Walheim 2000), are more often used.

Acquiring data for large-area compilations of forest ecosystem state and change is extremely challenging. Nevertheless, the increasing wealth of remotely sensed data from different space missions offers new possibilities (e.g., Araza et al. 2022, Dubayah et al. 2022). With such data, coupled with field reference data, models that predict variables of interest can be developed and subsequently applied across the areas of interest. During the last decade, many assessments of this kind have been conducted, focusing on indicators such as aboveground biomass (e.g., Araza et al. 2022). The results can be displayed both in terms of maps and in terms of formal predictions of population totals and means, using methods from model-based inference (cf., Gregoire 1998). However, although a large number of studies use remotely sensed data and model-based inference, confusion still often arises over how to assess and present uncertainties (e.g., Gregoire et al. 2016; Ståhl et al. 2024).

Choosing between design-based and model-based inference for ecosystem surveys based on remotely sensed data is often a choice between some form of laborious random sampling of field plots (for design-based inference) or purposive sampling of a limited number of plots (for model-based inference) for calibrating prediction models which are then applied across the area of interest. In large-area surveys, model-based inference is sometimes the only viable option due to budget constraints or limited accessibility to remote areas. Thus, several recent large-area surveys based on remotely sensed and field data explicitly have applied model-based inference or inference combining model and design known as hybrid inference (e.g., Chen et al. 2016, Saarela et al. 2016, Dubayah et al. 2022). In some cases, principles of model-based inference are applied implicitly, although no formal reference to the inferential framework is given (e.g., Gregoire et al. 2016).

An important distinguishing factor between model-based and design-based inference is that, in the former, the values of interest for the population elements as well as population totals and means are treated as random variables. They are assumed to be generated by a super-population model (e.g., Cassel et al. 1977, p. 80), which links one or more known auxiliary variables for each element with the target variable. The model asserts that the value of the variable of interest, for all the population elements, consists of a fixed part plus a random part (in standard additive models). In design-based inference, the values of the variable of interest for individual population elements, as well as totals and means, are treated as fixed

but unknown quantities. This implies several important differences between design-based and model-based inference, e.g. that design-bias and model-bias are defined differently (e.g., Cassel et al. 1977, Ståhl et al. 2024). Several studies using remotely sensed data for assessing ecosystem characteristics tend to apply model-based inference methodologies, but assess their results from a design-based viewpoint, which leads to confusing or incorrect assessments of uncertainties (e.g., Ståhl et al. 2024).

Many studies adopting model-based inference use the variance of a predictor of the quantity of interest (e.g., the population mean or total) as a measure of uncertainty. Although this does not take into account that the true value of the quantity is a random variable, which contributes to overall uncertainty, several studies suggest (e.g., McRoberts et al. 2018) that the variance of a predictor provides a fair approximation of overall uncertainty in large-area surveys. However, to account for all potential sources of uncertainty in model-based inference, the mean square error (MSE) rather than the variance would be the relevant uncertainty measure. Compared to the variance, the MSE acknowledges that the true value of the population quantity predicted is random, and it takes into account that predictors may be model-biased. Thus, thorough uncertainty assessments in model-based inference should use the MSE rather than the variance (e.g., Cassel et al. 1977).

The MSE discussed in this study is the MSE linked to formal model-based inference. It should not be confused with the MSE (or RMSE) assessed in many contemporary mainstream studies on remote sensing of forests, where pairwise comparisons at the level of population elements for which both model predictions and field data are available (e.g., Persson and Ståhl 2020). Whereas such studies are not irrelevant, they cannot be extended to produce uncertainty estimates for predictions for larger areas, such as predictions of population totals or means.

The objective of this study was to present a generic expression for the MSE in model-based inference and to partition the expression into its components. We describe how each of the components should be interpreted and how they can be estimated. Further, we demonstrate which components are important to include in comprehensive uncertainty assessments in different types of forest ecosystem surveys using remotely sensed data, through simulations mimicking conditions in Eastern Africa and Western USA.

The study focuses on formal model-based inference under ideal conditions with regard to data availability and possibilities to match different data sources. Real conditions are often more challenging, e.g. with partial availability of remotely sensed data, which may be poorly georeferenced. In such cases additional sources of uncertainty arise, which are only briefly discussed in this article.

To distinguish clearly between design-based and model-based inference, we use the term *prediction* for assessing a random quantity of interest in model-based inference and *estimation* for assessing a fixed quantity of interest in design-based inference[1]. Although the

[1] Note that fixed quantities exist also in model-based inference. For example, the variance, the MSE and the model-bias are fixed quantities, which thus are *estimated* rather than *predicted*.

focus of our study is model-based inference, comparisons with design-based inference are sometimes made.

2. Materials and Methods

2.1 Overview

In this chapter, we first define the MSE of a predictor in model-based inference and provide further arguments for why it is a relevant measure of overall uncertainty. A generic expression for the MSE is derived and its four components discussed. Secondly, we describe the data, methods, and simulations performed to estimate the MSE components in different hypothetical surveys mimicking conditions in Eastern Africa and Western USA.

2.2 The MSE of a predictor

The MSE of a predictor[2] is defined as the expected squared deviation between a predicted value and the true value of the quantity being predicted. This is a generic definition, valid for both model-based and design-based inference, although we speak of estimation rather than prediction in design-based inference. However, in model-based inference the true value is a random variable, which implies that the model-based MSE comprises different components compared to the design-based MSE.

In a survey, the target characteristics could be related to a single population element, to some domain of interest, or to the entire population, such as the population total or mean (generically denoted *Y* in the following). If the MSE is small, our predictor tends to be always close to the true value, which follows from the definition of MSE. Thus, the MSE can be considered as an intuitively straightforward measure to use for assessing overall uncertainty, for predictors at the level of individual population elements, domains, or totals or means.

In the following, we derive a general expression for the MSE in model-based inference and take a closer look at its components. By definition, the MSE of a predictor, $\hat{Y}$, of the (random) quantity *Y* is

$$\mathrm{MSE}\big(\hat{Y}\big) = E(\hat{Y} - Y)^2. \tag{1}$$

The right-hand side of the MSE expression can be expanded as

$$\mathrm{E}[\hat{Y} - Y]^2 = \mathrm{E}[(\hat{Y} - \mathrm{E}\big(\hat{Y}\big)) + (\mathrm{E}\big(\hat{Y}\big) - \mathrm{E}(Y)) + (\mathrm{E}(Y) - Y)]^2. \tag{2}$$

We expand the expression in this way as a preparatory step for breaking down the MSE into its components. Note that the terms have been deliberately paired, with the intention to let the contents within each of the parentheses remain intact while squaring the right-hand side of the expression. We thus obtain the following expression for the MSE:

[2] Sometimes abbreviated as MSEP, i.e. Mean Square Error of a Predictor

$$\mathrm{E}(\hat{Y} - \mathrm{E}(\hat{Y}))^2 + (\mathrm{E}(\hat{Y}) - \mathrm{E}(Y))^2 + \mathrm{E}(\mathrm{E}(Y) - Y)^2 + 2\mathrm{E}(\hat{Y} - \mathrm{E}(\hat{Y}))(\mathrm{E}(Y) - Y).$$

Two additional terms are obtained when squaring the expression, but each of them is zero. All the remaining terms have important meaning for the MSE of a predictor in model-based inference.

By definition, the term $\mathrm{E}(\hat{Y} - \mathrm{E}(\hat{Y}))^2$ is the variance of the predictor (e.g., Cassel et al. 1977). The term $(\mathrm{E}(\hat{Y}) - \mathrm{E}(Y))^2$ is the squared model-bias (ibid.). The term $\mathrm{E}(\mathrm{E}(Y) - Y)^2$ is the variance of the true target quantity (perhaps more intuitively seen if the order of the terms is reversed, but due to squaring, the order of the terms inside the parenthesis does not matter). Finally, the term $2\mathrm{E}(\hat{Y} - \mathrm{E}(\hat{Y}))(\mathrm{E}(Y) - Y) = -2\mathrm{E}(\hat{Y} - \mathrm{E}(\hat{Y}))(Y - \mathrm{E}(Y))$ is twice the negative of the covariance between the predictor and the true value.

The expectations are evaluated across an infinite number of realisations of populations from a superpopulation model, for a fixed sample of population elements used for calibrating models (e.g., Cassel et al. 1977). This may be compared to design-based inference, where probability theory applies due to selecting random samples of population elements from a single fixed population. Thus, note that the underlying assumptions which make estimators random variables in design-based inference and predictors random variables in model-based inference are entirely different (e.g., Cassel et al. 1977).

To summarize, we obtain the following expression for the MSE of a predictor in model-based inference[3]:

$$\mathrm{MSE}(\hat{Y}) = \mathrm{Var}(\hat{Y}) + \mathrm{Bias}^2(\hat{Y}) + \mathrm{Var}(Y) - 2\mathrm{Cov}(\hat{Y}, Y). \quad (3)$$

Because data are rarely available for direct empirical assessment of the MSE from a survey, it is important to understand how each of these components contribute to the MSE, and to find ways for estimating the influential terms. Direct empirical assessment of the MSE would only rarely be possible, for cases where several realisations of populations from the super-population model have been obtained, as well as data from each of these. This is seldom or never the case for ecosystems such as forests, where conditions remain relatively stable across long periods of time. Thus, any given forest can be considered as a single realisation from some superpopulation model, across a long time span.

However, assessments of the MSE components can be based on data from a single population realisation, similarly to how properties of design-based estimators can be assessed from a single sample from the population. We note that a generic albeit computationally demanding approach for estimating the MSE components would be to estimate the parameters of the model we assume to be a good proxy of the superpopulation model, based on data from the single population and sample available. Then, we use the estimated proxy super-population model for repeatedly simulating new populations and the associated model predictions of the target quantity of interest. The components of the MSE can then be empirically assessed based on the outcomes from the simulations. This could be seen as a

[3] In design-based inference, the MSE of an estimator $\hat{Y}$ is $\mathrm{MSE}(\hat{Y}) = \mathrm{Var}(\hat{Y}) + \mathrm{Bias}^2(\hat{Y})$, where $\mathrm{Bias}(\hat{Y})$ is the design-bias

parallel to bootstrapping from a single sample in surveys applying design-based inference (e.g., Shao 2003). In the case of model-based inference, however, the sample remains fixed whereas populations are randomly generated.

Other, more commonly applied, estimation approaches are described in the following sections for some of the MSE components, where we also discuss each of the components in more detail.

Component I: The variance of the predictor

The variance, $\mathrm{Var}(\hat{Y})$, of a predictor (at the level of an individual population element, a domain, or the population total or mean) expresses the variability of the predictor across different realisations of populations from the superpopulation model, conditional on a given sample of population elements (e.g. Cassel et al. 1977). For that sample, different realised populations lead to different predictions of the population characteristic of interest to us, which is the randomisation basis for this variance.

The variance of the predictor is defined as $\mathrm{E}(\hat{Y} - \mathrm{E}(\hat{Y}))^2$, i.e. it is a measure of the variability of the predictor in relation to the expected value of the predictor.

$\mathrm{Var}(\hat{Y})$ is sometimes confused with $\mathrm{Var}(\hat{Y} - Y)$, when "the variance" of predictors in model-based inference is discussed. However, note that

$$\mathrm{Var}(\hat{Y} - Y) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(Y) - 2\mathrm{Cov}(\hat{Y}, Y), \tag{4}$$

which is identical to $\mathrm{MSE}(\hat{Y})$ for a model-unbiased predictor (cf., Eq. 3). Thus, it is important to distinguish between $\mathrm{Var}(\hat{Y})$ and $\mathrm{Var}(\hat{Y} - Y)$ in model-based inference, although several studies (e.g., McRoberts et al. 2018) suggest that $\mathrm{Var}(\hat{Y})$ is the dominating component of $\mathrm{Var}(\hat{Y} - Y)$ in large-area surveys.

Estimation of $\mathrm{Var}(\hat{Y})$ is straightforward in case parametric models are applied, especially if a linear model is involved, such as a model linking aboveground biomass with a remotely sensed metric. For example, if we assume that a simple linear model adequately specifies the underlying superpopulation model (i.e., at the level of individual population elements), the model is

$$y_i = \alpha + \beta x_i + \epsilon_i, \tag{5}$$

where $y_i$is the biomass for the *i:*th population element, $x_i$is the remotely sensed metric for the same element, α and β are parameters, and ϵ_i is a random error term (with expected value zero). A standard predictor of the population total[4], τ, in model-based inference is

$$\hat{\tau} = \sum_{i=1}^{N} \hat{y_i} = \sum_{i=1}^{N} (\hat{\alpha} + \hat{\beta} x_i), \tag{6}$$

where *N* is the total number of population elements. The model parameters are estimated from the sample data, using some appropriate technique for parameter estimation such as

[4] We use the notation τ for the population total, rather the previously used "generic" *Y,* to avoid confusion with the values for individual population elements.

ordinary least squares regression (e.g. Chatterjee and Hadi 2013). In some cases of model-based inference, a distinction is made between the values of the elements which are actually measured and those that are predicted (e.g. Chambers and Clark 2012). In our case we assume that the population is large relative to the sample, and for simplicity we assume that model-based predictions are made for all population elements.

The variance of the predictor can be written as

$$\mathrm{Var}(\hat{t}) = \mathrm{Var}\left(\sum_{i=1}^{N} \hat{y}_i\right) = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Cov}(\hat{y}_i, \hat{y}_j) = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathrm{Cov}\left((\hat{\alpha} + \hat{\beta}x_i), (\hat{\alpha} + \hat{\beta}x_j)\right)$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} \left[\mathrm{Var}(\hat{\alpha}) + x_i \mathrm{Cov}(\hat{\alpha}, \hat{\beta}) + x_j \mathrm{Cov}(\hat{\alpha}, \hat{\beta}) + x_i x_j \mathrm{Var}(\hat{\beta})\right]. \quad (7)$$

Variances and covariances of parameter estimates can be estimated during the model fitting (e.g. Chatterjee and Hadi 2013), whereby a variance estimator is obtained by substituting the true variances and covariances with their estimated counterparts.

Equation (7) is commonly and more generally written in terms of linear algebra, e.g. as

$$\mathrm{Var}(\hat{t}) = \mathbf{1}^t \boldsymbol{X}\boldsymbol{C}\boldsymbol{X}^t \mathbf{1}\,, \quad (8)$$

where ***1*** is a *N x 1* column vector of unit numbers, ***X*** is a *N x p* matrix of remote sensing metrics (including an initial column of units), ***C*** is *p x p* variance-covariance matrix for the estimated model parameters, with *p* being the number of parameters in the model. Slightly different expressions are obtained depending on what the target quantity is; note that in (7) and (8) the predictor concerns the population total.

In case non-linear models are applied, Taylor approximation can be used to obtain approximations of (7) and (8) (e.g., Davidson and McKinnon 1993). In the case of non-parametric models, the approach outlined above is not applicable, but instead bootstrapping procedures can be used for estimating the variance (e.g., Esteban et al. 2019). Sometimes multiple models are involved. In case they are hierarchically nested, formulas applicable to variance estimation in hierarchical model-based inference can be utilised, as shown by Saarela et al. (2016) for the case of parametric models, and Fortin et al. (2024) and Saarela et al. (2025) for the case of non-parametric models and bootstrapping.

Component II: The squared model-bias

The model-bias is defined as $\mathrm{E}(\hat{Y}) - \mathrm{E}(Y)$, i.e. it is the difference between the expectations of predicted and true values, across an infinite number of realisations from the superpopulation model (e.g. Cassel et al. 1977). Note the difference between model-bias and design-bias, where the latter is defined as $\mathrm{E}(\hat{Y}) - Y$, with Y being fixed. Some studies, which apply model-based inference, confuse the two concepts (cf., Ståhl et al. 2024) and evaluate the design-bias of model-based predictions. This typically leads to observing that small true values are over-predicted and large true values are under-predicted (e.g., ibid).

A theoretical evaluation of model-bias should address the outcomes of predictors and true values across a large number of realisations from the superpopulation model. With only a single realisation at hand, realised true values should *not* be compared directly with the corresponding predicted values, but the model-bias should be assessed by studying residual terms versus predicted values of the target population quantity (e.g., Chatterjee and Hadi 2013). A relevant approach is to group the available data by predicted values and compute the average value of the residuals for each group. This gives an indication of whether the model is model-biased or model-unbiased for each group. In the absence of strong spatial autocorrelation of error terms, standard *t*-tests may be conducted to assess whether or not the mean of the residuals within a group significantly differs from zero.

However, the above procedure will not capture the major potential problem of model-bias in model-based inference, i.e. the model-bias that may arise when models are applied outside their range of applicability (e.g., Réjou-Méchain et al. 2019). Due to the scarcity of field data available for model training in large-area surveys (e.g., Duncanson et al. 2022), models trained on data from geographically restricted areas are often used across vast areas, assuming that the models are valid for these areas as well (e.g., ibid). Model-bias is likely to arise when models are applied in ecosystems that are different from the one where training data were acquired. We term this bias *model transfer bias*, and based on empirical data and simulations we will give examples of its magnitude.

A third type of model-bias follows from incorrect specification of the proxy superpopulation model. For example, if an important variable in the superpopulation model is left out from the proxy model, variable omission bias (cf., Cinelli and Hazlett 2020) may arise. However, we do not address variable omission bias further in this study. Instead, we assume that the models applied are adequately specified, in the sense that the correct auxiliary variables are included in the proxy superpopulation model.

Component III: The variance of the true value

The variance of the true value, $\mathrm{Var}(Y)$, is the variance of our population quantity of interest across an infinite number of realisations from the superpopulation model. With data from a single realisation, we will mostly be able to estimate a proxy of the superpopulation model, which can assist in estimating $\mathrm{Var}(Y)$. As previously shown, this model involves the relationship between one or more auxiliary variables and the dependent variable. It also involves properties of the model error terms (the ϵ_is), i.e. their magnitude and their correlation among population elements.

If a proxy superpopulation model is estimated from the data, it can be used for estimating $\mathrm{Var}(Y)$, either analytically (if the model is not too complex) or through simulation. For example, if we assume the same superpopulation model as in (5) and the target quantity is the population total, it can be expressed as

$$\tau = \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} (\alpha + \beta x_i + \epsilon_i)\,. \qquad (9)$$

Note that in the right-hand side of (9) all quantities but ϵ_i are fixed and thus $\mathrm{Var}(\tau)$ can be estimated as

$$\widehat{\mathrm{Var}}(\tau) = \sum_{i=1}^{N}\sum_{j=1}^{N}\widehat{\mathrm{Cov}}(\hat{\epsilon}_i, \hat{\epsilon}_j) = \sum_{i=1}^{N}\widehat{\mathrm{Var}}(\hat{\epsilon}_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N}\widehat{\mathrm{Cov}}(\hat{\epsilon}_i, \hat{\epsilon}_j)\,. \quad (10)$$

Estimation of (10) requires information about (i) the variance of the error terms at the level of individual population elements and any potential heteroscedasticity of this variance, and (ii) the extent of spatial autocorrelation among error terms. Residuals can be used as proxies of the error terms and used in the estimation of (10). To assess spatial autocorrelation among residuals, sample data must be available at both short and long distances between population elements, which may make accurate estimation of $\mathrm{Var}(\tau)$ challenging in practice.

Further, note that $\mathrm{Var}(\tau)$ will differ between cases where different sources of remotely sensed data have been used, and thus where different superpopulation models have been assumed. With weak correlation between auxiliary data and the response variable the error terms tend to be large and it is likely that their spatial autocorrelation will remain strong over long distances. Therefore, the variance of the true value may be large.

Component IV: The covariance between the predictor and the true value

This term, i.e. $\mathrm{Cov}(\hat{Y}, Y)$, is often overlooked in studies of model-based inference applied to remotely sensed data, although, e.g., Fortin et al. (2022) is an exception. The term emanates from the logic that, if the realised populations include a large portion of unusually large outcomes of the variable of interest, the predictions from a model estimated from such data would also tend to be larger than they would otherwise be, and vice versa for populations with unusually small outcomes from the superpopulation model. Thus, it is logical that this term is non-zero in many cases[5].

A simple further development of the covariance gives some additional insights

$$\mathrm{Cov}(\hat{Y}, Y) = \mathrm{Cov}(Y + \delta, Y) = \mathrm{Var}(Y) + \mathrm{Cov}(\delta, Y). \quad (11)$$

Here, δ is the difference between a predicted value and a true value. If the latter covariance, i.e. $\mathrm{Cov}(\delta, Y)$, is small, $\mathrm{Cov}(\hat{Y}, Y) \simeq \mathrm{Var}(Y)$ which leads to the conclusion that the MSE *can be smaller* than $\mathrm{Var}(\hat{Y})$, if the predictor is model-unbiased (cf., Eq. 3). This stands in contrast to design-based inference, where the MSE is always at least as large as the variance.

In model-unbiased large-area surveys it is thus possible that the MSE is smaller than the variance (cf., Fortin et al. 2022), although the difference would mostly be small (e.g. McRoberts et al. 2018). The reason is that the variance of a predictor assesses variability in relation to the average of predicted values across population realisations, whereas the MSE assesses variability in relation to the actual true values across realisations; these true values would mostly be closer to the predicted values than the average of the predicted values. However, in small-area surveys, as will be shown in the numerical results, the MSE is mostly much larger than the variance.

[5] However, note that there are cases where the covariance would be zero, e.g., for domain level prediction when there are no sample units in the target domain and no spatial autocorrelation among the error terms. Thus, the reasononing is valid especially for predictions of population totals and means in the presence of spatial autocorrelation of error terms.

We suggest that estimation of $\mathrm{Cov}(\hat{Y}, Y)$ generally requires estimation of a proxy super-population model from the sample data and using this proxy model for simulating a large number of populations from which Y is generated and $\hat{Y}$ predicted. Subsequently, the covariance is determined empirically from the simulation results. In specific cases (cf., Fortin et al. 2022), other possibilities for estimating this component may exist.

2.3 Monte-Carlo simulation study

We conducted a simulation study to assess the magnitude of the MSE components in different survey settings, where remotely sensed (RS) data were combined with field reference data for inferring aboveground biomass density (AGBD) through model-based inference. Different sources of RS data were applied in study areas with different sizes, using different sample sizes of field data. Further, models were transferred between populations to assess the magnitude of model transfer bias.

The methodology applied was Monte-Carlo simulation, where we, for each combination of study area and RS data source, estimated a proxy of the superpopulation model and applied it to generate a large number (3000) of population realisations. For each population and field data sample size, a new prediction model was estimated and the AGBD predicted for the target area. Thus, from each simulation a true AGBD value was obtained as well as a predicted value. Based on these, all the terms involved in the MSE could be empirically estimated in the conventional way based on the Monte-Carlo simulation outcomes (e.g., Papadopoulos and Yeung 2001).

With the proxy superpopulation models, the fixed model part for a specific population element was always the same in all the 3000 population outcomes. However, the random error realisations differed each time, leading to different realised true values of AGBD in each population. To mimic different degrees of spatial autocorrelation between the error terms (cf. Eq. 5), they were separated into two parts, i.e.

$$\epsilon_i = \epsilon_{a_i} + \epsilon_{b_i},$$

where ϵ_i is the error term for population element i, $\epsilon_{a_i}$is the part of the term that introduces spatial autocorrelation by assigning the same random value to an entire block of adjacent population elements (which contains element *i*), and ϵ_{b_i} is a random term that is different for every element. We applied this block approach to simulating spatially correlated errors, because other more mainstream methods (e.g., Haining et al. 1983, Rüttenauer 2022) turned out to be far too time-consuming to perform in Monte-Carlo simulations within large study areas.

To apply the block approach, we separated each study area into disjoint rectangles of equal size. By varying the relationship between ϵ_{a_i} and ϵ_{b_i} we mimicked populations with varying degrees of spatial autocorrelation. In the case of "no spatial autocorrelation" the entire random component was assigned to ϵ_{b_i}. In the other cases we assigned certain percentages of the total random error variance to each of the two components. The percentages of the

total variance allocated to ϵ_{a_i} were 25%, 50% and 75%, for "mild", "moderate" and "strong" spatial autocorrelation, respectively. All random error terms were assumed to follow normal distributions.

Because the study areas had slightly different sizes and shapes, the rectangles used for the block approach, described above, contained 120 x 20 elements in Eastern Africa (in total 162 rectangles) and 80 x 30 elements in Western USA (in total 260 rectangles). Intuitively, the approach can be seen as mimicking the case where a forest is divided into patches, within which conditions tend to be more similar than between patches.

Once AGBD values had been assigned to each population element according to the described procedure, a model with the same explanatory variables as the proxy superpopulation model was estimated based on sample data, acquired from samples allocated in a spatially systematic grid across the study area. The elements in the sample were always the same throughout all the population replicates. Then, the AGBD was predicted based on the estimated model.

The study area in Eastern Africa

Field data coupled with airborne laser scanning (ALS) data and Landsat 8 data were available from a miombo woodland study area located in the Liwale district in Tanzania, Eastern Africa, where aboveground biomass density had been assessed through model-assisted estimation (Næsset et al. 2016). A study area of size 34,920 ha[6] was tessellated into cells of size 30 m x 30 m. Landsat data were available for the same area and rescaled to the same resolution as the laser data. Canopy height distributions were derived from the ALS echoes within the respective cells. Separate distributions were created for the first and last echoes. A threshold value of 1.3 m above the ground surface was used and order statistics such as height percentiles, and mean and coefficient of variation, were derived. Canopy density measures were derived by dividing the height range between the 1.3m threshold and the 95 percentile into 10 equally sized height bins and calculating the proportion of hits about each threshold to the total number of hits. The Landsat data were normalised to correspond to surface reflectance values, in the range 0 to 1, for different wavelength bands (e.g., Saarela et al. 2025). All the RS data had undergone relevant pre-processing (cf., Anon 2005; Axelsson 2000; Hansen et al. 2013).

The motivation for involving different sources of RS data was to assess how the magnitude of the MSE components would differ depending on the goodness-of-fit of the proxy super-population model. ALS data are known to be strongly correlated with aboveground biomass (e.g., Næsset and Gobakken 2008), while optical satellite data are known to be only moderately correlated with aboveground biomass (e.g., Powell et al. 2010).

The AGBDs (Mg/ha) based on field data from the study area varied between 0 and 213, with a mean of 51 and standard deviation of 46.

[6] Slightly smaller than the study area in Næsset et al. (2016), due to the procedure we applied for simulating spatially correlated error terms

Our proxy superpopulation model based on ALS data was

$$AGBD_i = 16.463 + 0.985\ PL10_i^2 - 110.685\ DL1_i - 1.306\ meanPF0_i^2 + 217.976\ meanDF5_i^2 + 0.0493\ stdPF20_i^2 + \epsilon_i. \quad (12)$$

Here, *PL10* is the 10th height percentile of last echoes above 1.3 m , *DL1* is the canopy density above the first bin for last echoes, *meanPF0* is the mean value from 5x5 m cells of 0 percentiles of first echoes, *meanDF5* is the mean value from 5x5 m cells of vegetation density above the 5th bin of first echoes, and *stdPF20* is the standard deviation of 20th height percentiles of first echoes based on values from 5x5 m cells. For details, see Næsset et al. (2016).

The random error term, ϵ_i, was assumed to be normally distributed with standard deviation 29 Mg/ha.

Our proxy superpopulation model based on Landsat 8 data was

$$AGBD_i = -50.140 + 1.033\ TreeCover_i - 52.346\ \sqrt{Red_i} + 45.820\ \sqrt{Green_i} + \epsilon_i\ . \quad (13)$$

TreeCover represents the mean tree cover for the year 2010, obtained from Hansen et al. (2013); *Red* corresponds to spectral reflectance (0-1) in the red spectral band, with a wavelength range of 0.630–0.680 µm; *Green* refers to spectral reflectance (0-1) in the green spectral band with a wavelength range of 0.525–0.600 µm. In this case the error term, ϵ_i, was assumed to be normally distributed with standard deviation 31 Mg/ha.

The study area in Western USA

Field data coupled with data from the Global Ecosystem Dynamics Investigation (GEDI; Dubayah et al. 2022) and Landsat 8 data were available wall-to-wall for an artificial 56,160 ha large study area mimicking conditions in Oregon, Western USA, dominated by coniferous forests. Data for the study area were compiled through copula modelling (e.g. Ene et al. 2013), based on GEDI, Landsat and field data available from a previous study (Saarela et al. 2025). Copulas allow for empirical estimation of complex multivariate distributions from sample data, enabling random generation of multivariate data with similar distributions.

The resolution of both GEDI and Landsat data was the same as in the African study area, i.e. 30 m x 30 m, ensuring consistency in the simulation study. GEDI data were available as vegetation heights for different percentiles of the waveform pulse returns. Landsat data were normalised and available as surface reflectance values in the range 0 to 1, like in the case of data from the African study area.

The AGBD based on field data from the study area varied between 0 and 1906 Mg/ha, with a mean of 278 Mg/ha and standard deviation of 225 Mg/ha. Thus, the AGBDs were much larger in the US study area compared to the African study area.

Our proxy superpopulation model based on GEDI data was

$$AGBD_i = 30.799 + 17.648\ Rh50_i + \epsilon_i\ . \quad (14)$$

Here, *Rh*50 is the height at which cumulative 50% of the energy from a GEDI laser pulse was returned. The error terms, ϵ_i, were assumed to be normally distributed with standard deviation 149 Mg/ha. The larger magnitude of the error terms compared to the African models is a result of the substantially larger AGBDs in this study area.

The Landsat model for the US study area was

$$AGBD_i = 1132.4 - 8624.1\sqrt{Green_i} + 5699.9\sqrt{Blue_i} + \epsilon_i \,. \quad (15)$$

Green refers to spectral reflectance (0-1) in the green spectral band with a wavelength range of 0.525–0.600 µm; *Blue* corresponds to spectral reflectance (0-1) in the blue spectral band, with a wavelength range of 0.450–0.515 µm. In this case the error terms had a standard deviation of 150 Mg/ha.

Cases evaluated

Different analysis cases were specified, for which the magnitude of the MSE components were analysed through Monte-Carlo simulation. The cases were:

1) *Baseline case*: For each combination of study area and RS data source, we assessed the MSE components for the predicted mean AGBD for the entire study area, for a survey with 150 field sample plots and moderate spatial autocorrelation of the error terms.

2) *Variability due to study area size*: Using both RS data sources in each study area, we assessed the MSE components across different study area sizes, from very small (36 units, corresponding to 3.24 ha) to small (144 units, corresponding to 12.96 ha), intermediate (1024 units, corresponding to 92.16 ha), and large (the entire study area). This analysis was conducted using 150 field sample plots and moderate spatial autocorrelation of errors. For each study area size, the mean AGBD was about the same. We also assessed the magnitude of the MSE components for the case of a single population element, corresponding to assessing uncertainties for individual map elements in mapping.

3) *Variability due to field sample size*: For each of the RS data sources in each of the study areas, we assessed the MSE components across different field sample sizes, from small (50), to intermediate (150) and large (500). We present results for the entire study areas and for the case of moderate spatial autocorrelation of the error terms.

4) *Variability due to spatial autocorrelation of error terms*: For each of the data sources in each of the study areas, we assessed the MSE components under different assumptions of spatial autocorrelation of errors, from none to mild, moderate, and strong. We present results for the entire study area using 150 field sample plots.

5) *Effect of local model transfer*: For each of the RS data sources in each of the study areas, we evaluated the effect of local model transfer on the MSE components, especially the contribution of model bias to the MSE. The models were constructed as previously explained, but for each realised population they were only applied to the 50% of the population elements with the largest AGBDs. Thus, intuitively, we developed a model for average conditions but applied it only to dense forests.

6) *Effect of regional model transfer*: The Landsat models for Eastern Africa and Western USA were based on comparable and normalised data in terms of spatial reflectance in different wavelength bands. In this case, we applied the Landsat models developed in one study area in the other study area and assessed results in terms of the MSE, especially focussing on the contribution from the model bias.

3. Results

3.1 Baseline case

In Table 1, results for analysis case (1) are presented, i.e. baseline results at the level of the entire study areas for an intermediate sample of field plots and moderate spatial autocorrelation of the error terms.

In each of the study areas, for both types of RS data, the MSE was smaller than the variance of the AGBD predictor. This was due to the non-negligible value of the covariance between the predictor and the true value. The results in this specific case complied fairly well with Eq. (11), i.e. the variance of the true value and the covariance were about equally large. The MSEs for predictors based on ALS or GEDI data were smaller than the MSEs for predictors based on Landsat data, although the differences were small. The model-bias was always almost negligible. The difference in magnitude of MSEs between the two study areas was due to the huge differences in AGBD between the two study areas.

Table 1. Results for analysis case (1) in terms of MSE and its components (AGBD^2, Mg^2/ha^2)

Component	ALS Africa	Landsat Africa	GEDI USA	Landsat USA
Var($\hat{Y}$)	6.42	7.45	136.5	140.5
Bias^2	0.03	0.05	1.41	0.001
Var(Y)	1.94	2.35	29.5	27.6
-2Cov($Y, \hat{Y}$)	-3.81	-4.58	-62.7	-58.0
MSE	**4.58**	**5.29**	**104.6**	**110.1**

3.2 Variability due to study area size

In Figure 2, we present results where the size of the study area varied. Note the log-scale on the x-axis, and the difference in the maximum area size between the two study areas. The main observation is that at all area sizes except the largest, the main contribution to the MSE was from the variance of the true value. The small increase in MSE when moving from the smallest to the second smallest study area size in the case of Landsat in East Africa is probably a simulation artefact.

Including the case of mapping, i.e. results at the level of a single population element, would make Figure 2 difficult to read. Thus, results for this case are instead presented in Table 2. It can be seen that the variance of the true value contributes substantially to the total MSE at this scale; the other terms are negligible compared to Var(Y).

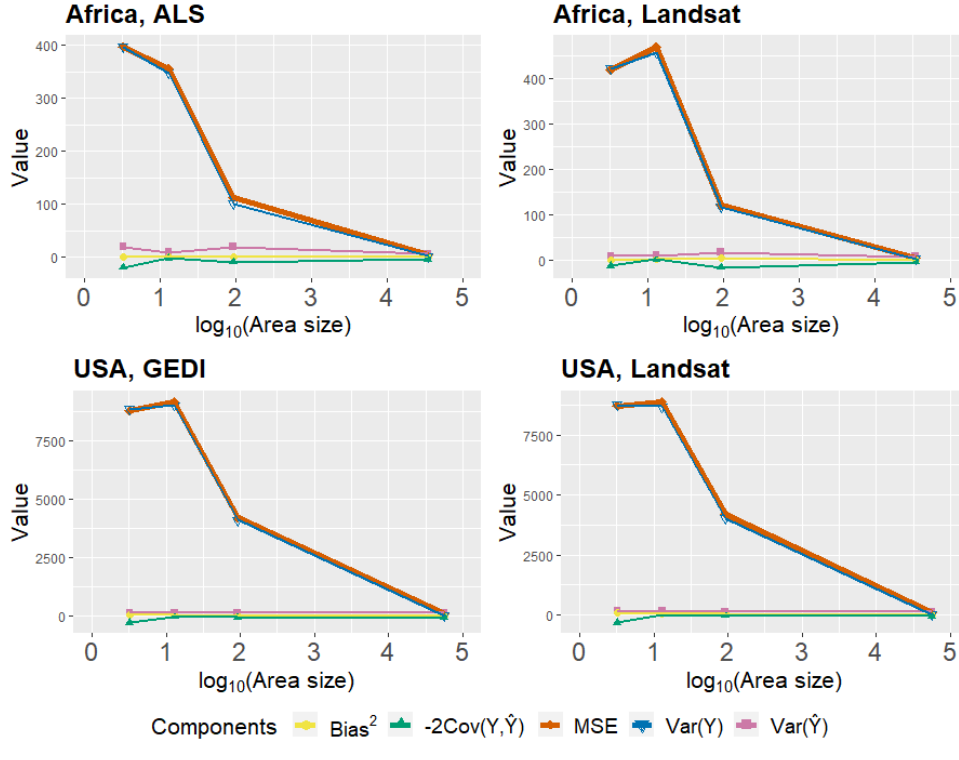


Figure 2. MSE and its components (AGBD², Mg²/ha²) for different study area sizes. Note the log-scale on the x-axis (log_{10}(area size in ha)).

Table 2. Results for analysis case (3) with an area size of one population element (i.e. 0.09 ha) in terms of MSE and its components (AGBD², Mg²/ha²)

Component	ALS Africa	Landsat Africa	GEDI USA	Landsat USA
$\text{Var}(\hat{Y})$	22.8	9.07	131.2	157.6
Bias^2	5.47	1.41	180.5	283.8
$\text{Var}(Y)$	744.8	859.8	18716.6	18799.3
$-2\text{Cov}(Y, \hat{Y})$	1.14	-14.1	-1.01	-386.0
MSE	**774.3**	**856.1**	**19027.2**	**18854.7**

3.3 Variability due to sample size

In Figure 3, we show the MSE and its components for different field sample sizes, when AGBD predictions were made for entire study areas. A clear trend of declining MSE with increasing sample size was obtained, driven by the decreasing variance of the AGBD predictor. All other MSE components remained about the same.

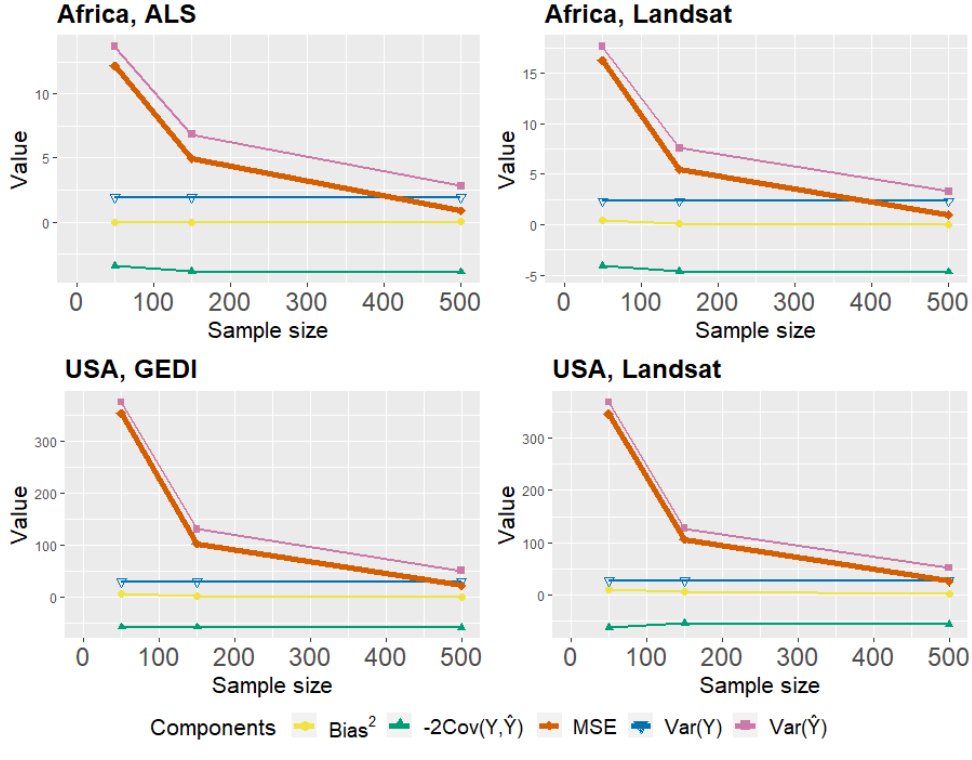


Figure 3. MSE and its components (AGBD², Mg²/ha²) for different field sample sizes.

3.4 Variability due to spatial autocorrelation of errors

In Figure 4, we present results in terms of MSE and its components when the spatial autocorrelation of errors varied. Overall, it can be observed that the spatial autocorrelation of errors substantially affected the results. With correlated errors, the MSEs decreased compared to the case of no spatial autocorrelation. One important reason is the stronger covariance between the predictor and the true value. Interestingly, the stronger the spatial autocorrelation, the larger the variance of the predictor, but the smaller the MSE of the predictor.

3.5 Effect of local model transfer

In this case the focus is the contribution of the model-bias to the MSE, when models are applied to a different population compared to the one used for model calibration. This case is similar to case (1) but the estimated model was applied only to the 50% population elements with the largest AGBD values in the study areas. In Table 3, results in terms of MSE and its components are shown. Compared to the results in Table 1, the MSEs increased substantially due to the squared model-bias terms. The proportion of the MSE that was due

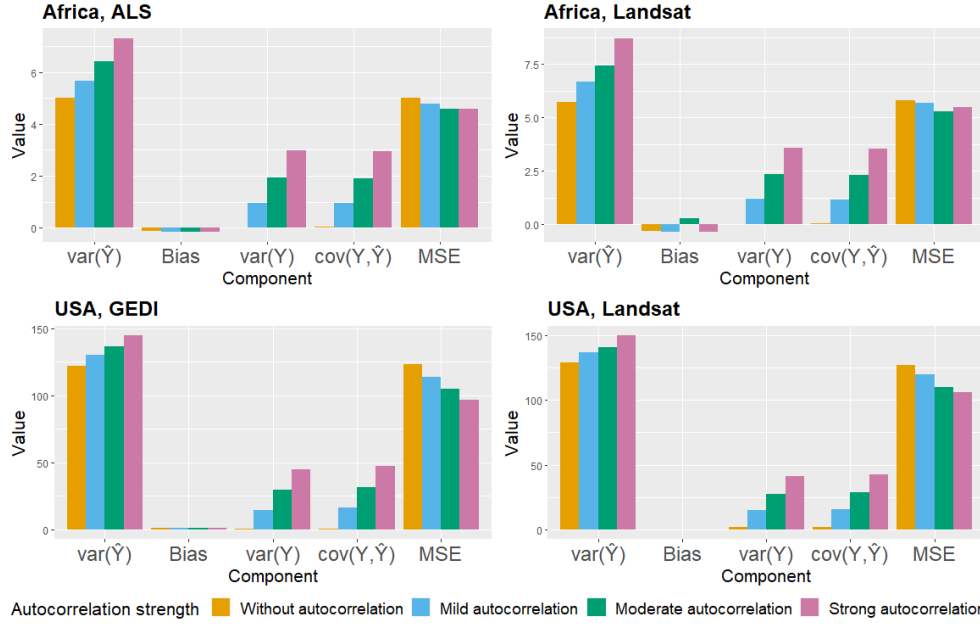


Figure 4. MSE and its components ($AGBD^2$, Mg^2/ha^2) for different magnitudes of spatial autocorrelation of error terms.

Table 3. MSE and its components in the case of local model transfer (AGBD², Mg²/ha²)

Component	ALS Africa	Landsat Africa	GEDI USA	Landsat USA
Var($\hat{Y}$)	9.16	8.88	203.1	214.1
Bias^2	222.9	427.5	6779.8	8416.2
Var(Y)	2.90	3.39	45.5	43.2
-2Cov($Y, \hat{Y}$)	-5.14	-5.70	-83.1	-84.0
MSE	**229.8**	**434.0**	**6945.3**	**8589.5**

to the squared model-bias ranged from 92.5% for ALS in Eastern Africa to 98.0% for Landsat in Western USA.

3.6 Effect of regional model transfer

In a final case, we assessed the effect of using a model developed in one region to another region. Thus, we applied the Landsat model developed in the USA to the study area in Africa, and vice versa. The results are shown in Table 4. Non-surprisingly, the contribution from the squared model-bias was profound, making up 94% (USA -> Africa) and 97% (Africa -> USA) of the MSE. The variance of the AGBD predictor also increased substantially.

4. Discussion

The results show that the contribution from the different components of the MSE differs substantially between different survey cases. The conclusion from previous studies (e.g., McRoberts et al. 2018) that the variance of a predictor is a fair approximation of overall MSE appears to be relevant for large-area surveys, especially if the spatial autocorrelation of error terms is weak and if model-bias can be ruled out.

However, the contribution of model-bias to the MSE appears to be a substantial problem, especially because it is notoriously difficult to evaluate whether or not a predictor is model-

Table 4. MSE and its components in the case of regional model transfer (AGBD², Mg²/ha²)

Component	Africa -> USA	USA -> Africa
Var($\hat{Y}$)	3358	970
Bias^2	122888	14448
Var(Y)	26.83	1.90
-2Cov($Y, \hat{Y}$)	19.12	-1.60
MSE	**126292**	**15419**

biased in real-world surveys. We concur with Renaud et al. (2022) and suggest that this issue merits substantial further investigation, especially because the scarcity of field data for model calibration in regional and global studies implies that prediction models are often extrapolated far beyond the populations from which calibration data were available. In such cases, it is likely that reporting only the variance of a predictor substantially underestimates uncertainty in terms of MSE. From the results obtained in our simulation study, the squared bias sometimes made up more than 90% of the MSE. Reporting only the variance of a predictor as its uncertainty in such cases would be severely misleading.

For cases where model-bias was not present, it is interesting to note that for large-area surveys the MSE of a predictor was typically smaller than the variance of the predictor, due to the negative contribution from the covariance between the predictor and the true value. The contribution from the covariance term was larger the stronger the spatial autocorrelation of errors was (cf., Fortin et al 2022). This stands in contrast to surveys applying design-based inference, where the MSE of an estimator is always at least as large as the variance of the estimator (e.g., Gregoire and Valentine 2007).

From studying the MSE components across different study area sizes it is clear that including the variance of the true value is imperative for small areas. At the level of individual population elements, i.e. mapping, it is by far the dominating MSE component and its magnitude can be assessed from model residuals (e.g., Saarela et al. 2020). For slightly larger but still "small" areas, including this term is typically a challenge because field data from which the actual spatial autocorrelation of errors can be estimated are normally lacking. However, these conclusions are far from new, and several different estimators or predictors have been proposed for small-area estimation in forest inventories (e.g., Breidenbach and Astrup 2012, Dettmann et al. 2022).

Non-surprisingly, the magnitude of the MSE also depends on the sample size of field data. This is well-known from previous studies (e.g., Dubayah et al. 2022). Especially, it is the variance of the predictor that decreases with increasing sample size. Whereas this is an important property of predictors in model-based inference, surveys that apply very large field data sample sizes may report very small uncertainties in terms of predictor variance, probably unrealistically small from the point of view of MSE being a more relevant uncertainty measure. In relative terms, the other MSE-components are likely to be more influential in such cases, not least the squared model-bias.

We pursued the study assuming idealised conditions regarding field and remotely sensed data. In practice (e.g., Zhang et al. 2014, Persson and Ståhl 2020), field data may contain errors, and geolocation inaccuracies may make matching of field and remotely sensed data challenging. Further, field plots often have different sizes and shapes compared to the remotely sensed data. A case sometimes discussed (e.g., Araza et al. 2022) is that the field plots are smaller than the population elements (pixels) for which remotely sensed data are available, so that the reference data can be seen as a sample of the conditions within a population element. When encountering multiple error sources, several studies on model-

based large-area assessment apply bootstrapping to quantify overall uncertainty, sometimes in seemingly ad-hoc manners assuming all error sources to be independent.

In formal application of model-based inference, we argue that it is important to understand what error sources are encountered and how they interact. A core issue is how different sources of error affect our possibility to correctly specify and estimate a proxy super-population model. For example, the case of field plots smaller than pixels would make the model parameter estimates more uncertain, but not biased[7] (e.g., Chatterjee and Hadi 2013). However, this increased uncertainty would be contained in the model parameter uncertainty estimates. On the other hand, the model error terms would appear larger than the error terms in the "real" superpopulation model, and corrections would be needed for avoiding overestimation of the variance of the true value. Analyses of the above kind would be needed for all "nuisance" error sources encountered.

Substantial confusion appears to exist among researchers and practitioners in forest remote sensing about how model-based inference should be conducted (e.g., Gregoire et al. 2016). Model-based and design-based perspectives are often mixed, which causes confusion (e.g., Ståhl et al. 2024). A mainstream approach is to apply modelling methods from model-based inference but evaluate results in a fixed population setting. This typically leads to the conclusion that small true values are overestimated and large true values underestimated, which is not a consequence of lack of data for modelling, but a consequence of mixing methods (e.g., Ståhl et al 2024).

We end by arguing, in line with Gregoire et al. (2016), that the current confusion around statistical methods is a severe obstacle for sound developments of forest assessments based on remotely sensed data.

5. Conclusions

In this article, we suggest that the MSE of a predictor provides comprehensive insight into uncertainties in model-based inference. We decompose the MSE into four components and discuss how they should be interpreted and be estimated.

In a Monte-Carlo simulation study mimicking conditions in Eastern Africa and Western USA, we demonstrate the magnitude of the different MSE components in different forest survey settings, focusing on predicting aboveground biomass density. We found that in small study areas the variance of the true value is the main component of the MSE. In larger areas, in case of no model-bias, the MSE of a predictor is typically slightly smaller than the variance of a predictor, due to the contribution from the covariance between the predictor and the true value in the MSE expression.

A main conclusion from the study is that model-bias may substantially increase the MSE of a predictor in case models are extrapolated beyond the population for which they were estimated. This is problematic, because model-bias is very challenging to assess in practice in

[7] Unless variables that are scale-dependent are used in the models

large-area studies. We suggest that further studies on this topic are required for making uncertainty estimates from large-area forest assessments based on remotely sensed data more reliable.

References

Angelstam, P., Roberge, J. M., Dönz-Breuss, M., Burfield, I. J., & Ståhl, G. (2004). Monitoring forest biodiversity: From the policy level to the management unit. *Ecological Bulletins*, 295-304.

Anon. (2005). TerraScan User's Guide.

Araza, A., De Bruin, S., Herold, M., Quegan, S., Labriere, N., Rodriguez-Veiga, P., ... & Lucas, R. (2022). A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps. *Remote Sensing of Environment*, *272*, 112917.

Axelsson, P. (2000). DEM generation from laser scanner data using adaptive TIN models. *International archives of photogrammetry and remote sensing*, *33*(4), 110-117.

Breidenbach, J., & Astrup, R. (2012). Small area estimation of forest attributes in the Norwegian National Forest Inventory. *European Journal of Forest Research*, *131*, 1255-1267.

Cassel, C. M., Särndal, C. E., & Wretman, J. H. (1977). *Foundations of inference in survey sampling*. John Wiley & Sons.

Chambers, R. L. & Clark, R. (2012). *An introduction to model-based survey sampling with applications*. Oxford University Press.

Chatterjee, S. & Hadi, A. S. (2013). *Regression analysis by example*. John Wiley & Sons.

Chen, Q., McRoberts, R. E., Wang, C., & Radtke, P. J. (2016). Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference. *Remote Sensing of Environment*, *184*, 350-360.

Cinelli, C. & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39-67.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics* (Vol. 63). New York: Oxford.

Dettmann, G. T., Radtke, P. J., Coulston, J. W., Green, P. C., Wilson, B. T., & Moisen, G. G. (2022). Review and synthesis of estimation strategies to meet small area needs in Forest inventory. *Frontiers in Forests and Global Change*, *5*, 813569.

Dubayah, R., Armston, J., Healey, S. P., Bruening, J. M., Patterson, P. L., Kellner, J. R., ... & Luthcke, S. (2022). GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, *17*(9), 095001.

Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., ... & Zgraggen, C. (2022). Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment*, *270*, 112845.

Ene, L. T., Næsset, E., & Gobakken, T. (2013). Model-based inference for k-nearest neighbours predictions using a canonical vine copula. *Scandinavian Journal of Forest Research*, *28*(3), 266-281.

Esteban, J., McRoberts, R. E., Fernández-Landa, A., Tomé, J. L., & Næsset, E. (2019). Estimating forest volume and biomass and their changes using random forests and remotely sensed data. *Remote Sensing*, *11*(16), 1944.

Fortin, M., van Lier, O., & Côté, J. F. (2022). Combining forest growth models and remotely sensed data through a hierarchical model-based inferential framework. *Canadian Journal of Forest Research*, *53*(2), 90-102.

Fortin, M., van Lier, O., Côté, J. F., Erdle, H., & White, J. (2024). A bootstrap-based approach to combine individual-based forest growth models and remotely sensed data. *Forestry: An International Journal of Forest Research*, *97*(4), 649-661.

Fridman, J., & Walheim, M. (2000). Amount, structure, and dynamics of dead wood on managed forestland in Sweden. *Forest ecology and management*, *131*(1-3), 23-36.

Gregoire, T. G. (1998). Design-based and model-based inference in survey sampling: appreciating the difference. *Canadian Journal of Forest Research*, *28*(10), 1429-1447.

Gregoire, T. G., & Valentine, H. T. (2007). *Sampling strategies for natural resources and the environment*. Chapman and Hall/CRC.

Gregoire, T. G., Næsset, E., McRoberts, R. E., Ståhl, G., Andersen, H. E., Gobakken, T., ... & Nelson, R. (2016). Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. *Remote Sensing of Environment*, *173*, 98-108.

Haining, R., Griffith, D. A., & Bennett, R. (1983). Simulating two-dimensional autocorrelated surfaces. *Geographical Analysis*, *15*(3), 247-255.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., ... & Townshend, J. R. (2013). High-resolution global maps of 21st-century forest cover change. *science*, *342*(6160), 850-853.

McRoberts, R. E., Næsset, E., Gobakken, T., Chirici, G., Condés, S., Hou, Z., ... & Walters, B. F. (2018). Assessing components of the model-based mean square error estimator for remote sensing assisted forest applications. *Canadian Journal of Forest Research*, *48*(6), 642-649.

Næsset, E., & Gobakken, T. (2008). Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. *Remote Sensing of Environment*, 112, 3079-3090.

Næsset, E., Ørka, H. O., Solberg, S., Bollandsås, O. M., Hansen, E. H., Mauya, E., ... & Gobakken, T. (2016). Mapping and estimating forest area and aboveground biomass in miombo woodlands in Tanzania using data from airborne laser scanning, TanDEM-X, RapidEye, and global forest maps: A comparison of estimated precision. *Remote Sensing of Environment*, *175*, 282-300.

Papadopoulos, C. E., & Yeung, H. (2001). Uncertainty estimation and Monte Carlo simulation method. *Flow Measurement and Instrumentation*, *12*(4), 291-298.

Penman, J., Gytarsky, M., Hiraishi, T., Krug, T., Kruger, D., Pipatti, R., ... & Wagner, F. (2003). Good practice guidance for land use, land-use change and forestry. *Good practice guidance for land use, land-use change and forestry.*

Persson, H. J., & Ståhl, G. (2020). Characterizing uncertainty in forest remote sensing studies. *Remote Sensing*, *12*(3), 505.

Powell, S. L., Cohen, W. B., Healey, S. P., Kennedy, R. E., Moisen, G. G., Pierce, K. B., & Ohmann, J. L. (2010). Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: A comparison of empirical modeling approaches. *Remote Sensing of Environment*, *114*(5), 1053-1068.

Réjou-Méchain, M., Barbier, N., Couteron, P., Ploton, P., Vincent, G., Herold, M., ... & Pélissier, R. (2019). Upscaling forest biomass from field to satellite measurements: sources of errors and ways to reduce them. *Surveys in Geophysics*, *40*, 881-911.

Renaud, J. P., Sagar, A., Barbillon, P., Bouriaud, O., Deleuze, C., & Vega, C. (2022). Characterizing the calibration domain of remote sensing models using convex hulls. *International Journal of Applied Earth Observation and Geoinformation*, *112*, 102939.

Rüttenauer, T. (2022). Spatial regression models: a systematic comparison of different model specifications using Monte Carlo experiments. *Sociological Methods & Research*, *51*(2), 728-759.

Saarela, S., Holm, S., Grafström, A., Schnell, S., Næsset, E., Gregoire, T. G., ... & Ståhl, G. (2016). Hierarchical model-based inference for forest inventory utilizing three sources of information. *Annals of Forest Science*, *73*(4), 895-910.

Saarela, S., Wästlund, A., Holmström, E., Mensah, A. A., Holm, S., Nilsson, M., ... & Ståhl, G. (2020). Mapping aboveground biomass and its prediction uncertainty using LiDAR and field data, accounting for tree-level allometric and LiDAR model errors. *Forest Ecosystems*, *7*, 1-17.

Saarela, S., Healey, S. P., Yang, Z., Roald, B. E., Patterson, P. L., Gobakken, T., ... & Ståhl, G. (2025). A Separable Bootstrap Variance Estimation Algorithm for Hierarchical Model-Based Inference of Forest Aboveground Biomass Using Data from NASA's GEDI and Landsat Missions. *Environmetrics*, *36*(1), e2883.

Shao, J. (2003). Impact of the bootstrap on sample surveys. *Statistical Science*, *18*(2), 191-198.

Schlamadinger, B., Bird, N., Johns, T., Brown, S., Canadell, J., Ciccarese, L., ... & Yamagata, Y. (2007). A synopsis of land use, land-use change and forestry (LULUCF) under the Kyoto Protocol and Marrakech Accords. *Environmental Science & Policy*, *10*(4), 271-282.

Ståhl, G., Gobakken, T., Saarela, S., Persson, H., Ekström, M., Healey, S. P., ... & McRoberts, R. E. (2024). Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time–And how this affects applications. *Forest Ecosystems*, 100164.

Trumbore, S., Brando, P., & Hartmann, H. (2015). Forest health and global change. *Science*, *349*(6250), 814-818.

Zhang, G., Ganguly, S., Nemani, R. R., White, M. A., Milesi, C., Hashimoto, H., ... & Myneni, R. B. (2014). Estimation of forest aboveground biomass in California using canopy height and leaf area index estimated from satellite data. *Remote Sensing of Environment*, *151*, 44-56.

Acta Universitatis Agriculturae Sueciae

Doctoral Thesis No. 2025:40

This thesis presents new methods to estimate plant abundance from presence/absence data assuming different types of spatial point processes for modelling the plant locations. Model-based and hybrid inference frameworks are applied. In addition, variance estimates are provided, and a broadened analysis of uncertainty is performed in a model-based inference context.

Léna Gozé received xer doctoral education at the Department of Forest Resource Management at the Swedish University of Agricultural Sciences (SLU), Umeå. In 2020, xe was awarded a Master of Science in Statistics from the University of Lille, France.

Acta Universitatis Agriculturae Sueciae presents doctoral theses from the Swedish University of Agricultural Sciences (SLU).

SLU generates knowledge for the sustainable use of biological natural resources. Research, education, extension, as well as environmental monitoring and assessment are used to achieve this goal.

ISSN 1652-6880
ISBN (print version) 978-91-8046-475-8
ISBN (electronic version) 978-91-8046-525-0