DISSERTATION

A COLLECTION OF STATISTICAL METHODS FOR APPLICATIONS IN TREE DEMOGRAPHY

Submitted by

Lane T. Drew

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2025

Doctoral Committee:

Advisor: Andee Kaplan

Yawen Guan
Matthew D. Koslovsky
Sarah Hart

ABSTRACT

A COLLECTION OF STATISTICAL METHODS FOR APPLICATIONS IN TREE DEMOGRAPHY

Ecological data are often characterized by complex spatial and temporal patterns, which can be difficult to analyze and interpret. In the context of tree demography, the study of forest dynamics at both the individual and stand level is crucial for understanding forest ecosystems and their response to environmental changes. In this dissertation, we present a collection of statistical methods and practical tools for addressing ecological questions, with a particular emphasis on tree demography. We introduce two novel modeling frameworks for identifying unique trees across multiple aerial scans and estimating individual tree growth as a function of spatial and spatio-temporal covariates, and a framework for estimating and simulating from location-dependent marked spatial point processes.

We first present a two-stage Bayesian spatial record linkage approach designed to identify unique trees across bi-temporal light detection and ranging (LiDAR) scans. This approach generalizes the linkage-averaging (LA) approach for record linkage to meaningfully propagate uncertainty from the linkage into a generic auxiliary data downstream modeling task. We introduce a novel approximate sampling scheme for the linkage that leverages the spatial structure of the data to achieve a degree of scalability in the model that allows it to be applied to spatial domain sizes that were previously too large to study. We apply this modeling framework with a generalized Michaelis–Menten style growth function to investigate the impact of key water and energy availability proxies on individual tree growth in a spruce-fir forest located on Snodgrass Mountain in the Gunnison National Forest of the Southern Rocky Mountains of Colorado, USA. We also provide a comprehensive set of numerical experiments on simulated data that mimics the characteristics of the empirical data from the case study.

We then present the **ldmppr** R package, which enables the efficient estimation, evaluation, and simulation of location-dependent marked spatial point processes characterized by regularity given a reference pattern and location-specific covariate surfaces. Originally motivated by our need to

simulate biologically realistic point patterns for our work in the previous chapter, this work addresses the need for a suitably flexible off-the-shelf method for working with location-dependent covariates in the mark distribution of a spatial point process. For example, we might consider the size of trees within a forest as a spatial point process where the size of a tree is a function of location-dependent covariates such as elevation, soil wetness measured by a topographic wetness index, and the aspect of the slope. We provide a detailed discussion of our modeling approach and the workflow for using the package, including a case study motivated by the empirical data from the previous chapter.

Finally, we present an alternative to the two-stage modeling framework introduced in the first chapter by utilizing a joint modeling approach for temporal record linkage to simultaneously identify unique individuals across multiple aerial scans and estimate individual tree growth as a function of spatial and spatio-temporal covariates. We develop a Bayesian hierarchical model that provides exact uncertainty quantification across both the linkage and the downstream modeling task concurrently. We incorporate a mechanistic downstream growth model based on an ordinary differential equation (ODE) that approximates the underlying growth process. We consider two alternative joint model formulations, and contrast the performance of the two approaches in a series of numerical experiments considering a wide range of observable scenarios as motivated by a dataset comprised of multi-temporal aerial scans of a spruce-fir forest near the Gothic Townsite located outside of Crested Butte, Colorado.

We conclude with a summary of the primary contributions of the work presented in this dissertation and the impact on the fields of tree demography, record linkage, and spatial point processes.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to my advisor, Dr. Andee Kaplan, for her unwavering support, guidance, and encouragement throughout my doctoral journey. Her patience and insight have been invaluable, and I am deeply appreciative of the opportunities she has provided me to grow both personally and professionally. I would also like to thank our primary collaborator, Dr. Ian Breckheimer, for sharing his time and expertise with me over the last several years, and for furnishing a rich collection of ecological challenges that have motivated much of the work presented in this dissertation. Additionally, I am grateful to my committee members, Dr. Yawen Guan, Dr. Matthew Koslovsky, and Dr. Sarah Hart, for their thoughtful feedback and support throughout this process. It has been a rare privilege to work with such a talented and dedicated group of scholars.

I would also like to acknowledge the support of my colleagues. Their wit and wisdom have made this journey all the more enjoyable, and I am thankful for the friendships that have blossomed along the way. They have enriched my life in countless ways, and I am grateful for the camaraderie we have shared.

Finally, I would like to thank my family and friends for their unwavering support and encouragement. My accomplishments are a testament to their love and belief in me, and I am eternally grateful for their presence in my life. The completion of this work represents the conclusion of one chapter (or several) of my life, and the beginning of the next adventure. In the words of Reginald Barclay, "Let's raise our glasses to the journey."

TABLE OF CONTENTS

Chapter 1

Introduction

In a world characterized by rapidly changing landscapes, the ability to efficiently analyze complex ecological data with robust uncertainty quantification is of paramount importance. In this dissertation, we present a collection of statistical methods and practical tools for addressing ecological questions, with a particular emphasis on tree demography. At the core, our work is motivated by the need to understand forest dynamics and the impact of environmental factors on tree growth at spatial scales that have previously been difficult to evaluate. The structure of this dissertation is outlined as follows.

In Chapter 2, we introduce a novel and highly scalable two-stage Bayesian spatial record linkage approach designed to identify unique trees across multiple light detection and ranging (LiDAR) scans that ensures proper posterior inference in the downstream modeling task. We provide details on the theoretical justifications for our approach and for the practical application of this modeling framework utilizing a general downstream task that depends on linking records across time points. We discuss the computational strategies that we employ to achieve scalability in the model, including the introduction of an approximate sampling scheme for the linkage that leverages the spatial structure of the data. We demonstrate the efficacy of the model through a case study considering individual tree growth estimation in a spruce-fir forest located on Snodgrass Mountain in the Gunnison National Forest of the Southern Rocky Mountains of Colorado, USA. We also provide a comprehensive set of numerical experiments on simulated data that mimic the characteristics of the empirical data from the case study.

In Chapter 3, we present the **ldmppr** R package (Drew and Kaplan, 2025), which enables the efficient estimation, evaluation, and simulation of location-dependent marked spatial point processes characterized by regularity in the point pattern given a reference pattern and location-specific covariate surfaces. Originally motivated by our need to simulate biologically realistic point patterns for our work in Chapter 2, this work addresses the need for a suitably flexible off-the-shelf method for investigating the impact of location-dependent covariates on the mark distribution of a spatial point process. For example, we might consider the size of trees within a forest as a spatial point

process where the size of a tree is a function of location-dependent covariates such as elevation, soil wetness measured by a topographic wetness index, and the aspect of the slope. We provide a detailed discussion of our modeling approach and the workflow for using the package, including a case study motivated by the empirical data from Chapter 2.

In Chapter 4, we explore an alternative to the two-stage modeling framework presented in Chapter 2 by utilizing a joint modeling approach to simultaneously identify unique individuals across multiple aerial scans and estimate individual tree growth as a function of spatial and spatio-temporal covariates. We develop a Bayesian hierarchical model that provides robust uncertainty quantification for the linkage and the downstream modeling task of interest simultaneously. We consider two alternative joint model formulations, and contrast the performance of the two approaches in a series of numerical experiments considering a wide range of observable scenarios as motivated by a dataset comprised of multi-temporal aerial scans of a spruce-fir forest located outside of Crested Butte, Colorado.

In Chapter 5, we summarize the primary contributions of the work presented in this dissertation and the impact on the fields of tree demography, record linkage, and spatial point processes. We also include a discussion of our planned and potential directions for future work that build upon the methods and tools introduced in this dissertation.

In the remainder of this initial chapter, we provide a brief overview of the fields of tree demography, record linkage, and spatial point processes with a focus on the relevant literature that motivates our work in the subsequent chapters. We first review key ecological concepts from tree demography, highlighting current methodological challenges that motivate our contributions in the remainder of this dissertation.

1.1 Tree demography

Tree demography is a subfield of ecology that focuses on the study of tree populations, their growth, survival, and reproduction over time. It encompasses the analysis of individual tree characteristics, population dynamics, and the influence of environmental factors on tree health and behavior. The study of tree demography is crucial for understanding forest ecosystems, biodiversity, and the impacts of climate change on tree species (Chave et al., 2005).

Research in tree demography is often a contrast between considering forest dynamics at the stand level, which involves investigating an aggregated population, and the health and behavior of individual trees within a forest. Historically, tree demography has relied on field surveys and measurements of individual trees, which can be time-consuming and logistically challenging, especially in large or remote forested areas (Saatchi et al., 2011; Wensel et al., 1987). Advances in remote sensing technology, particularly light detection and ranging (LiDAR) and structure-from-motion photogrammetry analysis utilizing high-resolution images from uncrewed aircraft systems (UAS, e.g., drones), have revolutionized the field by enabling the collection of both high-volume and high-resolution spatial data on tree structure and canopy characteristics (Hyyppä et al., 2008; Lefsky et al., 2002).

While these data collection methodologies have improved the ability to assess forest structure and dynamics at larger scales, they also introduce challenges related to data processing, segmentation, and the identification of individual trees across multiple time points (Koch et al., 2006). The segmentation of tree crowns from LiDAR data, for example, can be complex due to overlapping canopies, varying tree heights, and the presence of understory vegetation (Kaartinen et al., 2012). Additionally, the matching of individual trees across different scans or time points requires robust methods to provide uncertainty quantification when utilizing this data for modeling. Historical approaches to tree demography have often relied on heuristic methods for identifying unique individuals, or manual verification of tree matches, which can be labor-intensive and prone to error without a systematic approach to uncertainty quantification or propagation into the subsequent modeling tasks as in Ma et al. (2018). The disconnect in the data processing to modeling pipeline demonstrates a clear need for methodologies that can address these shortcomings and provide a more robust framework for tree demography research. The need to identify unique individuals across multiple time points over a broad spatial domain in the presence of varying sources of noise in the data made this application a compelling candidate for the development of a new modeling framework utilizing approaches from the record linkage and spatial point process literature that we introduce in the subsequent sections of this chapter.

1.2 Record linkage

The field of record linkage is primarily concerned with the process of identifying and linking records that refer to the same unique entity across different datasets (i.e., files) or within a single dataset in the absence of a unique identifying attribute and in the presence of noise (Gu et al., 2003; Winkler, 2006). Record linkage methods provide a probabilistic framework for matching potentially noisy records based on observed attributes, allowing for the incorporation of uncertainty in the matching process. Throughout this dissertation, we use the term record linkage to encompass both the identification of coreferent records within and between files (also commonly referred to as deduplication and entity resolution respectively). Examples of these applications include identifying unique patient records in healthcare systems (Padmanabhan et al., 2019), or unique individuals across repeated surveys over time (Steorts, 2015).

The earliest approaches to probabilistic record linkage were based on a model-based matching framework formalized by Fellegi and Sunter (1969), which performed matching between pairs of records directly according to a decision theoretic framework. While approaches that perform matching between pairs of records directly have remained popular (Sadinle, 2017), there has been a notable surge in the development of models built upon latent clustering structures known as latent entity models (Steorts et al., 2016; Liseo and Tancredi, 2011a). Latent entity models relate records indirectly through unobserved "true" entities, which are assumed to be the same across files. These models allow for simultaneous estimation of population size and the true latent field values associated with a record, and naturally fit within Bayesian hierarchical frameworks, whether in a two-stage or joint modeling approach.

We draw a distinction here between two-stage and joint modeling approaches that incorporate a record linkage component. The two record linkage modeling approaches differ in their approach to modularization. Bayarri et al. (2009) discuss the consequences of modularization in Bayesian hierarchical models, such that a modularized model is one that can be decomposed into two or more components that can be fit independently of one another. The rationale behind modularization is to allow for the independent development and validation of each component, which can be particularly useful in complex models where different components may require different modeling approaches or computational methods. This also prevents potentially malignant components from infecting

the entire model. Separating the individual components can prevent negative feedback loops from occurring between various components of the model, which can lead to overfitting or other issues in the model fitting process. However, modularization can also lead to a loss of information and may not fully capture the interdependencies between the components of the model, which can be particularly important in complex models where the components are highly interrelated.

In the context of record linkage, a two-stage approach performs the record linkage independently of the downstream modeling task, and the output from the record linkage model is used as input to the downstream task (Lahiri and Larsen, 2005; Kim and Chambers, 2012). This approach allows a high degree of flexibility in the choice of downstream task, as the record linkage model may be used as input to a variety of models. However, the linkage is crucially uninformed by any specific downstream task, which may constitute a meaningful loss of information in that process (Goldstein et al., 2012; Chambers et al., 2019). In contrast, a joint modeling approach allows for feedback between the linkage and downstream task, and provides the opportunity for direct uncertainty quantification across the entire modeling pipeline. However, the joint modeling approach is generally more complex and may require more sophisticated computational methods to fit the model. As a result, joint models can be difficult to generalize and have often been limited to use in scenarios where only two files are present (Gutman et al., 2013; Hof et al., 2017; Steorts et al., 2018). They may also suffer markedly from misspecification or poor fit in the downstream model, which two-stage models are insulated from as a result of modularization. A visual depiction of the difference between these two approaches is provided in Figure 1.1.

In Chapter 2, we present a two-stage modeling approach that utilizes a record linkage model to identify unique individuals across bi-temporal LiDAR scans paired with an individual tree growth model with robust uncertainty propagation through a generalization of the linkage-averaging (LA) approach of Sadinle (2018). In Chapter 4, we explore an alternative joint modeling approach that simultaneously identifies unique individuals across multiple aerial scans and estimates individual tree growth using a reframed version of the growth model introduced in Chapter 2 as a mechanistic model based on an ordinary differential equation. To facilitate the investigation and validation of both of these approaches, we rely on simulated data that mimic the characteristics of the empirical

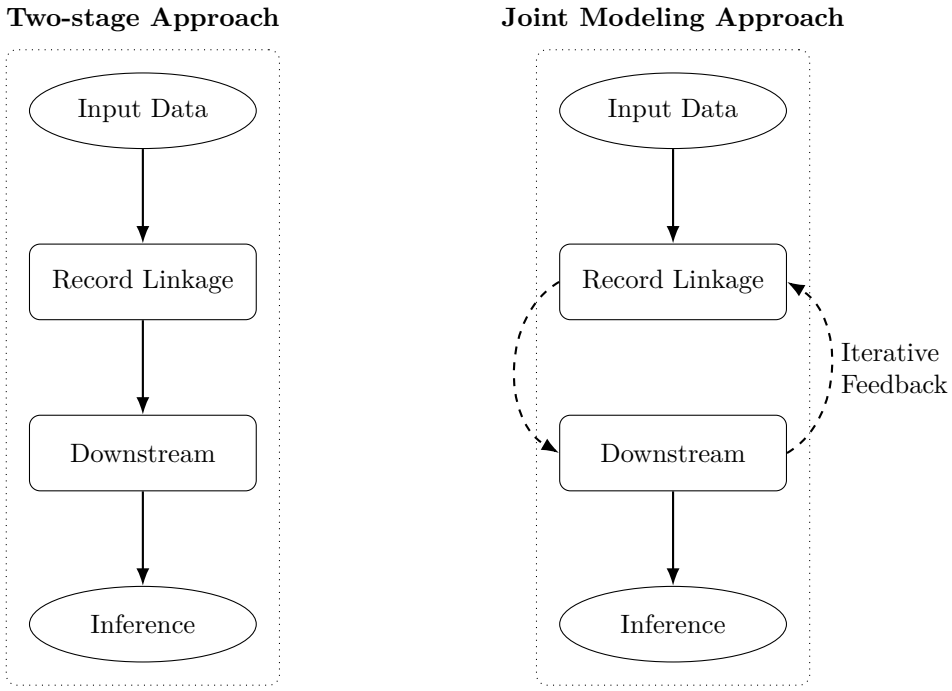

Figure 1.1: Comparison of the two-stage versus joint modeling approaches for incorporating record linkage into a modeling pipeline. The two-stage approach performs record linkage and the downstream modeling task sequentially, while the joint approach performs both tasks simultaneously with iterative feedback between the two components.

data from Chapter 2. We accomplish this by extending the available methods from the spatial point process literature to incorporate location-dependent marks, which we discuss in the next section.

1.3 Spatial point processes

Spatial point processes are a class of stochastic processes that model the spatial distribution of points in a given area or volume (Diggle, 2013). There is a rich literature on spatial point processes and their application to a wide variety of phenomena that can be modeled as a collection of points in space. The spatial distribution of these points is typically characterized by a spatial intensity function, which describes the expected number of points per unit area at each location in the space and is often denoted by $\lambda(\boldsymbol{x})$, where $\boldsymbol{x}$ is a point in the space of interest. The simplest spatial point processes are described as homogeneous such that the location of any individual point in the process is independent of the location of any other point in the process, and the intensity function is constant across the space.

The next layer of complexity in these processes is to incorporate spatially varying intensity functions, which allow for the modeling of heterogeneous point patterns, in addition to interpoint dependence. In point patterns with dependence between the points, they are commonly characterized by either clustering or inhibition (Baddeley et al., 2007). Clustering refers to the tendency of points to be located close together, while inhibition refers to the tendency of points to be regularly spaced.

Beyond characterizing the spatial distribution of points in a spatial point process, the natural extension of these models is to consider marks, which are the additional characteristics associated with a point in the process, such as the size or species of a tree. Marked spatial point processes are generally specified in terms of a spatial process for the locations and a mark process for the marks associated with each point in the process (Illian et al., 2008). However, due to computational constraints it is often assumed that the marks are independent of the underlying spatial process (Møller et al., 2016). This assumption is often violated in practice, as the marks of a point in a spatial point process may depend on the location of the points in the process. For example, the size of a tree within a forest is likely to depend on the location of the tree, as environmental and topographic factors influence the growth process. See Figure 1.2 for an example of a marked spatial point process with regularity in which the marks depend on a topographic covariate surface which changes over space.

In the context of tree demography, spatial point processes provide a powerful framework for modeling the distribution of trees within a forest stand. While the investigation of the dynamics of a point process may be of interest on its own, once a process has been characterized and estimated, it can also be used to simulate realizations from the process. Simulation is an incredibly powerful tool for validating the fit of point process models, and can be valuable for generating synthetic data for testing and validating models that depend on an underlying spatial process (Møller and Waagepetersen, 2003).

When considering application of the two-stage modeling framework presented in Chapter 2 to tree demography, we discovered the limited availability of methods for simulating marked spatial point processes with location-dependent marks that are characterized by regularity in the point pattern. To address this gap, we developed the **ldmppr** R package, which provides a flexible framework for estimating, validating, and simulating from marked spatial point processes with location-dependent

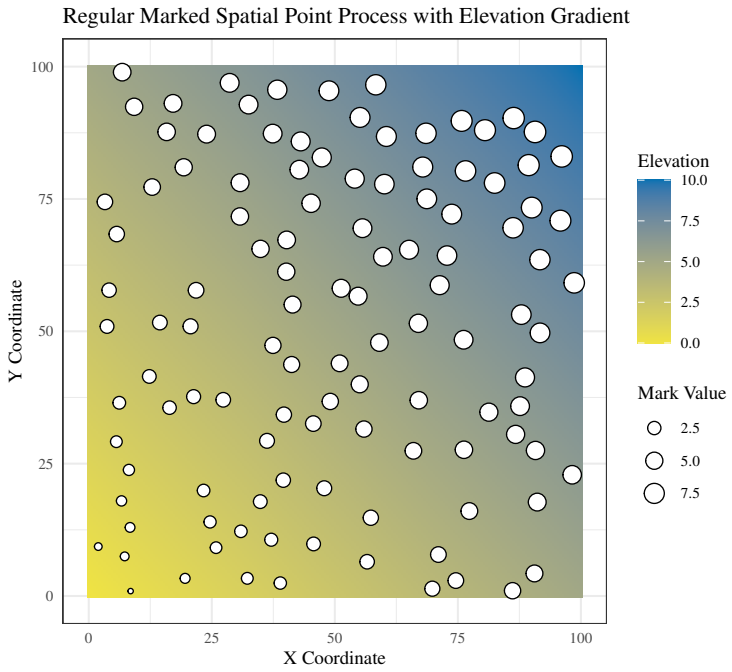

Figure 1.2: Example of a marked spatial point process with regularity in the point pattern in which the values of the marks depend on the elevation surface.

marks. We discuss the machinery underlying our modeling approach and its implementation in **ldmppr** in Chapter 3.

Chapter 2

A Bayesian Record Linkage Approach to Applications in Tree Demography Using Overlapping LiDAR Scans

2.1 Introduction

The characterization and quantification of forest dynamics have been areas of interest for ecologists for more than a century and have become increasingly important metrics for understanding the effects of climate change (Hyyppä et al. (2008)). Historical investigations of forest dynamics have relied on field surveys over limited spatial domains, which are generally time consuming and potentially difficult to perform (Saatchi et al. (2011); Wensel et al. (1987)). The advent and ongoing refinement of aerial laser scanning (ALS) technology has ushered in a new age of data collection in terms of scalability. The use of ALS data in the modeling of forest structure has become a standard approach, as it enables researchers to examine the health and behavior of forests at larger scales than has previously been possible by field survey (Babcock et al. (2016); Dalponte and Coomes (2016)). The obvious extension of these efforts is to functions that rely on repeat measurements over time. Despite improvements in the accuracy of ALS technology, there remains inherent uncertainty in both the scanning mechanism and subsequent post-processing of the data, as discussed by Huo and Lindberg (2020). In the existing literature, the mechanisms employed for identifying the unique individuals from scans across multiple time points are largely heuristic and rely on manual verification, as in Ma et al. (2018), and employ a two-stage modeling schema which fails to incorporate the uncertainty in the segmentation and matching procedures into the downstream task. To address these issues, we present an alternative two-stage framework utilizing a record linkage approach for spatial location data that is capable of efficiently identifying unique individuals across larger spatial domains while providing robust uncertainty quantification for the linkage that may then be propagated into the downstream modeling objective.

As our ability to collect and store data has exploded, so too has our need to engage in record linkage (also called deduplication or entity resolution). At its core, the field of record linkage is concerned with the resolution of unique records across overlapping files in the absence of a unique

identifier. In this chapter we use the term record linkage to encompass the process of identifying coreferent records both between and within files. Historically, files have represented repeated surveys over time (Steorts (2015)) or non-temporally linked overlapping databases, such as patient records, across different providers in the healthcare system (Padmanabhan et al. (2019)). The earliest approaches to probabilistic record linkage, as formalized by Fellegi and Sunter (1969), performed probabilistic matching between pairs of records according to a decision-theoretic framework. The field has developed consistently since its inception, and advances in Bayesian computational methods have given rise to a new class of probabilistic modeling approaches over the last 20 years, as discussed by Liseo and Tancredi (2011b). While methods that perform matching between pairs of records directly have remained useful (Sadinle (2017)), models built upon latent clustering structures have become increasingly popular alongside methods addressing alternative types of data (Steorts et al. (2016); Liseo and Tancredi (2011a)). Recently, record linkage has been successfully used to improve wildlife population inference from a series of sequential aerial photographs (Lu et al. (2022)). In a similar vein, we introduce a record linkage model for bi-temporal spatial location data derived from light detection and ranging (LiDAR) scans intended to improve tree demography inference.

While record linkage is an interesting and challenging endeavor on its own, it generally functions as the first step in the sequence of a statistical pipeline, as discussed by Kaplan et al. (2022). We implement our spatial record linkage model in a two-stage framework in which the record linkage and downstream task are performed sequentially. Our proposed approach performs the downstream modeling task using a randomly sampled subset of iterations from the posterior linkage structure, which propagates the uncertainty from the linkage into the second stage of the pipeline according to the linkage-averaging (LA) approach of Sadinle (2018). Crucially, in the LA framework, the linkage is not informed by the downstream task. Consequently, the output from the first stage record linkage model may be used as the input for a variety of downstream models, offering researchers a high degree of flexibility when adopting this modeling framework.

In our application, we pair the spatial record linkage model with an individual tree growth model, where we define growth as the change in canopy volume between surveys on an annual scale. The empirical dataset is comprised of LiDAR scans of Gunnison National Forest from 2015 and 2019, which were provided by the Rocky Mountain Biological Laboratory (RMBL). The individual

tree crown polygons and associated attributes were obtained from a 1/3 m resolution canopy height model using the `ITCsegment` (Dalponte and Coomes (2016)) algorithm in the **lidR** (version 3.4, Roussel et al. (2020)) package in R. We note that the record linkage and growth models are both designed to account for various sources of biological variation and measurement error in the data collection and post-processing procedures.

The remaining structure of the paper is as follows. In Section 2.2, we highlight the ecological hypotheses and empirical data that motivate our novel modeling approach. In Section 2.3, we introduce the relevant notation for the spatial record linkage model and downstream growth model. Additionally, we outline the computational strategies used to fit the model. In Section 2.4, we provide a discussion of the theoretical justification for the LA approach in a general auxiliary data task setting. In Section 2.5, we provide an analysis of the empirical data, and in Section 2.6, we examine the performance of the proposed modeling approach in a series of numerical experiments on simulated data. We conclude with a discussion and directions for future work in Section 2.7.

2.2 LiDAR derived individual tree characteristics from bi-temporal scans of Snodgrass Mountain

We apply our two-stage modeling approach to identify unique trees across time points and to estimate spatial patterns of tree growth as related to certain environmental drivers in a spruce-fir forest site located in the Southern Rocky Mountains (USA). The study site is a two square kilometer forested domain located on Snodgrass Mountain near the site of RMBL in the vicinity of Crested Butte, Colorado. The domain spans montane to lower subalpine mountain slopes at elevations from 2891-3395m and experiences a cold continental climate with persistent seasonal snowpack accounting for the majority of annual precipitation (Carroll et al. (2020)). Evergreen forests in the domain are dominated by Engelmann spruce (*Picea engelmannii*) and subalpine fir (*Abies lasiocarpa*), which account for more than 80% of the tree canopy. These forests also contain scattered lodgepole pine (*Pinus contorta*) and Rocky Mountain Douglas-fir (*Pseudotsuga menziesii subsp. glauca*). Deciduous quaking aspen (*Populus tremuloides*) forms large single-species stands on lower slopes of the study area, but these areas were excluded from the analysis due to the difficulty of assessing the growth of this species.

Forest structural data were collected for the study area via LiDAR in two intervals during 2015 and 2019. Both scans were collected from an airplane-based sensor in late summer before the drop of deciduous leaves. The laser scanner records discrete peaks of reflected energy at near-infrared wavelengths and uses the integrated Real-Time Kinematic (RTK) sensor position and estimated time-of-flight of laser pulses to locate laser reflections ("returns") in geographic space. The scanning process yields a dense (8-16 pts / m^2) cloud of 3-dimensionally located returns representing reflections from the ground, tree canopies, and other reflective surfaces. Details of each LiDAR dataset are provided in Table 2.1.

Table 2.1: LiDAR data collection attributes for the 2015 and 2019 scans provided by RMBL.

Scan Attribute	2015 Scan	2019 Scan
Acquisition Dates	August 7, 2015 and August 10, 2015	August 21 – September 24, 2019
Aircraft Used	Piper Navajo	Cessna Caravan
Sensor	Riegl (Leica) Q1560	Riegl (Leica) VQ1560i
Maximum Returns / Pulse	5	15
Target Pulse Density	Average 8 pulses/m^2	Average 2 pulses/m^2
Realized Point Density	9.4 points/m^2	9.4 points/m^2
Survey Altitude (AGL)	550 m	1159 m
Field of View	58°	58.5°

The LiDAR-derived point clouds were segmented and summarized to yield estimates of per-tree structural characteristics including tree top locations, maximum heights, and canopy volumes using functions in the R package **lidR** (version 3.4, Roussel et al. (2020)). Although numerous approaches exist to segment individual trees in LiDAR data (see Aubry-Kientz et al. (2019) for a recent comparison), for this analysis we adopted the commonly-used `ITCsegment` algorithm (Dalponte and Coomes (2016)). `ITCsegment` is a region-growing approach which iteratively incorporates points into candidate tree canopies starting at a set of seed locations. Seed locations (putative tree tops) were selected using a local maximum filter, identifying laser returns with high heights relative to a height-dependent local neighborhood. Canopy volumes were calculated by summing canopy heights for each segment using a 1 m resolution canopy surface model generated using the `pit-free` algorithm implemented in **lidR**. The segmentation and canopy surface model generation steps generate imperfect representations of individual tree locations and crown geometries. Errors in sensor geo-positioning and ranging measurements can lead to systematic spatial shifts in datasets collected at different times. Moreover, the location and trajectory of individual laser pulses differ

between scans, leading to small amounts of variability in estimated maximum heights and crown locations, as discussed by Poorazimy et al. (2022).

In Western North American conifer forests, tree growth is thought to be constrained by the (potentially interacting) availability of water and energy (Buechling et al. (2017); Heilman et al. (2022)). To investigate these constraints, we assembled estimates of key water and energy proxies across the domain from diverse remote sensing datasets. A 1 m resolution LiDAR-derived Digital Elevation Model (Goulden et al. (2020)) was used to compute topographic aspect "folded" about the north-south axis to distinguish high solar-radiation south-facing slopes from low solar-radiation north-facing slopes. We also computed a topographic wetness index (TWI) (Nobre et al. (2011)) as a water availability proxy. We augmented these topographic proxies with gridded climate data interpolated from weather station and microclimate sensors as well as satellite-derived maps of the persistence of seasonal snowpack (Breckheimer (2023)). Data for snowpack persistence and growing degree days was aggregated annually from 2015 to 2019, and the median observed values were used during modeling to capture the relative impact of these covariates over the period between LiDAR scans. Details pertaining to the covariates may be found in Table 2.2 along with a visualization of the derived tree geometries and raster images in Figure 2.1.

Table 2.2: Designations and details for the topographic covariates included in the analysis.

Growth Constraint	Covariate	Data Source	Resolution (m)
Energy	Folded Aspect	Goulden et al. (2020)	1
Energy	Growing Degree Days	Breckheimer (2023)	30
Water	HAS Wetness Index	Goulden et al. (2020); Nobre et al. (2011)	1
Water	Snowpack Persistence	Breckheimer (2023)	30

2.3 Models and notation

In this section we detail the spatial record linkage and growth models employed in our two-stage LA approach. We first define the necessary notation and present the proposed spatial record linkage model before developing the downstream individual tree growth model. We note that throughout this section, distributions identified with subscripts refer to truncated distributions over the specified bounds. For example, a truncated Inverse-Gamma distribution with parameters c and d over the range $[0, b]$ is denoted as Inverse-Gamma$_{[0,b]}(c, d)$. We finish this section with a discussion of the

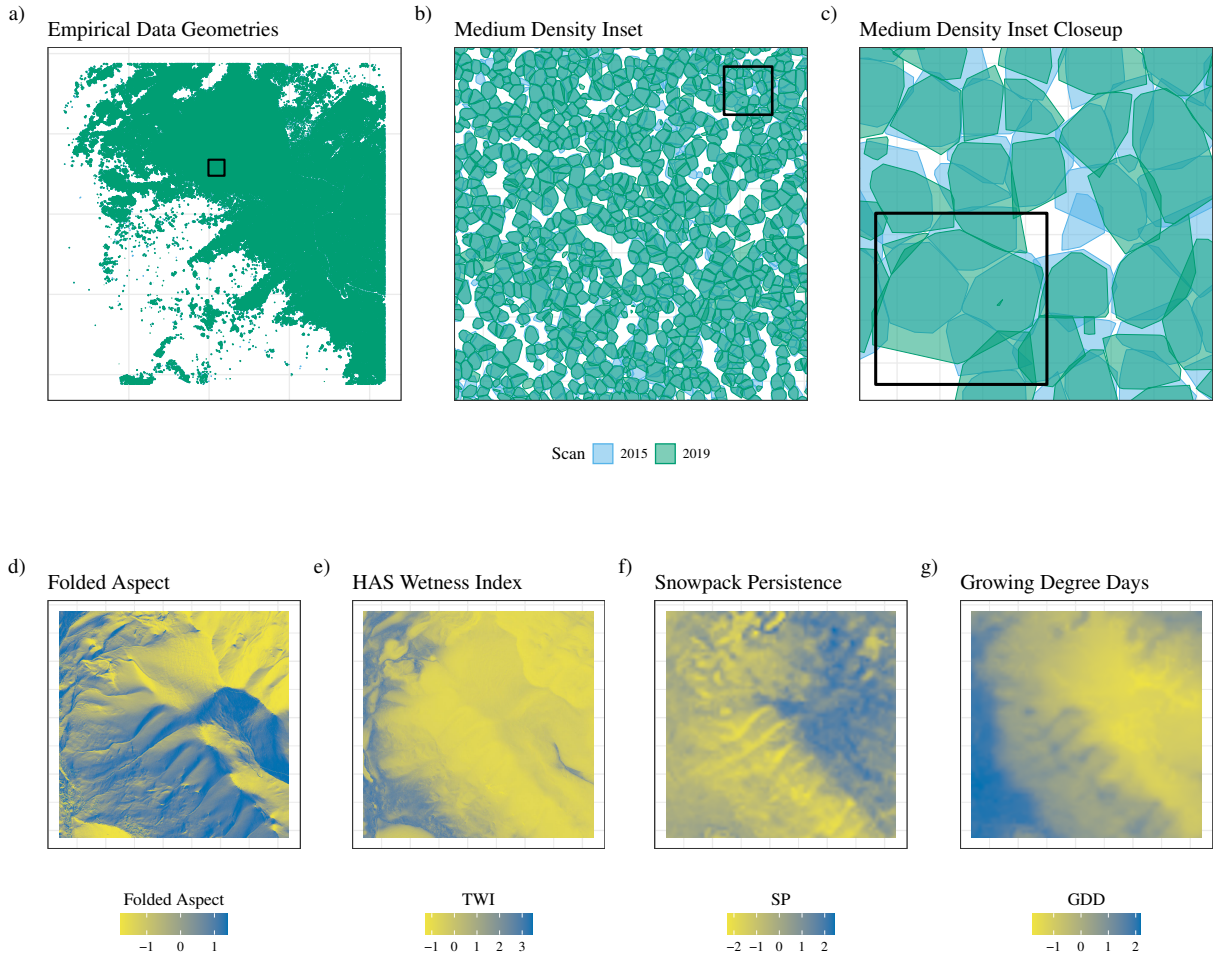

Figure 2.1: Plots (a) and (b) show the derived crown geometries from the 2015 and 2019 LiDAR scans performed by RMBL for the full datasets and a medium density inset (outlined in plot (a)). Plot (c) shows a closeup from the inset in (b), highlighting an instance in which multiple trees in the first file overlap with a single tree in the second file. Plots (d)–(g) show the scaled raster images for the topographic covariates of interest over the study domain.

computational strategies employed to facilitate scalability of the record linkage model to spatial domain sizes that are of practical interest.

2.3.1 Record linkage model

We begin by presenting the spatial record linkage model as a standalone component to introduce the model structure and to establish a baseline for inclusion in modeling pipelines with alternative downstream tasks. We provide a general model capable of handling two files, which is also capable of performing deduplication within files. We employ a Bayesian hierarchical structure based on latent

matching, as discussed by Steorts et al. (2016) and Liseo and Tancredi (2011a), such that records are linked to unobserved latent entities with true field values instead of being probabilistically matched to other records directly through comparison vectors, as in the work of Fellegi and Sunter (1969) and Sadinle (2017). In the spatial record linkage model, the value associated with the latent entity is the true unobserved location of the individual. We treat the observed data (i.e., location) as a noisy observation of the latent location and include provisions for different potential sources of noise. We consider error introduced as a function of translation and rotation in the data collection process, post-processing of the data, and due to biological mechanisms. We specify the data model and relevant notation as follows.

The model is constructed to handle two files, where the files are indexed by $i = 1, 2$ with relative size n_i for each file. The records within files are indexed from $j = 1, \ldots, n_i$ such that the total number of observed records is $n = \sum_{i=1}^{2} n_i$. We denote the observed location data for the jth record in file i as $\boldsymbol{y}_{ij}$, where $\boldsymbol{y}_{ij}$ is a numerical vector of length 2, that is, $\boldsymbol{y}_{ij} = (x, y)_{ij}$, corresponding to the spatial coordinates of the record in the (x, y)-plane. We note that in our application the term file is synonymous with a LiDAR scan and record with an individual identified tree such that n is the total number of individual trees detected across all scans.

We define the latent location vector as $\boldsymbol{s}_{j'}$, where $j' = 1, \ldots, N$ such that N is the maximum number of unique latent individuals in the population under consideration. The observed locations $\boldsymbol{y}_{ij}$ are modeled as noisy versions of the latent locations $\boldsymbol{s}_{j'}$. We assume that the record set $\boldsymbol{y}_{1j}$ exists in the same space as the latent locations, while the record set $\boldsymbol{y}_{2j}$ is modeled as a transformed version of the associated $\boldsymbol{s}_{j'}$, where we restrict the possible transformations considered to rotation and translation.

The linkage structure, which identifies the relationship between the observed records and the latents, is a vector of length n denoted as $\boldsymbol{\Lambda} = \{\lambda_{ij} : i = 1, 2; j = 1, \ldots, n_i\}$, where λ_{ij} is an integer indicating which $\boldsymbol{s}_{j'}$ the jth record in file i refers to. The linkage is implicitly dependent on the maximum latent population size N, as $\lambda_{ij} \in \{1, \ldots, N\}$. The specified linkage structure naturally defines a set of N clusters denoted $\mathcal{C}(\boldsymbol{\Lambda})$, which specify the records that are linked to each $\boldsymbol{s}_{j'}$ such that $\mathcal{C}(\boldsymbol{\Lambda}) = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$. The individual clusters are defined as the sets $\mathcal{C}_{j'} = \{(i, j) : \lambda_{ij} = j'\}$ for $j' = 1, \ldots, N$, and we note that the clusters may be empty or may contain records from the same

file in addition to records across files highlighting the capacity of the model for performing record linkage and deduplication simultaneously. In the context of our application, duplicates within files potentially occur during the LiDAR processing due to the segmentation algorithm such that a single individual may be erroneously split into multiple entities, as seen in panel (c) of Figure 2.1.

Previous record linkage modeling approaches are known to be sensitive to the specification of N, which functions as a hyperparameter in the model and quantifies our belief regarding the upper bound on the number of unique entities across all files (Steorts et al. (2016)). We specify $N = q \times \max(n_i)$ where the scale factor q is chosen to reflect the assumed degree of overlap between the files and such that $N \leq n$. Alternatively, a practitioner could specify $N = n$ to avoid making any a priori assumptions about the number of unique individuals across files. Although not explicitly a parameter in the model, we do effectively obtain an estimate of the number of unique individuals across files which is often of interest in studies investigating species abundance and provides some sense of the effective sample size for estimation of the downstream model parameters. Additional discussion regarding the specification of N is provided in Appendix A.1.

The data model, which describes the relationship between the observed point patterns and $\boldsymbol{s}$, allows us to model the variation produced by the underlying biological process (assumed to be tree growth in this application) separately from the error introduced in the LiDAR scanning and post-processing procedures by incorporating the image alignment framework introduced by Green and Mardia (2006). The model is specified as follows:

$$\boldsymbol{y}_{ij} | \boldsymbol{s}_{\lambda_{ij}}, \sigma^2, \boldsymbol{t}_i, \theta_i, D \sim \text{Normal}_{2,[D]}\big(\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D, \sigma^2 \boldsymbol{I}\big),$$

for $i = 1, 2$. The rotation, $\boldsymbol{R}(\theta_i)$, is the standard counterclockwise rotation matrix given by

$$\boldsymbol{R}(\theta_i) = \begin{bmatrix} \cos(\theta_i) & -\sin(\theta_i) \\ \sin(\theta_i) & \cos(\theta_i) \end{bmatrix}$$

and $\boldsymbol{t}_i$ is the two dimensional translation vector. We allow the rotation angle and translation to vary across files (i.e., scans). The rotation for each file is around the midpoint, denoted $\boldsymbol{\mu}_D$, of the spatial domain of interest D. We also note that the record set $\boldsymbol{y}_{1j}$ can be expressed in terms of the

rotation and translation framework with fixed $\theta_1 = 0$ and $\boldsymbol{t}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\mathsf{T}$, assuming that the records in the first file exist in the same space as the latent locations, $\boldsymbol{s}$, to reduce the effective number of parameters in the model. We adopt this expression to simplify notation going forward. The full spatial record linkage model is specified as follows:

$$
\begin{aligned}
\boldsymbol{y}_{ij} | \boldsymbol{s}_{\lambda_{ij}}, \sigma^2, \theta_i, \boldsymbol{t}_i, D &\overset{\text{ind}}{\sim} \text{Normal}_{2,[D]}(\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D, \sigma^2 \boldsymbol{I}), \\
\boldsymbol{s}_{j'} | N &\overset{\text{iid}}{\sim} \text{Uniform}(D^*), \\
\sigma^2 &\sim \text{Inverse-Gamma}_{[0,b_\sigma]}(c_\sigma, d_\sigma), \\
\lambda_{ij} | N &\overset{\text{iid}}{\sim} \text{Uniform}\{1, \ldots, N\}, \\
\theta_i &\propto \exp(\kappa \cos(\nu) \cos(\theta_i) + \kappa \sin(\nu) \sin(\theta_i)) \mathbb{I}\{|\theta_i| < b_\theta\}, \\
\boldsymbol{t}_i &\sim \text{Normal}_2(\boldsymbol{0}, \sigma_t^2 \boldsymbol{I}),
\end{aligned}
$$

where the prior for θ_i, the rotation parameter for file i, is the kernel of a truncated von Mises distribution, as discussed by Green and Mardia (2006).

As mentioned above, we model the observed locations as noisy transformations of the unobserved true $\boldsymbol{s}_{j'}$ according to a Gaussian noise process. We specify the underlying latent point process, $\boldsymbol{s}$, to follow a uniform distribution over a slightly expanded spatial domain D^* such that $D \subseteq D^*$, to allow for the possibility that the true location of an individual is outside of the observed spatial domain (i.e., the tree base is located outside of D, but the observed tree crowns are inside). We assume the simplest and least informative prior specification for $\boldsymbol{s}$, corresponding to complete spatial randomness with a fixed number of points. This prior allows the observed locations to provide the bulk of the information in determining the distribution of the $\boldsymbol{s}_{j'}$ and enables this distribution to vary adequately over space. In practice, users could specify more complicated latent process models, which may fit the observed data more precisely but with additional assumptions, computational cost, and complexity (see Leininger (2014) for an in-depth discussion of point process models in a Bayesian hierarchical framework). We provide the derivations of the joint posterior and full conditional distributions for the model alongside the algorithm for fitting the model in Appendix A.1.

Many latent record linkage modeling approaches employ a hit-miss mechanism for whether the observed records are a noisy distortion of the true latent value in a given field as in Steorts et al. (2016). In contrast, our model treats every observed record as a noisy copy of the latent, as we are dealing with spatial locations over a continuous domain. In the context of our motivating application, the observed locations are the tree crowns while the latent location is the tree base. Thus, the assumption that every observed location is a noisy observation of the truth has a clear physical interpretation in this case as well. We place a conjugate truncated Inverse-Gamma distribution prior on the measurement error parameter σ^2, where the upper bound corresponds to the maximum displacement that would be considered plausible based on the biological mechanisms of tree growth and the calibration of the LiDAR scanning equipment.

We note that our record linkage inspired modeling approach could be interpreted as a micro-clustering approach such that cluster sizes remain small, even as the number of records grows. Betancourt et al. (2022b) introduced a class of random partition models with microclustering behavior to achieve a similar goal, that is, guaranteeing that clusters remain small by assuming exchangeable sequences of clusters instead of exchangeable data points. In contrast to this approach, we obtain microclustering behavior in our model through the specification of N and the prior on σ^2 by guaranteeing a maximum number of unique latent locations and distance that observed locations may be from their associated true latent location, despite the fact that the prior for the linkage does not explicitly guarantee this property.

2.3.2 Downstream growth model

We now turn to the individual tree growth model we employ in this application. The model leverages the known allometric relationship between size and growth by using a flexible nonlinear function of the generalized Michaelis–Menten type to describe the annual individual growth-size curve while allowing for measurement error in the observed growth (measured by the change in canopy volume). Michaelis–Menten functions have been applied broadly across biological and ecological growth models, as described by López et al. (2000) and Bolker (2008). The generalized Michaelis–Menten function can take on a range of shapes from a logistic to a sigmoidal curve depending on the parameterization, where our specification may be seen in equation (2.1) below. The relative simplicity and flexibility of the function class combined with the clear biological

interpretations of the parameters make this model a compelling choice. Michaelis–Menten growth models have been found to be ideal for describing the relationship between diameter at breast height (DBH) and height for trees (Barbosa et al. (2019); Brahma et al. (2017)), which is analogous to the size-growth relationship between canopy volume and tree growth in our application. Although literature employing a Michaelis–Menten function in a measurement error model is scarce, the extension is natural and straightforward.

Our specification of the growth function incorporates topographically derived covariates, discussed in Section 2.2, allowing us to better understand the impact environmental drivers have on the growth of conifer species. We introduce relevant notation for the downstream growth model and clarify the relationship with the spatial record linkage model as follows.

We previously defined the set of clusters $\mathcal{C}(\mathbf{\Lambda})$ derived from the linkage structure of the spatial record linkage model, and we further restrict this set to the clusters for which growth is observed. We define the set of growth clusters $\mathcal{C}^G(\mathbf{\Lambda})$, with respect to several ecological conditions, which correspond to the individual trees identified by the linkage model for which a plausible change in canopy volume has occurred between the two time points. For notational clarity, we introduce the functions $\min^f()$ and $\max^f()$, which return the minimum and maximum file index, respectively, for a given cluster. We note the implicit ordering in the file indices relative to time such that the first file is the oldest and the second file is the most recent. We denote the observed canopy volume of the jth record in file i as v_{ij}, measured in cubic meters, where the file index is synonymous with the data collection time point associated with the file. The set of growth clusters may be defined accordingly, as $\mathcal{C}^G(\mathbf{\Lambda}) = \{\mathcal{C}_{j'} : \max^f(\mathcal{C}_{j'}) \neq \min^f(\mathcal{C}_{j'}) \;\&\; r_1 \cdot v^*_{\min^f(\mathcal{C}_{j'})} < v^*_{\max^f(\mathcal{C}_{j'})} < r_2 \cdot v^*_{\min^f(\mathcal{C}_{j'})}\}$ such that $\mathcal{C}^G(\mathbf{\Lambda}) \subseteq \mathcal{C}(\mathbf{\Lambda})$. Where $v^*_{\min^f(\mathcal{C}_{j'})}$ and $v^*_{\max^f(\mathcal{C}_{j'})}$, denote the summed volumes of the records associated with the minimum and maximum file indices in the cluster. This implicitly defines a procedure that merges the volumes of linked records within files as a result of deduplication from the linkage model. The hyperparameters r_1 and r_2 control the lower and upper bounds for the change in canopy volume for a growth cluster relative to the typical and biologically feasible growth behavior for the time interval between the observed records and such that $0 < r_1 < r_2$. For example, if we specify $r_1 = 0.9$ and $r_2 = 1.6$, then we would restrict our set of growth clusters to those that saw between a 10% loss and a 60% increase in canopy volume over the interval between measurements.

We exclude the clusters which do not satisfy the specified growth rate constraints from the set of growth clusters used to estimate the growth model parameters. We note that the excluded clusters from the linkage model correspond to changes in canopy volume due to abiotic factors or obvious errors in the linkage resulting in biologically implausible growth rates. While it is possible that a tree may experience a decline in canopy volume over time (e.g., during the mortality process), we have chosen to emphasize a method that focuses on the growth of healthy trees with the recognition that this restricts our understanding of the growth relationship conditional on the fact that a tree grew or experienced a small enough decline in canopy volume that the change could be attributed to errors in the LiDAR scanning and post-processing.

The row vector $\boldsymbol{x}_{s_c}$, of length $p+1$, contains p observed covariates at the latent location $\boldsymbol{s}_c$ for the growth cluster $\mathcal{C}_c^G$ with first element corresponding to the baseline growth rate asymptote. We note that topographic covariates (Folded Aspect, Growing Degree Days, HAS Wetness Index, and Snowpack Persistence in this application) are assumed to be centered and scaled across the entire surface of the domain of interest D prior to inclusion in the model.

For each growth cluster $\mathcal{C}_c^G \in \mathcal{C}^G(\boldsymbol{\Lambda})$, g_c is the observed annual growth for cluster c and is defined as a function of the first and last cumulative volume measurements for the linked record set $\mathcal{C}_c^G$ such that

$$g_c = \frac{v^*_{\max^f(c)} - v^*_{\min^f(c)}}{t(\max^f(c)) - t(\min^f(c))}.$$

The function $t()$ returns the year associated with the file index so that the difference in observed canopy volumes is scaled by the length of the interval between measurements to place the observed growth on an annual scale. We model the observed annual growth as a function of the true growth, which we specify as a Michaelis–Menten type function dependent upon the initial observed canopy volume and the environmental covariates associated with the record's latent location $\boldsymbol{s}_c$, while allowing for measurement error, such that

$$g_c|\gamma, \boldsymbol{\beta}, \tau, \boldsymbol{\Lambda}, \boldsymbol{x}_{\boldsymbol{s}_c}, \boldsymbol{v}^* \sim \text{Skewed t}(\mu_c, \tau, \delta, \omega) \quad \text{for } \mu_c = \frac{(\boldsymbol{x}_{s_c}\boldsymbol{\beta}){v^*_{\min^f(c)}}^{\alpha}}{\gamma^{\alpha} + {v^*_{\min^f(c)}}^{\alpha}}. \tag{2.1}$$

As mentioned above, a notable advantage of the generalized Michaelis–Menten style growth function is that the parameters of the true growth function have clear biological interpretations. The linear

component of the function adjusts the maximum growth asymptote as a function of the covariates at a given location. The parameter γ controls the size at which the growth rate saturates due to size scaling, establishing the inflection point of the growth curve. The parameter α controls the curvature of the growth function, where values of $\alpha > 1$ result in a sigmoidal curve and values of $\alpha \leq 1$ result in a shape more akin to a logistic curve. While the true growth is generally assumed to be nonnegative, our model allows for the observed growth to be negative as a function of measurement error.

The full growth model is defined as follows:

$$
\begin{aligned}
g_c|\gamma, \boldsymbol{\beta}, \tau, \boldsymbol{\Lambda}, \boldsymbol{x}_{\boldsymbol{s}_c}, \boldsymbol{v}^* &\overset{\text{ind}}{\sim} \text{Skewed t}(\mu_c, \tau, \delta, \omega), \\
\tau &\sim \text{Uniform}(0, b_\tau), \\
\delta &\sim \text{Normal}_{[-1,1]}(0, \sigma_\delta^2), \\
\omega &\sim \text{Gamma}(2, b_\omega), \\
\gamma &\sim \text{Uniform}(a_\gamma, b_\gamma), \\
\alpha &\sim \text{Beta}_{[c_\alpha, d_\alpha]}(a_\alpha, b_\alpha), \\
\beta_0 &\sim \text{Normal}(\mu_0, \sigma_0^2), \\
\beta_k &\overset{\text{ind}}{\sim} \text{Normal}(\mu_{\beta_k}, \sigma_{\beta_k}^2) \quad \text{for } k = 1, \ldots, p,
\end{aligned}
$$

where the specification of the hyperparameters is informed by the advice of domain science experts, while being adequately diffuse where appropriate. We model the observed growth, annual change in canopy volume, as a nonlinear function of the initial size and a set of environmental covariates at the latent location $\boldsymbol{s}_c$ of the individual with skewed measurement error derived from the LiDAR processing algorithm described in Section 2.2. We place a weak Uniform prior on γ with the lower and upper bounds specified as the minimum reasonable growth saturation value as a function of size and the maximum size observed in the first file, as the parameter corresponds to the size at which the growth rate reaches its half maximum. We specify a shifted and scaled Beta prior for the shape parameter α, which controls the curvature of the growth function, where the specified bounds limit the degree of possible curvature. A noninformative version of this prior takes $a_\alpha = b_\alpha = 1$ corresponding to a Uniform distribution over the specified range. We assume the

individual β_k coefficients are independent a priori and assign appropriately diffuse Normal priors with the understanding that all covariates have been centered and scaled prior to inclusion in the model. We place a Gamma prior on ω, where ω controls the kurtosis of the distribution. Finally, we specify a truncated Normal prior for the skewness parameter δ, where the truncation bounds follow the support of the parameter. In this model formulation we adopt a nonlinear regression skew-t error model, as presented by De la Cruz and Branco (2009), to account for the observed structure of our empirical data. For our specific skew-t density, we follow the formulation of Hansen (1994) and perform the appropriate location and scale adjustments such that μ_c and τ are the mean and variance of the distribution respectively when fitting the model. We do note, however, that this modeling framework may be applied more generally with alternative assumed error processes dependent upon the requirements of a given application. For example, we consider a normal error process in the simulation study that we present in Section 2.6.

Combining the spatial record linkage and downstream growth models, as seen in the plate diagram in Figure 2.2, we obtain the structure for the two-stage modeling approach. We would like to emphasize the distinction between the two models and the assumptions that are made in each. The linkage model is designed to be as flexible as possible to identify the relationship between the observed records and the latent locations, while the growth model is designed to estimate the growth of the trees. The models are specified to be used together, but the assumptions made in the linkage model are not necessarily identical to those made in the growth model. This two-stage approach allows the linkage to be used for a variety of downstream modeling objectives without the need to rerun the linkage model for each task. Following the LA procedure outlined by Sadinle (2018), we propose using a random sample of iterations from the marginal posterior of the linkage structure $\boldsymbol{\Lambda}$ and the latent spatial point process $\boldsymbol{s}$, obtained from the linkage stage, as inputs for the downstream model. The LA approach effectively marginalizes out the uncertainty from the linkage and the latent locations and provides equivalent inference for the growth model parameters compared to the marginal inference that would be obtained from a joint model under certain conditions. We discuss the requirements and justification for this approach in greater depth in Section 2.4.

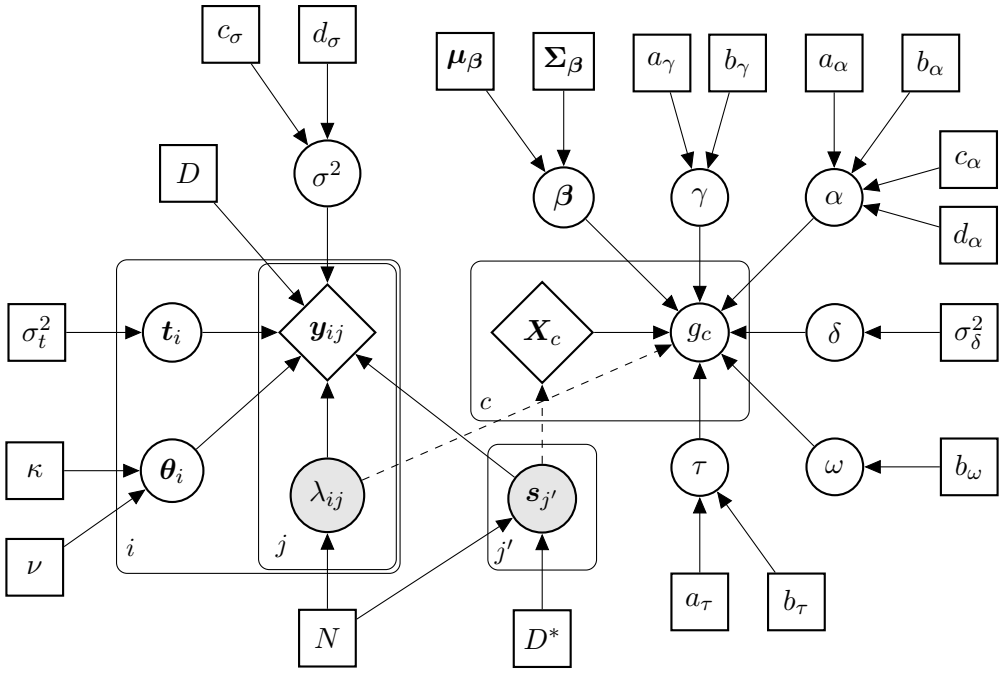

Figure 2.2: Plate diagram for the two-stage record linkage and downstream growth model. Where $i = 1, 2$ denotes the file index, $j = 1, \ldots, n_i$ denotes the record index within file i, $j' = 1, \ldots, N$ denotes the latent location index, and $c = 1, \ldots, |\mathcal{C}^G(\boldsymbol{\Lambda})|$ denotes the growth cluster index. The round nodes indicate parameters, while square nodes indicate hyperparameters. We note that solid arrows denote stochastic relationships, while dashed arrows identify the inputs from the record linkage model to the downstream growth model. This framework provides the structure for the LA approach discussed in detail in Section 2.4.

2.3.3 Computational strategies

The two-stage Bayesian hierarchical model framework that we propose has many strengths including interpretability, the ability to incorporate relevant prior knowledge, and robust uncertainty quantification across the entire modeling pipeline. However, Bayesian record linkage modeling approaches are known to carry additional computational overhead that can be prohibitive to the use of these models, in practice, when dealing with large datasets (see Steorts et al. (2014) and Marchant et al. (2021) for more complete discussions). We alleviate some of the computational expense associated with the use of a Markov chain Monte Carlo Gibbs sampling algorithm for our model through a few key mechanisms discussed below.

One of the most common approaches for improving the scalability of record linkage models is to exact some form of deterministic blocking for records to reduce the number of comparisons necessary. This mechanism is commonly employed across both Bayesian and frequentist record linkage implementations as a preprocessing step that invalidates certain linkages that are deemed to

be implausible (Steorts et al. (2014); Murray (2015)). While certain deterministic blocking schemes may impact the accuracy of the linkage and fail to adequately quantify the uncertainty associated with the procedure, we are able to take advantage of the spatial structure of our data and the biological limitations that invalidate certain links between observed records and $\boldsymbol{s}_{j'}$ as a function of Euclidean distance. As an alternative to blocking, we implement a sampling scheme for $\boldsymbol{\Lambda}$ in our Gibbs sampling algorithm that allows us to approximate the posterior linkage structure under the assumption that the observed location for an individual must be within a maximum distance of the true latent location $\boldsymbol{s}_{j'}$ that it is associated with. In contrast to blocking schemes which invalidate links as a function of comparisons between records, our approach limits the linkage structure directly. Absent the use of a blocking or approximation scheme, the time required to sample from the true posterior distribution of the linkage structure $\boldsymbol{\Lambda}$ increases quadratically with the number of records. Instead of considering the full set $\boldsymbol{s}$ of possible latent locations for each record when sampling the latent matching structure, we consider only $\boldsymbol{s}_{j'}$ within a bounding box around the observed record. Additionally, we impose the restriction that there must be at least two candidate $\boldsymbol{s}_{j'}$ within the bounding box; otherwise, we increase the size of the box iteratively until this condition is met to ensure a reasonable approximation of the cluster assignment probabilities. We note that the spatial bounding approach yields samples from an approximate posterior distribution; however, the $\boldsymbol{s}_{j'}$ removed from consideration have near zero probability associated with them as possible matches, and their removal allows us to maintain a consistent computational cost in the sampling of each individual λ_{ij}. In Figure 2.3 we consider the correlation between posterior similarity scores, which measure how often records are estimated to be coreferent, for bounding boxes of varying sizes. We see that the correlations between posterior similarity scores are close to 1 across bounding box sizes, demonstrating the accuracy of the spatial bounding box approach for approximating the true posterior distribution of the linkage structure for moderately dense subsets of the empirical data.

Functionally, this scheme improves both the speed and efficiency of the record linkage model, as seen in Figure 2.4, allowing the model to scale to much larger domains of interest, which is a clear limitation of alternative modeling approaches. As the size of the bounding box increases, we observe a near exponential increase in the average time required per iteration of the sampler run on a dense $300\,\mathrm{m}^2$ subset of our empirical data. Over this domain, we observe a 97.3 times speedup

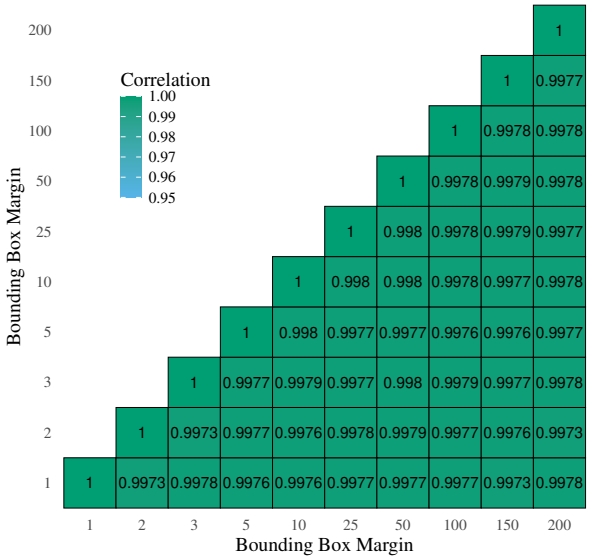

Figure 2.3: Correlation heatmap for the posterior similarity cluster scores for varying bounding box margins for an area of size $200\,\mathrm{m}^2$. The bounding box margin value specifies the distance to the boundary from the observed point. For example, a margin of 2 m corresponds to a $16\,\mathrm{m}^2$ bounding box centered at the observed location of the individual.

per iteration on average when using a bounding box of 3 m with the resulting linkage having a posterior similarity score correlation of 0.9984 compared to not using a bounding box. However, we note that the speed improvement for decreasing bounding box sizes is not universal as at some point we are required to expand the size of the box iteratively to meet the conditions established for guaranteeing a reasonable approximation of the cluster assignment probabilities.

To optimize the raw computation speed, the MCMC sampler is written in R and C++ using **Rcpp** (Eddelbuettel et al. (2023a)) and **RcppArmadillo** (Eddelbuettel et al. (2023b)) to improve scalability over a base R implementation. We implement the downstream growth model in **rstan** to take advantage of the speed and flexibility of the NUTS algorithm (Stan Development Team (2023)). Additionally, the optimized parallel computation available in **rstan** reduces the time required to fit the downstream growth model with minimal additional architecture required. The details of our Gibbs sampling algorithm for the spatial record linkage model may be found in Appendix A.1.

2.4 Linkage-averaging for parameters from auxiliary data models

We present a discussion of the theoretical justification for the LA approach, introduced by Sadinle (2018) for population size estimation, for a general downstream task with auxiliary data, that

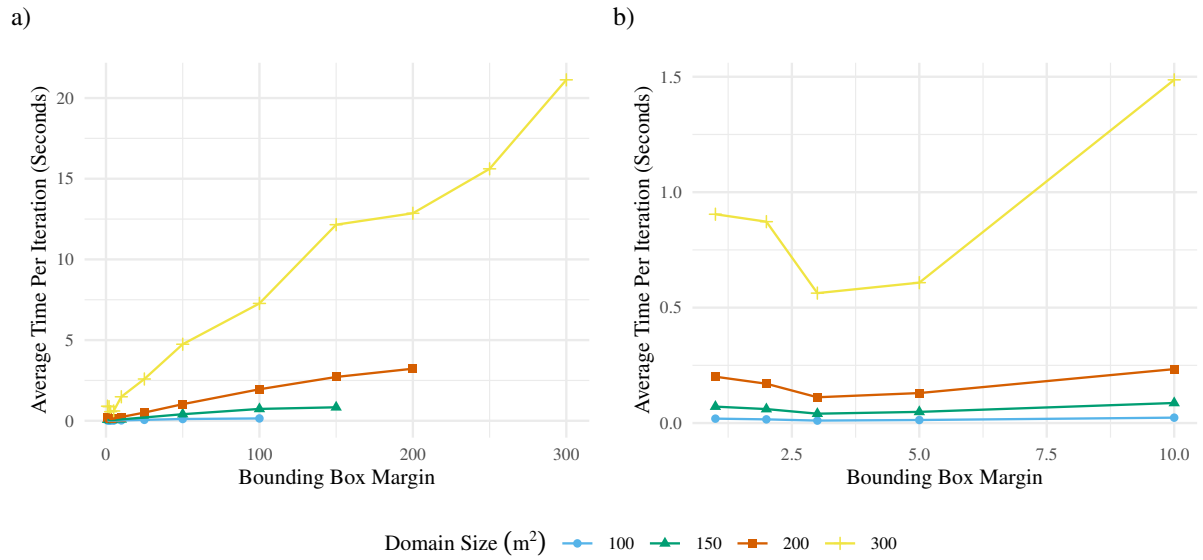

Figure 2.4: Record linkage model Gibbs sampler timing results per iteration for varying bounding box margins around each sampled point. We consider the timing for areas of size $100\,\mathrm{m}^2$, $150\,\mathrm{m}^2$, $200\,\mathrm{m}^2$, and $300\,\mathrm{m}^2$. Plot (a) shows the full timing results, while plot (b) shows an inset for smaller bounding box margins.

is, regression, when paired with a record linkage model that models the observed records as noisy versions of a set of true latent field values as in Steorts et al. (2016) and Liseo and Tancredi (2011a) and as defined in Section 2.3.1. We demonstrate that under the following two mild conditions, this LA approach may be reframed to provide proper Bayesian inference for the parameters of a more general downstream task.

Condition 1. Our beliefs regarding the linkage structure $\boldsymbol{\Lambda}$ and the true latent field values $\boldsymbol{s}$ are quantified by the joint posterior distribution $p_{\mathrm{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s}|\boldsymbol{y})$, arising from a record linkage model employing a latent matching structure, where the posterior is proportional to the product of the likelihood $\mathcal{L}_{\mathrm{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s}|\boldsymbol{y})$ and the joint prior $p(\boldsymbol{\Lambda}, \boldsymbol{s})$.

We note that this condition depends on the use of a record linkage model that employs a latent matching structure in which the fields are modeled directly, though it would be straightforward to adapt for an alternative linkage model construction.

Condition 2. If the true linkage structure $\boldsymbol{\Lambda}$ and latent field values $\boldsymbol{s}$ were known, the posterior $p_{\mathrm{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))$ arising from an auxiliary data model with likelihood $\mathcal{L}_{\mathrm{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))$ and joint prior $p(\boldsymbol{\Theta})$, which may be further decomposed depending on the structure of the model, would encapsulate our beliefs regarding the downstream model parameters.

The second condition describes the inferential process for $\boldsymbol{\Theta}$, the vector of parameters from the auxiliary data model, under the assumption that the true linkage structure and latent field values are known. The combination of these two conditions provide the basis for our underlying argument such that if these conditions hold, the following relationship

$$p_{\text{LA}}(\boldsymbol{\Theta}) = \text{E}_{\boldsymbol{\Lambda},\boldsymbol{s}|\boldsymbol{y}}\left[p_{\text{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))\right] = \sum_{\boldsymbol{\Lambda}}\sum_{\boldsymbol{s}} p_{\text{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))\mathcal{L}_{\text{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s}|\boldsymbol{y}),$$

is obvious to consider given its clear interpretation. We also demonstrate that $p_{\text{LA}}(\boldsymbol{\Theta})$ is a proper posterior distribution. In order to perform inference on $\boldsymbol{\Theta}$ and $(\boldsymbol{\Lambda}, \boldsymbol{s})$, given $\boldsymbol{y}$, we require a joint prior for $(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s})$ such that

$$p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s}) = p_{\text{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s}))p(\boldsymbol{\Lambda})p(\boldsymbol{s}),$$

which follows naturally from Conditions 1 and 2 above.

Theorem 2.4.1 (Bayesian validity of linkage-averaged auxiliary data model parameters joint posterior)**.** *The marginal posterior of* $\boldsymbol{\Theta}$ *under the likelihood* $\mathcal{L}_{\text{LRL}}(\boldsymbol{\Lambda}, \boldsymbol{s}|\boldsymbol{y})$ *of the latent record linkage model and joint prior* $p_{\text{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(s))p(\boldsymbol{\Lambda})p(\boldsymbol{s})$ *is* $p_{\text{LA}}(\boldsymbol{\Theta})$.

Theorem 2.4.1 establishes $p_{\text{LA}}(\boldsymbol{\Theta})$ as a valid posterior distribution. We provide the proof for 2.4.1 in Appendix A.2, as the details are similar to the proof of Sadinle (2018). We note that the proof of Sadinle (2018) holds specifically for population size estimation, in comparison to the result for a general downstream task with auxiliary data which we have established. In practice, we approximate the linkage-averaged posterior of $\boldsymbol{\Theta}$, $p_{\text{LA}}(\boldsymbol{\Theta})$, with a random sample $\boldsymbol{\Theta}^{(1,t)}, \ldots, \boldsymbol{\Theta}^{(l,t)} \sim p_{\text{AD}}(\boldsymbol{\Theta}|\mathcal{C}(\boldsymbol{\Lambda})^{(t)}, \boldsymbol{X}(\boldsymbol{s})^{(t)})$, for each $t = 1, \ldots, k$, such that

$$p_{\text{LA}}(\boldsymbol{\Theta}) \approx \frac{1}{kl}\sum_{t=1}^{k}\sum_{u=1}^{l} I(\boldsymbol{\Theta} = \boldsymbol{\Theta}^{(u,t)}),$$

where k and l are chosen to be sufficiently large to provide a reasonable approximation to the true posterior.

The spatial record linkage and downstream growth models detailed in Sections 2.3.1 and 2.3.2, respectively, clearly satisfy the construction discussed above with $\boldsymbol{\Theta} = (\alpha, \gamma, \boldsymbol{\beta}, \tau, \delta, \omega)$ and where

$\boldsymbol{X}(\boldsymbol{s})$ represents the auxiliary data component of the model. We note that an alternative to the LA approach is to model the record linkage and downstream task jointly, which allows the downstream task to inform the file linkage procedure. For example, Gutman et al. (2013) discuss a joint modeling approach based on multiple imputation that iteratively samples the unknown linking partition and the downstream model parameters. In their framework the unknown links are treated as missing data and imputed. While the joint modeling approach may potentially improve the linkage, it is often accompanied by a substantially increased computational burden, and the performance is sensitive to model misspecification for the downstream model. In contrast, the LA framework that we present provides equivalent marginal inference for the downstream model parameters as that obtained from a joint model, under the assumptions of Conditions 1 and 2, and allows more flexibility for the researcher to recycle the linkage for multiple downstream tasks of interest that may be implemented in parallel in a straightforward and efficient fashion.

2.5 Estimation of annual growth curves for Rocky Mountain conifer forests

In this section we return to the empirical data and related hypotheses regarding the annual growth behavior of Southern Rocky Mountain conifer forests presented in Section 2.2. We employ the two-stage LA approach for estimating the downstream growth model parameters using $k = 100$ randomly sampled iterations from the joint posterior distribution of the linkage structure $\boldsymbol{\Lambda}$ and latent locations $\boldsymbol{s}$ as the input for the growth model. For each pair $\left(\boldsymbol{\Lambda}^{(k)}, \boldsymbol{s}^{(k)}\right)$, we derive the set of growth clusters, $\mathcal{C}^G\left(\boldsymbol{\Lambda}^{(k)}\right)$, and the set of location-dependent covariates, $\boldsymbol{X}\left(\boldsymbol{s}^{(k)}_{\mathcal{C}^G\left(\boldsymbol{\Lambda}^{(k)}\right)}\right)$, which are then used to fit the growth model defined in Section 2.3.2. This procedure allows us to obtain estimates of the marginal posterior distributions of the growth model parameters of interest that are equivalent to the marginals obtained from a joint model for the linkage structure, $\boldsymbol{s}_{j'}$, and the growth model parameters, as discussed in Section 2.4. For this analysis we specify $r_1 = 0.9$ and $r_2 = 1.6$ such that we consider primarily positive growth with a maximum increase of 60% of the initial observed canopy volume over the four year study period. These cutoffs reflect typical growth behavior and disqualify implausible clusters arising from errors in the linkage and due to

environmental mechanisms like damage from extreme wind or lightning that do not reflect the biological mechanisms of tree growth.

In addition to the topographic covariates discussed in Section 2.2, we also include three inter-tree competition metrics. Fagerberg et al. (2022) highlight the importance of including competition indices in individual tree growth models for conifer species. For our application we consider relative spacing index (RSI), the ratio of the nearest neighbor distance to the average neighbor distance, larger neighbor volume (LNV), the summed canopy volumes of a tree's larger neighbors, and neighborhood density (ND), the density of individuals within the neighborhood of the individual. All competition metrics are calculated using the observed locations from the first scan (2015) within a 15 m neighborhood around each point such that all three are considered semi-distance-dependent competition indices. Ma et al. (2018) calculate LiDAR derived tree competition indices using a 15 m neighborhood, and we adopt the same neighborhood size for our analysis. To ensure that these metrics are accurate for all points considered, the downstream growth model is only fit to growth clusters located more than 15 m from the boundary of the study domain. We note that RSI and ND are measures of symmetric competition while LNV captures asymmetric competition among individuals to account for the variation possible across the range of competitive effects. An in depth discussion of competition indices and their construction may be found in Pommerening and Sánchez Meador (2018) and Contreras et al. (2011).

Given the size of the study domain (~2 km^2), we fix the rotation parameter for the second scan, θ_2, to be zero, as even a very small degree of rotation can have a large effect on points near the boundary of the domain. We allow for the possibility of scan-wide translation in this analysis and note that the choices of which image alignment components to include and the strength of their constituent priors will likely depend on the application. We select noninformative and weak priors, where appropriate, for the record linkage and downstream growth model parameters according to the outline in Section 2.3. We specify $q = 1.25$ in determining the maximum number of unique latent individuals N to provide flexibility across the range of point densities observed in the study area. The convergence of the record linkage model and downstream model variants is assessed by examining traceplots and Gelman–Rubin statistics (Gelman and Rubin (1992)) for each

approach. The full model specification details and select convergence diagnostics may be found in Appendix A.3.

In concert with the two-stage LA approach, we consider two alternative heuristic strategies for linking trees across scans, which we term nearest distance matching (NDM) and polygon overlap matching (POM). The NDM algorithm matches each tree crown location from the first scan with the closest point from the second scan. The POM approach uses the derived crown geometries from the LiDAR scans and traces the crown polygons from the 2015 scan forward and considers the overlap with the polygons from the 2019 scan, and the change in canopy volume is calculated as the difference between the estimated volumes. While these methods do not perform deduplication and fail to provide uncertainty quantification for the linkage, they represent simple and easy to implement strategies for identifying unique individuals across scans and obtaining growth estimates that are analogous to methods being used in practice. For example, Ma et al. (2018) use a more sophisticated heuristic matching algorithm coupled with manual review of marginal matches. We apply the same growth cluster restrictions for these methods as for the LA approach, so the set of derived growths is characteristically equivalent across the three linkage procedures. During model fitting, we consider four candidate growth models, three with the nonlinear Michaelis–Menten mean function, for each linkage strategy (Skewed t, Skew Normal, Normal, and Multiple Linear Regression with Normal Errors) and evaluate the model fit using the Continuous Ranked Probability Score (CRPS), as suggested by De la Cruz and Branco (2009) following the discussion of Gneiting and Raftery (2007). We use the scaled CRPS (sCRPS) presented by Bolin and Wallin (2023), which has been shown to be locally scale invariant and an improvement over the standard CRPS for model selection. The Skewed t model discussed in Section 2.3 was identified by the sCRPS metric as the top performing model for all 3 of the linkage schemes (additional details may be found in Appendix A.3). Figure 2.5 displays the 90% credible intervals for the covariate coefficients obtained from the three linkage approaches for the Skewed t model.

We see from the coverage plot that the POM approach results in drastically different estimates for the coefficients in terms of magnitude, and in some instances sign, coupled with high degrees of certainty in the estimates. In contrast, the NDM and LA approaches produce more similar estimates for the coefficients given that they are both distance based linkage approaches, although the NDM

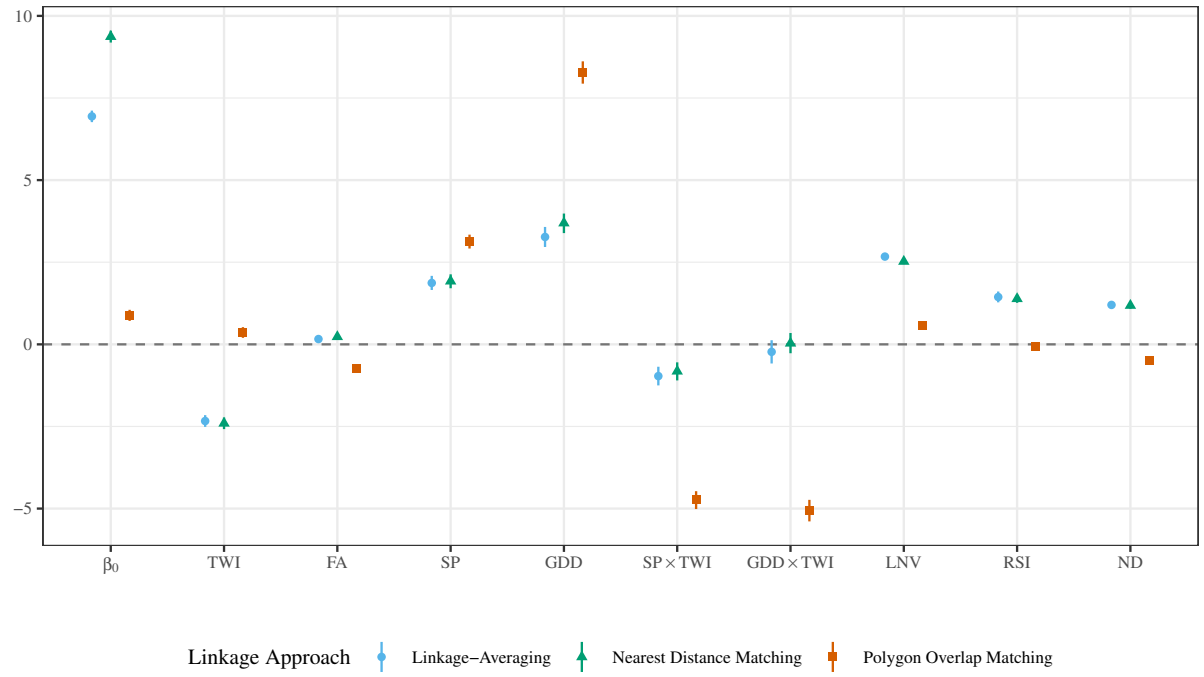

Figure 2.5: Comparison of the 90% credible intervals for the growth asymptote and the topographic and competition metric covariate coefficients obtained from the downstream model fit using data derived from the three different linkage approaches (LA, NDM, and POM).

is deterministic and so does not marginalize out the uncertainty from the linkage procedure when fitting the downstream task like the LA approach. We also note that the estimates for the growth asymptote β_0 are notably different across the three linkage approaches, even for models with similar estimates of the covariate coefficients. In Figure 2.6 we provide a comparison of the estimated size-dependent growth curves arising from the LA model under high and low growth scenarios as a function of the topographic covariates Snowpack Persistence and Growing Degree Days, which are measures of water and energy availability, respectively, as discussed in Section 2.2. The annual growth curve, μ_c (equation 2.1), is a function of size, where the growth asymptote is adjusted by the covariate values at the location of the tree. We consider the 20th and 80th quantiles of the empirical distribution for these covariates while holding all other covariates at their median values to highlight the marginal impact on growth for these individual covariates with 90% credible bands. In panel (c) we examine the growth curves for both covariates simultaneously to demonstrate the combined impact of the covariates on growth behavior in both suboptimal and optimal growth conditions.

Our analysis suggests the importance of including environmental variables related to growth conditions in addition to competition indices in modeling size-dependent individual tree growth over large spatial domains (Ford et al. (2017); Maes et al. (2019)). While the growth behavior is primarily constrained by size in our analysis, these additional metrics are influential in determining the growth behavior of forests across varied terrains and localized densities. Our work reinforces other studies on environmental constraints to conifer growth in the region which emphasize that both available energy and water from snowpack are important growth constraints (Berkelhammer et al. (2020); Carroll et al. (2020)). Somewhat surprisingly, we observed negative effects of a soil moisture proxy (HAS Wetness Index, Figure 2.1 panel (d) on tree growth, indicating that growth of some trees in our study domain may be limited in the wettest soils (Marks et al. (2020)). We also observe that the interaction of topographic proxies for energy and water availability may have an impact on growth when accounting for some collection of symmetric and asymmetric competition metrics. This inference for the downstream model is sensitive to the choice of linkage approach, particularly when considering the estimated growth asymptote.

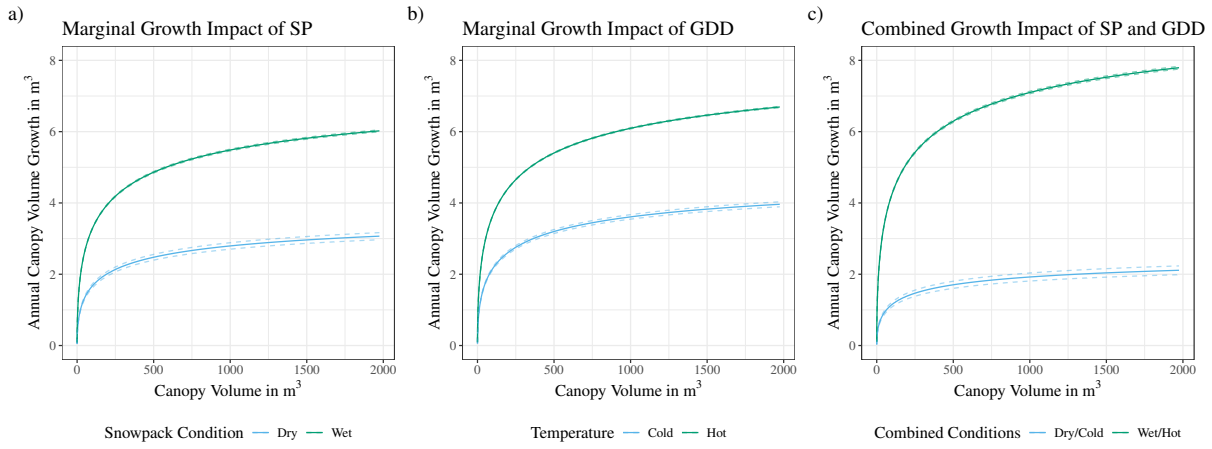

Figure 2.6: Comparison of the estimated growth curves for varying quantiles of covariates of interest where plot (a) is the growth curve for Snowpack persistence, plot (b) is the growth curve for Growing Degree Days, and plot (c) is the growth curve for Snowpack persistence and Growing Degree Days simultaneously. The plots demonstrate the change in growth behavior for low and high quantiles of the covariates while holding all other covariates at their median values, where the designations Dry/Cold and Wet/Hot correspond to the 20th and 80th quantiles respectively across the plots.

2.6 Simulation results

In this section we perform a sequence of simulation studies to examine the efficacy of our modeling framework. In Section 2.6.1, we introduce the data generation algorithm for producing biologically realistic simulated data sets. Subsequently, in Section 2.6.2, we assess the performance of the LA model as applied to a collection of simulated datasets under various scenarios modeled after the empirical data discussed in Section 2.2.

2.6.1 Data simulation

Historically, ecological surveys of forest growth dynamics have relied extensively on field measurement data for validating models using ALS data, as in Ma et al. (2018). These field surveys are often time consuming, expensive to perform, and provide a limited characterization of the model performance across a wide range of scenarios. Due to the scale and complexity of the study area we are considering, a validation dataset is unavailable. Instead, we gauge the efficacy of our model by considering the performance on simulated data across a variety of possible conditions as motivated by our empirical data. The majority of the existing simulation frameworks for marked point processes assume independence between the spatial point process and the mark distribution as noted by Guan and Afshartous (2007); however, our application necessitates location-dependent marks which are specified to be canopy volumes.

We address the disconnect between the available off-the-shelf methods and the requirements of our application through the use of a data simulation algorithm constructed to approximate the underlying marked spatial point process and the relevant biological mechanisms of forest populations, such as growth and recruitment, using three subjectively selected subsets of the empirical data with varying point densities as a basis. We initialize the procedure by simulating the latent point process $\boldsymbol{s}$ with a modification of the modeling scheme of Møller et al. (2016) for marked point processes in order to include topographically derived covariates in the simulation of the mark distribution. We consider three point densities motivated by the range of densities observed in the RMBL dataset. The point densities are 0.04, 0.06, and 0.08 individuals per square meter, which we describe as low, medium, and high, respectively, throughout the remainder of this section. We note that the simulated data is constructed to generate data from two files corresponding to the empirical data in our application.

One of the key innovations of the Møller et al. (2016) approach is to equate a marked spatial point process with a spatiotemporal point process by ordering the marks and mapping them to arrival times in a spatiotemporal process. For each selected density, we use the observed sizes from the 2015 dataset to generate the approximate arrival times of the points and then predict the mark associated with each point as a function of time, neighborhood characteristics, and topographic covariates in an iterative fashion until we obtain a point realization matching the intensity of the empirical reference pattern. We employ an embedded gradient boosted tree model, built using the `xgboost` package (Chen et al. (2023)) in R, to predict the marks given the set of derived features. The empirical data demonstrate a notable pattern of inhibition, or regularity, at the $100\,\mathrm{m}^2$ scale, so we include provisions for interpoint interaction as a function of size in the data generation procedure to capture this behavior. The interaction function for point patterns defined by regularity, such as a Strauss process, are often specified with a hard core radius such that points in the process cannot be within a certain radius of each other (Leininger (2014)). In our process we assume a soft core interaction radius such that we allow points to violate the hard core interaction radius with low probability. All of the simulated data is generated over $130\,\mathrm{m}^2$ areas and then restricted to the center $100\,\mathrm{m}^2$ area to account for possible edge effects in the point patterns. We use the raster images of the topographic covariates, provided by RMBL, to draw the location specific covariate values in order to simulate data that approximates the real data as closely as possible.

To accurately reflect the biological mechanism of juvenile recruitment (i.e., the seeding of offspring trees) over time, we generate the number and locations of potential recruits according to the realized parent point process. Each parent point is assigned a number of recruits based on its size, and the locations of recruits are modeled as arising from a t(1) distribution centered at the parent point. The marks for recruits are simulated from a heavily right-skewed scaled Beta distribution bounded by the minimum size observed in the distribution of the parent points in order to capture the fact that a relatively small number of recruits survive long enough to be identified. We note the LiDAR process used to collect the empirical data has a height detection threshold of approximately 2 m, and so empirical data for the size distribution of smaller trees was unavailable. However, the mechanics of recruitment are well studied, and we incorporate relevant domain knowledge in the construction of our mechanism allowing larger individuals to seed more

potential recruits using a sampling mechanism incorporating the individual's proportion of the total biomass contribution as the sampling weight. Johnson et al. (2021) provide a thorough discussion of the mechanisms involved in recruitment processes of conifer species which we leverage in our data generating process.

We proceed in generating the observed data from two time points by applying a simplified growth model to appropriate transformations of the latent point configuration $\boldsymbol{s}$. For purposes of illustration, we consider a growth model with the same mean structure as the model presented in Section 2.3.2. As noted before, our modeling approach may be adapted for different error processes, and so for simplicity we assume a standard Gaussian error process for this simulation study instead of the Skewed t process discussed previously. We introduce noise in the observed locations according to the data process of the spatial record linkage model in Section 2.3.1 such that each observed location is generated from a bivariate Normal distribution centered at the latent parent point $\boldsymbol{s}_{j'}$ with measurement error σ^2. Lastly, we translate and rotate the points to achieve the final spatial configurations for each file. Given the observed marks for the points from the first file, for each point we predict the growth from the first observed time to the second according to the Michaelis–Menten style mean function μ with measurement error τ^2. We note that growth is also predicted for the generated recruits but without the measurement error component, as these points are technically unobserved in the first file due to their sizes being below the detection threshold. The final set of observed data from two files is obtained by truncating the generated patterns to the center $100\,\mathrm{m}^2$ area. We apply the outlined simulation framework to produce a collection of datasets with known generating parameters and linkage structure as a baseline for assessing model performance, in the absence of a field inventory validation dataset, with varying levels of measurement error deemed to be plausible. A comparison of the simulated and empirical data for a $100\,\mathrm{m}^2$ medium density subset may be seen in Figure 2.7. A detailed discussion of the data generation algorithm and underlying assumptions may be found in Appendix A.4.

2.6.2 Simulation performance

In this section we assess the performance of the two-stage LA modeling approach on data generated using the simulation scheme discussed in Section 2.6.1. We are able to gauge the efficacy of both the record linkage model and the downstream growth model, given that the true linkage

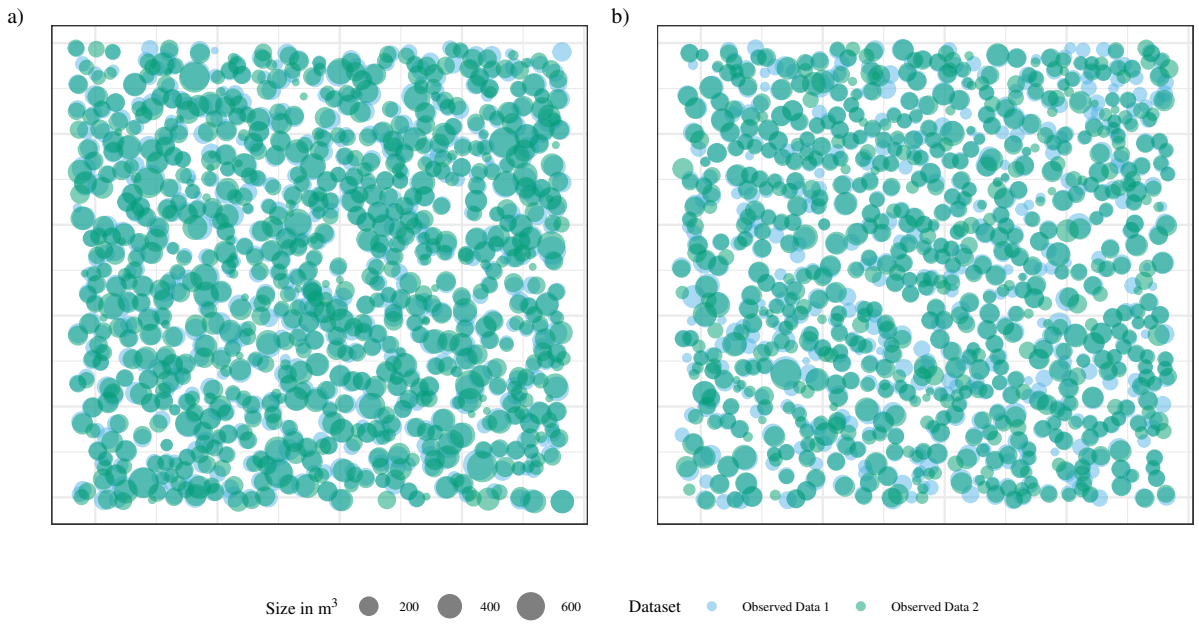

Figure 2.7: Comparison of the data from a medium density $100\,\mathrm{m}^2$ subset where plot (a) is the simulated data and plot (b) is the empirical data subset used to build the predictive model for the medium density validation data.

structure and parameter values used to generate the simulated data are known. We consider the performance of the model on 100 simulated datasets from low, medium, and high densities, which correspond to point intensities of approximately 0.04, 0.06, and 0.08, respectively, over $100\,\mathrm{m}^2$ areas. We first review the performance of the spatial record linkage model and then explore credible interval coverage rates for the growth model parameters across 100 simulated data sets for each density and with varying levels of noise in the observed locations.

For the linkage model, we consider the metrics precision and recall, which are standard evaluation criteria for classification tasks defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \& \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively. Precision measures the proportion of correctly identified matches out of all identified matches while recall measures the proportion of correctly identified matches out of all possible matches (Christen (2012)). We summarize the linkage performance for $\alpha = 1$ over 100 datasets with known true linkage in Figure 2.8. We consider three noise levels for the generating process for the

observed spatial locations corresponding to $\sigma = 0.25$, $\sigma = 0.35$, and $\sigma = 0.45$, which we term small, medium, and large, respectively. The model is run with $q = 1.25$ when determining N to match the value selected for the empirical data analysis. The results were similar for two alternative α values, which may be found in Appendix A.4.

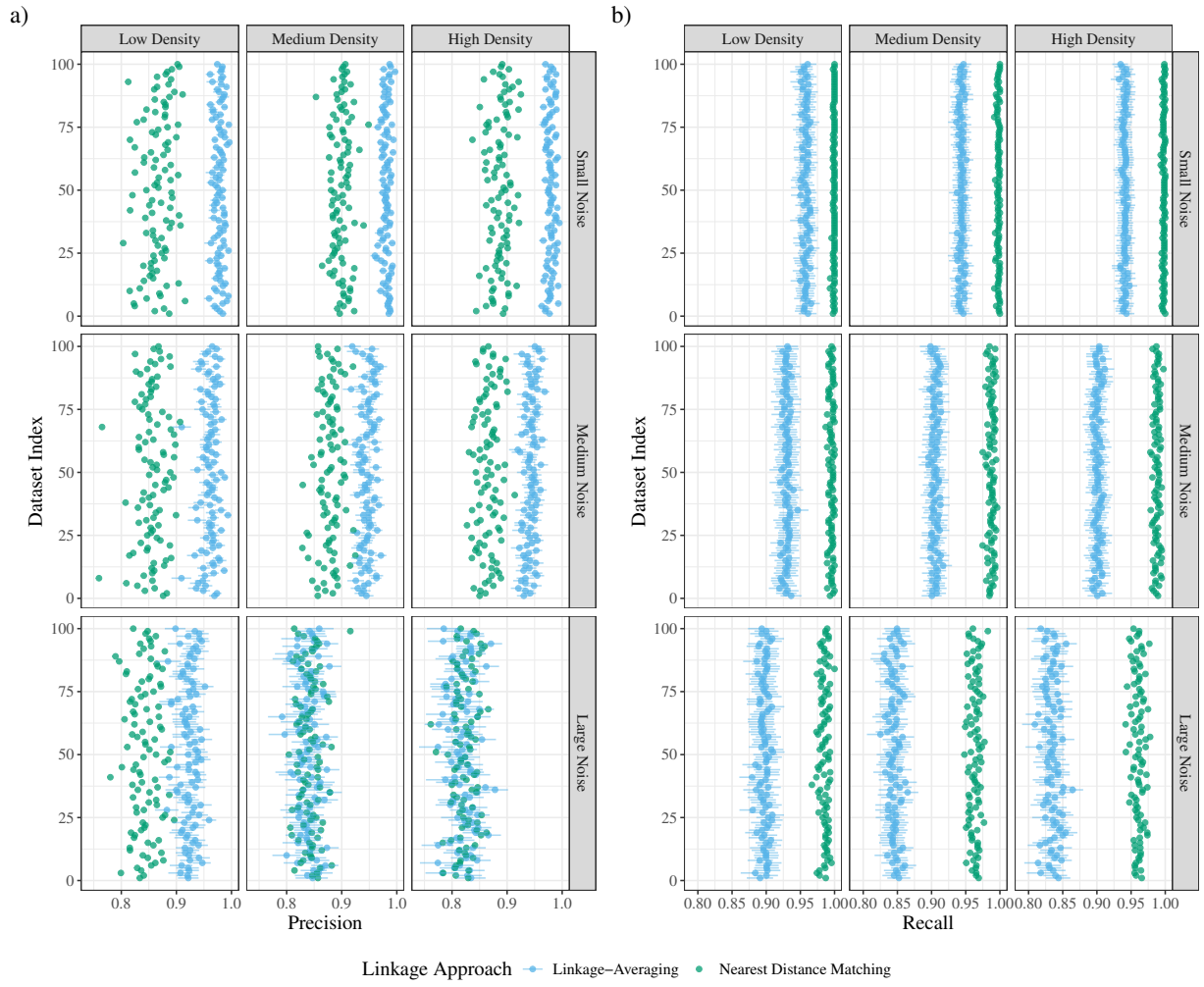

Figure 2.8: Plots comparing the precision (a) and recall (b) performance for the LA and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with $\alpha = 1$.

We note that the NDM algorithm only produces a single linkage estimate for each dataset and, consequently, a single estimate of precision and recall, while we obtain a full posterior distribution of the linkage for each dataset from the spatial record linkage model. We see from Figure 2.8 that the average precision performance for the record linkage model tends to be a notable improvement

over the NDM approach across the varying levels of density and noise. We note that the increase in precision for the record linkage model comes at a cost in the form of reduced recall compared to the NDM approach. The NDM approach matches each point from the first file with a point in the second file and, resultingly, captures more matches overall, but the accuracy of those matches is reduced relative to the LA approach. We note that incorrect links function much like random noise and may result in a form of attenuation bias, effectively driving the estimates of the downstream model covariate coefficients to zero. By prioritizing precision, we are able to draw more accurate conclusions about the impacts of the covariates on growth for the trees that are linked. However, this may result in additional bias when attempting to generalize the findings beyond the observed data.

We next consider the performance of the downstream growth model in terms of nominal coverage rates for 90% credible intervals for the growth model parameters of interest. We examine the same density, noise, and α settings as for the linkage model with growth parameter values $\gamma = 12$, $\boldsymbol{\beta} = \begin{bmatrix} 3 & 0.5 & -0.5 & 0.5 & -0.5 \end{bmatrix}^\mathsf{T}$, and $\tau^2 = 0.5$. Our prior specifications for overlapping parameters match those we use in the real data application to provide as direct an analogue between the two modeling scenarios as possible. The coverage results for the growth model parameters over the 100 datasets for each setting may be found in Table 2.3.

We note that the coverage rates for the true linkage model are consistently around the 90% nominal coverage rate, which serves as the gold standard for the growth model performance. Comparing the LA and NDM approaches, we see that the LA approach tends to outperform the NDM approach and often by a substantial margin. In the instances where the NDM coverage is closer to the nominal level, the coverage for the LA approach is generally more conservative due to the uncertainty propagation from the linkage stage of the modeling pipeline. These results are in line with our expectations regarding the performance of the different linkage approaches and provide evidence that the two-stage LA framework can reliably recover the parameters of interest from a downstream model. In particular, the LA approach consistently has better coverage for the growth asymptote parameter β_0, which may explain the differences that we observed in the estimates from the empirical analysis shown in Figure 2.5. We note that coverage rates for τ^2 are generally lower than the nominal level, most notably in the high noise scenarios for both the LA and

Table 2.3: Empirical coverage rates for 90% credible intervals for covariate parameters of the downstream growth model across the true linkage, LA, and NDM linkage approaches. We consider the coverage over 100 datasets for each setting with $\alpha = 1$ where the bolded coverages are closest to the nominal level (excluding the true linkage).

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.93	0.90	0.86	0.92	0.87	0.85	0.90	0.91
		LA	**0.91**	**0.89**	**0.90**	**0.95**	**0.94**	**0.88**	**0.90**	**0.89**
		NDM	0.70	0.38	0.75	**0.85**	0.80	0.72	0.45	0.17
	Medium	TL	0.85	0.83	0.92	0.89	0.92	0.88	0.91	0.89
		LA	**0.89**	**0.80**	**0.92**	**0.92**	**0.92**	**0.91**	**0.87**	**0.77**
		NDM	0.56	0.20	0.75	0.84	0.86	0.83	0.25	0.06
	Large	TL	0.90	0.90	0.92	0.93	0.94	0.89	0.90	0.92
		LA	**0.93**	**0.76**	**0.98**	1.00	0.95	**0.97**	**0.76**	**0.42**
		NDM	0.70	0.08	0.79	**0.85**	**0.87**	0.82	0.12	0.02
Medium	Small	TL	0.92	0.91	0.88	0.94	0.92	0.92	0.91	0.92
		LA	**0.91**	**0.89**	**0.88**	**0.95**	0.94	**0.93**	**0.92**	**0.75**
		NDM	0.43	0.36	0.64	0.83	**0.92**	0.79	0.36	0.23
	Medium	TL	0.90	0.92	0.89	0.94	0.97	0.93	0.91	0.90
		LA	**0.85**	**0.84**	**0.90**	0.98	0.98	**0.95**	**0.83**	**0.21**
		NDM	0.24	0.10	0.65	**0.91**	**0.86**	0.80	0.16	0.02
	Large	TL	0.86	0.84	0.85	0.90	0.88	0.90	0.89	0.87
		LA	**0.79**	**0.38**	**0.96**	1.00	0.99	0.98	**0.39**	0.00
		NDM	0.32	0.01	0.74	**0.94**	**0.87**	**0.85**	0.01	0.00
High	Small	TL	0.85	0.83	0.87	0.93	0.94	0.85	0.85	0.87
		LA	**0.84**	**0.79**	**0.88**	**0.94**	**0.95**	**0.88**	**0.89**	**0.67**
		NDM	0.36	0.30	0.74	0.95	0.82	0.66	0.29	0.17
	Medium	TL	0.87	0.85	0.90	0.84	0.91	0.83	0.88	0.88
		LA	**0.84**	**0.81**	**0.90**	0.97	1.00	**0.92**	**0.78**	0.06
		NDM	0.28	0.16	0.64	**0.96**	**0.89**	0.67	0.14	**0.08**
	Large	TL	0.86	0.81	0.83	0.92	0.97	0.88	0.86	0.92
		LA	**0.89**	**0.37**	**0.97**	1.00	1.00	1.00	**0.40**	0.00
		NDM	0.28	0.03	0.82	**0.99**	**0.96**	**0.83**	0.02	0.00

NDM approaches. This lower than nominal empirical coverage can be attributed to the incorrect links that are introduced by both the LA and NDM approaches, as compared to the true linkage structure, which results in an overestimation of the value of τ^2. However, even at higher noise levels, the empirical coverage for the covariate coefficients remains at or above the nominal level for the LA approach. This suggests that our interpretation of the impact of the covariates on growth is robust to the error in the linkage. We also note that, in the presence of large amounts of noise, we may encounter identifiability issues with the growth model parameters, as evidenced by the reduced coverage rates for β_0, γ, and τ^2 in the LA and NDM models. Coverage results for the growth model parameters for datasets with $\alpha = 2$ and $\alpha = 3$ were similar and may be found in Appendix A.4.

2.7 Discussion and future work

In this chapter we have established a two-stage modeling framework built around a record linkage model for spatial location data, which serves as the first step in a modeling pipeline constructed for bi-temporal location data. We demonstrated the efficiency and scalability of this approach for analyzing LiDAR derived individual tree characteristic data and provided a general schematic for using the LA approach for two-stage modeling to obtain equivalent inference for the downstream model parameters to the marginal inference obtained from a joint model for the linkage, latent spatial process, and downstream model. This framework enables researchers to investigate growth trends as a function of topographic information at a spatial scale that was previously difficult to achieve and provides flexibility in examining a variety of downstream models for different modeling objectives. It would also be straightforward to extend for use in a Bayesian model averaging construction, as introduced by Raftery et al. (1997), when considering a variety of candidate models for the same downstream modeling objective in lieu of a model selection procedure, as we employed in this application. Another natural extension of our record linkage model would be to applications with more than two files, though some care would need to be taken in specifying the prior for the linkage structure to ensure the correct identification of clusters containing records from multiple files. This extension could also be applied in the context of streaming data where the linkage structure is updated as new data is collected as discussed by Taylor et al. (2024).

We applied this two-stage framework to investigate individual-specific growth-size curves of conifer species on Snodgrass Mountain in the Southern Rocky Mountains of Colorado. We were able to quantify the impact of several key topographic covariates that serve as proxies for energy and water availability, which have been hypothesized to be limiting factors on growth for conifer species in this region. We demonstrated the effectiveness of our modeling approach in a series of numerical experiments on simulated data, in the absence of a ground truth dataset, and implemented a simulation framework for generating data arising from a bi-temporal process as a function of the data model from the linkage model and a general downstream growth model. This approach provides researchers with an alternative tool for model testing and validation in the absence of data with known linkage structure and degrees of measurement error from the various processes involved.

As an alternative to the image alignment framework of Green and Mardia (2006), one could consider an additional preprocessing step utilizing an image registration approach to transform the observed point clouds. The image registration literature is extensive, as discussed by Zitová and Flusser (2003), and includes a variety of approaches that may be utilized in the context of forestry data. For example, Ferraz et al. (2018) introduced an approach for generating a fused high-density vegetation point cloud using a time series of low-density LiDAR scans taken at different times from the NASA-JPL Airborne Snow Observatory. Their method similarly estimates a transformation matrix to align the point clouds, although it uses an iterative procedure and relies on the use of a collection of "tie objects" to estimate the transformation. However, in the context of our application, in addition to misalignment, we observe variability in the point clouds and derived digital surface models such that the estimated tree crowns have different shapes and sizes across scans due to the growth of the trees as seen in the inset of Figure 2.1 panel (c). As a result of the noise in the LiDAR data and the biological processes of tree growth, we have limited "tie objects" available in our study domain with fixed shapes and locations that would be usable in an image registration procedure. Consequently, we believe that the image alignment framework of Green and Mardia (2006), which enables a fully Bayesian implementation of the record linkage model that incorporates the uncertainty inherent in the LiDAR scanning process when identifying which records correspond to the same unobserved latent locations, is a more appropriate choice.

While our modeling approach is flexible and scalable, it can be sensitive to the specification of hyperparameters, like the maximum number of unique individuals across datasets, in facilitating the linkage. There is also a clear relationship with the amount of noise in the observed spatial locations and the performance of the linkage model, so care must be taken when considering the efficacy of a record linkage approach with extremely noisy data. Our modeling approach attempts to decompose the observed distortions in the data to more accurately address the possible sources of error, but these mechanisms are somewhat dependent on the spatial scale of the data (i.e., systematic rotation in a scan). We also note that, while the LA approach gives researchers a high degree of freedom, a joint modeling approach, where the downstream modeling objective influences the linkage, will likely lead to improved performance across the modeling pipeline if the downstream task can be well specified. In our future work, we plan to explore the viability of a joint modeling approach for similar problems which depend on multi-temporal spatial location data.

Chapter 3

ldmppr: Location-Dependent Marked Point Processes in R

3.1 Introduction

Point process models are a rich and complex class of models encompassing processes that occur over time, space, and potentially include additional information about the process in the form of marks. Marks, which are attributes associated with each point (i.e., size or type), add another layer of complexity, particularly when they depend on spatially or temporally varying covariates. The behavior of a point process is typically characterized by the relationship between points. For example, the simplest process is a homogeneous Poisson process, which assumes a constant intensity and no interaction between points and may be described as complete spatial randomness (CSR), where intensity describes the expected number of points per unit of area. More structurally involved processes may exhibit regularity (or inhibition), where points tend to repel each other, or clustering, where collections of points tend to occur in proximity to each other within the pattern. Capturing the dynamics of these processes can be challenging even without the inclusion of marks, and notably more so when this type of auxiliary information is present. Consequently, researchers often make the simplifying assumption that the marks associated with a point process are independent of the primary process itself. However, this assumption is often hard to justify from a scientific perspective, as in the case of trees within a forest where the sizes of the trees (i.e., canopy volumes) play a role in their distribution over space, and where the sizes themselves are likely to depend on location specific information such as elevation, soil moisture, or sunlight availability, which vary over space. Marked point processes have been used across a variety of fields including epidemiology, seismology, criminology, ecology, and the health sciences. Some examples of marked point processes include ambulance call locations with call severity and patient gender as marks (Bayisa et al., 2023), and locations of crime incidences with type of crime as a mark (Mohler et al., 2011).

In this chapter, we present **ldmppr** (Drew and Kaplan, 2025), an R package that provides a suite of tools for working with location-dependent marked point processes characterized by regularity

in the point pattern. While a wealth of R packages exist for working with point processes, such as **spatstat** (Baddeley and Turner, 2005) and **ptProcess** (Harte, 2010), their focus is often broad in scope and they fail to incorporate dependence in the mark distribution with a high degree of flexibility. In contrast, **ldmppr** is designed specifically for working with spatial marked point processes with dependence between the marks and locations that demonstrate inhibitory behavior in the spatial pattern. The package is structured to deliver a straightforward and modular workflow that simplifies the process of model estimation, evaluation, simulation, and visualization given a reference dataset and a set of corresponding covariate surfaces in the form of raster images. In addition to the included models, users can easily adapt the workflow to their specific needs by substituting appropriate components, while still taking advantage of the overall package structure.

While inhibitory marked point processes are commonly modeled using Gibbs processes, they are known to be computationally expensive and difficult to evaluate efficiently. Instead, we take a likelihood optimization approach that is computationally scalable, generally tractable, and enables a straightforward mechanism for simulating from and evaluating the goodness-of-fit of a model. To achieve this, we employ the mechanistic approach for equating a marked spatial point process with a spatio-temporal point process outlined by Møller et al. (2016), who demonstrated the utility of the self-correcting process introduced by Isham and Westcott (1979) for modeling marked point processes with regularity. We extend this framework to include location dependence in the mark distribution using a class of flexible non-linear models, improving the applicability of the original approach to a wider variety of biologically plausible and scientifically interesting processes.

The remainder of the paper is organized as follows. Section 3.2 provides an overview of the relevant point process theory and modeling framework underlying **ldmppr**. Section 3.3 describes the package structure and functionality. Section 3.4 showcases the standard workflow for the package on an example dataset, with an emphasis on forestry applications. Section 3.5 concludes with a discussion of the methodological contributions, strengths, limitations, and potential directions for future development of the package.

3.2 Mathematical background

We begin this section with a brief introduction of the theory of marked point processes to provide context for our modeling approach. We follow this with a discussion of our proposed approach for mapping a marked point process to a marked spatio-temporal process that preserves the dependence between marks and the process and the self-correcting model that we employ for point processes characterized by regularity.

3.2.1 Marked point processes

Marked point processes are an extension of point processes that include additional point specific information in the form of marks. Marks may represent continuous or discrete quantities (i.e., sizes or species of trees in a forest), and may be dependent or independent of the primary generating process. In a process with dependence, the distribution of the marks depends on the locations of the points, while independent marks occur independently of the points themselves. In practice, marked point processes may be modeled by a joint distribution for the points and their associated marks, or alternatively by the conditional distribution of the marks given the locations, often described by their first and second-order characteristics. The first order characteristic is the mark intensity function which describes the expected mark value per unit of area, while the second order characteristic, the mark correlation function, captures the dependence between marks at different times or locations (Schlather et al., 2004).

For purposes of illustration, we consider a marked spatial point process over a finite domain $\mathcal{S} \subset \mathbb{R}^2$ defined as $\mathbf{\Phi} = \{(\boldsymbol{U}_i, \boldsymbol{M}_i) : i = 0, \ldots, N\}$ with a two-dimensional nonnegative mark (though this may be extended to a higher dimensional mark space with additional real valued or discrete marks). We assume that the process is characterized by regularity such that points in the process tend to repel each other resulting in higher interpoint distances than would be observed in a pattern with CSR. We define $\boldsymbol{U}_i \in \mathcal{S}$ to be the spatial location of the i-th point and $\boldsymbol{M}_i = (M_{i1}, M_{i2}) \in \mathbb{R}^+ \times \mathbb{R}^+$ to be the mark vector associated with the i-th point, where $M_{i1} = M_1(\boldsymbol{U}_i)$ is taken to be a measure of size or age and $M_{i2} = M_2(\boldsymbol{U}_i)$ is a secondary mark that may be of additional scientific interest (i.e., height or dbh in forestry applications). We allow for the possibility that $M_{i1} = M_{i2}$, which may be of interest when utilizing our modeling approach to capture location dependence in the primary mark distribution. Additionally, $\boldsymbol{Z}(\boldsymbol{U}_i)$ represents a

vector of topographic (or location specific) covariates at location $\boldsymbol{U}_i$. We assume that the mark vector $\boldsymbol{M} = \{\boldsymbol{M}_i : i = 0, \ldots, N\}$ is ordered according to the first mark such that the M_{i1} are increasingly ordered continuous random variables with $0 \leq M_{01} \leq \cdots \leq M_{N1} \leq \tau < \infty$ and where τ is an unknown upper bound for the marks.

A common approach for modeling a process like $\boldsymbol{\Phi}$ with regularity in the spatial pattern, would be to adopt a Gibbs process model as in Møller and Waagepetersen (2003). However, these models typically include an intractable normalizing constant that makes them difficult to estimate efficiently and often rely on computationally expensive MCMC algorithms. Alternatively, assuming that we have a point process with a structure like $\boldsymbol{\Phi}$, we can use a likelihood based approach that allows for straight forward estimation, model evaluation, and simulation by extending the mechanistic framework introduced by Møller et al. (2016). The method maps a spatial marked point process onto a marked spatio-temporal process conditional on the history of the process by incorporating location dependence in the mark distribution and allowing for higher dimensional marks, as described in the following section.

3.2.2 Spatio-temporal process mapping with location-dependent marks

While marked point processes naturally capture the relationship between event locations and their associated attributes, directly modeling these dependencies can be challenging. By mapping a marked spatial point process onto a spatio-temporal process, we introduce a structured mechanism for estimation that preserves the dependence between marks and locations while leveraging established likelihood-based methods for estimation. Spatio-temporal point processes are a class of models that represent the occurrences of random events over time and space and may be extended to include additional marks that depend on space, time, or both domains simultaneously (Rathbun and Cressie, 1994). At their core, these processes are driven by a conditional intensity function

$$\lambda^*(t, \boldsymbol{y} \mid \mathcal{F}_t), \quad t > 0, \ \boldsymbol{y} \in \mathcal{S},$$

where $\mathcal{F}_t$ is the σ-algebra generated by the oldest point in the process $\boldsymbol{M}_N$, which we define as the anchor point for the process, and the points in the process $(T_i, \boldsymbol{Y}_i)$ with $T_i < t$ (a more thorough discussion may be found in Daley and Vere-Jones (1988)). The conditional intensity function

describes the instantaneous expected rate of events at time t and at location $\boldsymbol{y}$, conditional upon all of the points in the process occurring before time t.

This process may be described as a mechanistic model (Diggle, 2013), and for a marked point process of the form of $\boldsymbol{\Phi}$, defined in Section 3.2.1, we can equate the process with an infinite marked spatio-temporal point process $\{(T_1, \boldsymbol{Y}_1, X_1), (T_2, \boldsymbol{Y}_2, X_2), \dots\}$, where $\mathcal{T} \subset \mathbb{R}$ is the temporal domain. To facilitate this transformation, we introduce a mapping from the original marked point process notation $\boldsymbol{\Phi} = \{(\boldsymbol{U}_i, \boldsymbol{M}_i)\}$ to a spatio-temporal representation $\{(T_i, \boldsymbol{Y}_i, X_i)\}$, where times T_i are derived from primary marks M_{i1}, spatial locations are preserved as $\boldsymbol{Y}_i = \boldsymbol{U}_i$, and secondary marks X_i are conditionally associated with both spatial and temporal components. We define $T_i \stackrel{def}{=} T(M_{i1}) \in \mathcal{T}$ as the arrival time of the i-th point, derived from M_{i1} such that

$$T_i = 1 - \left[\frac{M_{i1} - \min(M_{i1})}{\max(M_{N1}) - \min(M_{i1})}\right]^{\delta}, \tag{3.1}$$

where δ controls the shape of the mapping function and $\delta = 1$ is a 1-to-1 linear mapping. Clearly, $T_0 = 0$ and $T_N = 1$ when this mapping function is applied, so the observed times occur on the interval $[0, 1]$ which ensures that the process is non-explosive (i.e., finite) over the observed interval. The choice of δ allows control over the mapping relationship between the marks of the original marked point process and the derived arrival times of the marked spatio-temporal process. $\boldsymbol{Y}_i \subset \mathcal{S}$ is the spatial location of the i-th point, such that $\boldsymbol{Y}_1 = \boldsymbol{U}_{N-1}, \dots, \boldsymbol{Y}_N = \boldsymbol{U}_0$, and $X_i = M_{i2}$ is the mark associated with the i-th point, where M_{i2} may depend on $\boldsymbol{Y}_i$ and $\boldsymbol{Z}(\boldsymbol{Y}_i)$. This representation allows us to generate a point pattern conditional on $(\boldsymbol{U}_N, \boldsymbol{M}_N)$, i.e., the point with the largest primary mark, where the behavior of the process for T_{N+1} is ignored as we exploit the relationship between the observed finite processes.

We define the conditional intensity function of the our marked spatio-temporal process as follows

$$\lambda^*(t, \boldsymbol{y}, x \mid \mathcal{F}_t), \quad t > 0,\ \boldsymbol{y} \in \mathcal{S},\ x \geq 0,$$

such that the intensity depends on the points arriving in the process before time t. The notation λ^* signifies the dependence on the history of the process, $\mathcal{F}_t$, which is suppressed in the notation going forward for concision. We note that under the assumption of conditional independence between the

arrival times and spatial locations, the conditional intensity function may be decomposed as follows

$$\lambda^*(t, \boldsymbol{y}, x) = \lambda^*(t) h_t^*(\boldsymbol{y}, x), \quad t > 0,\ \boldsymbol{y} \in \mathcal{S},\ x \geq 0,$$

where $\lambda^*(t)$ is the temporal intensity and $h_t^*(\boldsymbol{y}, x)$ is the likelihood for the marked spatial process. We note that $h_t^*(\boldsymbol{y}, x)$ may be decomposed further if we assume conditional independence between the locations and the secondary mark characteristic. This will extend the framework of Møller et al. (2016) to allow for spatio-temporal dependence in the secondary mark distribution and enable flexible estimation of the conditional mark distribution incorporating potentially complex location dependence.

We specify a general parametric model for the conditional intensity, i.e., $\lambda_{\boldsymbol{\theta}}^*(t, \boldsymbol{y}, x)$, where $\boldsymbol{\theta}$ is an unknown set of parameters and obtain the spatio-temporal log-likelihood conditional on $(\boldsymbol{U}_N, \boldsymbol{M}_N) = (\boldsymbol{u}_n, \boldsymbol{m}_n)$ such that

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \lambda_{\boldsymbol{\theta}}^*(t_i, \boldsymbol{y}_i, x_i) - \iiint_{(0,t_n)\times\mathcal{S}\times\mathbb{R}^+} \lambda_{\boldsymbol{\theta}}^*(t, \boldsymbol{y}, x)\, dt\, d\boldsymbol{y}\, dx\,.$$

To facilitate model fitting and simulation, we introduce the temporal-integrated intensity function

$$\Lambda^*(t) = \int_0^t \lambda^*(s)\, ds, \quad t > 0, \tag{3.2}$$

such that $S_i = \Lambda^*(T_i),\ i = 1, 2, \ldots$ form a unit rate Poisson process on $(0, \infty)$. This relationship allows us to easily simulate from the process once the model is estimated, which is a key component in assessing the goodness-of-fit of the estimated process. Details for the simulation algorithm are provided in Section 3.2.5.

3.2.3 Model exhibiting regularity

The framework we have introduced in this section applies to a general marked point process. We now turn our attention to processes exhibiting regularity in their spatial pattern as described in Section 3.2.1. Point processes may be described as self-correcting when the intensity function decreases as the number of events increases, resulting in an inhibitory effect that counterbalances event clustering and enforces regularity in the spatial pattern over time. We define a modified

version of the self-correcting model introduced by Isham and Westcott (1979) with the following intensity function

$$\lambda^*(t, \boldsymbol{y}, x) = \lambda^*_{\boldsymbol{\theta}_1}(t) h^*_{\boldsymbol{\theta}_2,t}(\boldsymbol{y}) g^*_{\boldsymbol{\theta}_3,t,\boldsymbol{y},\boldsymbol{z}(\boldsymbol{y})}(x) f^*_{\boldsymbol{\theta}_4}(t, \boldsymbol{y}).$$

This model incorporates an arrival time process ($\lambda^*_{\boldsymbol{\theta}_1}(t)$), spatial process ($h^*_{\boldsymbol{\theta}_2,t}(\boldsymbol{y})$), conditional mark process ($g^*_{\boldsymbol{\theta}_3,t,\boldsymbol{y}}(x)$), and spatio-temporal interaction ($f^*_{\boldsymbol{\theta}_4}(t, \boldsymbol{y})$). We discuss the formulations for each component as follows.

Arrival time process

The arrival time process T is modeled with intensity function $\lambda^*_{\boldsymbol{\theta}_1}(t)$, following Møller et al. (2016), given by

$$\lambda^*_{\boldsymbol{\theta}_1}(t) = \exp(\alpha_1 + \beta_1 t - \gamma_1 N(t)),$$

where $\boldsymbol{\theta}_1 = (\alpha_1, \beta_1, \gamma_1)$ such that $\alpha_1 \in \mathbb{R}$ is a baseline rate, $\beta_1 \in [0, \infty)$ is a log-linear function of t, and $\gamma_1 \in [0, \infty)$ is the scaling factor for $N(t)$ where $N(t) = \sum_{i=1}^{N} \mathbb{I}\{i \geq 0 : t_i < t\}$.

Spatial process

We model the spatial process with interaction function $\psi(r)$ such that

$$\psi_{\boldsymbol{\theta}_2}(r) = \mathbb{I}[r \leq \alpha_2](r/\alpha_2)^{\beta_2} + \mathbb{I}[r > \alpha_2], \quad r \geq 0,$$

where r is the Euclidean distance between two points (i.e., $r_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|$). This formulation is analogous to the pairwise interaction function in Diggle and Gratton (1984). The density for the process is given by

$$h^*_{\boldsymbol{\theta}_2,i}(\boldsymbol{y}_i) = \frac{1}{c^*_{\boldsymbol{\theta}_2,i}} \prod_{j:j<i} \psi_{\boldsymbol{\theta}_2}(r_{ij}), \tag{3.3}$$

where $c^*_{\boldsymbol{\theta}_2,i}$ is the normalizing constant for the spatial density, which is defined as

$$c^*_{\boldsymbol{\theta}_2,i} = \int_{\mathcal{S}} \prod_{j:j<i} \psi_{\boldsymbol{\theta}_2}(r_{ij})\, d\boldsymbol{y},$$

and $\boldsymbol{\theta}_2 = (\alpha_2, \beta_2) \in [0, \infty)^2$. The spatial density $h^*_{\boldsymbol{\theta}_2,i}$ defines an inhibitive circular region around each larger point $\boldsymbol{y}_j$, for all points with $t_j < t_i$, such that the strength of the interaction diminishes

at a polynomial rate for interpoint distances less than α_2 and disappears for distances greater than α_2.

Conditional mark process

We model the conditional mark process using a flexible non-linear model such that

$$g^*_{\boldsymbol{\theta}_3}(x_i \mid t_i, \boldsymbol{y}_i, \boldsymbol{z}(\boldsymbol{y}_i)) = \mathrm{G}_{x_i}(t_i, \boldsymbol{y}_i, \boldsymbol{z}(\boldsymbol{y}_i)),$$

where G is trained on the feature set $\{T_i, \boldsymbol{Y}_i, \boldsymbol{Z}(\boldsymbol{Y}_i)\}$, and $\boldsymbol{Z}(\boldsymbol{Y}_i)$ is the set of covariate values at the location $\boldsymbol{Y}_i$. We specify this component in generality, and note that any suitably flexible model may be chosen to model the conditional mark process (e.g., a random forest (Breiman, 2001) or gradient boosted tree model (Chen and Guestrin, 2016)). We assume that the mark process is conditionally independent given the times and locations and is parameterized by a non-overlapping set of parameters $\boldsymbol{\theta}_3$ when estimating the model. For example, in a forestry application, assuming conditional independence of marks (i.e., sizes) given spatial and temporal covariates implies that tree size at a given location and time is determined by local environmental factors and possible interpoint competition indices.

Spatio-temporal interaction

The final component of the model is the spatio-temporal interaction term, as introduced by Møller et al. (2016), which we define as

$$f^*_{\boldsymbol{\theta}_4}(t_i, \boldsymbol{y}_i) = \exp\left(-\alpha_4 \sum_{j:j<i} \mathbb{I}[\|\boldsymbol{y}_j - \boldsymbol{y}_i\| \leq \beta_4, t_i - t_j \geq \gamma_4]\right),$$

where $\boldsymbol{\theta}_4 = (\alpha_4, \beta_4, \gamma_4) \in [0, \infty)^3$ with $\beta_4 = \gamma_4 = 0$ if $\alpha_4 = 0$. This term connects the temporal and spatial process components of the full model and captures the dependence between these components. The form of the interaction term regulates the development of the process by limiting the occurrence of points that arrive in close proximity to each other in both space and time.

3.2.4 Parameter estimation

We now consider the procedure for estimating the parameters of the marked spatio-temporal process. We leverage the separability of the full log-likelihood under the assumption of conditional independence of the marks given the locations and arrival times and estimate the conditional mark process and the spatio-temporal process individually. This approach allows us to take advantage of highly flexible non-linear machine learning models to capture the conditional mark process, while maintaining the computational efficiency of maximum likelihood estimation for the spatio-temporal process. We note that both processes depend on the mapping from the original mark space to arrival times, which is controlled by the parameter δ in equation (3.1). However, we do not necessarily require the mappings to be identical for both processes, as the conditional mark process may exhibit different behavior than the spatial process. We provide an overview of the estimation procedure for the full model in practice in Section 3.3.

For the self-correcting model described in Section 3.2.3, we estimate the parameters of the process by optimizing the spatio-temporal component of the log-likelihood, which may be expressed as

$$L(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_4}) \propto \sum_{i=1}^{n} \left[\log \lambda^*_{\boldsymbol{\theta}_1}(t_i) + \log h^*_{\boldsymbol{\theta}_2,i}(\boldsymbol{y}_i) + \log f^*_{\boldsymbol{\theta}_4}(t_i, \boldsymbol{y}_i) \right] - \iint_{(0,t_n)\times\mathcal{S}} \lambda^*_{\boldsymbol{\theta}_1}(t) h^*_{\boldsymbol{\theta}_2,t}(\boldsymbol{y}) f^*_{\boldsymbol{\theta}_4}(t, \boldsymbol{y})\, dt\, d\boldsymbol{y}.$$

The parameter sets $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_4$ can be estimated using any suitable optimization algorithm that allows for bound constraints on the parameter space.

For the conditional mark process, we estimate the parameter set $\boldsymbol{\theta}_3$ using a flexible model that allows for the incorporation of spatial and temporal covariates. The set of features used in training the model which correspond to location specific covariates $\boldsymbol{Z}(\boldsymbol{Y}_i)$ may be derived from raster images or other spatial data sources and may include interpoint competition metrics, as discussed in Section 3.3.3.

We combine the estimated parameter values for the conditional mark process and the spatio-temporal process to obtain the full set of estimated parameters for the marked spatio-temporal process. This approach is modular and allows for a wide variety of models to be substituted for

the conditional mark process. In theory, this framework also holds for alternative spatio-temporal processes, such as those exhibiting clustering or other spatial patterns, though the specific form of the likelihood components would need to be adjusted accordingly.

3.2.5 Simulating from the spatio-temporal process

Utilizing the framework outlined in this section and estimates of the parameter sets $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4$, we can easily generate realizations from the marked spatio-temporal point process on $[0,1] \times \mathcal{S}$. Given $S_i = \Lambda^*(T_i),\ i = 1, 2, \ldots$ as defined in equation (3.2), we have that

$$T_1 = \frac{1}{\beta_1} \log\{1 + \beta_1(\gamma_1 - \alpha_1)S_1\}$$

and for $i = 2, 3, \ldots$

$$\begin{aligned} T_i = \frac{1}{\beta_1} \log\Big[&\exp(\beta_1 T_{i-1}) + \beta_1 \exp(\gamma_1 i - \alpha_1)S_i \\ &- \sum_{j=1}^{i-1} \{\gamma_1(i-j)\}\{\exp(\beta_1 T_j) - \exp(\beta_1 T_{j-1})\}\Big] \end{aligned}$$

such that we can generate a realization $t_1 < \cdots < t_n$ under the self-correcting model on $[0,1]$ as follows.

1. Generate a draw from a Poisson distribution with rate $\rho^*_{\widehat{\boldsymbol{\theta}}_1}(t) = \exp\{\widehat{\alpha}_1 + \widehat{\beta}_1 t_n\}$, where $t_n = 1$ to obtain the number of possible arrival times n^*.
2. For $i = 1, \ldots, n^*$, generate a draw from the uniform distribution on $(0, t_n)$.
3. Order the uniform draws such that $u_1 < \cdots < u_{n^*}$ to obtain the candidate arrival times $t^*_1, \ldots, t^*_{n^*}$.
4. For $i = 1, \ldots, n^*$, calculate the acceptance ratio

$$r^*_i = \frac{\lambda^*_{\widehat{\boldsymbol{\theta}}_1}(t^*_i)}{\exp\{\widehat{\alpha}_1 + \widehat{\beta}_1 t_n\}},$$

then generate a draw u^*_i from the uniform distribution on $(0,1)$ and accept t^*_i if $u^*_i < r^*_i$.

5. Given the set of accepted arrival times $\{t_i\}_{i=1}^n$, generate points $\boldsymbol{y}_i$ from the density $h^*_{\widehat{\boldsymbol{\theta}}_2,i}(\boldsymbol{y}_i)$ using rejection sampling with acceptance probability derived the unnormalized component of the spatial interaction in equation (3.3) and a uniform proposal distribution over $\mathcal{S}$.
6. Given the set of $\{t_i, \boldsymbol{y}_i\}_{i=1}^n$, thin the process such that each point is kept with probability $f^*_{\widehat{\boldsymbol{\theta}}_4}(t_i, \boldsymbol{y}_i)$ and the kept points are a realization of the spatio-temporal process.
7. For $i = 1, \ldots, n$, given $(t_i, \boldsymbol{y}_i, \boldsymbol{z}(\boldsymbol{y}_i))$, generate x_i from the density $g^*_{\widehat{\boldsymbol{\theta}}_3}(x_i \mid t_i, \boldsymbol{y}_i, \boldsymbol{z}(\boldsymbol{y}_i))$.

This modeling approach allows us to employ interpretable and computationally tractable likelihood methods for estimating and simulating from the spatio-temporal process and provides a high degree of flexibility when selecting and training a model for the conditional mark distribution as discussed in the following section.

3.3 Package structure

In this section, we introduce the core functionality for the **ldmppr** package and detail the key functions associated with spatio-temporal process and mark model estimation, model goodness-of-fit evaluation, simulation, and visualization. We note that the package is currently designed for working with marked point processes that can be mapped onto spatio-temporal processes where the pattern is characterized by regularity and the mark distribution is dependent on location specific covariate information.

3.3.1 Workflow

We begin by describing the standard workflow that we envision for using the package. We decompose the task of working with marked point processes into a handful of straightforward and manageable steps supported by an intuitive set of functions. The process may be outlined as follows and steps 1 & 2 may be performed in either order.

1. Estimate the parameters of a self-correcting point process given a reference dataset.
2. Train a mark model given the reference data set and topographic covariate surfaces in the form of rasters.
3. Check the fit of the estimated model using various non-parametric summaries for point processes and global envelope tests.

4. Simulate and visualize datasets from the fitted model.

We anticipate that users of the package will have a point process that they are interested in investigating, and we provide the tools to facilitate that exploration. We also note the modular structure of the package such that a user may provide their own estimated mark model in lieu of one of the currently available options in the package. In the remaining portions of this section, we detail the key functionality for each step given above.

3.3.2 Self-correcting model estimation

The first step in the model estimation procedure is to estimate the parameters of the self-correcting model detailed in Section 3.2.3 that captures the spatio-temporal process in the data. Given a reference dataset, we must define a mapping between our initial mark (i.e., size) and the arrival time in the process using the `power_law_mapping()` function. This function allows the user to specify how the marks are mapped to arrival times using the transformation given in equation (3.1) by specifying the value of δ in the `delta` argument. Once a mapping is established, we may proceed with estimating the parameters of the self-correcting process using the `estimate_parameters_sc()` function, where the arguments of the function are defined as follows.

Argument	Description
`data`	a matrix or data frame of times and locations (time, x, y).
`x_grid`	a vector of grid values for x.
`y_grid`	a vector of grid values for y.
`t_grid`	a vector of grid values for time.
`parameter_inits`	a vector of parameter initialization values.
`upper_bounds`	a vector of upper bounds for time, x, and y.
`opt_algorithm`	the NLopt algorithm to use for optimization.
`nloptr_options`	a list of named options for `nloptr` including `maxeval`, `xtol_rel`, `maxtime`.
`verbose`	`TRUE` or `FALSE`, whether to show optimization progress.

This function makes use of the R implementation of the NLopt library (Johnson, 2008) and returns an `nloptr` object from which the optimal solution may be obtained. The optimal choice of algorithm for the `nloptr()` function may depend on the dataset, but the default option is

`NLOPT_LN_SBPLX`, which is an implementation of the "Subplex" algorithm introduced by Rowan (1990) that incorporates explicit bound constraints. This algorithm is a more efficient and robust implementation of the original Nelder-Mead algorithm (Nelder and Mead, 1965). We have also found the "BOBYQA" algorithm of Powell (2009), which optimizes a bound-constrained objective function without requiring derivatives by iteratively building a quadratic approximation, to be a capable and efficient alternative. If the user wishes to test a set of different δ values for the mapping function, the `estimate_parameters_sc_parallel()` function may be used instead, making use of parallel computation if available. In practice, we expect the user to have some sense of the appropriate mapping relationship between sizes and arrival times, or some tuning may be necessary to obtain a reasonable fit. See Section 3.4 for an example of the effect of the choice of δ. We also note that the success of the optimization procedure may depend on the choice of initial values for the parameters and the `nloptr` options selected, where utilizing a large number of function evaluations and a reasonable tolerance for convergence may be necessary to obtain a good fit.

3.3.3 Mark model training

The second step in the model estimation procedure is to train the conditional mark model. We use the reference data with the mapped arrival times derived in the self-correcting model estimation step, or an alternate mapping, in concert with a set of covariate surfaces in the form of raster images. The raster images may be pre-processed using the `scale_rasters()` function, or may be provided in their raw form. To train the model, we use the `train_mark_model()` function, where the arguments of the function are provided as follows.

Argument	Description
`data`	a data frame containing named vectors x, y, size, and time.
`raster_list`	a list of raster objects.
`scaled_rasters`	`TRUE` or `FALSE`, whether the raster images have been pre-processed.
`model_type`	the machine learning model type (`xgboost` or `random_forest`).
`xy_bounds`	a vector of domain bounds (2 for x, 2 for y).
`save_model`	`TRUE` or `FALSE`, whether to save the generated model.
`save_path`	the path for saving the generated model.
`parallel`	`TRUE` or `FALSE`, whether to use parallelization in model training.

Argument	Description
`include_comp_inds`	`TRUE` or `FALSE`, whether to generate and use competition indices as covariates.
`competition_radius`	the distance for competition radius if `include_comp_inds` is `TRUE`.
`correction`	the type of correction to apply ("none", "toroidal", or "truncation").
`selection_metric`	the metric to use for identifying the optimal model ("rmse" or "mae").
`cv_folds`	the number of cross-validation folds to use in model training.
`tuning_grid_size`	the size of the tuning grid for hyperparameter tuning.
`verbose`	`TRUE` or `FALSE`, whether to show progress of model training.

The `train_mark_model()` function allows users to select between a random forest or gradient boosted tree model, using the **ranger** (Wright and Ziegler, 2017) and **xgboost** (Chen et al., 2024) engines respectively, where these models are effective at capturing potentially complex non-linear relationships. Users choose a model selection criteria, either root mean squared error (RMSE) or mean absolute error (MAE), and may employ cross validation and hyperparameter tuning using a grid design that optimizes the maximum entropy of the hyperparameter space. The function also allows users to incorporate a collection of interpoint competition metrics at a specified neighborhood size to capture additional trends that are not accounted for by the topographic covariates. The metrics include nearest neighbor distance, number of neighbors, average neighbor distance, nearest neighbor arrival time, sum of neighbor arrival times, and the ratio of nearest neighbor distance and arrival time. For an in depth discussion of competition indices and their construction, see Pommerening and Sánchez Meador (2018) and Contreras et al. (2011). Finally, users may select an edge correction mechanism when training the model to account for the possibility that the reference dataset provided is a subset of a larger dataset and that unobserved points are impacting the observed mark values (i.e., sizes).

3.3.4 Goodness-of-fit checks for the fitted model

Once the self-correcting model parameters are estimated and the mark model is trained, we turn our attention to assessing how well the fitted models capture the dynamics observed in the reference dataset. To accomplish this, we use the `check_model_fit()` function, which provides global envelope tests, using the **GET** package (Myllymäki et al., 2017), for a collection of standard

non-parametric marked point process summary functions. We include the L, F, G, J, E, and V functions as evaluation metrics and a combined global envelope test that provides a p value across the whole set. The L, F, G, and J functions capture the spatial dynamics of the point process, while the E and V functions capture the behavior of the mark process (see Møller and Waagepetersen (2003) and Schlather et al. (2004) for details on the summary functions). The arguments of the `check_model_fit()` function are defined as follows.

Argument	Description
`reference_data`	a `ppp` object for the reference dataset.
`t_min`	the minimum value for time.
`t_max`	the maximum value for time.
`sc_params`	a vector of parameter values $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \gamma_3)$.
`anchor_point`	a vector of (x, y) coordinates of the point to condition on.
`raster_list`	a list of raster objects.
`scaled_rasters`	`TRUE` or `FALSE`, whether the raster images have been pre-processed.
`mark_model`	a model object (typically from `train_mark_model`).
`xy_bounds`	a vector of domain bounds (2 for x, 2 for y).
`include_comp_inds`	`TRUE` or `FALSE`, whether to generate and use competition indices as covariates.
`competition_radius`	the distance for competition radius if `include_comp_inds` is `TRUE`.
`thinning`	`TRUE` or `FALSE`, whether to use the thinned or unthinned simulated values.
`correction`	the type of correction to apply ("none", "toroidal", or "truncation").
`n_sim`	the number of simulated datasets to generate.
`save_sims`	`TRUE` or `FALSE`, whether to save and return the simulated datasets.
`verbose`	`TRUE` or `FALSE`, whether to show progress of model checking.
`seed`	an integer value to set the seed for reproducibility.

The `check_model_fit()` function allows users to simulate a collection of datasets using the estimated parameters from the self-correcting process and trained mark model. The non-parametric summary functions are calculated for each dataset and the global envelopes are obtained across the whole collection of simulated datasets for each metric, as well as a combined test across all metrics. This allows the user to gauge whether the realizations from the estimated location-dependent marked point process accurately reflect the dynamics observed in the reference dataset across a variety of

different metrics. This includes the mark distribution as captured by the E and V functions, which Schlather et al. (2004) define as the conditional mean and variance of the mark associated with a typical random point given that another random point exists at distance r. Each individual metric includes a p value range indicating the compatibility of the reference dataset with the simulated datasets, and the combined test provides a single p value across the entire set of metrics allowing users to quickly gauge the overall fit of the model.

3.3.5 Simulation and visualization

When a user is satisfied with the estimated location-dependent marked point process obtained in the model estimation, training, and goodness of fit checking steps, they can proceed with simulating realizations from the process and visualizing the results. The `simulate_mpp()` function provides an efficient way to generate a realization from the process, where the arguments for the function are as follows.

Argument	Description
`sc_params`	a vector of parameter values $(\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \alpha_3, \beta_3, \gamma_3)$.
`t_min`	the minimum value for time.
`t_max`	the maximum value for time.
`anchor_point`	a vector of (x, y) coordinates of the point to condition on.
`raster_list`	a list of raster objects.
`scaled_rasters`	`TRUE` or `FALSE`, whether the raster images have been pre-processed.
`mark_model`	a model object (typically from `train_mark_model`).
`xy_bounds`	a vector of domain bounds (2 for x, 2 for y).
`include_comp_inds`	`TRUE` or `FALSE`, whether to generate and use competition indices as covariates.
`competition_radius`	the distance for competition radius if `include_comp_inds` is `TRUE`.
`correction`	the type of correction to apply ("none", "toroidal", or "truncation").
`thinning`	`TRUE` or `FALSE`, whether to thin the realization.

The function returns a `list` object containing the the simulated realization in a `ppp` object format as well as a `data frame`. With a realization of the process in hand, the user can easily visualize the marked point process object using the `plot_mpp()` function, which has the following arguments.

Argument	Description
`mpp_data`	a `ppp` object with marks or a data frame with columns (`x`, `y`, `size`).
`pattern_type`	the type of pattern to plot ("reference" or "simulated").

In the following section, we walk through the entire workflow described in Section 3.3.1 with an example forestry dataset.

3.4 Application

Equipped with an understanding of the package workflow and primary functionality, we now provide an example of using **ldmppr** to analyze a forest stand dataset comprised of canopy volumes (in cubic meters) and locations for conifer species in the Southern Rocky Mountains, obtained from Drew et al. (2024), that is included in the package. We incorporate four topographic covariates, in the form of raster surfaces, that have been previously found to be related to the processes of tree growth and that capture key environmental conditions like energy and water availability (Drew et al., 2025). The covariates included in this analysis are Southness Aspect, Topographic Wetness Index, Elevation, and Slope which are derived from a LiDAR based digital elevation model (DEM).

Following the steps outlined in Section 3.3, we begin by estimating the self-correcting model using $\delta = 1$ as the size to time mapping parameter. We specify the mapping as follows.

```
data("medium_example_data")
parameter_estimation_data <- medium_example_data %>%
  dplyr::mutate(time = power_law_mapping(size = size, delta = 1)) %>%
  dplyr::select(time, x, y)
```

Next, we define the grid values for the spatial and temporal components of the process, the initial parameter values, and the upper bounds for the optimization procedure.

```
x_grid <- seq(0, 50, length.out = 10)
y_grid <- seq(0, 50, length.out = 10)
t_grid <- seq(0, 1, length.out = 10)
parameter_inits <- c(2.5, 4.9, .015, 1.5, 10.5, .9, 3, .25)
upper_bounds <- c(1, 50, 50)
```

We estimate the parameters of the self-correcting process using the `estimate_parameters_sc()` function and obtain the optimal parameter estimates.

```
estimated_sc <- estimate_parameters_sc(
  data = parameter_estimation_data,
  x_grid = x_grid,
  y_grid = y_grid,
  t_grid = t_grid,
  parameter_inits = parameter_inits,
  upper_bounds = upper_bounds,
  opt_algorithm = "NLOPT_LN_BOBYQA",
  nloptr_options = list(
    maxeval = 300,
    xtol_rel = 1e-3
  ),
  verbose = FALSE
)
optimal_parameters <- estimated_sc$solution
cat(optimal_parameters)

#> 2.485842 5.051634 0.0163554 1.55796 11.38109 1.789767 2.992003 0.2333383
```

We obtain the estimated parameter set ($\alpha_1 = 2.4858$, $\beta_1 = 5.0516$, $\gamma_1 = 0.0164$, $\alpha_2 = 1.558$, $\beta_2 = 11.3811$, $\alpha_3 = 1.7898$, $\beta_3 = 2.992$, $\gamma_3 = 0.2333$) using the "BOBYQA" algorithm, described in Section 3.3.2. In practice, we recommend increasing the number of iterations and reducing the fractional tolerance level as time and resources allow to facilitate the best fit from the selected algorithm.

In addition to estimating the self-correcting process, we need to train the conditional mark model for canopy volume (m^3) given the arrival times, locations, location specific topographic covariates, and interpoint competition metrics in a 10 m neighborhood. We begin by loading the example raster images and preparing the model training data.

```
raster_paths <- list.files(system.file("extdata", package = "ldmppr"),
                           pattern = "\\.tif$", full.names = TRUE)
```

```
raster_paths <- raster_paths[grepl("_med\\.tif$", raster_paths)]
rasters <- lapply(raster_paths, terra::rast)
scaled_rasters <- scale_rasters(rasters)
model_training_data <- medium_example_data %>%
  dplyr::mutate(time = power_law_mapping(size, 1))
```

Next, to train the mark model, we opt for a gradient boosted tree model using the **xgboost** engine with 5-fold cross-validation and a hyperparameter tuning grid of size 200.

```
example_trained_mark_model <- train_mark_model(
  data = model_training_data,
  raster_list = scaled_rasters,
  scaled_rasters = TRUE,
  model_type = "xgboost",
  xy_bounds = c(0, 50, 0, 50),
  parallel = TRUE,
  include_comp_inds = TRUE,
  competition_radius = 10,
  correction = "none",
  selection_metric = "rmse",
  cv_folds = 5,
  tuning_grid_size = 200,
  verbose = FALSE
)
```

With the estimated parameters for the self-correcting process and a trained conditional mark model in hand, we use the `generate_mpp()` function to create a marked point process object for the reference dataset and define the anchor point M_n (i.e., the largest tree in the domain) to condition the process on.

```
reference_data <- generate_mpp(
  locations = medium_example_data[, c("x", "y")],
  marks = medium_example_data$size,
  xy_bounds = c(0, 50, 0, 50)
)
```

```
M_n <- as.matrix(medium_example_data[1, c("x", "y")])
```

We now assess how well the estimated process reflects the dynamics in the original dataset using the `check_model_fit()` function. In order to obtain valid p values at the $\alpha = .05$ level for the global envelope tests, Myllymäki et al. (2017) recommend using a minimum of 2500 realizations from the process.

```
example_model_fit <- check_model_fit(
  reference_data = reference_data,
  t_min = 0,
  t_max = 1,
  sc_params = optimal_parameters,
  anchor_point = M_n,
  raster_list = scaled_rasters,
  scaled_rasters = TRUE,
  mark_model = example_trained_mark_model$raw_model,
  xy_bounds = c(0, 50, 0, 50),
  include_comp_inds = TRUE,
  thinning = TRUE,
  correction = "none",
  competition_radius = 10,
  n_sim = 2500,
  save_sims = FALSE,
  verbose = FALSE,
  seed = 90211
)
```

Once we have run the `check_model_fit()` function, we can plot the combined global envelope test, or any of the individual tests for the summary functions (L, F, G, J, E, or V), as seen in Figure 3.1.

```
plot(example_model_fit$combined_env)
```

In this example, the estimated model does not actually provide an adequate fit to the reference dataset as evidenced by the small p value and the number of points from the reference data that fall

outside of the simulated envelopes (as highlighted in red). In particular, the estimated empty space function, $F(r)$, from the original data is not well captured by realizations from the fitted process. The $F(r)$ function is the cumulative distribution function of the distance from a fixed point to the nearest point in the process (Ripley, 1988). We also note that the function $V(r)$, which captures the variance of the mark associated with a typical random point given that another random point exists at distance r, is not well captured by the fitted process. This suggests that the estimated mark model may not be capturing the dynamics of the mark process well, or that the self-correcting process is not adequately capturing the spatial dynamics of the reference dataset.

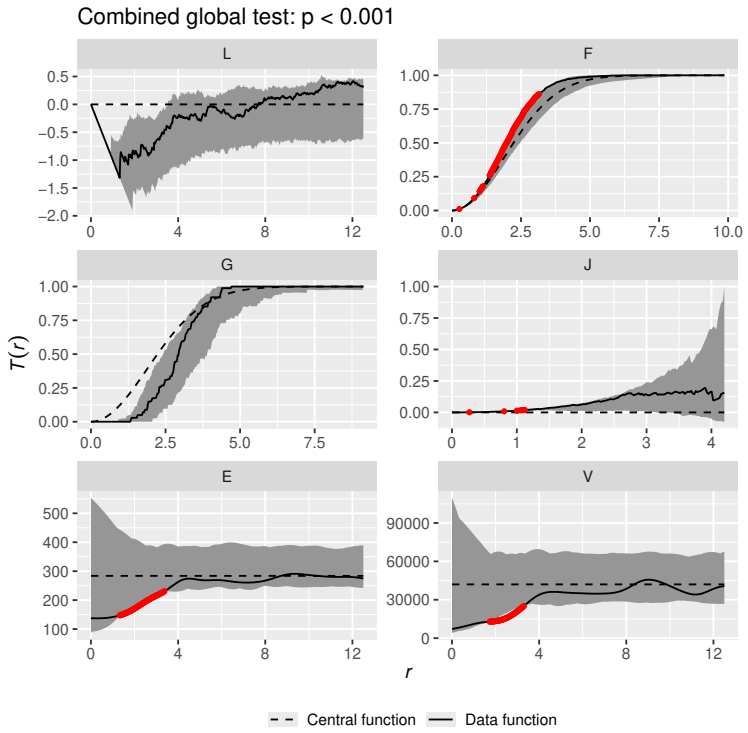

Figure 3.1: Combined global envelope test for the realizations from the fitted process. Solid black lines represent the reference process, dashed black lines represent a homogeneous Poisson process (or CSR), and the colored band represents the global envelope for the simulated datasets at the $\alpha = .05$ level. Reference values outside the envelope are highlighted in red and suggest a poor fit.

This result suggests that we may need to try an alternative value for δ in the size to arrival time mapping, or adjust the hyperparameters for the self-correcting model estimation step when using the `estimate_parameters_sc()` function.

Given that we received a less than optimal fit with our initial modeling attempt, we revisit both the self-correcting process and mark models to see if we can improve them. A standard approach for improving estimation of the self-correcting process is to increase the granularity of the grid used to estimate the model, while increasing the number of iterations and potentially reducing the tolerance threshold. We refit the model with a substantially finer grid below, using the same bounds and initial values for the model parameters as in the first fit.

```
x_grid <- seq(0, 50, length.out = 35)
y_grid <- seq(0, 50, length.out = 35)
t_grid <- seq(0, 1, length.out = 40)
estimated_sc_update <- estimate_parameters_sc(
  data = parameter_estimation_data,
  x_grid = x_grid,
  y_grid = y_grid,
  t_grid = t_grid,
  parameter_inits = parameter_inits,
  upper_bounds = upper_bounds,
  opt_algorithm = "NLOPT_LN_BOBYQA",
  nloptr_options = list(
    maxeval = 300,
    xtol_rel = 1e-4
  ),
  verbose = FALSE
)
```

We obtain the following improved parameter estimates from the optimization.

```
improved_optimal_parameters <- estimated_sc_update$solution
cat(improved_optimal_parameters)
```

```
#> 2.342867 5.120508 0.01371552 1.524611 10.70785 1.724994 3.15573 0.230967
```

We see that the improved parameter estimates set ($\alpha_1 = 2.3429$, $\beta_1 = 5.1205$, $\gamma_1 = 0.0137$, $\alpha_2 = 1.5246$, $\beta_2 = 10.7079$, $\alpha_3 = 1.725$, $\beta_3 = 3.1557$, $\gamma_3 = 0.231$) has shifted compared to the initial estimates.

Next, we retrain the mark model using two changes. First, we increase the size of the hyperparameter tuning grid to 300 to better explore the hyperparameter space and increase the number of cross-validation folds from 5 to 10, which typically reduces the variance of performance estimates to provide a more stable estimate of model generalization. Second, we adjust the size-time mapping relationship to use $\delta = .8$ instead of the $\delta = 1$ value that we used previously. We update the mapping in order to better capture the relationship between the marks and arrival times in the process, though we maintained the value of $\delta = 1$ for estimating the self-correcting process. This highlights the possibility that the mapping may need to be adjusted between the two processes to obtain the best fit to the data. We then retrain the mark model using the updated training data and adjusted model specification.

```
improved_model_training_data <- medium_example_data %>%
  dplyr::mutate(time = power_law_mapping(size, .8))
improved_example_trained_mark_model <- train_mark_model(
  data = improved_model_training_data,
  raster_list = scaled_rasters,
  scaled_rasters = TRUE,
  model_type = "xgboost",
  xy_bounds = c(0, 50, 0, 50),
  parallel = TRUE,
  include_comp_inds = TRUE,
  competition_radius = 10,
  correction = "none",
  selection_metric = "rmse",
  cv_folds = 10,
  tuning_grid_size = 300,
  verbose = FALSE
)
```

We proceed by checking the fit of our model using the improved parameter estimates for the self-correcting process and the retrained mark model. We use the same reference dataset and raster images as before, and again simulate 2500 realizations from the process to assess the fit of the updated model.

```
improved_example_model_fit <- check_model_fit(
  reference_data = reference_data,
  t_min = 0,
  t_max = 1,
  sc_params = improved_optimal_parameters,
  anchor_point = M_n,
  raster_list = scaled_rasters,
  scaled_rasters = TRUE,
  mark_model = improved_example_trained_mark_model$raw_model,
  xy_bounds = c(0, 50, 0, 50),
  include_comp_inds = TRUE,
  thinning = TRUE,
  correction = "none",
  competition_radius = 10,
  n_sim = 2500,
  save_sims = FALSE,
  verbose = FALSE,
  seed = 90211
)
```

Next, we check the combined global envelope test for the updated model to assess how well the estimated model captures the dynamics of the reference dataset.

```
plot(improved_example_model_fit$combined_env)
```

We see that the improved model provides a better fit to the reference dataset, as seen in Figure 3.2. Notably, the simulation envelopes contain the reference pattern across all six metrics. In addition, the p value for the combined global envelope test is $p = 0.303$, indicating that the estimated process is an improved fit to the reference dataset compared to the fit that we had initially achieved.

When evaluating the fit of a model, in addition to the non-parametric summary statistics and global envelope tests, it may also be useful to perform a visual comparison of a realization from the model and the reference dataset. To assess the agreement between the improved fitted model

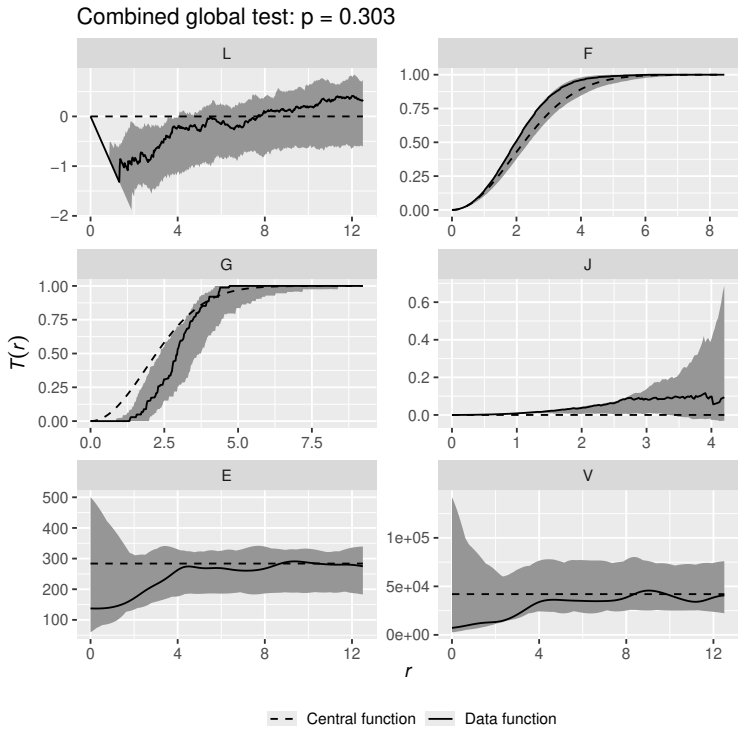

Figure 3.2: Combined global envelope test for the realizations from the improved fitted process. As in Figure 3.1, solid black lines represent the reference process, dashed black lines represent a homogeneous Poisson process (or CSR), and the colored band represents the global envelope for the simulated datasets at the $\alpha = .05$ level.

and the reference data, we simulate a realization from the improved model and compare it to the original reference dataset and a dataset simulated from the initial (poorly fitting) fitted model.

```
improved_simulated_mpp <- simulate_mpp(
  sc_params = improved_optimal_parameters,
  t_min = 0,
  t_max = 1,
  anchor_point = M_n,
  raster_list = scaled_rasters,
  scaled_rasters = TRUE,
  mark_model = improved_example_trained_mark_model$raw_model,
  xy_bounds = c(0, 50, 0, 50),
  include_comp_inds = TRUE,
  competition_radius = 10,
  correction = "none",
```

```
  thinning = TRUE
)
initial_simulated_mpp <- simulate_mpp(
  sc_params = optimal_parameters,
  t_min = 0,
  t_max = 1,
  anchor_point = M_n,
  raster_list = scaled_rasters,
  scaled_rasters = TRUE,
  mark_model = example_trained_mark_model$raw_model,
  xy_bounds = c(0, 50, 0, 50),
  include_comp_inds = TRUE,
  competition_radius = 10,
  correction = "none",
  thinning = TRUE
)
ref_plot <- plot_mpp(
  mpp_data = reference_data,
  pattern_type = "reference"
)
improved_sim_plot <- plot_mpp(
  mpp_data = improved_simulated_mpp$mpp,
  pattern_type = "simulated"
)
initial_sim_plot <- plot_mpp(
  mpp_data = initial_simulated_mpp$mpp,
  pattern_type = "simulated"
)
```

Figure 3.3 demonstrates that the improved model provides a more accurate representation of the reference dataset than we obtained from the initial model in terms of the spatial process and the conditional mark process. We also include histograms of the realized mark distributions to highlight the improvement in replicating the mark distribution of the reference process. This visualization

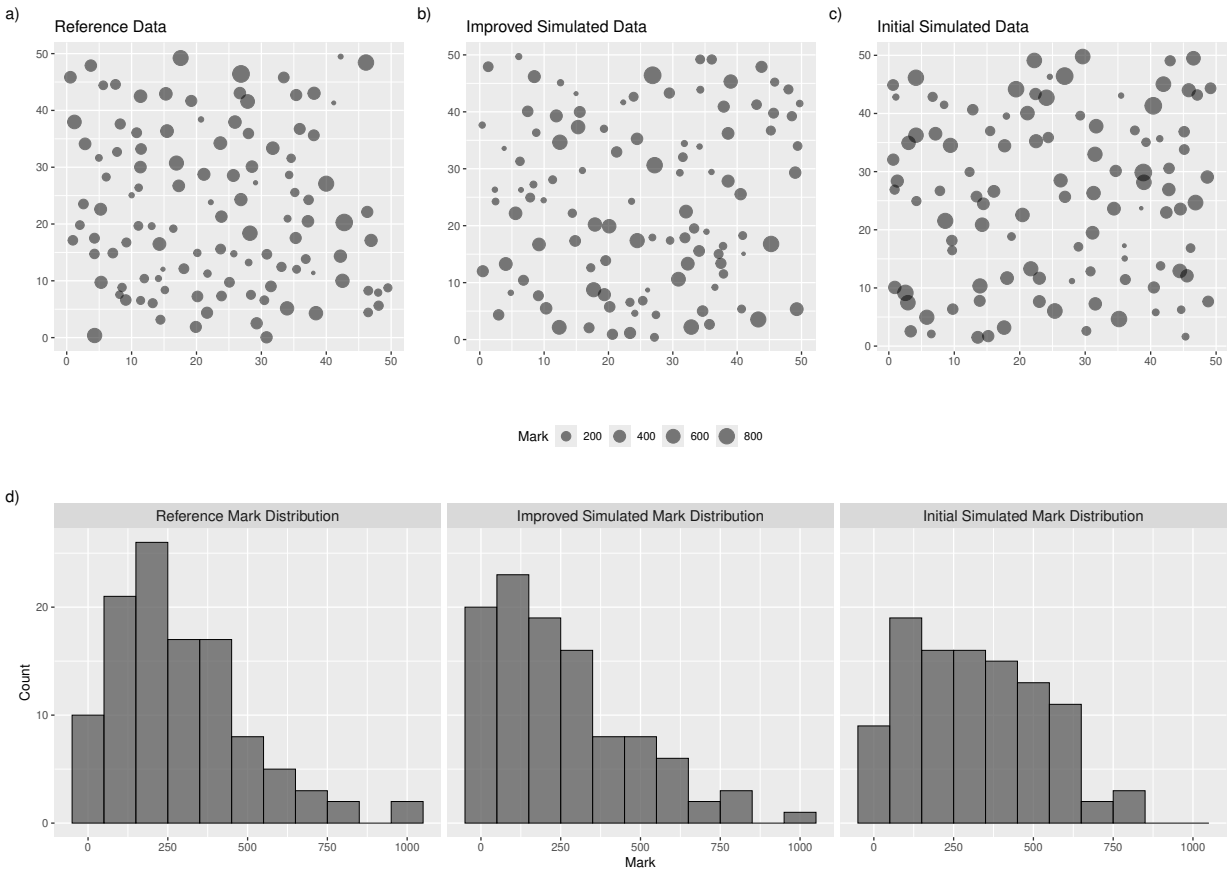

Figure 3.3: Plots (a) - (c) provide a comparison of a realization from the improved estimated process with the original reference dataset and a realization from the initial estimated process. Plot d) shows the corresponding observed mark distributions from the reference dataset and the simulation realizations.

provides additional evidence that the improved model captures the spatial and mark dynamics of the reference dataset more effectively, which results in the improved model being a better fit. This example highlights the utility of the package for working with a real marked point process of interest, and provides the intuition for using the main functionality of **ldmppr** to estimate, assess model fit, and simulate from a marked spatial point process.

3.5 Discussion

In this paper, we provide a novel framework for estimating location dependent marked point processes and introduce the **ldmppr** package, which contains a user-friendly modular suite of tools for model estimation, evaluation, simulation, and visualization for marked point processes with location dependence characterized by regularity in the spatial pattern. We outlined the typical

workflow for using the package and discussed the key functions and their arguments in detail before providing an example of using the package with a real forestry dataset. **ldmppr** simplifies the process of working with marked point processes and provides a likelihood-based estimation approach that is computationally feasible. While the framework presented applies to a broader range of marked point processes, the package in its current implementation is still somewhat limited in the types of patterns that it can address and may require some experimentation and iteration on the part of the user to obtain a satisfactory fit. As the package continues to develop, we would like to incorporate an additional model that can address point process data that demonstrates clustering behavior, as opposed to regularity, while still maintaining the focus on location dependent marks.

Chapter 4

Inferring Tree Growth from Linked Multi-temporal Remote Sensing Data with Exact Error Propagation at Scale

4.1 Introduction

The proliferation of remote sensing technologies, such as light detection and ranging (LiDAR) and structure-from-motion (SfM) photogrammetry, has revolutionized the field of ecology (Hyyppä et al., 2008; Lefsky et al., 2002). These technologies have facilitated the study of tree demography at unprecedented scales, allowing researchers to investigate growth dynamics of individual trees and the impact of changing environmental factors on tree health and physiological responses (Chave et al., 2005). However, the analysis of multi-temporal remote sensing data presents unique challenges, particularly in the identification of unique individuals across multiple scans. The presence of measurement error and changes in tree morphology and environmental conditions over time can complicate the process of linking individual trees across time points, leading to potential biases in downstream analyses (Dalponte and Coomes, 2016; Coomes et al., 2017).

In this chapter, we introduce a Bayesian joint modeling approach for simultaneous identification of unique individuals across multiple SfM-derived datasets and estimation of individual tree growth as a function of topographic, spatio-temporal, and individual-specific covariates. We incorporate a mix of covariates that capture environmental conditions related to water and energy availability, such as snowpack persistence, growing degree days, and a topographic wetness index, as well as topographic conditions, like elevation. Our approach builds upon the two-stage modeling framework presented by Drew et al. (2025), which utilized a record linkage model to identify unique individuals across bi-temporal LiDAR scans paired with an individual tree growth model. This approach provides robust uncertainty propagation from the record linkage to the downstream model through a generalization of the linkage-averaging (LA) approach of Sadinle (2018). Two-stage modeling pipelines are common in the record linkage literature, and refer to a schematic in which the record linkage step is performed independently of the downstream modeling task. Due to the high

computational cost associated with fitting Bayesian record linkage models at scale, the two-stage approach is often preferred in practice, as it allows for the record linkage model to be fit once and the output used as input for a variety of downstream tasks (Kaplan et al., 2022). This provides practitioners with a high degree of flexibility in the choice of downstream task, and there have been many efforts to improve the scalability of these models (Marchant et al., 2021; Taylor et al., 2024; Murray, 2015). While the two-stage approach provides a flexible framework for utilizing record linkage, it crucially does not allow for feedback between the linkage and the downstream task. The disconnection between the record linkage and downstream task can result in a loss of information and suboptimal propagation of uncertainty between the two stages of the modeling pipeline. Numerous methods have been proposed to improve the treatment of uncertainty propagation in two-stage approaches, however the linkage is always uninformed by the downstream task and potentially leaves improved performance on the table.

In contrast, a joint (single-stage) modeling approach allows for feedback between the linkage and the downstream task, providing an integrated framework that directly incorporates uncertainty from the linkage into the downstream modeling objective. While recovering unique identifiers for individuals is a core goal of record linkage, in certain settings it is not merely bookkeeping. For example, when modeling individual tree growth, the downstream estimand is only defined once a unique identifier for each individual tree is established as growth requires linking the same individual across multiple time points. A joint modeling approach allows for the estimation of individual tree growth parameters while simultaneously identifying unique individuals across multiple time points, providing a more comprehensive understanding of tree growth dynamics over time and reducing the potential for bias in the growth estimates due to incorrectly linked individuals. Historically, joint modeling approaches have been used successfully in the record linkage literature but have tended to focus on applications involving only two sets of records as they can be difficult to generalize due to their additional complexity and computational overhead (Gutman et al., 2013; Hof et al., 2017). The feedback between the linkage and the downstream task can also make it difficult to reuse the estimated linkage structure for alternative downstream tasks since the linkage is estimated conditional on a specific downstream objective. However, from an ideological perspective, the joint modeling approach enables exact uncertainty quantification across the entire modeling pipeline, and

Steorts et al. (2018) demonstrate that coupling can improve both linkage quality and downstream task performance relative to a two-stage approach. Joint modeling is thus particularly well-suited for applications where the linkage and downstream task are closely connected, as in tree demography, where identifying unique individuals is a prerequisite for estimating individual tree growth.

Motivated by the investigation of individual tree growth dynamics, we extend the bi-temporal spatial record linkage model introduced by Drew et al. (2025) to incorporate a latent clustering structure that relates records across an arbitrary number of time points. This modeling approach provides an integrated framework for analyzing multi-temporal remote sensing data, and is capable of resolving links both within and across files. We provide the formulation for a joint Bayesian hierarchical model that enables efficient estimation of the linkage structure and individual tree growth parameters under two variations of prior for the dependence structure in the model. We pair the spatial record linkage model with a reframed version of the Michaelis–Menten growth model employed by Drew et al. (2025), utilizing an ordinary differential equation (ODE) that describes the growth dynamics of individual trees over time. Bayesian hierarchical models incorporating mechanistic ODE models have been used in modeling animal movement (Hanks et al., 2011) and disease spread (Cook et al., 2023). However, to our knowledge this approach has not been employed previously in the context of temporal record linkage. In this chapter, we model the underlying individual tree canopy growth as a mechanistic ODE discretized over annual intervals, and embed that mechanism within the downstream growth model component of our joint model. This approach enables joint estimation of the latent growth, linkage, and measurement error process parameters via Markov chain Monte Carlo (MCMC), allowing us to share information across individuals and to properly propagate uncertainty from the mechanistic process into higher-level inference. We approximate the solution to the ODE using the Euler method, which allows us to generalize the model to an arbitrary number of observed time points and incorporate the growth model into a Bayesian hierarchical framework that is amenable to efficient estimation using MCMC methods. This method allows us to capture the temporal dependence structure for the observed canopy volumes associated with a unique individual across multiple time points, and enables easy imputation of missing data within the estimated clusters. Combining these two components, we introduce a

computationally efficient joint modeling framework that provides robust uncertainty quantification across the entire modeling pipeline.

There is a budding literature on temporal record linkage, which is primarily concerned with the problem of identifying unique individuals across multiple time points where differences between the records observed at different time points may be due to changes in the underlying individual, or due to noise in the data (Abdel Monem et al., 2025; Nanayakkara et al., 2018). As Li et al. (2011) note, the temporal record linkage problem is distinct from the traditional record linkage problem in that it requires the incorporation of temporal information into the matching process to account for the fact that the same individual may have different attributes at different time points. This is particularly important in the context of tree demography, where individual tree growth is a function of both spatial and temporal covariates, and the development of a unique individual over time can be related to an underlying growth process. Our proposed joint modeling approach incorporates both static information (i.e., location) and dynamic information (i.e., change in volume over time) into the record linkage process, allowing us to capture the temporal dynamics of individual tree growth while simultaneously identifying unique individuals across multiple time points. To our knowledge, this approach is the first temporal record linkage approach that utilizes an ODE as part of the model structure.

The remainder of this chapter is organized as follows. In Section 4.2, we provide a detailed description of the empirical data and ecological hypotheses that motivate our modeling approach. In Section 4.3, we introduce the joint modeling framework, including the record linkage model and the two specifications of the generalized multivariate growth model. We discuss the computational strategies used to fit the model and provide details on the MCMC sampling scheme. In Section 4.4, we present the results of a series of numerical experiments designed to evaluate the performance of the joint modeling approach under a variety of scenarios motivated by the empirical dataset introduced in Section 4.2. We consider the efficacy of the joint modeling approach in terms of the record linkage performance and the coverage rates for individual tree growth parameters. Finally, in Section 4.5, we summarize the key findings of our work and discuss our ongoing research and future work on this topic.

4.2 Empirical data and motivation

In this section, we introduce the empirical data that motivates our joint modeling approach and the motivating hypotheses that we intend to investigate. The data we consider is provided by the Rocky Mountain Biological Laboratory (RMBL), a high-elevation research station in Colorado, USA, and consists of four datasets containing individual tree characteristics from the Gothic Townsite vicinity. The data was collected across a four year period spanning 2021, 2022, 2023, and 2024, and details regarding the collection, structure, and processing of the data are provided in the remainder of this section.

4.2.1 Overview

The dataset contains annual estimates of the size and relative health of approximately 6000 conifer trees growing in a 105.3 ha area surrounding the campus of RMBL. The estimates were derived from radiometric measurements and high-resolution topography and canopy surface models produced using SfM photogrammetry analysis of images collected by an uncrewed aircraft system (UAS) from May 2021 through October 2024. Annual data are derived from four repeat drone flights performed each year during late spring through early fall, with the data from all flights in a given year composited into an annual maximum digital surface model. A digital elevation model (DEM) was subtracted from each annual surface model to create a single maximum canopy height surface of the study area for each year. This surface was then segmented into individual canopies using a region-growing algorithm. Canopies were classified into conifer and deciduous canopies using a high-resolution supervised classification map. To derive estimates of individual size, the canopy height model was processed to calculate crown volume for each individual in each year.

Forests in the study area contain a mix of Engelmann spruce (*Picea engelmannii*), subalpine fir (*Abies lasiocarpa*), and quaking aspen (*Populus tremuloides*). Although spruce and fir crowns are relatively discrete, the crowns of clonal aspen are irregular and closely spaced, making it difficult to identify and segment individual canopies. Because of this, we chose to focus on crowns of the two conifer species. We classified crowns into conifer and aspen categories using an existing 5 cm resolution vegetation classification map generated from 2021 multispectral UAS imagery. For each crown, we computed the proportion of cells classified as aspen or conifer and excluded crowns that

were estimated to be composed of more than 30% aspen pixels. Additional details for the data collection and processing steps are provided in Appendix B.1.

4.2.2 Covariates

Previous studies have shown that individual tree growth in Western North American conifer forests is associated with a variety of factors, including topography, climate conditions, and individual tree characteristics (Buechling et al., 2017; Heilman et al., 2022). We further investigate the findings of Drew et al. (2025), who demonstrated a link between individual tree growth and key water and energy availability proxies. We consider a similar relationship in the Gothic Townsite vicinity, where we have access to a high-resolution multi-temporal dataset. We examine the impact of topographic factors including elevation and topographic wetness index (TWI), which vary spatially over the study area but are constant over time. In addition, we include spatio-temporal environmental conditions derived from gridded climate data that were interpolated from a combination of weather station and microclimate sensors and satellite-derived maps of the persistence of seasonal snowpack. The derived covariates that we include are growing degree days (GDD) and snowpack persistence (SP), which are aggregated on a yearly basis across the study area from 2021–2024, and capture annual changes in energy and water availability, respectively.

In addition to these topographic and spatio-temporal covariates, we also consider the response of individual trees to environmental change, as measured by the normalized difference vegetation index (NDVI). NDVI is a measure between -1 and 1, which characterizes the amount of vegetation growth in an area such that a value of 0 corresponds to no vegetation growth, while vegetation dense areas will have values approaching 1 and values less than zero indicate a lack of dry land (Myneni et al., 1995). NDVI has been shown to be a useful proxy for vegetation health and growth (Pettorelli et al., 2005; Wang et al., 2004), and we anticipate that it may be a meaningful indicator of individual tree growth in the study area and capture a more individualized snapshot than some of the topographic covariates at the spatial scale of our dataset. We obtain NDVI estimates for each individual tree crown identified in each year, providing a yearly measure of individual tree health. Additional details regarding the calculation of yearly NDVI from the SfM data are provided in Appendix B.1.3.

As a final processing step in preparing the data, we identified 10 sub-areas within the study area that are geographically distinct and comprise relatively homogeneous populations of conifers with unique microclimates. These areas were selected based on the presence of a sufficient number of conifer trees and the availability of covariate data. The derived areas may be seen in Figure 4.1. We are interested in assessing whether the growth dynamics of individual trees can be meaningfully captured across these micro-environments, and whether the collection of covariates we have selected are sufficient to identify trends across relatively homogeneous populations of trees. We note that the selection of these areas is not intended to be exhaustive, but rather to provide a representative sample of the Gothic Townsite vicinity that captures the diversity of microclimates and topographic conditions present in the study area. Additional information about the individual areas may be found in Appendix B.1.7.

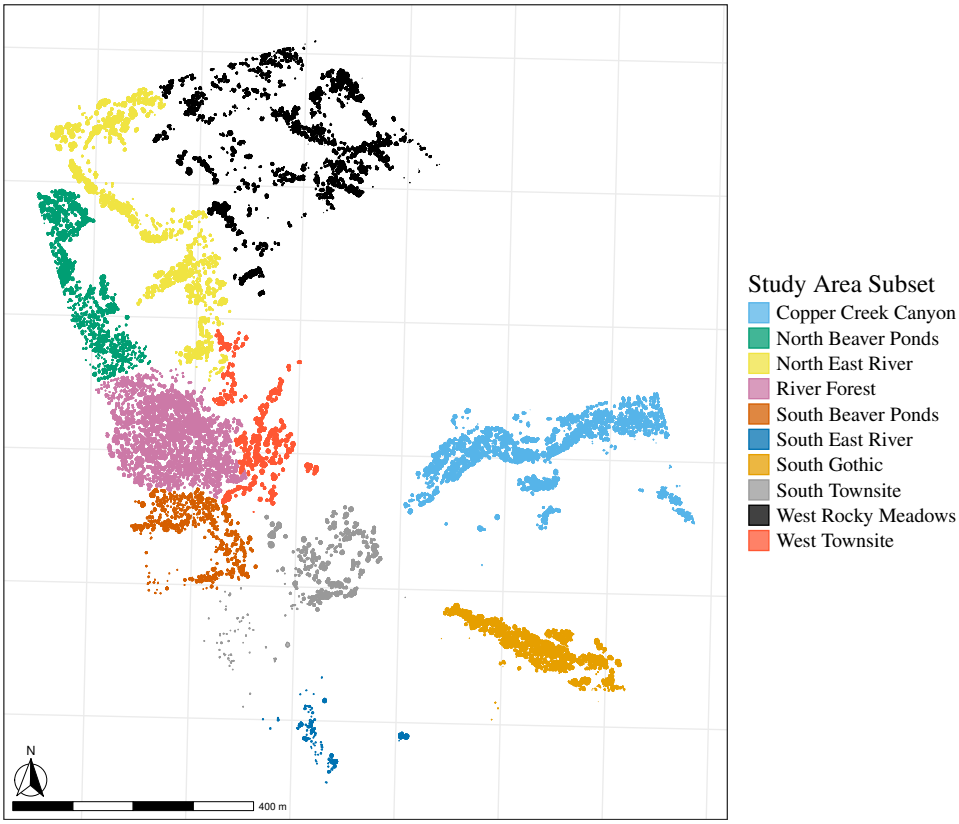

Figure 4.1: Map of the Gothic Townsite vicinity showing the 10 areas selected for our analysis with derived crown geometries across all 4 study years.

4.3 Models and notation

In this section, we define the relevant notation and components of the joint modeling framework. We begin by introducing the notation used to describe the record linkage model, which functions to identify unique individuals across multiple time points. We then introduce the two specifications of the generalized multivariate growth model, which describes the growth dynamics of individual trees over time. We follow this with an explicit formulation of the joint model that combines the record linkage and growth model components. We note that throughout this section, distributions identified with subscripts refer to distributions truncated over the bounds provided. For example, an Inverse-Gamma distribution with parameters c and d that is truncated over the range $[0, b]$ is denoted as Inverse-Gamma$_{[0,b]}(c, d)$. Finally, we discuss the computational methods employed to fit the model at scale.

4.3.1 Record linkage model

For clarity, we present the record linkage component of the joint modeling framework first, as it serves as the foundation for the subsequent growth model. We note that in the context of record linkage, the term file is synonymous with a specific dataset, and we use the term record to refer to an individual observation within a file, and we will use these terms interchangeably. In the context of the Gothic Townsite dataset, a record corresponds to an individual tree crown, and a file corresponds to the set of tree crowns observed in a single year. The spatial record linkage model that we introduce in this section is a modification of the model introduced by Drew et al. (2025), that maintains the latent entity structure of that model, but removes the ability to account for rotation and translation, while including a hit-miss mechanism that allows the locations of observed records to exactly match the latent locations that they are associated. Hit-miss mechanisms were originally introduced by Copas and Hilton (1990) to account for instances in which observed records are exact matches, and we follow the outline of Steorts et al. (2016) in adapting this mechanism to our setting.

These adaptations in the model reflect the structure of our motivating empirical dataset, as introduced in Section 4.2, such that we observe the locations of individual tree crowns at a high spatial resolution, but only up to the resolution of the data collection process (i.e., estimated tree crown locations are precise up to the size of a pixel). Consequently, it is possible to observe the

locations of individual tree crowns at their "true" locations, which we refer to as hits, or at a distorted location, which we classify as a miss. With high-resolution data, the cost of the additional machinery to account for rotation and translation outweighs the benefit, as a large proportion of the records are expected to be hits, and the distortion introduced by the data collection process is expected to be small relative to the scale of the spatial domain of the data.

In this formulation of the spatial record linkage model, the latent field values correspond to the true unobserved locations of the unique individuals (i.e., the locations of the tree bases) that exist across multiple time points. We treat the observed data (i.e., tree crown locations) as potentially noisy versions of the latent field values, and we restrict our attention to error introduced as a result of the data processing and biological mechanisms, which is anticipated to be small relative to the scale of the spatial domain of the data. We specify the necessary notation and model structure as follows.

The model is designed to handle an arbitrary number of files, m, which are indexed by $i = 1, \ldots, m$ with corresponding file size n_i. The records within files are indexed by $j = 1, \ldots, n_i$, such that the total number of records across all files is $n = \sum_{i=1}^{m} n_i$. We denote the observed record j in file i as $\boldsymbol{y}_{ij} \in D$, such that $\boldsymbol{y}_{ij} = (x, y)_{ij}$ is the spatial location of the record in the (x, y)-plane of the spatial domain D. The latent entities are denoted as $\boldsymbol{s}_{j'}$ for $j' = 1, \ldots, N$, where N is the maximum number of unique individuals across all files, and $\boldsymbol{s}_{j'} \in \mathbb{R}^2$ is the true location of the j'-th unique individual in the (x, y)-plane.

We define the linkage structure $\boldsymbol{\Lambda}$, which identifies the latent entity each record is associated with, as a vector of length n such that $\boldsymbol{\Lambda} = \{\lambda_{ij} : i = 1, \ldots, m, j = 1, \ldots, n_i\}$, where $\lambda_{ij} \in \{1, \ldots, N\}$ is an integer identifying the latent entity associated with record j in file i. $\boldsymbol{\Lambda}$ naturally defines a set of clusters $\mathcal{C}(\boldsymbol{\Lambda}) = \{\mathcal{C}_{j'}\}_{j'=1}^{N}$, where each cluster corresponds to a unique $\boldsymbol{s}_{j'}$, and the records associated with that individual are those for which $\lambda_{ij} = j'$ for some $j' \in \{1, \ldots, N\}$ such that $\mathcal{C}_{j'} = \{\lambda_{ij} : \lambda_{ij} = j'\}$. Owing to the structure of the model, we note that clusters may be empty, or may contain multiple records from a single file or multiple files. This highlights the capability of the model to perform both within-file and across-file record linkage, as well as the ability to resolve links between records that are not present in all files. In the context of our empirical data, this means that we can identify unique individuals (i.e., trees) across multiple years of data, even if some

trees are not observed in all years and resolve possible erroneous splitting behavior arising during the canopy segmentation procedure.

We note that previous work in the record linkage literature has demonstrated that latent entity models can be sensitive to the choice of N (Steorts et al., 2016). Addressing the sensitivity to N can be addressed by treating it as a parameter in the model and placing an appropriate prior as in Betancourt et al. (2022a), or by treating N as a hyperparameter and utilizing prior knowledge about the degree of overlap between the files to specify an appropriate value. In this work, we follow the approach of Drew et al. (2025) in treating N as a fixed hyperparameter, and we specify $N = q \times \max(n_i)$, where q is chosen to reflect a conservative estimate of the degree of overlap between the files.

As mentioned previously, this formulation of the model incorporates a hit-miss mechanism that allows the observed record locations to exactly match the latent entity locations that they are associated with up to the resolution of the data collection process. We introduce the latent distortion vector $\boldsymbol{z}$ as a vector of length n such that $\boldsymbol{z} = \{z_{ij} : i = 1, \ldots, m, j = 1, \ldots, n_i\}$, where $z_{ij} \in \{0, 1\}$ is an indicator variable that indicates whether record j in file i is a hit ($z_{ij} = 0$) or a miss ($z_{ij} = 1$). A hit indicates that the observed record location $\boldsymbol{y}_{ij}$ is exactly equal to the latent entity location $\boldsymbol{s}_{\lambda_{ij}}$, while a miss indicates that the observed record location is a distortion of the latent entity location. In high spatial resolution remote sensing data, we expect that the vast majority of records will be hits, such that the distortion probability $\theta \equiv \theta_{ij} = P(z_{ij} = 1)$ is expected to be small. Considering the high-resolution of our data, we assume that this distortion probability is constant across records, though in alternative applications this could be adjusted to reflect some known characteristic of the individual. We note that the distortion vector $\boldsymbol{z}$ is not observed, and we treat it as a latent variable in the model.

We specify our data model, which describes the relationship between the observed records and the latent entities, as follows. We assume that the observed record locations, $\boldsymbol{y}_{ij}$, are generated from a two component mixture distribution corresponding to a hit or miss such that

$$\mathbf{y}_{ij} \mid z_{ij}, \mathbf{s}_{\lambda_{ij}}, \sigma^2, D \sim (1 - z_{ij})\mathbb{I}\{\boldsymbol{y}_{ij} \in B(\boldsymbol{s}_{\lambda_{ij}})\}|B(\boldsymbol{s}_{\lambda_{ij}})|^{-1} + z_{ij}\text{Normal}_{2,[D]}(\mathbf{s}_{\lambda_{ij}}, \sigma^2\mathbf{I}),$$

for $i = 1, \ldots, m$, and where $B(\boldsymbol{s}_{\lambda_{ij}})$ is a bounding shape centered at $\boldsymbol{s}_{\lambda_{ij}}$. The first component of the mixture corresponds to a hit such that $\mathbf{y}_{ij}$ is uniformly distributed over the bounding shape $B(\boldsymbol{s}_{\lambda_{ij}})$, and where $|B(\boldsymbol{s}_{\lambda_{ij}})|^{-1}$ is the normalizing constant of the uniform density over the bounding shape. The inclusion of the bounding shape $B(\boldsymbol{s}_{\lambda_{ij}})$ allows us to account for the fact that the observed record locations are not exact, but rather observed up to the resolution of the data collection process. $B(\boldsymbol{s}_{\lambda_{ij}})$ is typically assumed to be a square or circle with a side length or diameter corresponding to the resolution of the data (i.e., pixel size). The full record linkage model is then specified as follows,

$$
\begin{aligned}
\mathbf{y}_{ij} \mid z_{ij}, \mathbf{s}_{\lambda_{ij}}, \sigma^2, D &\overset{ind}{\sim} (1 - z_{ij})\mathbb{I}\{\boldsymbol{y}_{ij} \in B(\boldsymbol{s}_{\lambda_{ij}})\}|B(\boldsymbol{s}_{\lambda_{ij}})|^{-1} + z_{ij}\text{Normal}_{2,[D]}(\mathbf{s}_{\lambda_{ij}}, \sigma^2\mathbf{I}) \\
\mathbf{s}_{j'} \mid N &\overset{iid}{\sim} \text{Uniform}\,(D^*) \\
\sigma^2 &\sim \text{Inverse-Gamma}_{[0,\sigma^2_{\max}]}\,(c_\sigma, d_\sigma) \\
\lambda_{ij} \mid N &\overset{iid}{\sim} \text{Uniform}\,\{1, \ldots, N\} \\
z_{ij} \mid \theta &\overset{iid}{\sim} \text{Bernoulli}(\theta) \\
\theta &\sim \text{Beta}(a_\theta, b_\theta).
\end{aligned}
$$

We note that the latent locations, $\boldsymbol{s}_{j'}$, are assumed a priori to be uniformly distributed over the spatial domain D^*, which is a slightly increased spatial domain such that $D \subset D^* \subset \mathbb{R}^2$, allowing for the possibility that the true location of an individual may be slightly outside the observed domain D (i.e., the true tree base location is located out of the study area, while the observed crowns are within). The choice of a Uniform prior over D^* for the latent locations provides a high degree of flexibility in the model, allowing the observed locations to maximally inform the posterior distribution of the latent locations. In practice, it would also be possible to specify an alternative prior distribution for the latent locations, though this would require additional assumptions about the underlying distribution of the individuals in the spatial domain and increased computational overhead. Leininger (2014) provides a thorough discussion of point process models in a Bayesian hierarchical context.

4.3.2 Growth model

In this section, we present two specifications of the generalized multivariate growth model that we employ as the downstream model in our joint modeling framework. We extend the generalized Michaelis–Menten style growth model presented in Drew et al. (2025) by reframing it as an ordinary differential equation (ODE) that describes the growth dynamics of an individual over time. This mechanism allows us to utilize the Euler method to approximate the solution of the ODE to establish an underlying latent growth process. We are also able to incorporate a temporal dependence structure between the observed size values, which provides additional flexibility in capturing the growth dynamics of individuals over time. While Michaelis–Menten style models were initially introduced in the context of enzyme kinetics, they have been generalized to describe a wide variety of growth processes (López et al., 2000; Bolker, 2008). Goudar et al. (2004) discuss the framing of these models as ODEs, which we can think of as an analog to dynamic models in the context of ecology and biology that are used to describe the evolution of a system over time (Bolker, 2008).

As mentioned in Section 4.3.1, the linkage structure $\boldsymbol{\Lambda}$ implicitly defines a set of clusters $\mathcal{C}(\boldsymbol{\Lambda}) = \{\mathcal{C}_{j'}\}_{j'=1}^{N}$, where each cluster corresponds to a unique individual (i.e. tree) across multiple time points. In connection to the downstream growth model, we further restrict the set of clusters to those that contain at least one record, which we denote as $\mathcal{C}^G(\boldsymbol{\Lambda}) = \{c : c = j' \ \& \ |\mathcal{C}_{j'}| > 0\}$, such that $\mathcal{C}^G(\boldsymbol{\Lambda}) \subseteq \mathcal{C}(\boldsymbol{\Lambda})$. We refer to $\mathcal{C}^G(\boldsymbol{\Lambda})$ as the set of growth clusters, and we denote the number of growth clusters as $C = |\mathcal{C}^G(\boldsymbol{\Lambda})|$. In the remainder of this section, we index the growth clusters by $c = 1, \ldots, C$, and we denote the records associated with growth cluster c as $\boldsymbol{V}_c = \{V_{c,0}, V_{c,1}, \ldots, V_{c,m-1}\}$, where $V_{c,t}$ is the observed volume of individual c at time t. The number of time points matches the number of files, m, though, we index the time points by $t = 0, 1, \ldots, m-1$ to reflect the fact that the first time point corresponds to the initial time of observation. We note that a growth cluster may contain records from all files, or may contain records from only a subset of files, depending on the linkage structure $\boldsymbol{\Lambda}$. In practice, this means that some individuals may not be observed in all files (i.e., time points) and the growth model must be able to accommodate this missing data structure. We also note the growth clusters have a one-to-one correspondence with a set of latent locations $\boldsymbol{s}_c$ for $c = 1, \ldots, C$, such that $\boldsymbol{s}_c$ is the true location of individual c in

the (x, y)-plane. The latent locations $\boldsymbol{s}_c$ are assumed to be the same as the latent entity locations $\boldsymbol{s}_{\lambda_{ij}}$ for all records associated with growth cluster c.

In the growth model used by Drew et al. (2025), the growth of an individual between two time points is described by a generalized Michaelis–Menten style model, which models volume as a function of location-dependent covariates in a curvilinear relationship, such that

$$g_c = \frac{(\mathbf{x}(\mathbf{s}_c)\boldsymbol{\beta})V_c^{\alpha}}{\gamma^{\alpha} + V_c^{\alpha}},$$

where g_c is the growth of individual c on an annualized basis such that

$$g_c(t) = \frac{V_{c,t+1} - V_{c,t}}{(t+1) - t}.$$

The growth function incorporates location-specific covariates in the vector $\mathbf{x}(\mathbf{s}_c)$, which modulate the maximum asymptote of the curve. The vector $\boldsymbol{\beta}$ contains the linear coefficients associated with the covariates, and includes an intercept term β_0 that represents the baseline growth asymptote. V_c represents the measured volume of the individual at the initial time point. The remaining parameters of the growth model are γ, which determines the half-saturation point of the curve, and α, which controls the curvature of the growth function. Rewriting this model as an ODE, we obtain

$$\frac{dV(\boldsymbol{s})}{dt} = \frac{(\mathbf{x}_t(\mathbf{s})\boldsymbol{\beta})V^{\alpha}}{\gamma^{\alpha} + V^{\alpha}}.$$

While the ODE is not solvable in closed form, we can approximate the solution using the Euler method, which allows us to express the growth of an individual c at time t using a discrete approximation of the form

$$u_{c,t+1} = u_{c,t} + h \cdot \frac{(\mathbf{x}_t(\mathbf{s}_c)\boldsymbol{\beta})u_{c,t}^{\alpha}}{\gamma^{\alpha} + u_{c,t}^{\alpha}}, \tag{4.1}$$

for the latent growth process $u_{c,t}$, where h is the temporal step size. We note that the latent growth process $\boldsymbol{u}_c$ relies on an initial value $u_{c,0}$, which is the volume of the individual at the initial time point, t_0. Given the initial value, covariates, and growth function parameters, we can deterministically compute the trajectory of the latent growth process over time.

Figure 4.2 shows an example growth curve in black, and demonstrates the quality of the Euler approximation at varying step size values, where the step size in our application corresponds to yearly measurements (i.e., $h = 1$). Recognizing that the initial conception of the generalized Michaelis–Menten tree growth model presented by Drew et al. (2025) can be reframed as an ODE, we are able to use the Euler approximation, defined in (4.1), to extend the model to accommodate the change in volume of an individual over time given measurements at m discrete time points $(V_{c,0}, V_{c,1}, \ldots, V_{c,m-1})$ as a function of the underlying latent growth process with some additional error structure. The multi-temporal structure of this modeling approach allows us to incorporate both spatial, temporal, and even spatio-temporal covariates into the model, which can be used to capture the finer scale behavior of individuals across both homogeneous and heterogeneous landscapes. This approach allows us to model growth over multiple time points assuming a latent growth process with temporal dependence and measurement error.

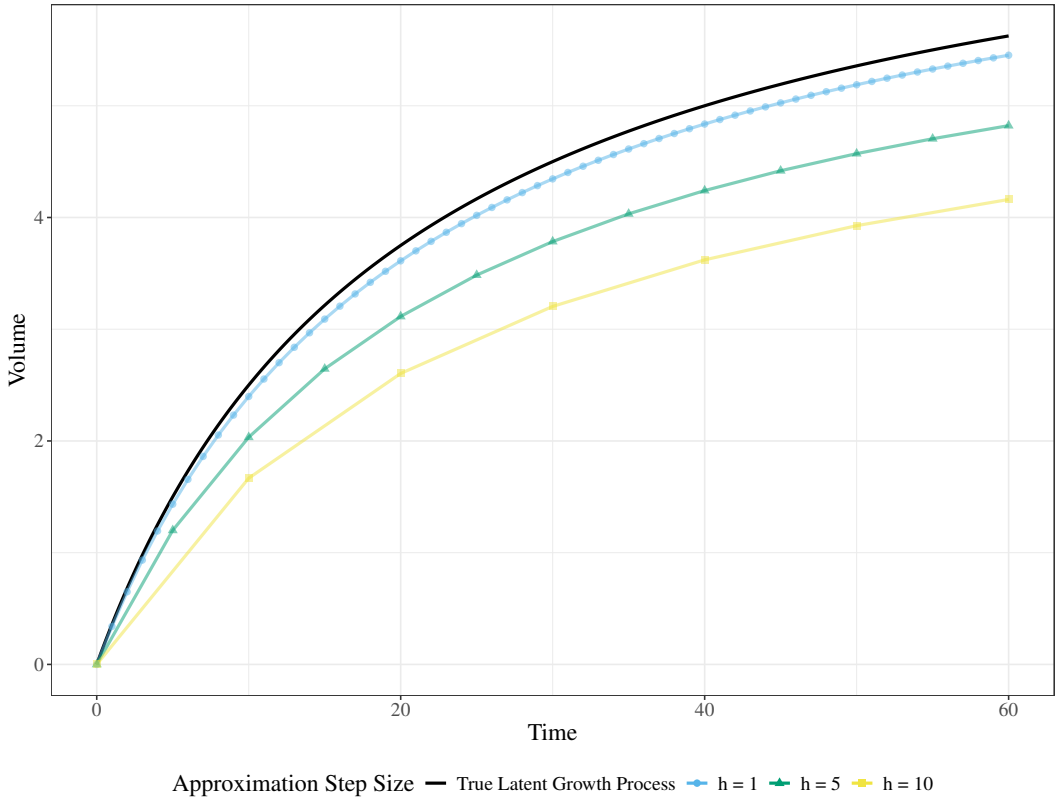

Figure 4.2: Illustration of the Euler approximation of the solution to the ODE. The black line shows the true underlying growth curve, while remaining lines show the Euler approximation for varying step sizes to demonstrate the quality of the approximation for sufficiently small intervals between volume observations.

We present two alternative formulations of the generalized multivariate growth model that are based on the Euler approximation of the Michaelis–Menten ODE. The two model specifications differ in the way that they incorporate the temporal dependence structure between the observed size values and may be framed in terms of a choice of prior specification for the initial value of the latent growth process, $u_{c,0}$. We characterize the two models as the full dependence and reduced dependence models, respectively. The full dependence model jointly the models the observed size values from m files as follows

$$\boldsymbol{V}_c \sim \text{Normal}(\boldsymbol{u}_c, \tau^2\mathbf{I} + \nu^2\boldsymbol{\Sigma}),$$

where $\boldsymbol{V}_c = [V_{c,0}\ V_{c,1}\ \cdots\ V_{c,m-1}]^\mathsf{T}$ is the vector of observed size values for individual c across m time points. The mean structure is given by $\boldsymbol{u}_c = [u_{c,0}\ u_{c,1}\ \cdots\ u_{c,m-1}]^\mathsf{T}$, where $u_{c,0}$ is the initial value of the latent growth process for individual c, and $u_{c,t}$ is the value of the latent growth process at time t following the Euler approximation introduced in (4.1). The covariance structure is given by $\tau^2\mathbf{I} + \nu^2\boldsymbol{\Sigma}$, where τ^2 is the measurement error variance, ν^2 is the variance of the temporal dependence structure, and $\boldsymbol{\Sigma}$ is defined by a squared exponential kernel such that

$$\boldsymbol{\Sigma}_{i,j} = \exp\left(-\frac{|t_i - t_j|^2}{2\rho^2}\right).$$

In this formulation, the covariance matrix $\boldsymbol{\Sigma}$ captures the temporal dependence structure between the observed size values, where ρ is a parameter that controls the range of dependence between the observed size values. The full dependence model allows for the observed size values to be correlated across time points, which is particularly useful when the observed size values are expected to exhibit temporal dependence due to underlying biology and environmental factors. The full dependence model is specified as follows,

$$\begin{aligned}
\boldsymbol{V}_c \mid u_{c,0}, \boldsymbol{\beta}, \alpha, \gamma, \tau^2, \nu^2, \rho, \boldsymbol{X}(\boldsymbol{s}_c) &\overset{ind}{\sim} \text{Normal}(\boldsymbol{u}_c, \tau^2\mathbf{I} + \nu^2\boldsymbol{\Sigma}) \\
u_{c,0} &\overset{ind}{\sim} \text{Normal}(\mu_{u_{0c}}, \sigma^2_{u_0}) \\
\mu_{u_{0c}} &\overset{iid}{\sim} \text{Normal}(\tilde{\mu}, \tau^2_{\mu_{u_0}}) \\
\sigma^2_{u_0} &\sim \text{Inverse-Gamma}(c_{\sigma_{u_0}}, d_{\sigma_{u_0}}) \\
\boldsymbol{\beta} &\sim \text{Normal}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)
\end{aligned}$$

$$
\begin{aligned}
\alpha &\sim \text{Beta}_{[c_\alpha, d_\alpha]}(a_\alpha, b_\alpha) \\
\gamma &\sim \text{Gamma}_{[\gamma_{\min}, \gamma_{\max}]}(c_\gamma, d_\gamma) \\
\tau^2 &\sim \text{Inverse-Gamma}_{[0, \tau^2_{\max}]}(c_\tau, d_\tau) \\
\nu^2 &\sim \text{Inverse-Gamma}_{[0, \nu^2_{\max}]}(c_\nu, d_\nu) \\
\rho &\sim \text{Uniform}(0, r).
\end{aligned}
$$

Under the full dependence model, the dependence structure is shared across all observed size values for an individual, which we posit will allow for the model to capture the temporal dynamics of growth more effectively. The initial value of the latent growth process $u_{c,0}$ is modeled hierarchically with a Normal prior distribution, where $\mu_{u_{0c}}$ is the initial value for individual c, and $\sigma^2_{u_0}$ controls how much $u_{c,0}$ can deviate from its mean. The hyperparameters $\tilde{\mu}$ and $\tau^2_{\mu_{u_0}}$ control the prior distribution of the initial value across individuals. For an empirically-motivated specification, $\tilde{\mu}$ can typically be specified the mean of the observed initial values, $\bar{V_0}$, across all unique individuals, and $\tau^2_{\mu_{u_0}}$ may be specified as the variance of the observed initial values, $\text{Var}(\boldsymbol{V}_0)$. We note that the empirically-motivated specification of the prior for $u_{c,0}$ is not required, but it may help to provide a reasonable starting point for the model that is grounded in the observed data.

In contrast, the reduced dependence model assumes a modified mean and covariance structure for the observed size values, such that $V_{c,0}$ is modeled independently of the remaining observed size values. The vector of observed size values, $\boldsymbol{V}_c = [V_{c,1} \ \cdots \ V_{c,m-1}]^\mathsf{T}$ is then jointly modeled as

$$
\boldsymbol{V}_c \sim \text{Normal}(\boldsymbol{u}_c, \tau^2\mathbf{I} + \nu^2\boldsymbol{\Sigma}),
$$

where $\boldsymbol{u}_c = [u_{c,1} \ u_{c,2} \ \cdots \ u_{c,m-1}]^\mathsf{T}$, again following the recursive Euler approximation for the mean structure such that $u_{c,1}$ is a function of $u_{c,0}$. Instead of modeling $V_{c,0}$ jointly with the other observed sizes, we incorporate it into the prior for $u_{c,0}$ to root the latent growth process in the observed data. The reduced dependence model is specified as follows,

$$
\begin{aligned}
\boldsymbol{V}_c \mid u_{c,0}, \boldsymbol{\beta}, \alpha, \gamma, \tau^2, \nu^2, \rho, \boldsymbol{X}(\boldsymbol{s}_c) &\overset{ind}{\sim} \text{Normal}(\boldsymbol{u}_c, \tau^2\mathbf{I} + \nu^2\boldsymbol{\Sigma}) \\
u_{c,0} \mid V_{c,0} &\overset{ind}{\sim} \text{Normal}(V_{c,0}, \sigma^2_u)
\end{aligned}
$$

$$
\begin{aligned}
V_{c,0} &\overset{ind}{\sim} \text{Normal}(\eta_c, \sigma_V^2) \\
\boldsymbol{\beta} &\sim \text{Normal}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\
\alpha &\sim \text{Beta}_{[c_\alpha, d_\alpha]}(a_\alpha, b_\alpha) \\
\gamma &\sim \text{Gamma}_{[\gamma_{\min}, \gamma_{\max}]}(c_\gamma, d_\gamma) \\
\tau^2 &\sim \text{Inverse-Gamma}_{[0, \tau^2_{\max}]}(c_\tau, d_\tau) \\
\nu^2 &\sim \text{Inverse-Gamma}_{[0, \nu^2_{\max}]}(c_\nu, d_\nu) \\
\rho &\sim \text{Uniform}(0, r).
\end{aligned}
$$

While this formulation of the model does not allow for the initial value of the latent growth process to be explicitly correlated with the subsequent observed size values, it uses an empirically-motivated approach to connect the latent growth process with the observed data by incorporating the initial volume value $V_{c,0}$ into the prior for $u_{c,0}$. An additional trade-off is that we are required to specify a prior for $V_{c,0}$, which depends on the mean parameter η_c and the variance parameter σ_V^2. In general, these two hyperparameters may be specified based on prior knowledge about the expected initial volume of the individuals, or allowed to be weakly informative by increasing the value of σ_V^2 to allow the model to learn the initial volume from the data.

Both model formulations are designed to accommodate missing volume values in the observed data. Missing volume measurements can potentially arise from a variety of scenarios including error in the estimated linkage structure, mortality due to environmental factors like extreme wind or lightning, or as a function of size such that trees near the lower detection threshold for the data processing algorithm may be missing from some subset of scans over time. Given the nature of the data, it is expected that some individuals will not be observed in all time points, and the models are designed to handle this eventuality by treating the missing values as latent variables in the model. Under the assumption that the missing values are missing at random, the models can still learn the growth dynamics of the individuals over time, even in the presence of missing data, and impute the missing values depending on the observed data and the linkage structure.

The primary differences between the full dependence (FD) and reduced dependence (RD) growth models can be attributed to the priors placed on the initial latent growth value $u_{c,0}$ and the covariance structure of the observed volume values. While the full dependence model jointly models

the observed volume values with a shared covariance, the reduced dependence model decouples the initial observed volume measurement $V_{c,0}$ from the subsequent observed volume values. The reduced dependence model is less computationally burdensome, but at the cost of a weaker coupling between the initial and subsequent observed volume values. We note that in both models, $u_{c,0}$ connects the observed volume measurements through the latent growth process. We also note that the full dependence model may struggle from additional identifiability issues when the number of missing data points in a given cluster is large relative to the number of observed data points. We provide a brief summary of the key distinctions between the two downstream model formulations in Table 4.1.

Table 4.1: Key differences between the two downstream growth model variants.

Aspect	Full Dependence Model	Reduced Dependence Model
V_c Likelihood	$(V_{c,0}, \ldots, V_{c,m-1}) \sim N(\boldsymbol{u}_c, \tau^2 I + \nu^2 \Sigma)$	$V_{c,0} \sim N(\eta_c, \sigma_V^2)$; $(V_{c,1}, \ldots, V_{c,m-1}) \sim N(\boldsymbol{u}_c, \tau^2 I + \nu^2 \Sigma)$
$u_{c,0}$ Prior	$u_{c,0} \sim N(\mu_{u_{c,0}}, \sigma_{u0}^2)$	$u_{c,0} \mid V_{c,0} \sim N(V_{c,0}, \sigma_u^2)$
Cov. Dim.	$m \times m$	$(m-1) \times (m-1)$
Best for	Full dependence across all time points	Speed/stability when $\boldsymbol{V}_0$ is accurate
Weaknesses	Heavier MCMC burden, potential identifiability issues if m is large	Loss of shared information between initial and subsequent volume values when data is missing

4.3.3 Joint model specification

From the discussion of the models in Section 4.3.1 and Section 4.3.2, there is a clear relationship between the linkage and downstream modeling objective. The record linkage model identifies the unique individuals across multiple time points, while the growth model describes the growth dynamics of those individuals over time. While a two-stage approach can be used to propagate the uncertainty from the record linkage model into the growth model, this approach does not allow the growth model to inform the linkage. From an ideological perspective, we might prefer to model the linkage and downstream task simultaneously, as this provides exact uncertainty quantification across the entire modeling pipeline and allows the growth model to meaningfully inform the linkage structure. Figure 4.3 illustrates the joint modeling framework, where the possible priors for $u_{c,0}$ are shown with dotted lines. We include the priors for both the full and reduced dependence models,

to highlight that the primary difference between the two in terms of the model structure is the specification of the prior for the initial latent growth value for a cluster $u_{c,0}$.

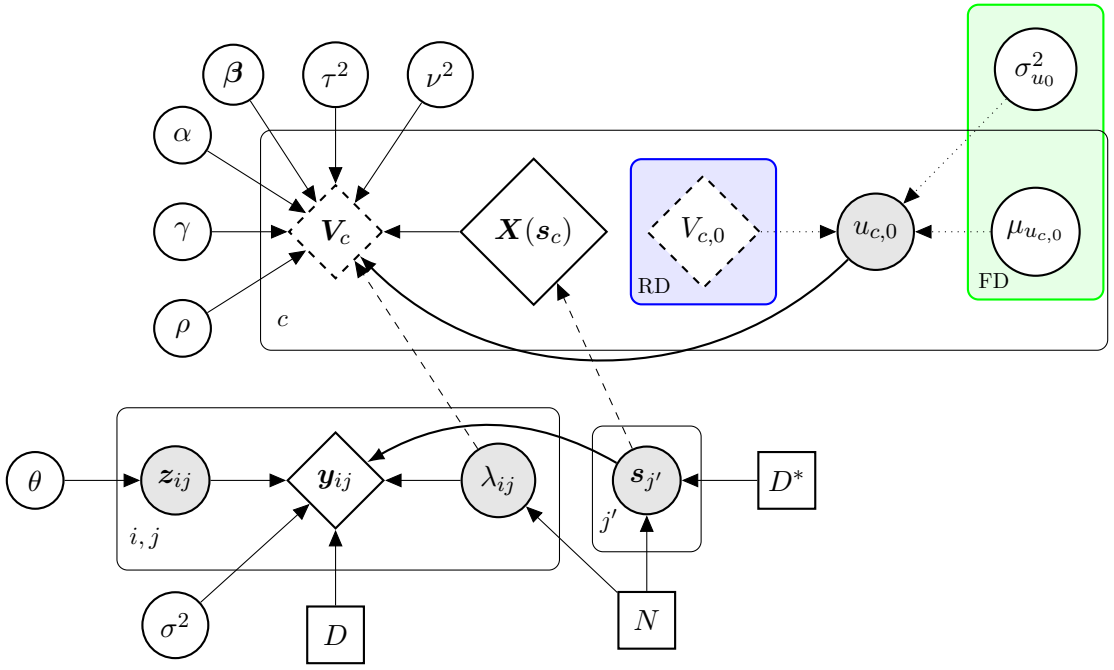

Figure 4.3: Plate diagram for the joint model. Where $i = 1, \ldots, m$ denotes the file index, $j = 1, \ldots, n_i$ denotes the record index within file i, $j' = 1, \ldots, N$ denotes the latent location index, and $c = 1, \ldots, C$ denotes the growth cluster index. Solid diamond nodes indicate data, dashed diamond nodes indicate possible missing data, round nodes indicate parameters, and square nodes indicate hyperparameters. Solid arrows denote stochastic relationships, while dashed arrows identify the inputs from the record linkage model to the downstream growth model. Dotted arrows signify a possible prior for the latent growth process initial value $u_{c,0}$. The prior for the full dependence model (FD) is highlighted on the green plate, while the prior for the reduced dependence model (RD) is highlighted on the blue plate.

In general, the joint model is specified as a hierarchical Bayesian model that combines the record linkage model and the growth model into a single framework. The joint model allows sampling from the posterior distributions of the parameters of both models simultaneously, which provides a more coherent framework for modeling the linkage and growth dynamics of individuals over time. The joint model is specified in generality as follows,

$$p(\boldsymbol{\Lambda}, \boldsymbol{s}, \boldsymbol{u}_0, \theta, \beta, \alpha, \gamma, \tau^2, \nu^2, \rho \mid \boldsymbol{y}, \boldsymbol{V}) \propto \underbrace{p(\boldsymbol{y} \mid \boldsymbol{\Lambda}, \boldsymbol{s}, \sigma^2)}_{\text{Linkage Likelihood}} \times \underbrace{p(\boldsymbol{V} \mid \boldsymbol{\Lambda}, \boldsymbol{s}, \boldsymbol{u}, \tau^2, \nu^2, \rho)}_{\text{Conditional Growth Likelihood}}$$

$$\times \underbrace{p(\boldsymbol{\Lambda}, \boldsymbol{s}, \theta, \sigma^2)\ p(\boldsymbol{u}_0 \mid \cdot)\ p(\boldsymbol{\beta}, \alpha, \gamma, \tau^2, \nu^2, \rho, \cdots)}_{\text{Priors}},$$

where the prior for $\boldsymbol{u}_0$, and remaining unspecified priors, depend on the choice of the full or reduced dependence model, as described in the previous section. The full specifications for the joint model variants are provided in Appendix B.2.

In both versions of the joint model, the linkage structure identifies the growth clusters and the latent locations associated with those clusters provide the spatio-temporal covariates for the growth model. The temporal dependence structures allow the growth model to capture departures from the latent growth process that are not explained by the covariates, while the measurement error variance τ^2 captures the uncertainty in the observed size values. We provide a visual illustration of the full dependence joint model in Figure 4.3, which shows the relationships between the observed record locations, latent entity locations, and the growth clusters. Details for the full conditional distributions of the joint model variants are provided in Appendix B.2. In the following section, we discuss the inference procedure for the joint model, and the computational strategies that we employ to allow the model to scale to datasets of reasonable size.

4.3.4 Inference and computational strategies

The joint modeling framework presented in Section 4.3.3 allows us to perform inference on the record linkage structure and the growth dynamics of individuals simultaneously. While this approach has the potential to provide higher quality estimates of the linkage structure and downstream model parameters, it also introduces additional computational challenges due to the increased complexity of the model compared to a two-stage approach. The feedback loop between the two components of the model is directly responsible for the improved inferential capabilities of the joint model, but also requires us to sample from a joint posterior distribution that is notably more complex than the individual components. In this section, we discuss the statistical and computational strategies that we employ to perform inference on the joint model.

In order to perform inference on the joint model, we use an MCMC approach to sample from the joint posterior distribution of the model parameters. While Bayesian approaches to record linkage tend to carry high computational costs (see Steorts et al. (2014) and Marchant et al. (2021) for in-depth discussions), the joint model allows us to leverage the structure of the model to reduce the

computational burden. Typically, record linkage applications address this computational burden by reducing the number of potential links between records by invalidating comparisons that are deemed impossible to be matches a priori through blocking. In traditional blocking schemes, the blocking structure is defined based on a set of blocking variables that are used to group records into blocks, where comparisons are only made between records within the same block (Steorts et al., 2014; Murray, 2015).

In lieu of a traditional blocking scheme, we use the spatial structure of our data to define a spatial bounding box that is based on the latent locations of the records. Instead of evaluating the probability of all possible latent locations when sampling λ_{ij} for a given record, we restrict our attention to latent locations that are within a certain distance of the record's observed location. This approach limits the linkage structure directly and provides an approximation of the true posterior linkage structure that is high quality, while also being computationally efficient. To ensure that the approximation is sufficiently accurate, we require the sampler to consider at least two candidate latent locations for each record, which maintains the stochastic nature of the record linkage model while ensuring the model explores the space of possible links in an efficient fashion. Drew et al. (2025) demonstrated that this approach resulted in an approximately 97.3 times speed up per iteration in the two-stage record linkage model with near-identical inference on $\mathbf{\Lambda}$, and we expect similar performance improvements in the joint model variants.

In addition, the hit-miss mechanism included in the record linkage model allows us to take advantage of the spatial structure of the data to further reduce the number of potential candidate links for a given record. When a record is deemed to be a hit, the model can immediately assign the record to a corresponding latent location relative to the resolution of the data (i.e., record locations are accurate to the pixel level). This effectively reduces the cost of sampling the cluster assignments for records that are hits, as the model can immediately assign the record to a latent within the range of a pixel without needing to sample using the spatial bounding box approximation.

To facilitate sampling from the posterior distributions of both the full and reduced dependence models in the growth component, we introduce a latent Gaussian process, $\boldsymbol{w}_c$, for each individual c such that the observed sizes values are conditionally independent given the latent Gaussian process.

The updated specification for the observed size values $\boldsymbol{V}_c$ in both models is then given by

$$\boldsymbol{V}_c \mid \boldsymbol{u}_c, \boldsymbol{w}_c, \tau^2 \overset{ind}{\sim} \text{Normal}(\boldsymbol{u}_c + \boldsymbol{w}_c, \tau^2 \mathbf{I})$$
$$\boldsymbol{w}_c \overset{iid}{\sim} \text{Normal}(\mathbf{0}, \nu^2 \boldsymbol{\Sigma}),$$

where $\boldsymbol{w}_c$ is a latent Gaussian process that captures the temporal dependence structure between the observed size values, and $\boldsymbol{u}_c$ implicitly depends on $u_{c,0}$, the spatio-temporal covariate vector associated with the latent location of the cluster $\boldsymbol{X}(\boldsymbol{s}_c)$, and the remaining generalized Michaelis–Menten growth function parameters. This construction allows us to sample more efficiently from the conditional posterior distributions of the variance parameters τ^2 and ν^2, and simplifies the process of dealing with missing values. In practice, this means that when a size observation is missing in a given cluster c, we can easily impute the missing value by sampling from the conditional posterior distribution. Details for the MCMC algorithms utilized to fit the joint model variants are provided in Appendix B.3.

In addition to these stochastic strategies, we use a number of computational strategies to optimize the efficiency of the MCMC sampler. To improve the speed and reduce the overhead of the joint MCMC sampler schemes, we implement the models in a hybrid R and C++ framework. We use **Rcpp** (Eddelbuettel et al., 2023a) to interface between R and C++, and **RcppArmadillo** (Eddelbuettel et al., 2023b) to take advantage of the linear algebra capabilities of the Armadillo library in C++. In order to facilitate rapid identification of the high density regions of the posterior distribution, we implement an initialization scheme for the linkage structure and the growth clusters that identifies viable clusters of records with regard to the resolution of the data. We pair this initialization scheme with adaptive Metropolis–Hastings (MH) proposals using the approach introduced by Rosenthal (2011) for our MH-within-Gibbs sampling steps, and a tempering scheme that allows the sampler to warm up by upweighting the prior information to prevent runaway feedback loops in the parameter estimates due to poor initialization (Geyer and Thompson, 1995). Details for the initialization scheme are provided in Appendix B.4.

4.4 Numerical experiments

In this section, we present a series of numerical experiments to evaluate the performance of the joint model variants. We compare the full and reduced dependence joint models in terms of their ability to accurately recover the true linkage structure and downstream growth model parameters on simulated data motivated by our empirical dataset introduced in Section 4.2. We consider a range of scenarios to capture the behavior of the two model variants across a wide array of possible conditions. We begin this section with a discussion of the data generating scheme, followed by details for the simulation settings, and a discussion of the results of the numerical experiments with emphasis on both the linkage performance and downstream model parameter coverage and their implications for the inferential potential of the joint modeling framework.

4.4.1 Data generation

To ground the simulation study in our empirical data, we focus on a subset of the River Forest area as the touchpoint for our simulated data. This area is the most densely populated region within our study area, and is expected to be representative of the population dynamics in our study area. To emulate the dynamics in our empirical data, we incorporate a set of spatio-temporal covariates in our data generating mechanism. Figure 4.4 shows the River Forest area and the $100\,\mathrm{m}^2$ subset that we use as a template for our simulated data alongside the covariate surfaces for the entire study area that we leverage to generate the data.

Utilizing the **ldmppr** package in R (Drew and Kaplan, 2025), we begin by estimating a location-dependent marked point process model for the subset of the River Forest area data for the year 2021. This modeling approach relies on an extension of the framework introduced by Møller et al. (2016) that maps a marked point process onto a spatio-temporal process. While marked point processes are often difficult to model, especially with location dependence in the mark distribution, this approach enables straightforward and flexible estimation and simulation. We estimate the generating parameters for the underlying spatio-temporal marked point process and train an `XGBoost` model (Chen et al., 2024) to predict location-dependent sizes using the observed data and raster surfaces.

We note that our empirical data tends to exhibit a pattern of inhibition within the spatial distribution of trees across the study area, which makes the **ldmppr** package a natural choice for modeling the point process. Additional details for the point process estimation and mark model

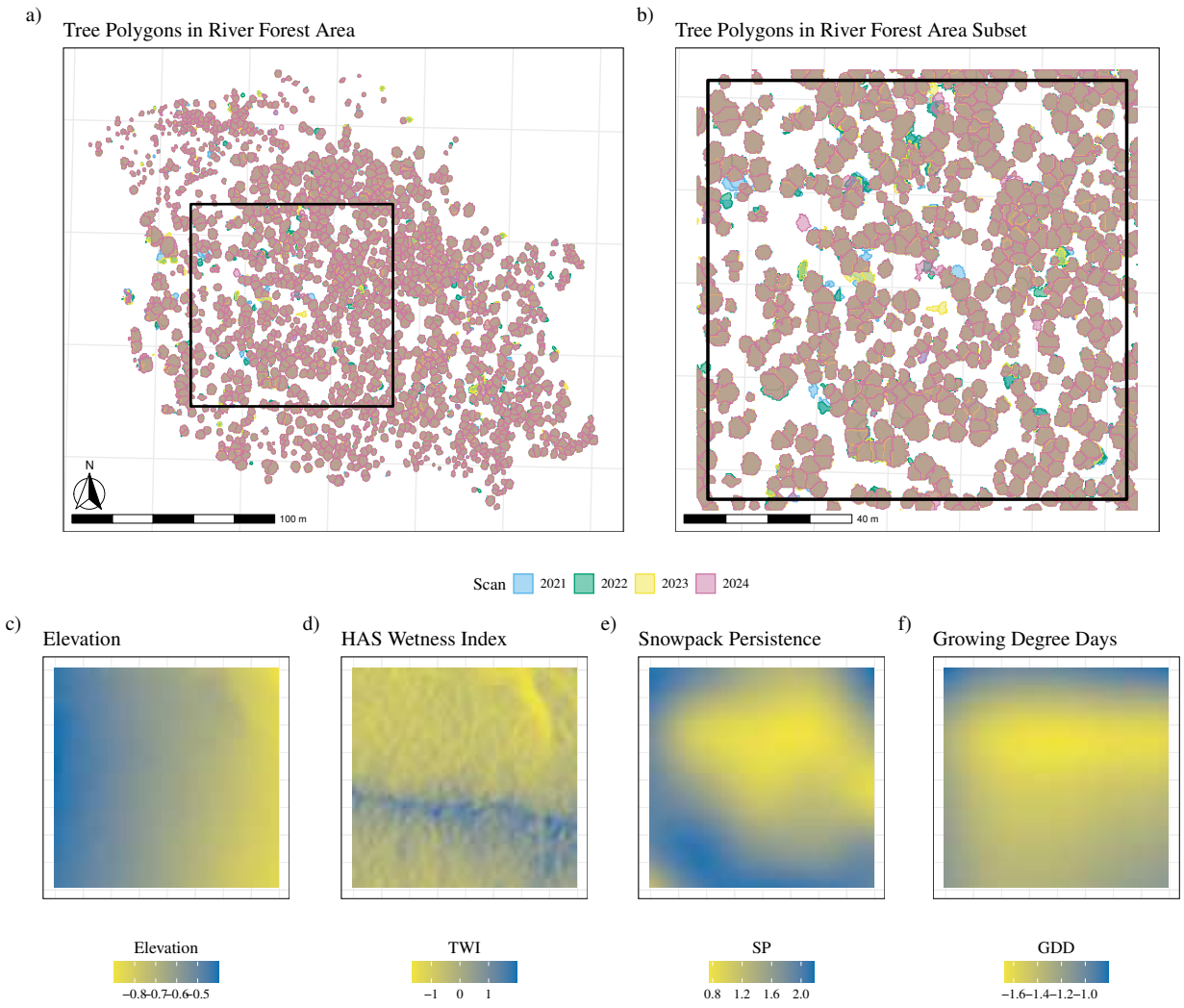

Figure 4.4: Plot (a) shows the derived crown polygons for the River Forest area by scan year. Plot (b) shows the 100 m^2 subset of the River Forest area that we use as a template for our simulated data. Plots (c)–(f) show the covariate raster surfaces for the River Forest subset, normalized over the entire study area, that we use to generate the data, including Elevation, TWI, Snowpack Persistence (2021), and Growing Degree Days (2021).

training steps are provided in Appendix B.5. We generate the simulated data by sampling from the estimated point process model and the trained mark model, which provides us with a set of simulated latent individuals. We proceed by generating data for the observed files according to either the full or reduced dependence joint model specifications introduced in Section 4.3.3. This scheme gives us flexibility in adjusting the generating model parameters to evaluate the performance of the joint model variants on simulated data that is similar to the Gothic Townsite dataset.

4.4.2 Simulation settings

In this section, we outline the conditions for our simulation study and the various mechanisms that we aim to investigate. We consider a broad range of simulation settings to evaluate the performance of the joint model variants across settings designed to mimic conditions that might be observed in the natural world. In particular, we consider the impact of point density, hit-miss distortion probability, and signal-to-noise ratio in the growth model on the performance of the joint model variants. We fix the number of files under consideration to 4, matching the configuration in our empirical data, and generate 100 datasets for each simulation setting. In addition to the full and reduced dependence joint model variants, we also provide results for both formulations in a two-stage approach, where the record linkage model is fit to the observed location data, and the growth model is fit to the identified growth clusters of records utilizing the linkage-averaging approach introduced in Section 4.1. We also provide results for the downstream growth model fit with the true linkage (TL) structure, which serves as our oracle for the growth model.

The River Forest subset represents the highest density of individuals in a sub-area across our study area with a value of approximately $470\,\text{ha}^{-1}$. In our simulation study, we treat this as our medium density setting ($400\,\text{ha}^{-1}$), and additionally consider a low density setting ($300\,\text{ha}^{-1}$), and a high density setting ($500\,\text{ha}^{-1}$). We consider three levels of hit-miss distortion probability, which we denote as low (0.15), medium (0.25), and high (0.35). We obtained an estimate for the distortion probability from our empirical data using the linkage derived from our joint model intialization scheme, which yielded a value of approximately 0.25. To evaluate the impact of the signal-to-noise ratio in the growth model, we consider five scenarios that vary the relationships between τ^2, ν^2, and ρ. An overview of the simulation conditions is provided in Table 4.2, where starred levels indicate the standard level for that factor.

4.4.3 Results

In this section, we present the results of our numerical experiments to evaluate the performance of the joint model variants against the two-stage implementations and the downstream model fit with the true linkage. All results in this section are from models fit to thinned versions of the simulated datasets, such that we remove 5% of the observed data points from each file to simulate the presence of missing data. This is done to evaluate the robustness of the joint model variants

Table 4.2: Overview of the three primary factors varied in the simulation study, including point density, hit-miss distortion probability, and signal-to-noise ratio. Standard levels for each factor are highlighted with a star.

Factor	Levels / Settings
Point Density	low: ≈ 300 trees/ha *medium: ≈ 400 trees/ha (River Forest reference spec) high: ≈ 500 trees/ha
Hit-Miss Distortion Probability	small: 0.15 *medium: 0.25 large: 0.35
Signal-to-Noise Ratio	*(1) $\tau^2 = 1.0,\ \nu^2 = 0.5,\ \rho = 0.75$ (2) $\tau^2 = 1.0,\ \nu^2 = 1.0,\ \rho = 0.75$ (3) $\tau^2 = 1.0,\ \nu^2 = 0.5,\ \rho = 0.5$ (4) $\tau^2 = 1.0,\ \nu^2 = 0.5,\ \rho = 1.0$ (5) $\tau^2 = 1.5,\ \nu^2 = 1.0,\ \rho = 1.0$

to missing data, and to ensure that the models are able to handle missing data in a principled manner. To reduce the number of possible combinations of different settings, we focus on each variant of interest across all levels while holding the remaining conditions at their standard levels. We highlight the most notable results from the numerical experiments, with a particular emphasis on the coverage performance for the $\boldsymbol{\beta}$ parameters in the growth model, as these parameters are the primary component of interest in describing the impact of spatial, temporal, and spatio-temporal covariates on individual tree growth. We note that all models are fit to data with a generating mechanism that matches the model.

Linkage performance

As the quality of the linkage is expected to have a notable impact on the performance of the downstream growth model, we begin by evaluating the linkage performance of the joint model variants against the two-stage model variants. In particular, we consider the metrics of precision and recall to assess the quality of the linkage structure identified by each model variant. Precision is defined as the proportion of true matches among all identified matches, while recall is defined as the proportion of true matches that were identified by the model (Christen, 2012). These metrics provide a balanced view of the linkage performance, as they account for both false positives and false negatives in the linkage structure. We present the results for the linkage performance across

varying density levels in Figure 4.5, while holding the remaining factors identified in Table 4.2 at their standard levels. Results for the remaining factors are provided in Appendix B.6.

We observe that the full dependence joint model variant consistently outperforms the reduced dependence joint model and both variants of the two-stage models. We note that while the two-stage models perform well in terms of precision, they tend to have lower recall compared to the full dependence joint model, particularly in the high density setting. The full dependence joint model shows a slight decrease in precision compared to the two-stage variants, however balances this with an improvement in recall. This suggests that the full dependence joint model may be better able to capture the entirety of the true linkage structure, particularly in high density settings where the feedback loop between the record linkage and growth model components is most beneficial. In contrast, the reduced-dependence joint model shows a broader range of precision–recall performance, most notably as density increases, which highlights the additional linkage uncertainty that arises from the reduced dependence structure relative to the full dependence structure. We note that the linkage performance results for the joint model variants are consistent across the varying hit-miss distortion probabilities and signal-to-noise ratio scenarios, with the full dependence joint model consistently outperforming the reduced dependence joint model and the two-stage models. We also note that the performance of all model variants is generally above 0.9 for both precision and recall.

Downstream model parameter coverage

In addition to the linkage performance, we evaluate the coverage of the downstream growth model parameters across the various levels of the factors outlined in Table 4.2. We focus on the coverage of the $\boldsymbol{\beta}$ parameters, as these parameters are of primary interest in our empirical data analysis. Table 4.3 provides a summary of the coverage for the model variants across varying density levels. We observe that the full dependence joint model tends to outperform the reduced dependence model in capturing the parameters of the dependence structure, particularly in higher density settings. In contrast to the two-stage models, the joint model variants show robust coverage for the $\boldsymbol{\beta}$ coefficients suggesting that the feedback loop between the record linkage and growth model components is beneficial for accurately estimating the impact of covariates on the growth asymptote. In instances where the model variants have higher coverage rates than the true linkage model, we note that this is likely due to the joint and two-stage model variants providing more

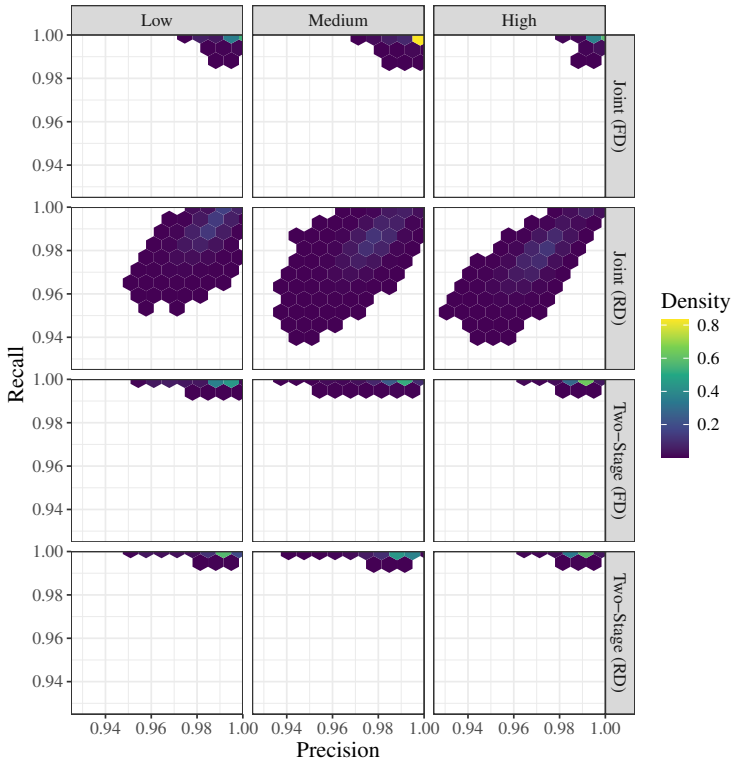

Figure 4.5: Precision and recall joint densities for the full dependence (FD) and reduced dependence (RD) joint and two-stage model variants across varying density levels. The precision and recall values are combined across 100 simulated datasets for each model variant and density level.

conservative estimates of the parameters, which results in wider credible intervals that capture the true parameter values more frequently but at lower resolution.

Table 4.4 provides a summary of the coverage for the model variants across varying hit-miss distortion probabilities. We observe a noticeable increase in the coverage rates for the joint model variants as the hit-miss distortion probability decreases, particularly for the full dependence joint model. Both joint model variants show robust coverage for the $\boldsymbol{\beta}$ coefficients compared to the two-stage implementations, and coverage levels closer to the nominal level suggesting that the credible intervals are not overly conservative.

Table 4.5 provides a summary of the coverage for the model variants across varying signal-to-noise ratio scenarios. The scenarios that we highlight in this experiment are designed to explore the relationship between the measurement error variance τ^2, the process noise variance ν^2, and the temporal correlation parameter ρ. We vary the direct relationship between the two variance components, τ^2 and ν^2, and consider the impact of varying levels of temporal correlation on the

Table 4.3: 90% empirical coverage rates by density for both variants of the joint, two-stage, and true linkage models. The coverage rates are averaged across 100 simulated datasets for each density level.

		Empirical Coverage by Parameter									
Model	Density	α	β_0	β_1	β_2	β_3	β_4	γ	ν^2	ρ	τ^2
Joint (FD)	Low	0.75	0.99	1.00	0.99	1.00	0.98	0.66	0.11	0.14	0.38
	Med	0.87	1.00	0.99	0.98	1.00	0.99	0.46	0.07	0.08	0.15
	High	0.87	0.99	1.00	1.00	1.00	1.00	0.51	0.01	0.02	0.03
Joint (RD)	Low	0.82	0.93	0.98	0.93	0.98	0.89	0.75	0.16	0.26	0.19
	Med	0.84	0.97	0.97	0.90	0.92	0.88	0.89	0.00	0.14	0.03
	High	0.75	0.94	0.93	0.94	0.90	0.74	0.87	0.06	0.08	0.08
Two-Stage (FD)	Low	0.64	0.97	1.00	0.99	0.86	0.12	0.75	0.27	0.22	0.48
	Med	0.78	1.00	1.00	1.00	0.61	0.06	0.71	0.13	0.17	0.38
	High	0.73	0.97	1.00	1.00	0.50	0.03	0.67	0.07	0.22	0.37
Two-Stage (RD)	Low	0.86	0.99	1.00	0.99	0.80	0.09	0.76	0.54	0.34	0.69
	Med	0.79	0.97	1.00	0.99	0.42	0.04	0.90	0.47	0.40	0.70
	High	0.81	0.97	1.00	0.96	0.29	0.02	0.85	0.34	0.25	0.63
True Linkage (FD)	Low	0.65	0.98	0.99	0.99	0.98	0.93	0.74	0.77	0.32	0.37
	Med	0.72	0.96	0.98	0.95	0.97	0.99	0.75	0.78	0.36	0.44
	High	0.67	0.95	0.95	0.95	0.97	0.96	0.75	0.84	0.36	0.43
True Linkage (RD)	Low	0.71	0.98	1.00	1.00	0.99	1.00	0.80	0.97	0.61	0.65
	Med	0.69	0.98	0.99	0.97	0.99	0.97	0.91	0.88	0.62	0.74
	High	0.69	0.96	0.98	0.95	0.95	0.96	0.84	0.96	0.61	0.72

coverage rates for the model parameters. The results from the table suggest that the growth model may suffer from identifiability issues when priors are weakly specified. However, the coverage for the $\boldsymbol{\beta}$ coefficients remains robust across the joint model variants in spite of the potential issues with determining the shape of the growth curve.

Across all of the simulation settings, we observe that the joint model variants consistently outperform the two-stage approach in terms of coverage for the $\boldsymbol{\beta}$ coefficients in the growth model, which are the primary parameters of interest in describing environmental impacts on individual tree growth. We note that the joint model variants struggle to reliably capture the dependence structure due to potential identifiability constraints in the structure of the growth model. Effectively, as the amount of noise in the data increases, it becomes difficult for the models to pin down the exact shape of the growth function. We believe that this may be addressed through the use of stronger

Table 4.4: 90% empirical coverage rates by hit-miss distortion probability for both variants of the joint, two-stage, and true linkage models. The coverage rates are averaged across 100 simulated datasets for each hit-miss distortion level.

		Empirical Coverage by Parameter									
Model	Hit-Miss	α	β_0	β_1	β_2	β_3	β_4	γ	ν^2	ρ	τ^2
Joint (FD)	Small	0.80	0.94	1.00	0.97	0.98	0.99	0.59	0.41	0.32	0.35
	Medium	0.87	1.00	0.99	0.98	1.00	0.99	0.46	0.07	0.08	0.15
	Large	0.87	1.00	1.00	1.00	1.00	1.00	0.46	0.00	0.00	0.05
Joint (RD)	Small	0.81	0.96	0.97	0.95	0.98	0.94	0.87	0.33	0.27	0.27
	Medium	0.84	0.97	0.97	0.90	0.92	0.88	0.89	0.00	0.14	0.03
	Large	0.76	0.93	0.95	0.92	0.87	0.72	0.87	0.00	0.02	0.00
Two-Stage (FD)	Small	0.75	0.96	1.00	0.98	0.84	0.29	0.69	0.37	0.34	0.52
	Medium	0.78	1.00	1.00	1.00	0.61	0.06	0.71	0.13	0.17	0.38
	Large	0.77	1.00	1.00	1.00	0.59	0.02	0.73	0.13	0.17	0.40
Two-Stage (RD)	Small	0.78	0.96	1.00	0.99	0.66	0.25	0.89	0.61	0.53	0.76
	Medium	0.79	0.97	1.00	0.99	0.42	0.04	0.90	0.47	0.40	0.70
	Large	0.81	0.99	1.00	1.00	0.18	0.01	0.90	0.41	0.37	0.69
True Linkage (FD)	Small	0.73	0.96	0.99	0.94	0.98	0.98	0.73	0.84	0.42	0.54
	Medium	0.72	0.96	0.98	0.95	0.97	0.99	0.75	0.78	0.36	0.44
	Large	0.72	0.96	0.98	0.94	0.97	0.98	0.70	0.83	0.28	0.39
True Linkage (RD)	Small	0.69	0.98	1.00	0.98	0.98	0.97	0.89	0.92	0.64	0.74
	Medium	0.69	0.98	0.99	0.97	0.99	0.97	0.91	0.88	0.62	0.74
	Large	0.69	0.98	0.99	0.98	0.99	0.97	0.86	0.94	0.67	0.77

priors on the growth model parameters, which would help to constrain the range of possible growth functions that the model can fit to the data. This dynamic may also explain the relatively solid coverage performance of the reduced dependence joint model for the downstream growth model parameters, in spite of somewhat poorer performance in terms of the linkage. The erroneous or missing links in the two-stage are likely to be more impactful on the downstream growth model parameters as the record linkage component of the two-stage approach is agnostic with regard to the downstream task, which may lead to bias in the estimates of the downstream model parameters. The joint model variants are better protected from linkage errors that result in implausible growth conditions, and in particular the full dependence joint model benefits from the shared dependence structure across all data points associated with a unique cluster (or individual).

Table 4.5: 90% empirical coverage rates by signal-to-noise ratio for both variants of the joint, two-stage, and true linkage models. The coverage rates are averaged across 100 simulated datasets for each signal-to-noise ratio setting.

		Empirical Coverage by Parameter									
Model	SNR	α	β_0	β_1	β_2	β_3	β_4	γ	ν^2	ρ	τ^2
Joint (FD)	(1)	0.87	1.00	0.99	0.98	1.00	0.99	0.46	0.07	0.08	0.15
	(2)	0.88	0.99	0.99	1.00	1.00	1.00	0.47	0.12	0.10	0.10
	(3)	0.89	1.00	1.00	1.00	1.00	1.00	0.51	0.07	0.04	0.05
	(4)	0.85	1.00	1.00	1.00	1.00	1.00	0.48	0.04	0.05	0.49
	(5)	0.90	1.00	1.00	1.00	1.00	1.00	0.47	0.01	0.05	0.19
Joint (RD)	(1)	0.84	0.97	0.97	0.90	0.92	0.88	0.89	0.00	0.14	0.03
	(2)	0.82	0.99	0.98	0.96	0.93	0.85	0.91	0.10	0.18	0.10
	(3)	0.78	0.95	0.99	0.96	0.94	0.79	0.87	0.03	0.11	0.06
	(4)	0.77	0.97	0.94	0.94	0.96	0.89	0.86	0.04	0.20	0.06
	(5)	0.79	0.98	0.97	0.96	0.94	0.84	0.91	0.03	0.19	0.07
Two-Stage (FD)	(1)	0.78	1.00	1.00	1.00	0.61	0.06	0.71	0.13	0.17	0.38
	(2)	0.78	0.95	1.00	0.99	0.72	0.07	0.64	0.55	0.27	0.44
	(3)	0.77	0.97	1.00	0.99	0.63	0.09	0.74	0.54	0.11	0.22
	(4)	0.77	0.99	1.00	0.98	0.60	0.05	0.73	0.05	0.25	0.72
	(5)	0.79	0.99	1.00	1.00	0.80	0.06	0.59	0.21	0.53	0.75
Two-Stage (RD)	(1)	0.79	0.97	1.00	0.99	0.42	0.04	0.90	0.47	0.40	0.70
	(2)	0.83	0.99	1.00	1.00	0.55	0.04	0.90	0.78	0.41	0.68
	(3)	0.82	0.98	1.00	0.99	0.43	0.03	0.90	0.85	0.11	0.50
	(4)	0.79	0.97	1.00	1.00	0.42	0.01	0.90	0.16	0.59	0.86
	(5)	0.84	0.99	1.00	1.00	0.66	0.03	0.86	0.59	0.74	0.82
True Linkage (FD)	(1)	0.72	0.96	0.98	0.95	0.97	0.99	0.75	0.78	0.36	0.44
	(2)	0.76	0.94	0.99	0.95	0.97	0.98	0.67	0.89	0.50	0.55
	(3)	0.74	0.97	0.98	0.96	0.97	0.99	0.73	0.97	0.31	0.43
	(4)	0.71	0.95	0.99	0.95	0.96	0.98	0.72	0.56	0.39	0.61
	(5)	0.74	0.94	0.99	0.94	0.96	0.99	0.55	0.63	0.55	0.69
True Linkage (RD)	(1)	0.69	0.98	0.99	0.97	0.99	0.97	0.91	0.88	0.62	0.74
	(2)	0.77	0.98	0.99	0.98	0.98	0.97	0.86	0.84	0.68	0.72
	(3)	0.73	0.98	0.99	0.97	0.99	0.97	0.87	0.93	0.51	0.70
	(4)	0.65	0.98	0.99	0.97	0.99	0.97	0.83	0.81	0.72	0.81
	(5)	0.76	0.98	0.99	0.98	0.99	0.97	0.85	0.81	0.77	0.78

4.5 Discussion and future work

In Section 4.1 we posed two primary challenges associated with using multi-temporal remote sensing data for tree demography. To understand changes in tree populations over time, we need to reliably identify unique individuals across scans and propagate the uncertainty from the linkage procedure into the downstream inference on growth dynamics. The joint modeling approach introduced in this chapter provides a framework for unifying a record linkage and downstream modeling task as a single Bayesian hierarchical model to provide robust uncertainty quantification across the entire modeling pipeline and improved inference for the downstream model parameters of interest compared to a two-stage approach. In contrast to two-stage approaches (e.g., Drew et al. (2025); Sadinle (2018)), by allowing feedback between the linkage and downstream submodels, the joint model yields more coherent posterior inference.

We provided a novel combination of a record linkage model designed to accommodate a generic number of files with a mechanistic growth model that describes the growth dynamics of individual trees as a function of topographic and environmental conditions that vary over space and time. We demonstrated the capabilities of the joint modeling approach on simulated data calibrated to reflect the Gothic Townsite dataset, and saw that the full dependence joint model variant outperformed the two-stage approach in terms of linkage performance. The coverage rates for both joint model variants tended to show improvement compared to the two-stage approach for the primary growth model parameters of interest. These results mirror findings from previous work on joint modeling in the record linkage literature (Steorts et al., 2018), which have shown that joint modeling approaches can provide improved inference for downstream model parameters of interest compared to strictly two-stage approaches. Given the results of our numerical experiments, we recommend the use of the full dependence joint model variant for applications where the linkage is important in addition to the downstream model inference. If the linkage is not of primary interest, the reduced dependence joint model variant may be a viable alternative option as it is more computationally efficient and provides reasonable coverage for the growth model parameters of interest.

Despite these strengths, there are a few potential limitations to the joint modeling approach. In particular, the joint model variants are computationally intensive and require a significant amount of time to fit on large datasets (measured in terms of the total number of records across files). While

we have implemented several strategies to improve the efficiency of the MCMC sampler, there is still room for improvement. One potential avenue for improvement is the incorporation of a split/merge mechanism for sampling the linkage structure paired with locally balanced proposals as introduced in Zanella (2020). This would potentially allow the sampler to explore the space of possible linkage structures more efficiently and enable the model to more easily break up clusters of records that may not be coreferent. Additionally, a more sophisticated prior for the linkage structure that ensures microclustering behavior, as in Betancourt et al. (2022b), might limit the sensitivity of the joint modeling approach to misspecification in the downstream model by enforcing reasonable cluster sizes relative to the number of files.

We also note that the joint model variants can suffer from identifiability issues such that there is some sensitivity to the choice of hyperparameters, and the performance of the model can be particularly impacted in settings where the signal-to-noise ratio is weak. Some care must be taken when selecting the appropriate joint model variant and its relevant hyperparameters to ensure that the model is able to accurately recover the linkage structure and downstream model parameters consistently. On a related note, the generalized Michaelis–Menten style growth model that we have used in the joint model variants can suffer from identifiability issues in settings where the growth dynamics are not well defined, which can lead to poor performance in the model.

In our future work, our primary objective is to fit the joint model variants to the full Gothic Townsite dataset to gauge the impact of the spatial (elevation and TWI), spatio-temporal (snowpack persistence and growing degree days), and individual-specific (NDVI) covariates on the growth dynamics of the population in the study area. We plan to employ a multi-stage MCMC approach to recover global parameter distributions for the downstream model parameters of interest using an approach similar to the one introduced by Johnson et al. (2022). The scalability issues associated with the joint modeling approach limit the size of the dataset that it may be applied to, and the incorporation of a multi-stage MCMC approach would enable submodels to be fit on spatially disparate subsets of the dataset in parallel and then combined in a second stage to recover global parameter distributions. This approach provides a nice balance between localized inference at the submodel level and global inference across the entire study area. Additionally, we intend to explore the impact of misspecification in the application of the two joint model variants. We want to assess

the sensitivity of the joint model variants to mismatched generating models in terms of their ability to recover the linkage structure and downstream growth model parameters of interest, which will allow us to further delineate the utility of the two proposed dependence structures. Finally, we would like to explore modifications of the sampling approach for the linkage structure, including a split/merge mechanism and alternative priors, to facilitate increased scalability for the joint modeling approach through improved mixing in the sampler.

Chapter 5

Conclusion

Accurately tracking individual-tree growth over large, evolving landscapes is critical for understanding forest carbon dynamics, but has historically been hampered by fragmented remote-sensing data and uncertain links across years. In this dissertation, we present a collection of statistical methods and practical tools that turn those challenges into opportunities for richer inference. We explored the use of Bayesian hierarchical models for the analysis of multi-temporal remote sensing data through the integration of record linkage and downstream modeling tasks. We introduced two novel modeling approaches that leverage the strengths of Bayesian hierarchical modeling to provide robust uncertainty quantification and improved inference for the parameters of interest in the downstream modeling objectives. Motivated by the challenges of working with large heterogeneous spatial datasets in the absence of a gold standard dataset for assessing the accuracy of record linkage and downstream modeling tasks, we developed a framework for working with marked spatial point processes with location-dependent marks to allow us to characterize the spatial distribution of individual trees as a function of environmental conditions and topographic features. We provide a brief summary of the contributions of each chapter below, and conclude by discussing possible avenues for future research on the work presented in this dissertation.

In Chapter 2, we introduced a highly flexible and scalable two-stage modeling approach for bi-temporal LiDAR data that first identifies unique individuals across files, and then propagates the uncertainty from the linkage task to the downstream modeling task in a principled manner using a generalization of the linkage-averaging approach. We provided the theoretical justification for this approach, and introduced a novel sampling strategy that allows for efficient sampling of the linkage structure. While previous approaches to study individual tree growth were limited in the spatial scales they could address, the two-stage approach we introduced is capable of scaling to datasets containing hundreds of thousands of unique individuals, and provides insight about the impact of environmental conditions on individual tree growth dynamics across large spatial extents.

In Chapter 3, we introduced the R package **ldmppr** to provide a set of tools for working with marked spatial point processes with location-dependent marks characterized by regularity in the

point pattern. Our work in Chapter 2, and the lack of a ground truth dataset with known linkage and growth, motivated us to develop a framework for estimating and simulating from marked spatial point processes, e.g., trees within a forest stand, to provide a way to both characterize the spatial distribution of individual trees and to simulate from a fitted model to generate biologically realistic synthetic datasets for model testing and validation. We extended the approach of Møller et al. (2016) to allow for location dependence in the mark distribution through the incorporation of a suitably flexible mark model. We provided the mathematical framework for this approach, and detailed its implementation in the **ldmppr** package. The package is designed with a user-friendly workflow that enables users to fit marked spatial point process models to their data, and to simulate from the fitted model to generate synthetic datasets.

In Chapter 4, we introduced a Bayesian joint modeling approach that unifies the record linkage and downstream modeling tasks into a single hierarchical model capable of identifying unique individuals across an arbitrary number of files, and modeling the growth dynamics of individual trees as a function of topographic and environmental conditions. We detailed a novel spatio-temporal record linkage model designed to provide exact uncertainty quantification across the entire modeling pipeline. We introduced two alternative formulations of the joint model that differ in the assumptions made about the dependence structure between observations in the downstream model, and investigated their performance on simulated data calibrated to reflect the Gothic Townsite dataset.

The collection of methods and tools presented in this dissertation provides a comprehensive framework for working with multi-temporal remote sensing data. We enable researchers to accurately track individual tree growth over time at varying levels of spatial resolution, and to characterize the impact of environmental conditions on the growth dynamics. The two-stage modeling approach provides a flexible and scalable solution for working with large datasets, while the joint modeling approach provides a more coherent framework for integrating record linkage and downstream modeling tasks where interdependence between the linkage and downstream task are likely. The **ldmppr** package provides a set of tools for working with marked spatial point processes with location-dependent marks, and enables researchers to better understand the spatial dynamics of their data and to generate synthetic datasets for model testing and validation.

However, the work presented in this dissertation is not without its limitations. In particular, the two-stage modeling approach can struggle to accurately capture the dependence structure between observations in the downstream model, and can be sensitive to the amount of noise in the data. The joint modeling approach, while more coherent, can be computationally intensive and may struggle with identifiability issues in the growth model. The **ldmppr** package is currently limited to marked spatial point processes characterized by regularity in the point pattern, and does not yet support marked spatial point processes characterized by clustering behavior, which are another common feature of ecological data. In light of these limitations, we outline some possible avenues for future research below.

5.0.1 Marked spatial point processes with location-dependent marks future work

In its current iteration, the **ldmppr** package provides a set of tools for working with marked spatial point processes with location-dependent marks characterized by regularity in the point pattern. We would like to extend the package to include marked spatial point processes characterized by clustering behavior. We believe that this is a worthwhile extension of the general framework that we have provided for incorporating location-dependence into marked point processes, and would facilitate adoption of the package for a wider range of applications.

5.0.2 Spatio-temporal record linkage future work

In both Chapter 2 and Chapter 4, we introduced Bayesian hierarchical models for spatio-temporal record linkage that utilize a latent matching structure to identify unique individuals across files. The specifications of these models depended on specificaiton of a hyperprior for the number of possible unique individuals across files, which can be difficult to specify in practice. We would like to explore the incorporation of a more sophisticated mechanism for determining the number of unique individuals across files, such as the prior introduced by Betancourt et al. (2022b) which has a microclustering property that maintains the growth of cluster sizes at a sublinear rate as the number of records increases. The microclustering behavior would enable the model to adapt to an increasing number of records, while targeting appropriate cluster sizes. This would allow for more

flexibility in the model and could potentially improve the performance of the record linkage task in settings where the number of unique individuals is not known a priori.

We would also like to explore the incorporation of a split/merge mechanism for sampling the linkage structure paired with locally balanced proposals as introduced by Zanella (2020). While we have implemented a novel strategy for approximately sampling the linkage structure using a spatial bounding box, the incorporation of a split/merge mechanism would allow the sampler to explore the space of possible linkage structures more efficiently and enable the model to more easily break up clusters of records that may not be coreferent. This would be particularly useful in settings where the number of unique individuals is large, and the linkage structure is complex. One of our primary objectives for future work is to improve the scalability of the spatio-temporal record linkage model to enable inference at larger spatial scales, such as entire forest stands. This would enable researchers to get a more comprehensive picture of the evolution of individual trees over time, and to better understand the impact of environmental conditions on individual tree growth dynamics.

Bibliography

Abdel Monem, R. I., Hassanein, E. E., and El Qutaany, A. Z. (2025). Temporal record linkage for heterogeneous big data records. *Egyptian Informatics Journal*, 30:100642.

Aubry-Kientz, M., Dutrieux, R., Ferraz, A., Saatchi, S., Hamraz, H., Williams, J., Coomes, D., Piboule, A., and Vincent, G. (2019). A Comparative Assessment of the Performance of Individual Tree Crowns Delineation Algorithms from ALS Data in Tropical Forests. *Remote Sensing*, 11(9):1086.

Babcock, C., Finley, A. O., Cook, B. D., Weiskittel, A., and Woodall, C. W. (2016). Modeling forest biomass and growth: Coupling long-term inventory and LiDAR data. *Remote Sensing of Environment*, 182:1–12.

Baddeley, A., Bárány, I., and Schneider, R. (2007). Spatial point processes and their applications. *Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004*, pages 1–75.

Baddeley, A. and Turner, R. (2005). spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42.

Barbosa, R., Ramírez-Narváez, P., Fearnside, P., Villacorta, C., and Carvalho, L. (2019). Allometric models to estimate tree height in northern Amazonian ecotone forests. *Acta Amazonica*, 49:81–90.

Bayarri, M. J., Berger, J. O., and Liu, F. (2009). Modularization in Bayesian analysis, with emphasis on analysis of computer models. *Bayesian Analysis*, 4(1):119–150.

Bayisa, F. L., Ådahl, M., Rydén, P., and Cronie, O. (2023). Regularised Semi-parametric Composite Likelihood Intensity Modelling of a Swedish Spatial Ambulance Call Point Pattern. *Journal of Agricultural, Biological and Environmental Statistics*, 28(4):664–683.

Berkelhammer, M., Still, C. J., Ritter, F., Winnick, M., Anderson, L., Carroll, R., Carbone, M., and Williams, K. H. (2020). Persistence and Plasticity in Conifer Water-Use Strategies. *Journal of Geophysical Research: Biogeosciences*, 125(2):e2018JG004845.

Betancourt, B., Sosa, J., and Rodríguez, A. (2022a). A prior for record linkage based on allelic partitions. *Computational Statistics & Data Analysis*, 172:107474.

Betancourt, B., Zanella, G., and Steorts, R. C. (2022b). Random Partition Models for Microclustering Tasks. *Journal of the American Statistical Association*, 117(539):1215–1227.

Bolin, D. and Wallin, J. (2023). Local scale invariance and robustness of proper scoring rules. *Statistical Science*, 38(1):140–159.

Bolker, B. M. (2008). *Ecological Models and Data in R*. Princeton University Press.

Brahma, B., Sileshi, G. W., Nath, A. J., and Das, A. K. (2017). Development and evaluation of robust tree biomass equations for rubber tree (Hevea brasiliensis) plantations in India. *Forest Ecosystems*, 4(1):14.

Breckheimer, I. (2023). Integrated Snow and Air Temperature Metrics for Ecological Applications.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC Press LLC, London, UNITED KINGDOM.

Buechling, A., Martin, P. H., and Canham, C. D. (2017). Climate and competition effects on tree growth in Rocky Mountain forests. *Journal of Ecology*, 105(6):1636–1647.

Carroll, R. W. H., Gochis, D., and Williams, K. H. (2020). Efficiency of the Summer Monsoon in Generating Streamflow Within a Snow-Dominated Headwater Basin of the Colorado River. *Geophysical Research Letters*, 47(23):e2020GL090856.

Chambers, R., Salvati, N., Fabrizi, E., and da Silva, A. D. (2019). Domain estimation under informative linkage. *Statistical Theory and Related Fields*, 3(2):90–102.

Chave, J., Andalo, C., Brown, S., Cairns, M., Chambers, J., Eamus, D., Folster, H., Fromard, F., Higuchi, N., and Kira, T. (2005). Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, 145(1):87–99.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2023). *Xgboost: Extreme Gradient Boosting*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., Li, Y., and Yuan, J. (2024). *Xgboost: Extreme Gradient Boosting*.

Christen, P. (2012). Data matching : Concepts and techniques for record linkage, entity resolution, and duplicate detection. In *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*, Data-Centric Systems and Applications. Springer, Berlin.

Contreras, M. A., Affleck, D., and Chung, W. (2011). Evaluating tree competition indices as predictors of basal area increment in western Montana forests. *Forest Ecology and Management*, 262(11):1939–1949.

Cook, J. D., Williams, D. M., Walsh, D. P., and Hefley, T. J. (2023). Bayesian forecasting of disease spread with little or no local data. *Scientific Reports*, 13(1):8137.

Coomes, D. A., Dalponte, M., Jucker, T., Asner, G. P., Banin, L. F., Burslem, D. F. R. P., Lewis, S. L., Nilus, R., Phillips, O. L., Phua, M.-H., and Qie, L. (2017). Area-based vs tree-centric approaches to mapping forest carbon in Southeast Asian forests from airborne laser scanning data. *Remote Sensing of Environment*, 194:77–88.

Copas, J. B. and Hilton, F. J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society. Series A, Statistics in society*, 153(3):287–320.

Daley, D. J. and Vere-Jones, D. (1988). *An Introduction to the Theory of Point Processes*. Springer, New York, NY, 1 edition.

Dalponte, M. and Coomes, D. A. (2016). Tree-centric mapping of forest carbon density from airborne laser scanning and hyperspectral data. *Methods in Ecology and Evolution*, 7(10):1236–1245.

De la Cruz, R. and Branco, M. D. (2009). Bayesian analysis for nonlinear regression model under skewed errors, with application in growth curves. *Biometrical Journal*, 51(4):588–609.

Diggle, P. J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Chapman and Hall/CRC, New York, 3 edition.

Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo Methods of Inference for Implicit Statistical Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 46(2):193–212.

Drew, L. and Kaplan, A. (2025). *Ldmppr: Estimate and Simulate from Location Dependent Marked Point Processes*.

Drew, L., Kaplan, A., and Breckheimer, I. (2024). Data from "A Bayesian Record Linkage Approach to Applications in Tree Demography Using Overlapping LiDAR Scans".

Drew, L., Kaplan, A., and Breckheimer, I. (2025). A Bayesian record linkage approach to applications in tree demography using overlapping LiDAR scans. *The Annals of Applied Statistics*, 19(3):2027–2052.

Eddelbuettel, D., Francois, R., Allaire, JJ., Ushey, K., Kou, Q., Russell, N., Ucar, I., Bates, D., and Chambers, J. (2023a). *Rcpp: Seamless R and C++ Integration*.

Eddelbuettel, D., Francois, R., Bates, D., Ni, B., and Sanderson, C. (2023b). *RcppArmadillo: 'rcpp' Integration for the 'Armadillo' Templated Linear Algebra Library*.

Fagerberg, N., Olsson, J.-O., Lohmander, P., Andersson, M., and Bergh, J. (2022). Individual-tree distance-dependent growth models for uneven-sized Norway spruce. *Forestry: An International Journal of Forest Research*, 95(5):634–646.

Fellegi, I. P. and Sunter, A. B. (1969). A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Ferraz, A., Saatchi, S., Bormann, K. J., and Painter, T. H. (2018). Fusion of NASA Airborne Snow Observatory (ASO) Lidar Time Series over Mountain Forest Landscapes. *Remote Sensing*, 10(2):164.

Ford, K. R., Breckheimer, I. K., Franklin, J. F., Freund, J. A., Kroiss, S. J., Larson, A. J., Theobald, E. J., and HilleRisLambers, J. (2017). Competition alters tree growth responses to climate at individual and stand scales. *Canadian Journal of Forest Research*, 47(1):53–62.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Geyer, C. J. and Thompson, E. A. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goldstein, H., Harron, K., and Wade, A. (2012). The analysis of record-linked data using multiple imputation with data value priors. *Statistics in Medicine*, 31(28):3481–3493.

Goudar, C. T., Harris, S. K., McInerney, M. J., and Suflita, J. M. (2004). Progress curve analysis for enzyme and microbial kinetic reactions using explicit solutions based on the Lambert W function. *Journal of Microbiological Methods*, 59(3):317–326.

Goulden, T., Hass, B., Brodie, E., Chadwick, K. D., Falco, N., Maher, K., Wainwright, H., and Williams, K. (2020). NEON AOP Survey of Upper East River CO Watersheds: LAZ Files, LiDAR Surface Elevation, Terrain Elevation, and Canopy Height Rasters.

Green, P. J. and Mardia, K. V. (2006). Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 93(2):235–254.

Gu, L., Baxter, R. A., Vickers, D., and Rainsford, C. P. (2003). Record linkage: Current practice and future directions.

Guan, Y. and Afshartous, D. R. (2007). Test for independence between marks and points of marked point processes: A subsampling approach. *Environmental and Ecological Statistics*, 14(2):101–111.

Gumbel, E. J. (1954). Statistical Theory of Extreme Values and Some Practical Applications. A Series of Lectures. Technical Report PB175818, National Bureau of Standards, Washington, D. C. Applied Mathematics Div.

Gutman, R., Afendulis, C. C., and Zaslavsky, A. M. (2013). A Bayesian Procedure for File Linking to Analyze End-of-Life Medical Costs. *Journal of the American Statistical Association*, 108(501):34–47.

Hanks, E. M., Hooten, M. B., Johnson, D. S., and Sterling, J. T. (2011). Velocity-Based Movement Modeling for Individual and Population Level Inference. *PLOS ONE*, 6(8):e22795.

Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35(3):705.

Harte, D. S. (2010). PtProcess: An R package for modelling marked point processes indexed by time. *Journal of Statistical Software*, 35(8):1–32.

Heilman, K. A., Dietze, M. C., Arizpe, A. A., Aragon, J., Gray, A., Shaw, J. D., Finley, A. O., Klesse, S., DeRose, R. J., and Evans, M. E. K. (2022). Ecological forecasting of tree growth: Regional fusion of tree-ring and forest inventory data to quantify drivers and characterize uncertainty. *Global Change Biology*, 28(7):2442–2460.

Hof, M. H., Ravelli, A. C., and Zwinderman, A. H. (2017). A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112(520):1504–1515.

Huo, L. and Lindberg, E. (2020). Individual tree detection using template matching of multiple rasters derived from multispectral airborne laser scanning data. *International Journal of Remote Sensing*, 41(24):9525–9544.

Hyyppä, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., and Maltamo, M. (2008). Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests. *International Journal of Remote Sensing*, 29(5):1339–1366.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. John Wiley & Sons.

Isham, V. and Westcott, M. (1979). A self-correcting point process. *Stochastic Processes and their Applications*, 8(3):335–347.

Johnson, D. J., Magee, L., Pandit, K., Bourdon, J., Broadbent, E. N., Glenn, K., Kaddoura, Y., Machado, S., Nieves, J., Wilkinson, B. E., Almeyda Zambrano, A. M., and Bohlman, S. A. (2021). Canopy tree density and species influence tree regeneration patterns and woody species diversity in a longleaf pine forest. *Forest Ecology and Management*, 490:119082.

Johnson, D. S., Brost, B. M., and Hooten, M. B. (2022). Greater Than the Sum of its Parts: Computationally Flexible Bayesian Hierarchical Modeling. *Journal of Agricultural, Biological and Environmental Statistics*, 27(2).

Johnson, S. G. (2008). *The NLopt Nonlinear-Optimization Package*.

Kaartinen, H., Hyyppä, J., Yu, X., Vastaranta, M., Hyyppä, H., Kukko, A., Holopainen, M., Heipke, C., Hirschmugl, M., Morsdorf, F., Næsset, E., Pitkänen, J., Popescu, S., Solberg, S., Wolf, B. M., and Wu, J.-C. (2012). An International Comparison of Individual Tree Detection and Extraction Using Airborne Laser Scanning. *Remote Sensing*, 4(4):950–974.

Kaplan, A., Betancourt, B., and Steorts, R. C. (2022). A Practical Approach to Proper Inference with Linked Data. *The American Statistician*, 76(4):384–393.

Kim, G. and Chambers, R. (2012). Regression Analysis under Probabilistic Multi-Linkage. *Statistica Neerlandica*, 66(1):64–79.

Koch, B., Heyder, U., and Weinacker, H. (2006). Detection of individual tree crowns in airborne LiDAR data. *Photogrammetric Engineering and Remote Sensing*, 72:357–363.

Kuhn, M. and Wickham, H. (2020). *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.*

Lahiri, P. and Larsen, M. D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100(469):222–230.

Lefsky, M. A., Cohen, W. B., Harding, D. J., Parker, G. G., Acker, S. A., and Gower, S. T. (2002). Lidar remote sensing of above-ground biomass in three biomes. *Global Ecology and Biogeography*, 11(5):393–399.

Leininger, T. J. (2014). *Bayesian Analysis of Spatial Point Patterns*. PhD thesis, Duke University.

Li, P., Dong, X. L., Maurino, A., and Srivastava, D. (2011). Linking temporal records. *Proc. VLDB Endow.*, 4(11):956–967.

Liseo, B. and Tancredi, A. (2011a). Bayesian Estimation of Population Size via Linkage of Multivariate Normal Data Sets. *Journal of Official Statistics*, 27(3):491–505.

Liseo, B. and Tancredi, A. (2011b). Some advances on Bayesian record linkage and inference for linked data.

López, S., France, J., Gerrits, W. J. J., Dhanoa, M. S., Humphries, D. J., and Dijkstra, J. (2000). A generalized Michaelis-Menten equation for the analysis of growth. *Journal of Animal Science*, 78(7):1816–1828.

Lu, X., Hooten, M. B., Kaplan, A., Womble, J. N., and Bower, M. R. (2022). Improving Wildlife Population Inference Using Aerial Imagery and Entity Resolution. *Journal of Agricultural, Biological and Environmental Statistics*.

Ma, Q., Su, Y., Tao, S., and Guo, Q. (2018). Quantifying individual tree growth and tree competition using bi-temporal airborne laser scanning data: A case study in the Sierra Nevada Mountains, California. *International Journal of Digital Earth*, 11(5):485–503.

Maes, S. L., Perring, M. P., Vanhellemont, M., Depauw, L., Van den Bulcke, J., Brūmelis, G., Brunet, J., Decocq, G., den Ouden, J., Härdtle, W., Hédl, R., Heinken, T., Heinrichs, S., Jaroszewicz, B., Kopecký, M., Máliš, F., Wulf, M., and Verheyen, K. (2019). Environmental drivers interactively affect individual tree growth across temperate European forests. *Global Change Biology*, 25(1):201–217.

Marchant, N. G., Kaplan, A., Elazar, D. N., Rubinstein, B. I. P., and Steorts, R. C. (2021). D-blink: Distributed End-to-End Bayesian Entity Resolution. *Journal of Computational and Graphical Statistics*, 30(2):406–421.

Marks, C. O., Yellen, B. C., Wood, S. A., Martin, E. H., and Nislow, K. H. (2020). Variation in Tree Growth along Soil Formation and Microtopographic Gradients in Riparian Forests. *Wetlands*, 40(6):1909–1922.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, 106(493):100–108.

Møller, J., Ghorbani, M., and Rubak, E. (2016). Mechanistic spatio-temporal point process models for marked point processes, with a view to forest stand data. *Biometrics*, 72(3):687–696.

Møller, J. and Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, New York.

Murray, J. S. (2015). Probabilistic Record Linkage and Deduplication after Indexing, Blocking, and Filtering. *Journal of Privacy and Confidentiality*, 7(1).

Myllymäki, M., Mrkvička, T., Grabarnik, P., Seijo, H., and Hahn, U. (2017). Global envelope tests for spatial processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(2):381–404.

Myneni, R. B., Hall, F. G., Sellers, P. J., and Marshak, A. L. (1995). The interpretation of spectral vegetation indexes. *IEEE Transactions on Geoscience and Remote Sensing*, 33(2):481–486.

Nanayakkara, C., Christen, P., and Ranbaduge, T. (2018). Temporal graph-based clustering for historical record linkage.

Nelder, J. A. and Mead, R. (1965). A Simplex Method for Function Minimization. *The Computer Journal*, 7(4):308–313.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., Waterloo, M., and Saleska, S. (2011). Height Above the Nearest Drainage – a hydrologically relevant new terrain model. *Journal of Hydrology*, 404(1):13–29.

Padmanabhan, S., Carty, L., Cameron, E., Ghosh, R. E., Williams, R., and Strongman, H. (2019). Approach to record linkage of primary care data from Clinical Practice Research Datalink to other health-related patient data: Overview and implications. *European Journal of Epidemiology*, 34(1):91–99.

Pettorelli, N., Vik, J. O., Mysterud, A., Gaillard, J.-M., Tucker, C. J., and Stenseth, N. C. (2005). Using the satellite-derived NDVI to assess ecological responses to environmental change. *Trends in Ecology & Evolution*, 20(9):503–510.

Pommerening, A. and Sánchez Meador, A. J. (2018). Tamm review: Tree interactions between myth and reality. *Forest Ecology and Management*, 424:164–176.

Poorazimy, M., Ronoud, G., Yu, X., Luoma, V., Hyyppä, J., Saarinen, N., Kankare, V., and Vastaranta, M. (2022). Feasibility of Bi-Temporal Airborne Laser Scanning Data in Detecting Species-Specific Individual Tree Crown Growth of Boreal Forests. *Remote Sensing*, 14(19):4845.

Powell, M. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Technical Report, Department of Applied Mathematics and Theoretical Physics*.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, 92(437):179–191.

Rathbun, S. L. and Cressie, N. (1994). A space-time survival point process for a longleaf pine forest in southern georgia. *Journal of the American Statistical Association*, 89(428):1164–1174.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge University Press, Cambridge.

Rosenthal, J. S. (2011). Optimal Proposal Distributions and Adaptive MCMC. In *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC.

Roussel, J.-R., Auty, D., Coops, N. C., Tompalski, P., Goodbody, T. R. H., Meador, A. S., Bourdon, J.-F., de Boissieu, F., and Achim, A. (2020). lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sensing of Environment*, 251:112061.

Rowan, T. H. (1990). *Functional Stability Analysis of Numerical Algorithms*. PhD thesis, University of Texas at Austin, USA.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and Morel, A. (2011).

Benchmark map of forest carbon stocks in tropical regions across three continents. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24):9899–9904.

Sadinle, M. (2017). Bayesian Estimation of Bipartite Matchings for Record Linkage. *Journal of the American Statistical Association*, 112(518):600–612.

Sadinle, M. (2018). Bayesian propagation of record linkage uncertainty into population size estimation of human rights violations. *The Annals of Applied Statistics*, 12(2):1013–1038.

Schlather, M., Ribeiro, P. J., and Diggle, P. J. (2004). Detecting Dependence between Marks and Locations of Marked Point Processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(1):79–93.

Stan Development Team (2023). RStan: The R interface to Stan.

Steorts, R. C. (2015). Entity Resolution with Empirically Motivated Priors. *Bayesian Analysis*, 10(4):849–875.

Steorts, R. C., Hall, R., and Fienberg, S. E. (2016). A Bayesian Approach to Graphical Record Linkage and Deduplication. *Journal of the American Statistical Association*, 111(516):1660–1672.

Steorts, R. C., Tancredi, A., and Liseo, B. (2018). Generalized Bayesian Record Linkage and Regression with Exact Error Propagation. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*, pages 297–313, Berlin, Heidelberg. Springer-Verlag.

Steorts, R. C., Ventura, S. L., Sadinle, M., and Fienberg, S. E. (2014). A Comparison of Blocking Methods for Record Linkage. In Domingo-Ferrer, J., editor, *Privacy in Statistical Databases*, Lecture Notes in Computer Science, pages 253–268, Cham. Springer International Publishing.

Taylor, I., Kaplan, A., and Betancourt, B. (2024). Fast Bayesian Record Linkage for Streaming Data Contexts. *Journal of Computational and Graphical Statistics*, 33(3):833–844.

Tinkham, W. T. and Woolsey, G. A. (2024). Influence of Structure from Motion Algorithm Parameters on Metrics for Individual Tree Detection Accuracy and Precision. *Remote Sensing*, 16(20):3844.

Wang, J., , P. M., R., , K. P., P., and and Kettle, W. D. (2004). Relations between NDVI and tree productivity in the central Great Plains. *International Journal of Remote Sensing*, 25(16):3127–3138.

Wensel, L., Meerschaert, W., and Biging, G. (1987). Tree height and diameter growth models for Northern California conifers. *Hilgardia*, 55(8):1–20.

Winkler, W. E. (2006). Overview of Record Linkage and Current Research Directions. Research Report Series RR2006/02, U.S. Census Bureau, Washington, DC.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Zanella, G. (2020). Informed Proposals for Local MCMC in Discrete Spaces. *Journal of the American Statistical Association*, 115(530):852–865.

Zhang, W., Qi, J., Wan, P., Wang, H., Xie, D., Wang, X., and Yan, G. (2016). An easy-to-use airborne LiDAR data filtering method based on cloth simulation. *Remote Sensing*, 8(501).

Zitová, B. and Flusser, J. (2003). Image registration methods: A survey. *Image and Vision Computing*, 21(11):977–1000.

Appendix A

A Bayesian Record Linkage Approach to Applications in Tree Demography Using Overlapping LiDAR Scans Supplementary Material

A.1 Model specification and implementation details

A.1.1 Model specification

We present the joint posterior distribution of the record linkage model and the full conditional distributions, which provide the basis for the MCMC algorithm used to sample from the posterior distribution. In the following sections, the rotation and translation parameters indexed by i take values $i = 1, 2$ where $\theta_1 = 0$ and $\boldsymbol{t}_1 = \begin{bmatrix} 0 & 0 \end{bmatrix}^\mathsf{T}$ are assumed to be fixed.

The joint posterior distribution of the model is

$$
\begin{aligned}
&[\boldsymbol{\Lambda}, \boldsymbol{s}, \sigma^2, \theta_2, \boldsymbol{t}_2 \mid \boldsymbol{y}] \\
&\propto \prod_{i=1}^{2}\prod_{j=1}^{n_i}[\boldsymbol{y}_{ij} \mid \boldsymbol{s}_{\lambda_{ij}}, \sigma^2, \theta_i, \boldsymbol{t}_i, D] \times \prod_{j'=1}^{N}[\boldsymbol{s}_{j'}] \times [\boldsymbol{\Lambda} \mid N] \times [\sigma^2] \times [\theta_2] \times [\boldsymbol{t}_2] \\
&\propto \left\{\prod_{i=1}^{2}\prod_{j=1}^{n_i}(\sigma^2)^{-1}\exp\Big(-\frac{1}{2\sigma^2}(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D))^\mathsf{T}\right. \\
&\qquad\qquad \left.(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D))\Big)\right\} \\
&\times |D|^{-N}\frac{d_\sigma^{c_\sigma}}{\Gamma(c_\sigma)}(\sigma^2)^{-c_\sigma-1}\exp\Big(-\frac{d_\sigma}{\sigma^2}\Big)I\left\{\sigma^2 < b_\sigma\right\}\frac{N!}{N^n} \\
&\times \exp\Big(\kappa\cos(\nu)\cos(\theta_2) + \kappa\sin(\nu)\sin(\theta_2)\Big)\exp\Big(-\frac{1}{2\sigma_t^2}\boldsymbol{t}_2^\mathsf{T}\boldsymbol{t}_2\Big) \\
&\propto \exp\Big(-\frac{1}{\sigma^2}\Big\{\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D))^\mathsf{T}\right. \\
&\qquad\qquad \left.(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D)) + d_\sigma\Big\}\Big) \\
&\times (\sigma^2)^{-n-c_\sigma-1}I\left\{\sigma^2 < b_\sigma\right\}\exp\Big(\kappa\cos(\nu)\cos(\theta_2) + \kappa\sin(\nu)\sin(\theta_2)\Big)\exp\Big(-\frac{1}{2\sigma_t^2}\boldsymbol{t}_2^\mathsf{T}\boldsymbol{t}_2\Big),
\end{aligned}
$$

where $n = \sum_{i=1}^{2} n_i$.

The full conditional distributions for the model parameters are

$$\sigma^2 \mid \boldsymbol{s}, \boldsymbol{\Lambda}, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{y} \sim \text{Inverse-Gamma}_{[0,b_\sigma]}\Bigg(n + c_\sigma,$$
$$\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D))^\mathsf{T}$$
$$(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D)) + d_\sigma\Bigg)$$

$$\boldsymbol{s}_{j'} \mid \boldsymbol{\Lambda}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{T}, \boldsymbol{y} \sim \text{N}_{2,[D^*]}\Bigg(\frac{1}{n_{j'}}\sum_{i=1}^{2}\sum_{(j):\lambda_{ij}=j'}\Big[\boldsymbol{R}(\theta_i)^\mathsf{T}(\boldsymbol{y}_{ij} - \boldsymbol{t}_i - \boldsymbol{\mu}_D) + \boldsymbol{\mu}_D\Big], \frac{\sigma^2}{n_{j'}}\boldsymbol{I}\Bigg)$$

$$P(\lambda_{ij} = \ell \mid \boldsymbol{\Lambda}_{-(ij)}, \sigma^2, \boldsymbol{s}, \theta_i, \boldsymbol{t}_i, \boldsymbol{y}) \propto \exp\Bigg(-\frac{1}{2\sigma^2}(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_\ell - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D))^\mathsf{T}$$
$$(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i)(\boldsymbol{s}_\ell - \boldsymbol{\mu}_D) + \boldsymbol{t}_i + \boldsymbol{\mu}_D))\Bigg)$$

$$\theta_2 \mid \boldsymbol{s}, \boldsymbol{\Lambda}, \sigma^2, \boldsymbol{t}_2, \boldsymbol{y} \propto \exp\Bigg((\kappa\cos(\nu) + \boldsymbol{S}_{2_{11}} + \boldsymbol{S}_{2_{22}})\cos(\theta_2) + (\kappa\sin(\nu) - \boldsymbol{S}_{2_{12}} + \boldsymbol{S}_{2_{21}})\sin(\theta_2)\Bigg)$$
$$\text{where } \boldsymbol{S}_2 = \frac{1}{2\sigma^2}\sum_{j=1}^{n_2}(\boldsymbol{y}_{2j} - \boldsymbol{t}_2 - \boldsymbol{\mu}_D)(\boldsymbol{s}_{\lambda_{2j}} - \boldsymbol{\mu}_D)^\mathsf{T}$$

$$\boldsymbol{t}_2 \mid \boldsymbol{s}, \boldsymbol{\Lambda}, \sigma^2, \theta_2, \boldsymbol{y} \sim \text{N}_2\Bigg(\Bigg(\frac{\sigma_t^2}{n_2\sigma_t^2 + \sigma^2}\Bigg)\sum_{j=1}^{n_2}\{\boldsymbol{y}_{2j} - [\boldsymbol{R}(\theta_2)(\boldsymbol{s}_{\lambda_{2j}} - \boldsymbol{\mu}_D) + \boldsymbol{\mu}_D]\}, \frac{\sigma_t^2\sigma^2}{n_2\sigma_t^2 + \sigma^2}\boldsymbol{I}\Bigg).$$

We note that the conditional distribution of the linkage structure reflects the possibility of duplicate records within files (or scans).

A.1.2 Considerations when specifying N

We provide additional insight into the specification of the hyperparameter N, which controls the maximum number of unique latent individuals in the model and consequently the number of observed clusters. As noted in Section 3.1, the choice of N can have a large impact on the linkage behavior of the model. If N is too small, the model may not be able to accurately capture the linkage structure, leading to poor performance as a result of over linking. If N is too large, the model may be overly flexible and may not be able to accurately capture the linkage structure due to under linking. For example, specifying $N = max(n_i)$ would result in a model that guarantees no under linking, but may functionally result in forcing the model to link points that shouldn't be. In contrast, specifying $N = \sum_{i=1}^{2} n_i$ would result in a model that guarantees no over linking, but may result in the model failing to link points that should be linked.

In practice, the choice of N should be informed by the expected number of unique individuals in the dataset. In the absence of prior knowledge about the degree of overlap between the files, we recommend specifying N more conservatively so as not to induce over linking of unrelated records. See the included comparisons of the linkage performance of the model under different values of N in the simulation study in Appendix A.4.2.

A.1.3 Gibbs sampler algorithm

The algorithm below describes the MCMC Gibbs sampler algorithm for sampling from the joint posterior distribution of the spatial record linkage model. As above, the rotation and translation parameters are only sampled for file 2.

1. Define initial values for $\sigma^{2(0)}$, $\{\boldsymbol{s}_{j'}^{(0)}\}_{j'=1}^{N}$, $\theta_2^{(0)}$, $\boldsymbol{t}_2^{(0)}$, and $\boldsymbol{\Lambda}^{(0)}$.
2. Set $k = 1$.
3. Update $\boldsymbol{\Lambda}^{(k)}$ using a Gibbs sampling step for each $\lambda_{ij}^{(k)}$, making use of the Gumbel max trick from Gumbel (1954),

$$\begin{aligned}
\eta_{ij\ell}^{(k)} &= -\frac{1}{2\sigma^2}(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i^{(k-1)})(\boldsymbol{s}_\ell^{(k-1)} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i^{(k-1)} + \boldsymbol{\mu}_D))^{\mathsf{T}} \\
&\quad (\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i^{(k-1)})(\boldsymbol{s}_\ell^{(k-1)} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i^{(k-1)} + \boldsymbol{\mu}_D)) \\
z_\ell &\overset{iid}{\sim} \text{Gumbel}(0,1) \\
\lambda_{ij}^{(k)} &= \arg\max_{\ell=1,\ldots,N} \eta_{ij\ell}^{(k)} + z_\ell
\end{aligned}$$

4. Update $\boldsymbol{s}^{(k)}$ using Gibbs sampling for each $\boldsymbol{s}_{j'}^{(k)}$,

$$\boldsymbol{s}_{j'}^{(k)} \sim \text{N}_{2,[D^*]}\left(\frac{1}{n_{j'}^{(k)}}\sum_{i=1}^{2}\sum_{(j):\lambda_{ij}^{(k)}=j'}\left[\boldsymbol{R}(\theta_i^{(k-1)})^{\mathsf{T}}(\boldsymbol{y}_{ij} - \boldsymbol{t}_i^{(k-1)} - \boldsymbol{\mu}_D) + \boldsymbol{\mu}_D\right], \frac{\sigma^{2(k-1)}}{n_{j'}^{(k)}}\boldsymbol{I}\right)$$

5. Update $\sigma^{2(k)}$ using Gibbs sampling,

$$\begin{aligned}
\sigma^{2(k)} \sim \text{Inverse-Gamma}_{[0,b_\sigma]}\Bigg(& n + c_\sigma, \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i^{(k-1)})(\boldsymbol{s}_{\lambda_{ij}^{(k)}}^{(k)} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i^{(k-1)} + \boldsymbol{\mu}_D))^{\mathsf{T}} \\
& (\boldsymbol{y}_{ij} - (\boldsymbol{R}(\theta_i^{(k-1)})(\boldsymbol{s}_{\lambda_{ij}^{(k)}}^{(k)} - \boldsymbol{\mu}_D) + \boldsymbol{t}_i^{(k-1)} + \boldsymbol{\mu}_D)) + d_\sigma\Bigg).
\end{aligned}$$

6. Update ${\theta_2}^{(k)}$ using Metropolis–Hastings. Propose $\theta_2^* \sim \mathrm{N}(\theta_2^{(k-1)}, \sigma^2_{\theta,\text{tune}})$, noting that the proposal is symmetric. Calculate the M-H ratio as follows

$$mh_{\theta_2} = \frac{[\theta_2^*, \boldsymbol{s}^{(k)}, \boldsymbol{\Lambda}^{(k)}, {\sigma^2}^{(k)}, \boldsymbol{t}_2^{(k-1)} \mid \boldsymbol{y}]}{[\theta_2^{(k-1)}, \boldsymbol{s}^{(k)}, \boldsymbol{\Lambda}^{(k)}, {\sigma^2}^{(k)}, \boldsymbol{t}_2^{(k-1)} \mid \boldsymbol{y}]}.$$

Set $\theta_2^{(k)} = \theta_2^*$ with probability $\min(1, mh_{\theta_2})$; otherwise, set $\theta_2^{(k)} = \theta_2^{(k-1)}$.

7. Update $\boldsymbol{t}_2^{(k)}$ using Gibbs sampling,

$$\boldsymbol{t}_2^{(k)} \sim \mathrm{N}_2\Bigg(\bigg(\frac{\sigma_t^2}{n_2\sigma_t^2 + {\sigma^2}^{(k)}}\bigg)\sum_{j=1}^{n_2}(\boldsymbol{y}_{2j} - [\boldsymbol{R}(\theta_2^{(k)})(\boldsymbol{s}_{\lambda_{2j}}^{(k)} - \boldsymbol{\mu}_D) + \boldsymbol{\mu}_D], \frac{\sigma_t^2 {\sigma^2}^{(k)}}{n_2\sigma_t^2 + {\sigma^2}^{(k)}}\boldsymbol{I}\Bigg).$$

8. Save $\{\boldsymbol{s}_{j'}^{(k)}\}_{j'=1}^{N}$, ${\sigma^2}^{(k)}$, $\theta_2^{(k)}$, $\boldsymbol{t}_2^{(k)}$, and $\boldsymbol{\Lambda}^{(k)}$.
9. Set $k = k + 1$ and return to Step 3. Iterate this algorithm through steps 3-8 until the sample size is large enough to adequately approximate the joint posterior distribution.

For the Metropolis–Hastings within Gibbs step for updating θ_2, we use an adaptive algorithm that targets an acceptance ratio of 0.44 as discussed in Chapter 4 of Brooks et al. (2011).

A.2 Proof of Theorem 4.1

We provide the proof of Theorem 4.1 (Bayesian validity of linkage-averaged auxiliary data model parameters joint posterior), which follows a similar structure to the proof of Theorem 4.1 in Sadinle (2018).

Proof. The joint posterior of $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{s}$ is

$$p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) \propto \mathcal{L}_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) p_{\mathrm{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(s)) p(\boldsymbol{\Lambda}) p(\boldsymbol{s})$$

such that the inverse proportionality constant is

$$\sum_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{s}} \sum_{\boldsymbol{\Theta}} \mathcal{L}_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) p_{\mathrm{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(s)) p(\boldsymbol{\Lambda}) p(\boldsymbol{s}) = \sum_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{s}} \mathcal{L}_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y})$$

since $\sum_{\boldsymbol{\Theta}} p_{\mathrm{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(s)) = 1$. As $p_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) \propto \mathcal{L}_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) p(\boldsymbol{\Lambda}) p(\boldsymbol{s})$ with inverse proportionality constant $\sum_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{s}} \mathcal{L}_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) p(\boldsymbol{\Lambda}) p(\boldsymbol{s})$, it follows that

$$p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) = p_{\mathrm{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(\boldsymbol{s})) p_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}).$$

Then,

$$p(\boldsymbol{\Theta} \mid \boldsymbol{y}) = \sum_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{s}} p(\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) = \sum_{\boldsymbol{\Lambda}} \sum_{\boldsymbol{s}} p_{\mathrm{AD}}(\boldsymbol{\Theta} \mid \mathcal{C}(\boldsymbol{\Lambda}), \boldsymbol{X}(s)) p_{\mathrm{L}}(\boldsymbol{\Lambda}, \boldsymbol{s} \mid \boldsymbol{y}) = p_{\mathrm{LA}}(\boldsymbol{\Theta}).$$

□

A.3 Empirical data analysis details

A.3.1 Empirical model specifications

Following the general structure introduced in Section 3.1, in our analysis of the RMBL dataset we specified the record linkage model as follows

$$\begin{aligned}
\boldsymbol{y}_{ij} \mid \boldsymbol{s}_{\lambda_{ij}}, \sigma^2, \theta_i, \boldsymbol{t}_i, D &\overset{ind}{\sim} \mathrm{Normal}_{2,[D]}\left(\boldsymbol{R}(\theta_i)\left(\boldsymbol{s}_{\lambda_{ij}} - \boldsymbol{\mu}_D\right) + \boldsymbol{t}_i + \boldsymbol{\mu}_D, \sigma^2 \boldsymbol{I}\right) \\
\boldsymbol{s}_{j'} \mid N &\overset{iid}{\sim} \mathrm{Uniform}\left(D^*\right) \\
\sigma^2 &\sim \text{Inverse-Gamma}_{[0,3.175]}\left(0.0001, 0.0001\right) \\
\lambda_{ij} \mid N &\overset{iid}{\sim} \mathrm{Uniform}\left\{1, \ldots, N\right\} \\
\boldsymbol{t}_2 &\sim \mathrm{Normal}_2\left(\boldsymbol{0}, .0001^2 \boldsymbol{I}\right),
\end{aligned}$$

where D^* extends the bounds of D by 1 meter in each direction and with $\theta_1 = \theta_2 = 0$. We specified the prior on σ^2 to be uninformative and chose the upper bound relative to the maximum reasonable displacement between a tree top and its base. We selected $\sigma_t^2 = .0001^2$ to limit the amount of translation possible relative to the expectation that these shifts would be on the scale of 10-30 centimeters based on the calibration of the ALS equipment.

Following the general structure introduced in Section 3.2, we specified the downstream growth model as follows

$$
\begin{aligned}
g_c \mid \gamma, \boldsymbol{\beta}, \tau^2, \boldsymbol{\Lambda}, \boldsymbol{x}_{\boldsymbol{s}_c}, \boldsymbol{v}^* &\overset{ind}{\sim} \text{Skewed t}\left(\mu_c, \tau, \delta, \omega\right) \\
\tau &\sim \text{Uniform}\left(0, 100\right) \\
\delta &\sim \text{Normal}_{[-1,1]}\left(0, .1^2\right) \\
\omega &\sim \text{Gamma}\left(2, 0.2\right) \\
\gamma &\sim \text{Uniform}\left(200, 2073.17\right) \\
\alpha &\sim \text{Beta}_{[.5,5]}\left(1, 1\right) \\
\beta_0 &\sim \text{Normal}\left(0, 20^2\right) \\
\beta_k &\overset{ind}{\sim} \text{Normal}\left(0, 2.5^2\right), \text{ for } k = 1, \ldots, 9.
\end{aligned}
$$

The hyperparameters for this model were chosen to be uninformative where possible. In specifying the range for γ, we relied on the input of our subject matter expert for the lower bound and specified the upper bound as the largest size (canopy volume) observed in the 2015 dataset. For the lower bound on α, we found when this value was unconstrained the sampler would push the value towards 0 so we specified a reasonable lower bound based on our expectations regarding the shape of the growth function.

A.3.2 Empirical model convergence diagnostics

In this section, we provide select convergence diagnostics for the two-stage record linkage model fit to the RMBL dataset.

To assess the convergence of the record linkage model, we inspected traceplots for the continuous model parameters and then considered the number of unique estimated individuals from the linkage shown in Figure A.1. We see that the chains converge to a stable distribution, and after burn-in we calculate a corresponding Gelman–Rubin statistic of $\hat{R} = 1.10071$.

A.3.3 Model selection criteria

In this section, we provide the model selection details for the empirical data analysis mentioned in Section 5. Figure A.2 shows the replicated densities from the fitted models for each linkage

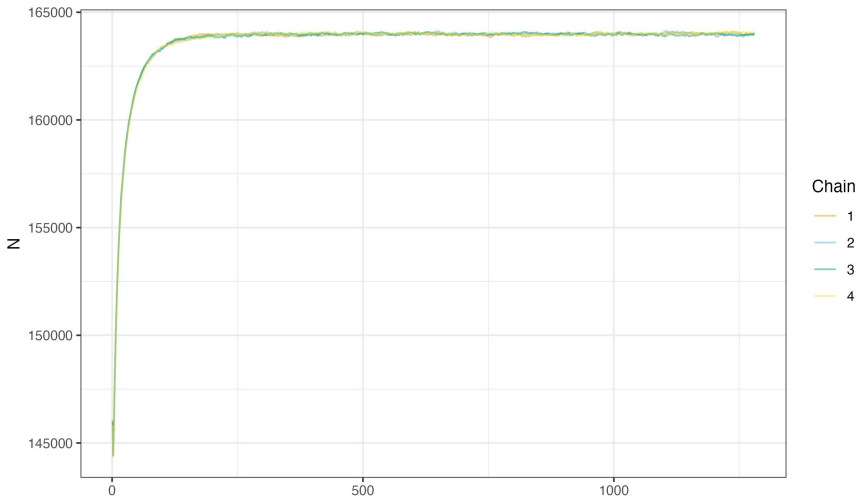

Figure A.1: Traceplot for the number of unique individuals from the four empirical linkage model chains after thinning.

approach. We note that the replicated densities are only for a single iteration from each model, but are representative of the general behavior of the models.

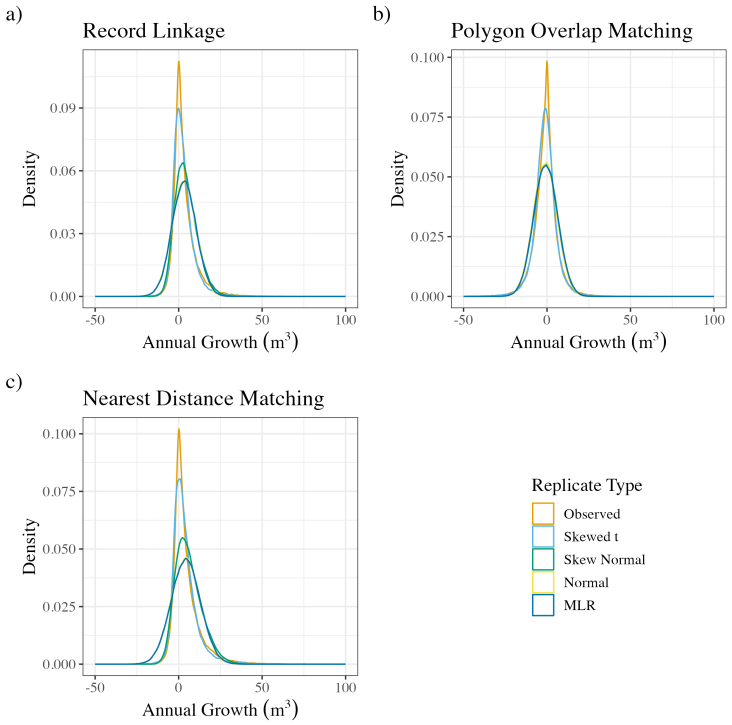

Figure A.2: Plots comparing the observed density of annual growth to the replicated densities from the Skewed t, Skew Normal, Normal, and MLR models for each linkage approach.

In Table A.1, we present the sCRPS for each model variant for the Linkage-Averaging approach fit to the RMBL dataset. Values of sCRPS closer to 0 indicate a better fit for the model. Results for the POM and NDM linkage approaches were similar as seen in Table A.2. All three linkage approaches selected the Skewed t as the best fitting model in terms of sCRPS. We note that the sCRPS presented for the LA approach is the average sCRPS across the 100 different downstream model fits.

Table A.1: sCRPS scores by model variant for the Linkage-Averaging approach.

	sCRPS Value	
Model	Mean	SE
Skewed t	**-1.9766**	0.0003
Skew Normal	-2.0198	0.0003
Normal	-1.9939	0.0002
MLR	-1.9966	0.0002

Table A.2: sCRPS scores by model variant for the Polygon Overlap Matching and Nearest Distance Matching approaches.

		sCRPS Value	
Linkage	Model	Estimate	SE
POM	Skewed t	**-1.9774**	0.0016
	Skew Normal	-1.9851	0.0014
	Normal	-1.9870	0.0014
	MLR	-1.9889	0.0014
NDM	Skewed t	**-2.0562**	0.0032
	Skew Normal	-2.1033	0.0027
	Normal	-2.0756	0.0021
	MLR	-2.0768	0.0021

A.4 Simulation study details

A.4.1 Data simulation algorithm

In this section, we provide the details of our algorithm for generating the simulated datasets used in our simulation study.

To facilitate the generation of a spatially dependent marked point process, we combined the likelihood based approach of Møller et al. (2016) with a nested predictive model. Using the `tidymodels` [Kuhn and Wickham (2020)] framework in R and the `XGBoost` engine, we constructed models for the low, medium, and high density subsets that predict canopy volume as a function of spatial covariates, time, and neighborhood dynamics in a radius of 15 m. We engineered a collection of relevant features to capture inter-point interactions in concert with the location-specific covariates derived from the raster images provided by RMBL. The raster images were centered and scaled on $130\,\mathrm{m}^2$ areas, and then the predictive models were built on a reduced $100\,\mathrm{m}^2$ area to guarantee that the points near the boundary have accurate values for the neighborhood characteristic covariates.

The predictive models for canopy volume were tuned, tested, and validated using training and test data splits (80/20) and 5-fold cross validation. The resulting models have high predictive accuracy, and may overfit the observed data, but serve as a reasonable basis for the data generation algorithm described in detail below. For each density we also generated a predictive model for size-dependent interaction radii. We fit a cubic linear regression model with canopy radius as the response and canopy volume as the predictor based on the documented allometric relationship between canopy area and canopy volume. These models allow older (i.e., larger) points to have larger interaction radii, reflecting the biological mechanisms of tree growth and interaction.

We outline the data generation algorithm for the underlying point process and the observed files as follows.

1. Specify the spatial domain D, assumed to be a $100m^2$ area, and the expanded domain of interest D^*, where D^* is taken to be an expansion of D of size $130\,\mathrm{m}^2$ obtained by extending the boundaries of D by 15 m in each direction. Specify N, the number of latent points in the expanded domain D^*. We consider three density settings motivated by the empirical data, which correspond to $N = 800, 1251, 1321$ in $130\,\mathrm{m}^2$ areas for the low, medium, and high density settings. Obtain the marks from the empirical dataset with appropriate density, where we note that the marks in this application are canopy volumes. We denote the marks as V_i for $i \in \{1, \ldots, N\}$, and once selected, we order the V_i in descending order from largest to smallest. Generate the ages, $T_0, \ldots, T_{N-1}$, from the marks such that $T_i = V_{(1)} - V_{(i)}$ with $T_0 = 0$.

2. For $i = 1$, generate a point $\boldsymbol{s}_1$ uniformly on D^* and predict the size based on the location and interaction radius based on the predicted size using the predictive models discussed previously. Set $i = 2$.

3. For $i = 2, \ldots, N$ generate a point $\boldsymbol{s}_i$ uniformly on D^* and derive the covariate values from the generated location and collection of previously accepted points $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{i-1}$. Predict the size and interaction radius of $\boldsymbol{s}_i$ and determine if $\boldsymbol{s}_i$ is within the interaction radius of any of the previously generated points. If $\boldsymbol{s}_i$ is not within the interaction radius of any of the previously generated points, accept $\boldsymbol{s}_i$. If $\boldsymbol{s}_i$ is within the interaction radius of any of the previously generated points, generate a Uniform$(0, 1)$ random variable U_k for each point of overlap $(1, \ldots, k)$, and accept the location of $\boldsymbol{s}_i$ if all $U_k < .01$. If $\boldsymbol{s}_i$ is rejected, generate a new location uniformly on D^* and repeat the procedure above. We allow points to violate the hardcore interaction radius to more accurately reflect the mechanics in the empirical data. This procedure is iterated until $i = N$. We note that each obtained point has a location, mark, and interaction radius associated with it at the end of this step.

4. The points simulated in steps 2 and 3 serve as the latent parent points, and we next turn to the mechanism for obtaining recruits, which are points that are under the observable size threshold in the empirical data collection procedure. To obtain recruits, we consider the total observable canopy volume in D^*, i.e., $\lfloor \sum_{i=1}^{N} V_i \rfloor$, and we generate an additional set of unobservable latent points of cardinality equal to the integer part of the total canopy volume of the observable latents (i.e., recruits). The points are generated by sampling from the set of observable latents, with replacement, using the proportion of canopy volume contribution over the domain D^* as the sampling weight. The recruit locations are then sampled from bivariate t(1) distributions centered at the sampled latent points. This mechanism allows older (larger) points to generate more potential recruits reflecting the known biological mechanisms of recruitment.

   The sizes for each point are sampled from a truncated Beta distribution scaled from zero to the minimum predicted size, $V_{(1)}$, of the observable latents. We note that the detection threshold for the empirical data corresponds to a height of approximately 2 m, and we define the canopy volume threshold to be the minimum predicted value in the generated data. As discussed

in the procedure for the observable latent points, we allow for the possibility that recruits violate the interaction radius of parent trees by a similar mechanism. Points that violate the interaction radius of parents points are removed, unless they satisfy the constraint described in step 3. Recruit points that fall outside of the expanded domain D^* are also discarded.

5. To generate the observed data for the first file $\boldsymbol{Y}_1^*$, take the observable latent points $\boldsymbol{s}_i$ for $i = 1, \ldots, N$ and for each i, obtain $\boldsymbol{y}_{1i}^* \sim \text{Normal}_2(\boldsymbol{s}_i, \sigma^2 \boldsymbol{I})$. Each $\boldsymbol{y}_{1i}^*$ represents a noisy observed value of $\boldsymbol{s}_i$ to mimic post-processing error specified in varying levels of small ($\sigma = .25$), medium ($\sigma = .35$), and large ($\sigma = .45$).
6. To generate the observed data for the second file, we consider the set of all observable and unobservable latent points, $\boldsymbol{s}^*$ of size N^*, on the expanded domain D^*, and perform the following process to mimic the annual growth over a one year period.
   a. We define a growth function of the Michaelis–Menten form such that

$$\mu_i = \frac{a_i v_{i,t_1}^{\alpha}}{\gamma^{\alpha} + v_{i,t_1}^{\alpha}} \text{ where } a_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i},$$

   and generate the expected annual growth for each individual of non-zero size given its current size (canopy volume) $v_{i,t}$ and covariate values for southness, slope, TWI, and DEM at its spatial location. For $i = 1, \ldots, N$, obtain $g_i \sim \text{Normal}(\mu_i, \tau^2)$. For $i = N + 1, \ldots, N^*$, obtain $g_i = \mu_i$. Calculate the updated size for each individual by adding the estimated growth to the current size. We note that it is possible for the observed growth to be negative for the parent latent points as a function of measurement error, though it is unlikely for the true growth to be negative.
   b. Update the growth of individuals from the original set of all observable and unobservable latents after completing the growth cycle. In this application, we perform this cycle once to mimic the annual growth cycle over a one year period, but the procedure may be iterated successively to mimic the annual cycle over a multi-year span.
   c. Obtain the set of points with sizes above the specified detection threshold, and denote the set of points as $\boldsymbol{s}_2^*$ with size N_2^*.

d. For all individuals in the set $\boldsymbol{s}_2^*$, i.e., those recruited into the observable population after the first observed time, generate $y_{2i}^* \sim \text{Normal}_2(\boldsymbol{R}(\theta_2)(\boldsymbol{s}_{2i}^* - \boldsymbol{\mu}_D) + \boldsymbol{t}_2 + \boldsymbol{\mu}_D, \sigma^2\boldsymbol{I})$. This procedure applies rotation and translation to the latent locations of all individuals in the second file as functions of the parameters θ and $\boldsymbol{t}$, and then adds noise to mimic post-processing error as described in the procedure for the first file. The rotation parameter is assumed to be on the order of a fraction of a degree (i.e., $\theta_2 = .005$) and the translation is assumed to be on the magnitude of $30cm$ at most (i.e., $\boldsymbol{t}_2^\mathsf{T} = \begin{bmatrix} 0.025 & 0.025 \end{bmatrix}^\mathsf{T}$).

7. After obtaining the two datasets, $\boldsymbol{Y}_1^*$ and $\boldsymbol{Y}_2^*$ of sizes N_1^* and N_2^* respectively, truncate the observable domain to D and obtain the final observed datasets $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ of sizes n_1 and n_2.

A.4.2 Additional simulation results

In this section, we provide some additional model specification details and simulation results for the two-stage models. We compare the performance of the linkage-averaging and nearest distance matching approaches compared to the true linkage. We provide the precision and recall comparison plots for $\alpha = 2$ and $\alpha = 3$ in Figure A.3 and Figure A.4 respectively. We note that the precision and recall plots for $\alpha = 1$ are shown in Section 6, and the results are quite similar to what we see for the alternative values of α. We also include precision and recall comparisons for the same datasets fit with $q = 1.1$ such that $N = 1.1 \times \max(n_i)$ as a point of comparison for $\alpha = 1$, $\alpha = 2$, and $\alpha = 3$ in Figure A.5, Figure A.6, and Figure A.7 respectively. We see that the specification of N impacts the linkage behavior of the record linkage model such that increasing N results in higher precision, but lower recall which may impact the downstream analysis.

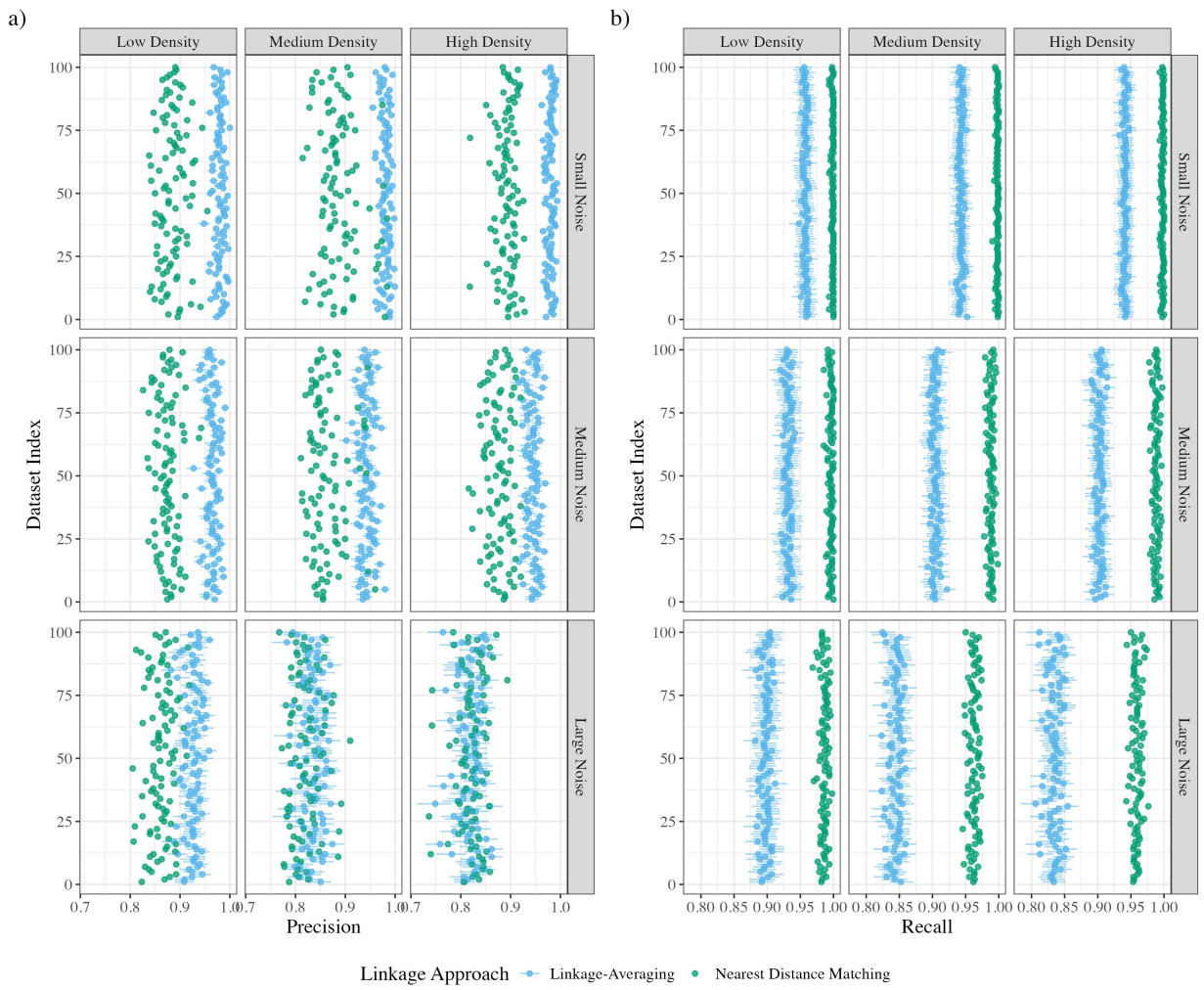

Figure A.3: Plots comparing the precision a) and recall b) performance for the LA and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with $\alpha = 2$.

To evaluate the performance of the downstream growth model, we employ a Gaussian measurement error growth model of the following form

$$
\begin{aligned}
g_c \mid \gamma, \boldsymbol{\beta}, \tau^2, \boldsymbol{\Lambda}, \boldsymbol{x}_{\boldsymbol{s}_c}, \boldsymbol{v}^* &\overset{ind}{\sim} \text{Normal}(\mu_c, \tau^2) \\
\tau &\sim \text{Uniform}\,(0, 100) \\
\gamma &\sim \text{Uniform}\,(a_\gamma, 1500) \\
\alpha &\sim \text{Beta}_{[0,5]}\,(1, 1) \\
\beta_0 &\sim \text{Normal}(0, 20) \\
\beta_k &\overset{ind}{\sim} \text{Normal}(0, 2.5), \text{ for } k = 1, \ldots, 4,
\end{aligned}
$$

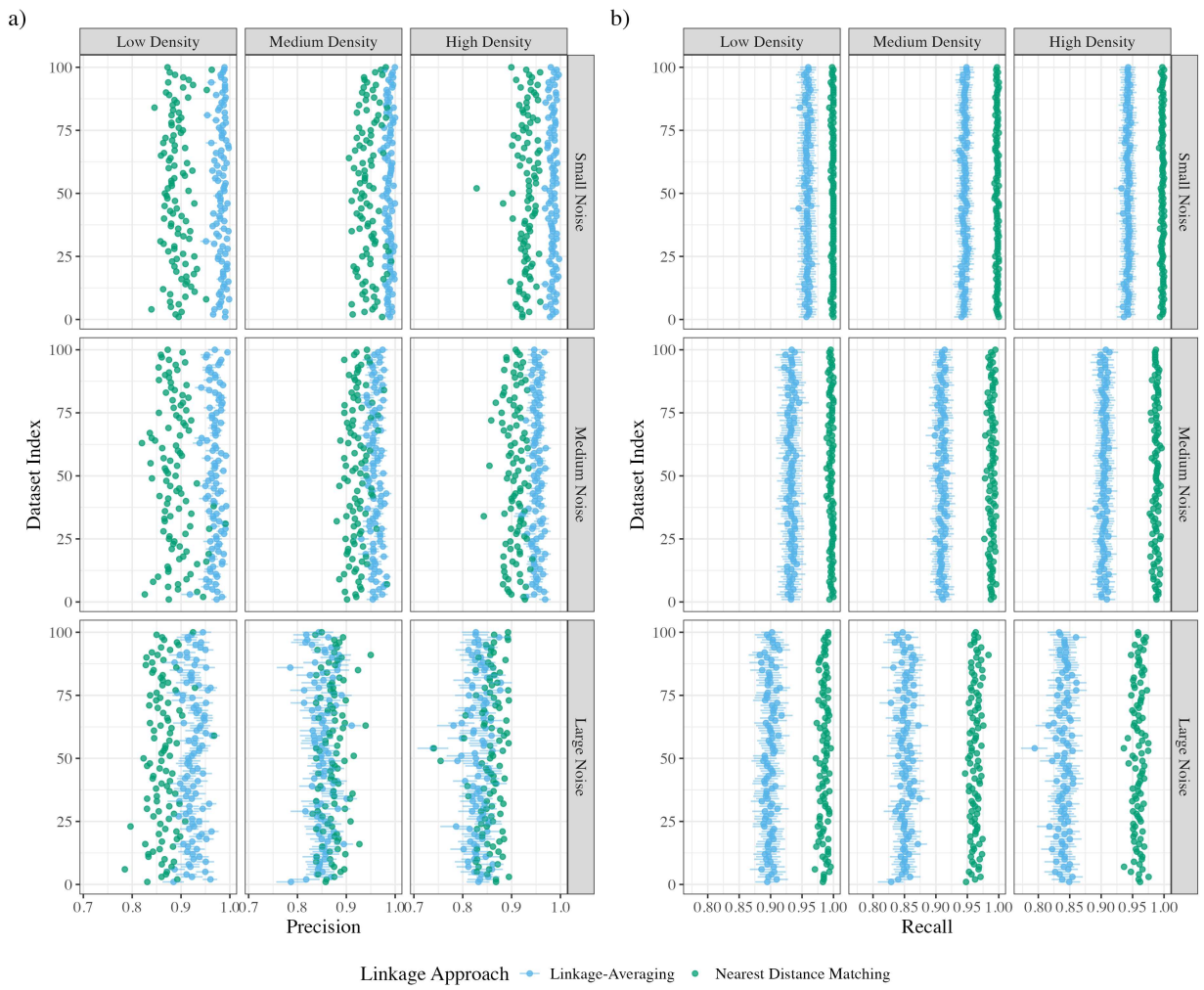

Figure A.4: Plots comparing the precision a) and recall b) performance for the LA and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with $\alpha = 3$.

similar to the model presented in the empirical data analysis, but with a simplified error process for purposes of illustration. Table A.3 and Table A.4 show the coverage results for the $\alpha = 2$ and $\alpha = 3$ simulation studies respectively, which mirror the results for $\alpha = 1$ shown in Section 6 with $q = 1.25$ such that $N = 1.25 \times \max(n_i)$. We note that the coverage rates for the growth model using the true linkage are consistently around the 90% nominal coverage rate, which serves as the gold standard for the growth model performance. Comparing the LA and NDM approaches, we see that the LA approach tends to outperform the NDM approach and often by a substantial margin. In the instances where the NDM coverage is closer to the nominal level, the coverage for the LA approach is generally more conservative due to the uncertainty propagation from the linkage stage of the

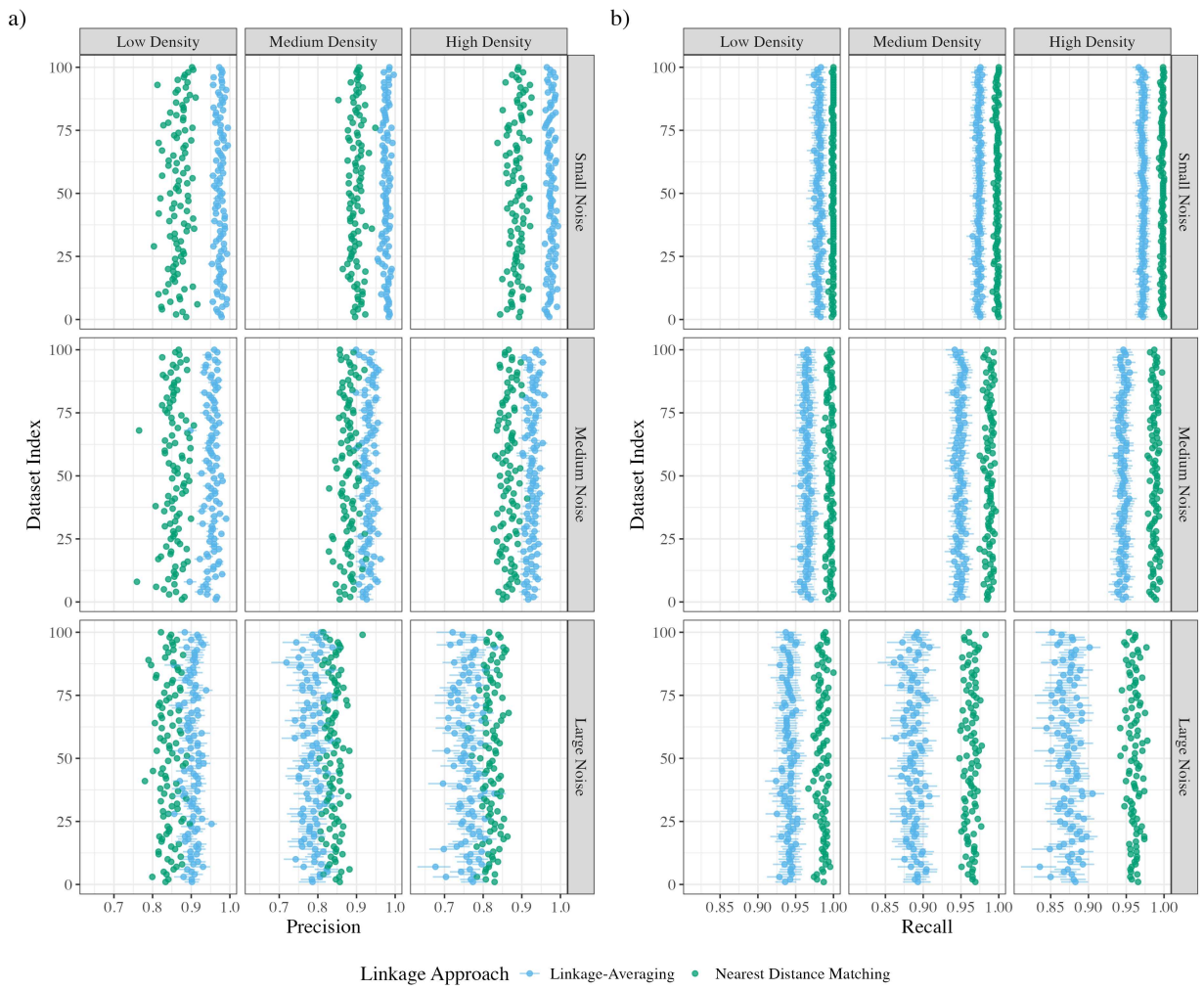

Figure A.5: Plots comparing the precision a) and recall b) performance for the LA with $q = 1.1$ and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with $\alpha = 1$.

modeling pipeline. These results are in line with our expectations regarding the performance of the different linkage approaches, and provide evidence that the two-stage linkage-averaging framework can reliably recover the parameters of interest from a downstream model. They also highlight the trend that while the LA and NDM approaches may have similar coverage for the covariate coefficients, the coverage for the growth asymptote β_0 is much better for the LA model.

In the following tables, we show the simulation results for the same datasets fit with $q = 1.1$ such that $N = 1.1 \times \max(n_i)$. The results for $\alpha = 1$, $\alpha = 2$, and $\alpha = 3$ are presented in Table A.5, Table A.6, and Table A.7 respectively. We note that the true linkage and NDM models do not depend on the hyperparameter N and so the coverages for those methods match those from the tables above.

Table A.3: Empirical coverage table for $N = 1.25 \times \max(n_i)$ and $\alpha = 2$.

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.88	0.88	0.87	0.93	0.94	0.90	0.86	0.87
		LA	**0.89**	**0.92**	**0.94**	**0.96**	0.97	**0.93**	**0.90**	**0.84**
		NDM	0.62	0.28	0.85	0.83	**0.86**	0.82	0.47	0.15
	Medium	TL	0.91	0.89	0.87	0.94	0.94	0.88	0.91	0.89
		LA	**0.95**	**0.90**	**0.94**	0.98	0.97	0.96	**0.94**	**0.78**
		NDM	0.58	0.24	0.82	**0.89**	**0.87**	**0.86**	0.39	0.12
	Large	TL	0.85	0.90	0.90	0.91	0.90	0.87	0.87	0.87
		LA	**0.98**	**0.82**	1.00	0.98	1.00	1.00	**0.86**	**0.38**
		NDM	0.51	0.08	**0.86**	**0.87**	**0.87**	**0.86**	0.15	0.04
Medium	Small	TL	0.93	0.92	0.87	0.89	0.90	0.91	0.89	0.88
		LA	**0.95**	**0.90**	**0.93**	**0.95**	0.95	**0.93**	**0.94**	**0.85**
		NDM	0.60	0.41	0.71	0.84	**0.87**	0.83	0.52	0.25
	Medium	TL	0.86	0.87	0.89	0.87	0.85	0.83	0.87	0.94
		LA	**0.99**	**0.96**	**0.99**	0.98	0.99	0.99	**0.98**	**0.12**
		NDM	0.37	0.21	0.70	**0.94**	**0.87**	**0.86**	0.33	0.05
	Large	TL	0.94	0.94	0.90	0.92	0.92	0.86	0.89	0.89
		LA	**1.00**	**0.25**	**0.99**	1.00	0.99	1.00	**0.91**	0.00
		NDM	0.11	0.00	0.79	**0.95**	**0.89**	**0.89**	0.04	0.00
High	Small	TL	0.89	0.89	0.86	0.87	0.87	0.93	0.88	0.85
		LA	**0.94**	**0.95**	**0.92**	**0.91**	**0.93**	**0.94**	**0.93**	**0.61**
		NDM	0.51	0.36	0.74	0.95	**0.87**	**0.86**	0.48	0.11
	Medium	TL	0.91	0.85	0.91	0.94	0.94	0.89	0.92	0.86
		LA	**0.98**	**0.96**	**1.00**	**0.99**	0.99	**0.98**	**1.00**	**0.06**
		NDM	0.28	0.19	0.77	**0.99**	**0.90**	**0.82**	0.31	0.01
	Large	TL	0.88	0.89	0.91	0.90	0.90	0.86	0.91	0.92
		LA	**0.99**	**0.20**	**0.99**	**1.00**	1.00	1.00	**0.92**	0.00
		NDM	0.04	0.00	0.79	**1.00**	**0.99**	**0.92**	0.02	0.00

Table A.4: Empirical coverage table for $N = 1.25 \times \max(n_i)$ and $\alpha = 3$.

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.92	0.88	0.88	0.85	0.87	0.88	0.83	0.92
		LA	**0.95**	**0.87**	**0.91**	**0.90**	0.91	**0.92**	**0.89**	**0.90**
		NDM	0.65	0.35	0.85	0.85	**0.90**	0.84	0.53	0.26
	Medium	TL	0.91	0.91	0.91	0.86	0.95	0.89	0.90	0.95
		LA	**0.97**	**0.94**	**0.98**	0.94	0.98	**0.94**	**0.98**	**0.84**
		NDM	0.47	0.25	0.81	**0.93**	**0.96**	0.82	0.36	0.14
	Large	TL	0.95	0.89	0.92	0.89	0.85	0.86	0.87	0.89
		LA	**0.99**	**0.86**	0.98	0.99	0.99	1.00	**0.95**	**0.41**
		NDM	0.32	0.07	**0.90**	**0.87**	**0.90**	**0.92**	0.14	0.04
Medium	Small	TL	0.92	0.87	0.85	0.92	0.92	0.91	0.89	0.87
		LA	**0.98**	**0.91**	**0.94**	0.94	0.93	**0.94**	**0.95**	**0.71**
		NDM	0.55	0.38	0.76	**0.92**	**0.92**	0.81	0.55	0.17
	Medium	TL	0.88	0.86	0.86	0.92	0.92	0.88	0.93	0.91
		LA	**1.00**	**0.90**	**0.97**	0.98	0.98	0.99	**1.00**	**0.13**
		NDM	0.43	0.20	0.76	**0.91**	**0.88**	**0.83**	0.43	0.04
	Large	TL	0.91	0.88	0.93	0.91	0.90	0.91	0.88	0.92
		LA	**0.98**	**0.26**	**1.00**	1.00	1.00	1.00	**0.98**	0.00
		NDM	0.10	0.06	0.72	**0.95**	**0.91**	**0.87**	0.10	**0.01**
High	Small	TL	0.89	0.91	0.94	0.96	0.93	0.93	0.91	0.88
		LA	**0.95**	**0.90**	**0.97**	0.97	**0.95**	**0.95**	**0.95**	**0.68**
		NDM	0.62	0.42	0.82	**0.91**	**0.85**	0.78	0.66	0.28
	Medium	TL	0.90	0.93	0.84	0.90	0.91	0.92	0.91	0.88
		LA	**1.00**	**0.91**	**0.98**	0.99	0.99	1.00	**1.00**	**0.07**
		NDM	0.42	0.21	0.81	**0.94**	**0.86**	**0.84**	0.44	0.04
	Large	TL	0.89	0.90	0.88	0.89	0.90	0.93	0.91	0.92
		LA	**1.00**	**0.12**	0.99	**1.00**	1.00	**0.99**	**1.00**	0.00
		NDM	0.07	0.00	**0.84**	**1.00**	**0.98**	0.79	0.08	0.00

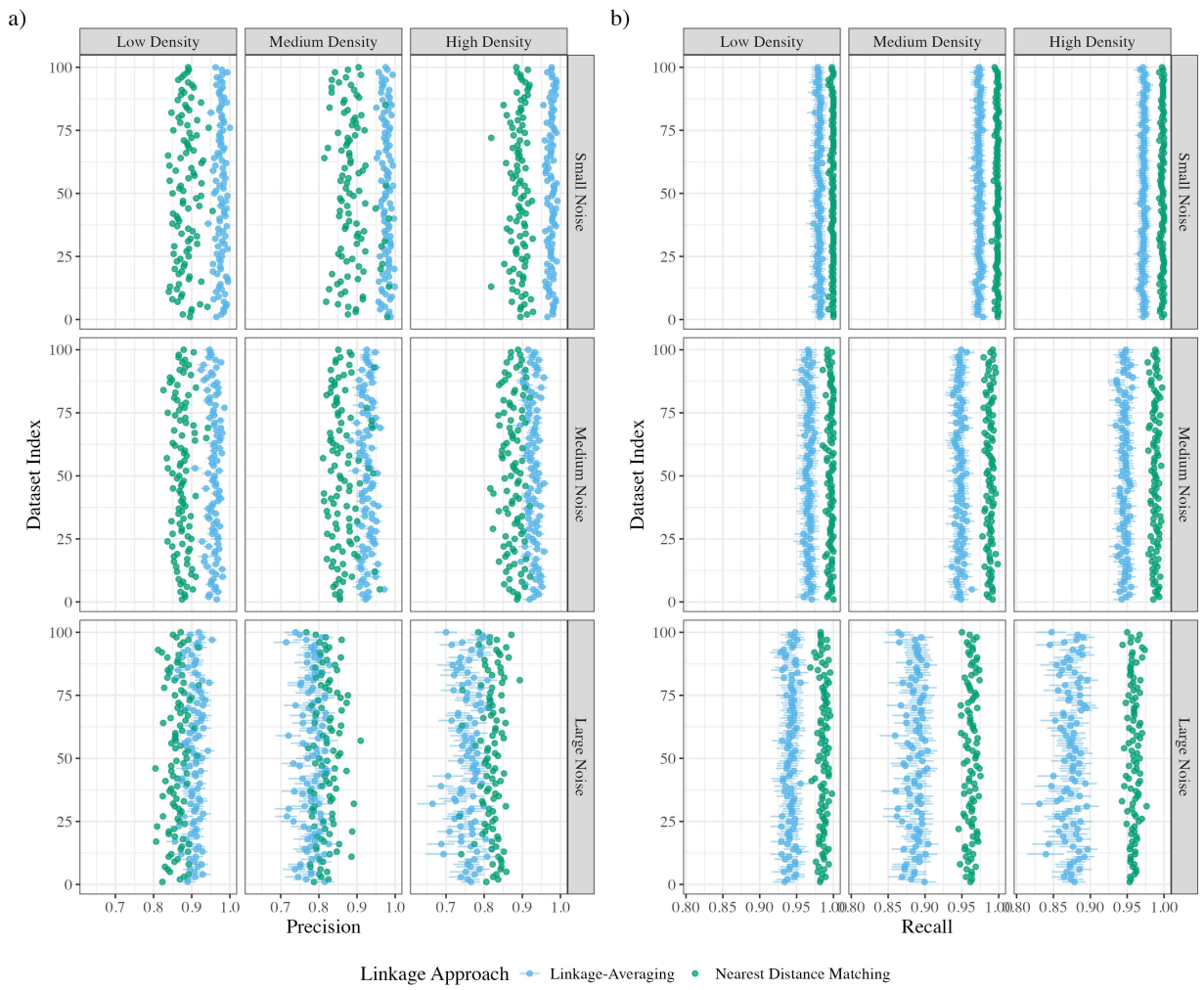

Figure A.6: Plots comparing the precision a) and recall b) performance for the LA with $q = 1.1$ and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with $\alpha = 2$.

We note that the loss in precision observed in the linkage with $q = 1.1$ results in reduced coverage for the downstream growth model parameters compared to the linkage with $q = 1.25$.

Table A.5: Empirical coverage table for $N = 1.1 \times \max(n_i)$ and $\alpha = 1$.

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.93	0.90	0.86	0.92	0.87	0.85	0.90	0.91
		LA	**0.91**	**0.88**	**0.85**	**0.94**	**0.92**	**0.89**	**0.89**	**0.86**
		NDM	0.70	0.38	0.75	0.85	0.80	0.72	0.45	0.17
	Medium	TL	0.85	0.83	0.92	0.89	0.92	0.88	0.91	0.89
		LA	**0.86**	**0.80**	**0.93**	**0.90**	0.95	**0.94**	**0.86**	**0.70**
		NDM	0.56	0.20	0.75	0.84	**0.86**	0.83	0.25	0.06
	Large	TL	0.90	0.90	0.92	0.93	0.94	0.89	0.90	0.92
		LA	**0.93**	**0.67**	**0.99**	0.98	0.99	**0.98**	**0.60**	**0.25**
		NDM	0.70	0.08	0.79	**0.85**	**0.87**	**0.82**	0.12	0.02
Medium	Small	TL	0.92	0.91	0.88	0.94	0.92	0.92	0.91	0.92
		LA	**0.89**	**0.84**	**0.83**	**0.94**	0.93	**0.91**	**0.83**	**0.67**
		NDM	0.43	0.36	0.64	0.83	**0.92**	0.79	0.36	0.23
	Medium	TL	0.90	0.92	0.89	0.94	0.97	0.93	0.91	0.90
		LA	**0.86**	**0.76**	**0.89**	0.98	0.98	**0.96**	**0.80**	**0.06**
		NDM	0.24	0.10	0.65	**0.91**	**0.86**	0.80	0.16	0.02
	Large	TL	0.86	0.84	0.85	0.90	0.88	0.90	0.89	0.87
		LA	**0.88**	**0.18**	**0.99**	1.00	1.00	1.00	**0.34**	0.00
		NDM	0.32	0.01	0.74	**0.94**	**0.87**	**0.85**	0.01	0.00
High	Small	TL	0.85	0.83	0.87	0.93	0.94	0.85	0.85	0.87
		LA	**0.84**	**0.78**	**0.88**	**0.95**	**0.95**	**0.88**	**0.85**	**0.63**
		NDM	0.36	0.30	0.74	**0.95**	0.82	0.66	0.29	0.17
	Medium	TL	0.87	0.85	0.90	0.84	0.91	0.83	0.88	0.88
		LA	**0.85**	**0.80**	**0.91**	0.97	1.00	**0.90**	**0.74**	0.01
		NDM	0.28	0.16	0.64	**0.96**	**0.89**	0.67	0.14	**0.08**
	Large	TL	0.86	0.81	0.83	0.92	0.97	0.88	0.86	0.92
		LA	**0.91**	**0.18**	**0.98**	1.00	1.00	1.00	**0.32**	0.00
		NDM	0.28	0.03	**0.82**	**0.99**	**0.96**	**0.83**	0.02	0.00

Table A.6: Empirical coverage table for $N = 1.1 \times \max(n_i)$ and $\alpha = 2$.

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.88	0.88	0.87	0.93	0.94	0.90	0.86	0.87
		LA	**0.88**	**0.88**	**0.92**	**0.96**	0.97	**0.90**	**0.89**	**0.84**
		NDM	0.62	0.28	0.85	0.83	**0.86**	0.82	0.47	0.15
	Medium	TL	0.91	0.89	0.87	0.94	0.94	0.88	0.91	0.89
		LA	**0.95**	**0.90**	**0.95**	0.98	0.97	0.96	**0.94**	**0.75**
		NDM	0.58	0.24	0.82	**0.89**	**0.87**	**0.86**	0.39	0.12
	Large	TL	0.85	0.90	0.90	0.91	0.90	0.87	0.87	0.87
		LA	**0.97**	**0.75**	1.00	0.98	1.00	0.98	**0.81**	**0.20**
		NDM	0.51	0.08	**0.86**	**0.87**	**0.87**	**0.86**	0.15	0.04
Medium	Small	TL	0.93	0.92	0.87	0.89	0.90	0.91	0.89	0.88
		LA	**0.96**	**0.89**	**0.94**	**0.94**	0.96	**0.94**	**0.95**	**0.76**
		NDM	0.60	0.41	0.71	0.84	**0.87**	0.83	0.52	0.25
	Medium	TL	0.86	0.87	0.89	0.87	0.85	0.83	0.87	0.94
		LA	**0.99**	**0.91**	**0.98**	0.99	0.99	0.98	**0.96**	**0.05**
		NDM	0.37	0.21	0.70	**0.94**	**0.87**	**0.86**	0.33	**0.05**
	Large	TL	0.94	0.94	0.90	0.92	0.92	0.86	0.89	0.89
		LA	**1.00**	**0.07**	**0.99**	1.00	1.00	1.00	**0.83**	0.00
		NDM	0.11	0.00	0.79	**0.95**	**0.89**	**0.89**	0.04	0.00
High	Small	TL	0.89	0.89	0.86	0.87	0.87	0.93	0.88	0.85
		LA	**0.95**	**0.96**	**0.91**	**0.90**	**0.93**	0.95	**0.92**	**0.56**
		NDM	0.51	0.36	0.74	0.95	**0.87**	**0.86**	0.48	0.11
	Medium	TL	0.91	0.85	0.91	0.94	0.94	0.89	0.92	0.86
		LA	**0.99**	**0.94**	**1.00**	1.00	1.00	0.99	**1.00**	**0.04**
		NDM	0.28	0.19	0.77	**0.99**	**0.90**	**0.82**	0.31	0.01
	Large	TL	0.88	0.89	0.91	0.90	0.90	0.86	0.91	0.92
		LA	**0.99**	**0.01**	**0.99**	**1.00**	1.00	1.00	**0.88**	0.00
		NDM	0.04	0.00	0.79	**1.00**	**0.99**	**0.92**	0.02	0.00

Table A.7: Empirical coverage table for $N = 1.1 \times \max(n_i)$ and $\alpha = 3$.

			Empirical Coverage by Parameter							
Density	Noise	Linkage Approach	α	β_0	β_1	β_2	β_3	β_4	γ	τ^2
Low	Small	TL	0.92	0.88	0.88	0.85	0.87	0.88	0.83	0.92
		LA	**0.94**	**0.86**	**0.90**	**0.89**	0.91	**0.91**	**0.88**	**0.90**
		NDM	0.65	0.35	0.85	0.85	**0.90**	0.84	0.53	0.26
	Medium	TL	0.91	0.91	0.91	0.86	0.95	0.89	0.90	0.95
		LA	**0.97**	**0.94**	**0.97**	**0.92**	**0.95**	**0.94**	**0.94**	**0.79**
		NDM	0.47	0.25	0.81	0.93	0.96	0.82	0.36	0.14
	Large	TL	0.95	0.89	0.92	0.89	0.85	0.86	0.87	0.89
		LA	**0.96**	**0.72**	0.98	0.98	0.99	0.99	**0.88**	**0.23**
		NDM	0.32	0.07	**0.90**	**0.87**	**0.90**	**0.92**	0.14	0.04
Medium	Small	TL	0.92	0.87	0.85	0.92	0.92	0.91	0.89	0.87
		LA	**0.98**	**0.90**	**0.93**	0.95	0.95	**0.93**	**0.94**	**0.63**
		NDM	0.55	0.38	0.76	**0.92**	**0.92**	0.81	0.55	0.17
	Medium	TL	0.88	0.86	0.86	0.92	0.92	0.88	0.93	0.91
		LA	**0.99**	**0.77**	**0.95**	0.98	0.97	0.99	**1.00**	**0.08**
		NDM	0.43	0.20	0.76	**0.91**	**0.88**	**0.83**	0.43	0.04
	Large	TL	0.91	0.88	0.93	0.91	0.90	0.91	0.88	0.92
		LA	**0.98**	0.01	**1.00**	1.00	1.00	1.00	**0.96**	0.00
		NDM	0.10	**0.06**	0.72	**0.95**	**0.91**	**0.87**	0.10	**0.01**
High	Small	TL	0.89	0.91	0.94	0.96	0.93	0.93	0.91	0.88
		LA	**0.95**	**0.87**	**0.96**	0.95	**0.94**	**0.95**	**0.95**	**0.59**
		NDM	0.62	0.42	0.82	**0.91**	0.85	0.78	0.66	0.28
	Medium	TL	0.90	0.93	0.84	0.90	0.91	0.92	0.91	0.88
		LA	**1.00**	**0.84**	**0.99**	1.00	1.00	1.00	**1.00**	0.02
		NDM	0.42	0.21	**0.81**	**0.94**	**0.86**	**0.84**	0.44	**0.04**
	Large	TL	0.89	0.90	0.88	0.89	0.90	0.93	0.91	0.92
		LA	**1.00**	**0.01**	0.99	**1.00**	1.00	**0.99**	**1.00**	0.00
		NDM	0.07	0.00	**0.84**	**1.00**	**0.98**	0.79	0.08	0.00

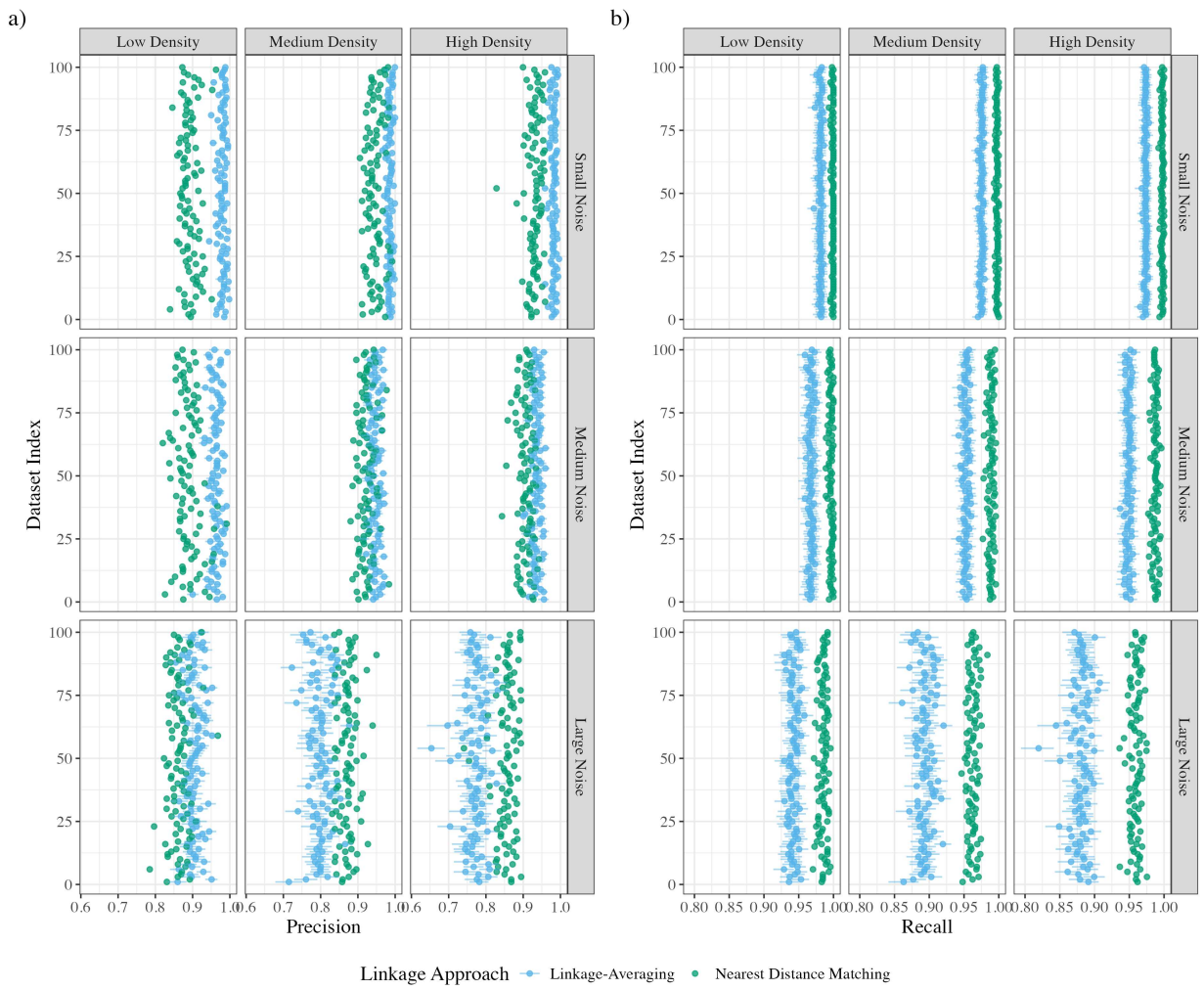

Figure A.7: Plots comparing the precision a) and recall b) performance for the LA with $q = 1.1$ and NDM linkage approaches on 100 simulated datasets for each density and noise combination with known true linkage and with $\alpha = 3$.

Appendix B

Inferring Tree Growth from Linked Multi-temporal Remote Sensing Data with Exact Error Propagation at Scale Supplementary Material

B.1 Empirical data collection and processing

B.1.1 Drone platform and sensor

Images for structure analysis were collected using P1 and Altum camera systems concurrently mounted on a DJI Matrice 350 RTK UAS platform. The P1 camera (DJI Corporation, Shenzhen, China) collects 45 megapixel red-green-blue (RGB) images, and the Altum sensor (AgEagle Corporation, Wichita, KS, USA), collects 3.2 megapixel images in blue, green, red, red-edge, and near-infrared wavelengths. Table B.1 shows the sensor settings used for each sensor.

Table B.1: Comparison of sensor attributes and settings for the P1 and Altum sensors.

Sensor Attribute / Setting	P1 Sensor	Altum Sensor
Bands	Red, Green, Blue	Blue (475 nm), Green (560 nm), Red (668 nm), Red Edge (717 nm), Near-IR (842 nm)
Image pixel size	8192×5460	2064×1544
Image bit depth	8 bit	16 bit
Capture Rate (images/sec)	1.43	0.91
Effective ground sample distance at flight altitude of 110 m (cm)	1.38	4.74
Side overlap at 28 m flightline spacing (%)	75.18	71.40
Forward overlap at 13 m/s (%)	87.91	80.49
Shutter speed	1/1200 sec	varying
ISO	100	varying

B.1.2 Flight planning and image acquisition

We used a programming script to generate terrain-following waypoint flights covering the entire study area at a flightline spacing of 28 meters, maintaining a nominal above-terrain distance of 110 meters, yielding images with 1.38 cm (P1) or 4.74 cm (Altum) ground resolution. Each sensor was set to capture at its maximum rate, yielding between 4600 – 4800 images per flight for the

P1 sensor, and between 1800 – 2000 images for the Altum. All flights were performed with a Real-Time Kinematic (RTK) GPS system with a base station mounted on a known monument. RTK geopositions representing the locations of the camera pupil were automatically added to the image metadata for the P1 images. These RTK positions were supplemented with the geopositions of six UAS-visible ground control points (bucket lids) distributed throughout the study area. Independent testing indicated average horizontal errors of 5-7 cm and average vertical errors of 5-11 cm when compared to known monuments.

Flights were performed in clear-sky or entirely cloudy conditions in late morning (approximately 10:00 am to 12:00 pm local time) to minimize the effects of shadows. Each automated flight took between 52 and 65 minutes to complete, including battery swaps. Table B.2 shows the acquisition dates and sky conditions of all UAS data contributing to the analysis.

Table B.2: Sky conditions during image acquisition dates.

Date	Sky Conditions
6/29/2021	Cloudy
7/26/2021	Sunny
8/31/2021	Cloudy
11/9/2021	Sunny
7/1/2022	Sunny
8/4/2022	Sunny
8/18/2022	Sunny
10/17/2022	Sunny
7/11/2023	Sunny
8/1/2023	Sunny
8/21/2023	Sunny
10/25/2023	Sunny
6/24/2024	Sunny
7/17/2024	Sunny
8/6/2024	Sunny
8/26/2024	Cloudy

B.1.3 Structure-from-motion image processing

Images from the P1 sensor captured in each flight were used to derive canopy structure properties after processing using structure-from-motion photogrammetry in Agisoft Metashape software (version 1.7.2). Our photogrammetry workflow is based on the results of Tinkham and Woolsey (2024), and used "high" dense cloud reconstruction settings and mild depth filtering to generate a point cloud

from images. This point cloud from each flight was then rasterized to produce a Digital Surface Model (DSM) within Metashape at 5 cm resolution. To reduce interpolation errors, DSM heights were only generated for pixels with dense cloud data. Each resulting DSM was then manually checked for horizontal and vertical alignment against a set of 5 known monuments.

To generate normalized difference vegetation index (NDVI) estimates for each year, we selected a single representative peak-season flight and co-aligned P1 and Altum multispectral data from that flight in Metashape. After radiometric calibration using the Altum DLS2 built-in sun sensor and MicaSense calibration target, georeferencing with 6 ground-control points, and generating a DSM from the combined Altum-P1 data, we generated a 5 cm resolution NDVI orthomosaic from the Altum Red and Near-Infrared bands. We then exported this to QGIS and checked for alignment with the structural data.

B.1.4 Post-processing of canopy surface models

We composited the four surface models collected in each year by taking the maximum non-missing value for each 5 cm pixel. To reduce noise and improve performance of the canopy segmentation, we aggregated the resulting map to 25 cm by taking the median of 5 cm cells. The resulting maps represent canopy heights above mean sea level. To produce a map of height above the ground surface, we created a single reference Digital Elevation Model using the Cloth Simulation Filter (Zhang et al., 2016) to classify ground points derived from an SfM dense cloud generated from UAS data collected on June 29th, 2021. Ground points were interpolated into a continuous raster map using the `TIN` algorithm implemented in the R package **lidR** (Roussel et al., 2020). Finally, we subtracted the DEM from the DSM canopy surface for each year, generating a single canopy height model (CHM) for every year from 2021 – 2024.

B.1.5 Segmentation and filtering by vegetation type

We delineated individual tree crowns by segmenting the annual CHM's using the `ITCSegment` algorithm (Dalponte and Coomes, 2016). `ITCSegment` is a region-growing algorithm that starts at focal pixels that are local height maxima (i.e., tree tops), and then iteratively adds adjacent pixels until stopping criteria are met. To be included in a crown, pixels needed to be at least 1 m in height and at least 0.2 times the maximum height of the crown. Tree tops were delineated using a local

maximum filter approach which finds maxima in a moving window with a radius that increases with tree height.

B.1.6 Crown size extraction

Finally, as a measure of the size of the crown which integrates both height and horizontal extent, we calculated the canopy volume of each segmented crown. The conifer trees of interest in the study area have largely conical crowns. This simplifies the problem of volume estimation because canopies extend nearly to the ground surface and have exterior surfaces that are well-represented by the CHM. Within each segmented crown polygon, we calculated canopy volume by multiplying the height of each 0.25 m cell in a crown by its horizontal extent and then summing the values for all pixels in a crown.

B.1.7 Sub-area characteristics

In this section, we provide summary characteristics of each of the 10 sub-areas used in the analysis. Table B.3 shows counts of records by year, footprint area (ha), and individual specific covariate ranges for each sub-area.

B.2 Full conditional distributions for the joint model variants

Following the specifications for the full and reduced dependence downstream growth model introduced in the previous section, we provide the joint models based on each formulation.

The full dependence joint model is specified as

$$
\begin{aligned}
\mathbf{y}_{ij} \mid z_{ij}, \mathbf{s}_{\lambda_{ij}}, \sigma^2, D &\overset{ind}{\sim} (1 - z_{ij})\mathbb{I}\{\boldsymbol{y}_{ij} \in B(\boldsymbol{s}_{\lambda_{ij}})\}|B(\boldsymbol{s}_{\lambda_{ij}})|^{-1} \\
&\quad + z_{ij}\text{Normal}_{2,[D]}(\mathbf{s}_{\lambda_{ij}}, \sigma^2\mathbf{I}) \\
\mathbf{s}_{j'} \mid N &\overset{iid}{\sim} \text{Uniform}\,(D^*) \\
\sigma^2 &\sim \text{Inverse-Gamma}_{[0,\sigma^2_{\max}]}\,(c_\sigma, d_\sigma) \\
\lambda_{ij} \mid N &\overset{iid}{\sim} \text{Uniform}\,\{1, \ldots, N\} \\
z_{ij} \mid \theta &\overset{iid}{\sim} \text{Bernoulli}(\theta) \\
\theta &\sim \text{Beta}(a_\theta, b_\theta)
\end{aligned}
$$

Table B.3: Summary of sub-area characteristics: counts of records by year, footprint area (ha), and individual specific covariate ranges.

Area	Records by year (2021-2024)	Footprint (ha, hull)	Max height (m) (median [min–max])	Canopy volume (m^3) (median [min–max])	Crown NDVI (median [min–max])
Copper Creek Canyon	1242, 1250, 1257, 1236; Total: 4985	10.096	17.8 (2.0–40.3)	111.4 (0.2–1417.9)	0.78 (0.25–0.97)
North Beaver Ponds	815, 797, 789, 803; Total: 3204	2.948	12.6 (2.0–34.4)	53.6 (0.4–1007.5)	0.84 (0.22–0.95)
North East River	1079, 1075, 1076, 1086; Total: 4316	7.716	14.9 (2.0–34.3)	87.8 (0.4–954.6)	0.83 (0.36–0.97)
River Forest	1543, 1551, 1550, 1531; Total: 6175	4.111	19.6 (2.1–32.7)	123.5 (0.3–1071.8)	0.81 (0.37–0.93)
South Beaver Ponds	505, 499, 497, 497; Total: 1998	2.895	14.6 (2.1–28.5)	66.9 (0.2–556.8)	0.83 (0.09–0.95)
South East River	83, 41, 49, 46; Total: 219	2.221	11.3 (2.0–20.3)	46.2 (1.0–783.7)	0.84 (0.50–0.93)
South Gothic	618, 610, 601, 587; Total: 2416	3.864	19.8 (2.1–33.8)	161.0 (0.4–985.3)	0.81 (0.37–0.97)
South Townsite	322, 286, 280, 281; Total: 1169	6.672	15.8 (2.0–29.1)	114.2 (0.8–1261.6)	0.83 (0.31–0.95)
West Rocky Meadows	1378, 1326, 1341, 1292; Total: 5337	12.074	12.9 (2.0–33.2)	55.6 (0.1–1067.4)	0.77 (0.23–0.96)
West Townsite	362, 361, 363, 353; Total: 1439	4.125	20.6 (2.2–32.9)	182.2 (2.2–1219.1)	0.81 (0.30–0.96)

$$
\begin{aligned}
\boldsymbol{V}_c \mid \mathcal{C}_c^G(\boldsymbol{\Lambda}), u_{c,0}, \boldsymbol{w}_c, \boldsymbol{\beta}, \alpha, \gamma, \tau^2, \boldsymbol{X}(\boldsymbol{s}_c) &\overset{ind}{\sim} \text{Normal}(\boldsymbol{u}_c + \boldsymbol{w}_c, \tau^2\mathbf{I}) \\
u_{c,0} &\overset{ind}{\sim} \text{Normal}(\mu_{u_{c,0}}, \sigma^2_{u_0}) \\
\mu_{u_{c,0}} &\overset{iid}{\sim} \text{Normal}(\tilde{\mu}, \tau^2_{\mu_{u_0}}) \\
\sigma^2_{u_0} &\sim \text{Inverse-Gamma}(c_{\sigma_{u_0}}, d_{\sigma_{u_0}}) \\
\boldsymbol{w}_c &\overset{iid}{\sim} \text{Normal}(\mathbf{0}, \nu^2\boldsymbol{\Sigma}) \\
\boldsymbol{\beta} &\sim \text{Normal}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \\
\alpha &\sim \text{Beta}_{[c_\alpha, d_\alpha]}(a_\alpha, b_\alpha) \\
\gamma &\sim \text{Gamma}_{[\gamma_{\min}, \gamma_{\max}]}(c_\gamma, d_\gamma) \\
\tau^2 &\sim \text{Inverse-Gamma}_{[0, \tau^2_{\max}]}(c_\tau, d_\tau) \\
\nu^2 &\sim \text{Inverse-Gamma}_{[0, \nu^2_{\max}]}(c_\nu, d_\nu) \\
\rho &\sim \text{Uniform}(0, r).
\end{aligned}
$$

The reduced dependence joint model is specified as

$$
\begin{aligned}
\mathbf{y}_{ij} \mid z_{ij}, \mathbf{s}_{\lambda_{ij}}, \sigma^2, D &\overset{ind}{\sim} (1 - z_{ij})\mathbb{I}\{\boldsymbol{y}_{ij} \in B(\boldsymbol{s}_{\lambda_{ij}})\}|B(\boldsymbol{s}_{\lambda_{ij}})|^{-1} \\
&\quad + z_{ij}\text{Normal}_{2,[D]}(\mathbf{s}_{\lambda_{ij}}, \sigma^2\mathbf{I}) \\
\mathbf{s}_{j'} \mid N &\overset{iid}{\sim} \text{Uniform}\,(D^*) \\
\sigma^2 &\sim \text{Inverse-Gamma}_{[0, \sigma^2_{\max}]}\,(c_\sigma, d_\sigma) \\
\lambda_{ij} \mid N &\overset{iid}{\sim} \text{Uniform}\,\{1, \ldots, N\} \\
z_{ij} \mid \theta &\overset{iid}{\sim} \text{Bernoulli}(\theta) \\
\theta &\sim \text{Beta}(a_\theta, b_\theta) \\
\boldsymbol{V}_c \mid \mathcal{C}_c^G(\boldsymbol{\Lambda}), u_{c,0}, \boldsymbol{w}_c, \boldsymbol{\beta}, \alpha, \gamma, \tau^2, \boldsymbol{X}(\boldsymbol{s}_c) &\overset{ind}{\sim} \text{Normal}(\boldsymbol{u}_c + \boldsymbol{w}_c, \tau^2\mathbf{I}) \\
u_{c,0} \mid \mathcal{C}_c^G(\boldsymbol{\Lambda}), V_{c,0} &\overset{ind}{\sim} \text{Normal}(V_{c,0}, \sigma^2_u) \\
V_{c,0} &\overset{ind}{\sim} \text{Normal}(\eta_c, \sigma^2_V) \\
\boldsymbol{w}_c &\overset{iid}{\sim} \text{Normal}(\mathbf{0}, \nu^2\boldsymbol{\Sigma}) \\
\boldsymbol{\beta} &\sim \text{Normal}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)
\end{aligned}
$$

$$
\begin{aligned}
\alpha &\sim \text{Beta}_{[c_\alpha, d_\alpha]}(a_\alpha, b_\alpha) \\
\gamma &\sim \text{Gamma}_{[\gamma_{\min}, \gamma_{\max}]}(c_\gamma, d_\gamma) \\
\tau^2 &\sim \text{Inverse-Gamma}_{[0, \tau^2_{\max}]}(c_\tau, d_\tau) \\
\nu^2 &\sim \text{Inverse-Gamma}_{[0, \nu^2_{\max}]}(c_\nu, d_\nu) \\
\rho &\sim \text{Uniform}(0, r).
\end{aligned}
$$

We present the full conditional distributions for the record linkage component of the joint model variants below. We note that these full conditionals are identical for both the full and reduced dependence joint model variants.

The full conditional for σ^2 is given by

$$
\begin{aligned}
&\left[\sigma^2 \mid \mathbf{s}, \mathbf{\Lambda}, \mathbf{z}, \mathbf{y}\right] \\
&\sim \text{Inverse-Gamma}_{[0, \sigma^2_{\max}]}\left(c_\sigma + \sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}, d_\sigma + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbb{I}\{z_{ij}=1\}(\mathbf{y}_{ij} - \mathbf{s}_{\lambda_{ij}})^\mathsf{T}(\mathbf{y}_{ij} - \mathbf{s}_{\lambda_{ij}})\right)
\end{aligned}
$$

The full conditional for $\boldsymbol{s}_{j'}$ is given by

$$
\begin{aligned}
&\left[\mathbf{s}_{j'} \mid \mathbf{\Lambda}, \mathbf{z}, \sigma^2, \mathbf{y}, D^*\right] \\
&\overset{\text{ind}}{\sim} \begin{cases}
P\left(\mathbf{s}_{j'} = \mathbf{y}_{ij}\right) = \dfrac{1}{n_{j'}} & \text{for } i,j \text{ such that } \lambda_{ij} = j' \text{ and } z_{ij} = 0 \\
\text{N}_{D^*}\left(\dfrac{1}{n_{j'}}\displaystyle\sum_{i=1}^{m}\sum_{(j):\lambda_{ij}=j'} \mathbf{y}_{ij}, \dfrac{\sigma^2}{n_{j'}}\mathbf{I}\right) & \text{for } i,j \text{ such that } \lambda_{ij} = j',\ z_{ij} = 1, \text{ and } n_{j'} > 0 \\
\text{Uniform}\left(D^*\right) & \text{for } i,j \text{ such that } \lambda_{ij} = j',\ z_{ij} = 1, \text{ and } n_{j'} = 0
\end{cases}
\end{aligned}
$$

The full conditional for z_{ij} is given by

$$\left[z_{ij} \mid \mathbf{\Lambda}, \mathbf{s}, \sigma^2, \theta, \mathbf{y}\right] \overset{\text{ind}}{\sim} \text{Bernoulli}(p_{ij}), \text{ where}$$

$$p_{ij} = \begin{cases} 1 & \text{if } \mathbb{I}\{\boldsymbol{y}_{ij} \notin B(\boldsymbol{s}_{\lambda_{ij}})\} \\ \dfrac{\frac{\theta}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}_{ij}-\mathbf{s}_{\lambda_{ij}})^\mathsf{T}(\mathbf{y}_{ij}-\mathbf{s}_{\lambda_{ij}})\right)}{\frac{\theta}{2\pi\sigma^2}\exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}_{ij}-\mathbf{s}_{\lambda_{ij}})^\mathsf{T}(\mathbf{y}_{ij}-\mathbf{s}_{\lambda_{ij}})\right) + (1-\theta)|B(\mathbf{s}_{\lambda_{ij}})|^{-1}} & \text{if } \mathbb{I}\{\boldsymbol{y}_{ij} \in B(\boldsymbol{s}_{\lambda_{ij}})\} \end{cases}$$

The full conditional for λ_{ij} is given by

$$\begin{aligned} &P(\lambda_{ij} = \ell \mid \boldsymbol{\Lambda}_{-(ij)}, \mathbf{s}, \mathbf{z}, \sigma^2, \theta, \mathbf{y}) \\ &\propto (1-z_{ij})\mathbb{I}\{\boldsymbol{y}_{ij} \in B(\boldsymbol{s}_\ell)\}|B(\boldsymbol{s}_\ell)|^{-1} + z_{ij}\left((2\pi\sigma^2)^{-1}\exp\left(-(2\sigma^2)^{-1}(\mathbf{y}_{ij}-\mathbf{s}_\ell)^\mathsf{T}(\mathbf{y}_{ij}-\mathbf{s}_\ell)\right)\right) \\ &\quad \times \left[\exp\left(-\frac{1}{2\tau^2}(\boldsymbol{V}_c - \boldsymbol{u}_c - \boldsymbol{w}_c)^\mathsf{T}(\boldsymbol{V}_c - \boldsymbol{u}_c - \boldsymbol{w}_c)\right)\right]^{\mathbb{I}\{c=\ell\}} \\ &\propto \begin{cases} 1 & \text{if } z_{ij} = 0 \\ \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}_{ij}-\mathbf{s}_\ell)^\mathsf{T}(\mathbf{y}_{ij}-\mathbf{s}_\ell)\right)\exp\left(-\frac{1}{2\tau^2}(\boldsymbol{V}_c - \boldsymbol{u}_c - \boldsymbol{w}_c)^\mathsf{T}(\boldsymbol{V}_c - \boldsymbol{u}_c - \boldsymbol{w}_c)\right)^{\mathbb{I}\{c=\ell\}} & \text{if } z_{ij} = 1 \end{cases} \end{aligned}$$

The full conditional for θ is given by

$$[\theta \mid \mathbf{z}] \sim \text{Beta}\left(a_\theta + \sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}, b_\theta + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(1-z_{ij})\right).$$

We note the relationships between $\mathbf{y}$, $\mathbf{s}$, $\mathbf{z}$, and $\boldsymbol{\Lambda}$ are such that when we condition one parameter on the remaining values, we obtain certain mixture distributions. For instance, if $z_{ij} = 0$, then $\mathbf{y}_{ij} = \mathbf{s}_{\lambda_{j'}}$ for some j'.

The full conditional distributions for the parameters in the full dependence joint model are as follows.

The full conditional of $\mathbf{V}_{c,m}$ is:

$$\left[\mathbf{V}_{c,m} \mid u_{c,0}, \mathbf{w}_c, \boldsymbol{\beta}, \gamma, \alpha, \tau^2, \nu^2, \rho, \mathbf{X}_(\mathbf{s}_c)\right] \sim \text{Normal}\left(\mathbf{u}_{c,m} + \mathbf{w}_{c,m}, \tau^2\mathbf{I}\right).$$

The full conditional of $u_{c,0}$ is:

$$\left[u_{c,0} \mid V_{c,0}, \mathbf{w}_c, \boldsymbol{\beta}, \gamma, \alpha, \tau^2, \mathbf{V}_c, \mathbf{X}_(\mathbf{s}_c)\right] \propto \left[\exp\left(-\frac{1}{2\tau^2}\left(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c\right)^\mathsf{T}\left(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c\right)\right)\right] \times \left[\exp\left(-\frac{1}{2\sigma^2_{u_0}}(u_{c,0} - \mu_{u_{c,0}})^2\right)\right].$$

The full conditional of $\mathbf{w}_c$ is:

$$\left[\mathbf{w}_c \mid \mathbf{V}_c, u_{c,0}, \tau^2, \nu^2\rho\right] \sim \text{Normal}\left(\left[\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\nu^2}\boldsymbol{\Sigma}^{-1}\right]^{-1}\left[\frac{1}{\tau^2}\left(\mathbf{V}_c - \mathbf{u}_c\right)\right], \left[\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\nu^2}\boldsymbol{\Sigma}^{-1}\right]^{-1}\right).$$

The full conditional of $\mu_{u_{c,0}}$ is:

$$\left[\mu_{u_{c,0}} \mid u_{c,0}, \sigma^2_{u0}\right] \sim \text{Normal}\left(\left[\frac{1}{\sigma^2_{u_0}} + \frac{1}{\tau^2_{\mu_{u_0}}}\right]^{-1}\left[\frac{u_{c,0}}{\sigma^2_{u_0}} + \frac{\tilde{\mu}}{\tau^2_{\mu_{u_0}}}\right], \left[\frac{1}{\sigma^2_{u_0}} + \frac{1}{\tau^2_{\mu_{u_0}}}\right]^{-1}\right).$$

The full conditional of $\sigma^2_{u_0}$ is:

$$\left[\sigma^2_{u_0} \mid u_{c,0}, \mu_{u_{c,0}}\right] \sim \text{Inverse-Gamma}\left(c_{\sigma_{u_0}} + \frac{1}{2}|\mathcal{C}^G(\boldsymbol{\Lambda})|, d_{\sigma_{u_0}} + \frac{1}{2}\sum_{c \in \mathcal{C}^G(\boldsymbol{\Lambda})}(u_{c,0} - \mu_{u_{c,0}})^2\right).$$

For sampling $\boldsymbol{\beta}$, γ, and α we define

$$\mathbf{V}^*_c = \begin{bmatrix} V_{c,1} - u_{c,0} \\ V_{c,2} - u_{c,1} \\ \vdots \\ V_{c,m-1} - u_{c,m-2} \end{bmatrix}, \quad \text{and } \mathbf{A}_c = \text{Diag}\left(\begin{bmatrix} h_1 * \frac{u^\alpha_{c,0}}{\gamma^\alpha + u^\alpha_{c,0}} \\ h_2 * \frac{u^\alpha_{c,1}}{\gamma^\alpha + u^\alpha_{c,1}} \\ \vdots \\ h_{m-1} * \frac{u^\alpha_{c,m-2}}{\gamma^\alpha + u^\alpha_{c,m-2}} \end{bmatrix}\right).$$

Then, we can write the full conditional distribution of $\boldsymbol{\beta}$ as

$$\begin{aligned}&\left[\boldsymbol{\beta} \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \alpha, \gamma, \tau^2\right] \\ &\sim \text{Normal}\left(\left[\frac{1}{\tau^2}\sum_c \boldsymbol{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}\mathbf{A}_c\mathbf{X}(\mathbf{s}_c) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}\right]^{-1}\left[\frac{1}{\tau^2}\sum_c \mathbf{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}(\mathbf{V}^*_c - \mathbf{w}_c) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right],\right. \\ &\qquad \left.\left[\frac{1}{\tau^2}\sum_c \mathbf{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}\mathbf{A}_c\mathbf{X}(\mathbf{s}_c) + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}\right]^{-1}\right).\end{aligned}$$

for $c \in \mathcal{C}^G(\mathbf{\Lambda})$.

The full conditional distribution of γ is:

$$\begin{aligned}
&\left[\gamma \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \boldsymbol{\beta}, \alpha, \tau^2\right] \\
&\propto \prod_{c \in \mathcal{C}^G(\mathbf{\Lambda})} \left[\exp\left(-\frac{1}{2\tau^2}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)^\mathsf{T}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)\right)\right] \\
&\quad \times \left[\gamma^{a_\gamma - 1}\exp\left(-\frac{\gamma}{b_\gamma}\right)\mathbb{I}\{\gamma_{\min} \leq \gamma \leq \gamma_{\max}\}\right].
\end{aligned}$$

The full conditional distribution of α is:

$$\begin{aligned}
&\left[\gamma \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \boldsymbol{\beta}, \alpha, \tau^2\right] \\
&\propto \prod_{c \in \mathcal{C}^G(\mathbf{\Lambda})} \left[\exp\left(-\frac{1}{2\tau^2}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)^\mathsf{T}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)\right)\right] \\
&\quad \times \left[(\alpha - c_\alpha)^{a_\alpha - 1}(d_\alpha - \alpha)^{b_\alpha - 1}\mathbb{I}\{c_\alpha \leq \alpha \leq d_\alpha\}\right].
\end{aligned}$$

The full conditional distribution of τ^2 is:

$$\begin{aligned}
&\left[\tau^2 \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \boldsymbol{\beta}, \alpha, \gamma\right] \\
&\sim \text{Inverse-Gamma}_{[0,\tau^2_{\max}]}\left(c_\tau + \frac{1}{2}|\mathcal{C}^G(\mathbf{\Lambda})|, d_\tau + \frac{1}{2}\sum_{c \in \mathcal{C}^G(\mathbf{\Lambda})}(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c)^\mathsf{T}(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c)\right).
\end{aligned}$$

The full conditional distribution of ν^2 is:

$$\begin{aligned}
&\left[\nu^2 \mid \{\boldsymbol{w}_c\}, \rho\right] \\
&\sim \text{Inverse-Gamma}_{[0,\nu^2_{\max}]}\left(c_\nu + \frac{m}{2}|\mathcal{C}^G(\mathbf{\Lambda})|, d_\nu + \frac{1}{2}\sum_{c \in \mathcal{C}^G(\mathbf{\Lambda})}\mathbf{w}_c^\mathsf{T}\mathbf{\Sigma}^{-1}\mathbf{w}_c\right).
\end{aligned}$$

The full conditional distribution of ρ is:

$$\left[\rho \mid \{\boldsymbol{w}_c\}, \nu^2\right] \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}|\mathcal{C}^G(\boldsymbol{\Lambda})|} \exp\left(-\frac{1}{2\nu^2} \sum_{c \in \mathcal{C}^G(\boldsymbol{\Lambda})} \mathbf{w}_c^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \mathbf{w}_c\right) \mathbb{I}\{0 < \rho < r\}.$$

The full conditional distributions for the parameters in the reduced dependence joint model are as follows.

The full conditional of $\mathbf{V}_{c,m}$ is:

$$\left[\mathbf{V}_{c,m} \mid u_{c,0}, \mathbf{w}_c, \boldsymbol{\beta}, \gamma, \alpha, \tau^2, \nu^2, \rho, \mathbf{X}_(\mathbf{s}_c)\right] \sim \text{Normal}\left(\mathbf{u}_{c,m} + \mathbf{w}_{c,m}, \tau^2 \mathbf{I}\right).$$

The full conditional of $V_{c,0_m}$ is:

$$[V_{c,0_m} \mid u_{c,0}] \sim \text{Normal}\left(\left[\frac{1}{\sigma_u^2} + \frac{1}{\sigma_V^2}\right]^{-1} \left[\frac{u_{c,0}}{\sigma_u^2} + \frac{\eta_c}{\sigma_V^2}\right], \left[\frac{1}{\sigma_u^2} + \frac{1}{\sigma_V^2}\right]^{-1}\right).$$

The full conditional of $u_{c,0}$ is:

$$\begin{aligned}\left[u_{c,0} \mid V_{c,0}, \mathbf{w}_c, \boldsymbol{\beta}, \gamma, \alpha, \tau^2, \mathbf{V}_c, \mathbf{X}_(\mathbf{s}_c)\right] \propto &\left[\exp\left(-\frac{1}{2\tau^2}\left(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c\right)^{\mathsf{T}}\left(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c\right)\right)\right] \\ &\times \left[\exp\left(-\frac{1}{2\sigma_u^2}(u_{c,0} - V_{c,0})^2\right)\right].\end{aligned}$$

The full conditional of $\mathbf{w}_c$ is:

$$\left[\mathbf{w}_c \mid \mathbf{V}_c, u_{c,0}, \tau^2, \nu^2\rho\right] \sim \text{Normal}\left(\left[\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\nu^2}\boldsymbol{\Sigma}^{-1}\right]^{-1}\left[\frac{1}{\tau^2}\left(\mathbf{V}_c - \mathbf{u}_c\right)\right], \left[\frac{1}{\tau^2}\mathbf{I} + \frac{1}{\nu^2}\boldsymbol{\Sigma}^{-1}\right]^{-1}\right).$$

For sampling $\boldsymbol{\beta}$, γ, and α we define

$$\mathbf{V}_c^* = \begin{bmatrix} V_{c,1} - u_{c,0} \\ V_{c,2} - u_{c,1} \\ \vdots \\ V_{c,m-1} - u_{c,m-2} \end{bmatrix}, \quad \text{and } \mathbf{A}_c = \text{Diag}\left(\begin{bmatrix} h_1 * \frac{u_{c,0}^{\alpha}}{\gamma^{\alpha} + u_{c,0}^{\alpha}} \\ h_2 * \frac{u_{c,1}^{\alpha}}{\gamma^{\alpha} + u_{c,1}^{\alpha}} \\ \vdots \\ h_{m-1} * \frac{u_{c,m-2}^{\alpha}}{\gamma^{\alpha} + u_{c,m-2}^{\alpha}} \end{bmatrix}\right).$$

Then, we can write the full conditional distribution of $\boldsymbol{\beta}$ as

$$\begin{aligned}
&\left[\boldsymbol{\beta} \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \alpha, \gamma, \tau^2\right] \\
&\sim \text{Normal}\left(\left[\frac{1}{\tau^2}\sum_c \boldsymbol{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}\mathbf{A}_c\mathbf{X}(\mathbf{s}_c) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\left[\frac{1}{\tau^2}\sum_c \mathbf{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}(\mathbf{V}_c^* - \mathbf{w}_c) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right],\right. \\
&\qquad \left.\left[\frac{1}{\tau^2}\sum_c \mathbf{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}\mathbf{A}_c\mathbf{X}(\mathbf{s}_c) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\right)
\end{aligned}$$

for $c \in \mathcal{C}^G(\boldsymbol{\Lambda})$.

The full conditional distribution of γ is:

$$\begin{aligned}
&\left[\gamma \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \boldsymbol{\beta}, \alpha, \tau^2\right] \\
&\propto \prod_{c \in \mathcal{C}^G(\boldsymbol{\Lambda})}\left[\exp\left(-\frac{1}{2\tau^2}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)^\mathsf{T}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)\right)\right] \\
&\quad \times \left[\gamma^{a_\gamma - 1}\exp\left(-\frac{\gamma}{b_\gamma}\right)\mathbb{I}\{\gamma_{\min} \leq \gamma \leq \gamma_{\max}\}\right].
\end{aligned}$$

The full conditional distribution of α is:

$$\begin{aligned}
&\left[\gamma \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \boldsymbol{\beta}, \alpha, \tau^2\right] \\
&\propto \prod_{c \in \mathcal{C}^G(\boldsymbol{\Lambda})}\left[\exp\left(-\frac{1}{2\tau^2}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)^\mathsf{T}(\mathbf{V}_c^* - \mathbf{A}_c\mathbf{X}_{\mathbf{s}_c}\boldsymbol{\beta} - \mathbf{w}_c)\right)\right] \\
&\quad \times \left[(\alpha - c_\alpha)^{a_\alpha - 1}(d_\alpha - \alpha)^{b_\alpha - 1}\mathbb{I}\{c_\alpha \leq \alpha \leq d_\alpha\}\right].
\end{aligned}$$

The full conditional distribution of τ^2 is:

$$\begin{aligned}
&\left[\tau^2 \mid \{\boldsymbol{V}_c\}, \{\boldsymbol{u}_{c,0}\}, \{\boldsymbol{w}_c\}, \boldsymbol{\beta}, \alpha, \gamma\right] \\
&\sim \text{Inverse-Gamma}_{[0,\tau^2_{\max}]}\left(c_\tau + \frac{1}{2}|\mathcal{C}^G(\boldsymbol{\Lambda})|, d_\tau + \frac{1}{2}\sum_{c \in \mathcal{C}^G(\boldsymbol{\Lambda})}(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c)^\mathsf{T}(\mathbf{V}_c - \boldsymbol{u}_c - \mathbf{w}_c)\right).
\end{aligned}$$

The full conditional distribution of ν^2 is:

$$\begin{aligned}&\left[\nu^2 \mid \{\boldsymbol{w}_c\}, \rho\right] \\ &\sim \text{Inverse-Gamma}_{[0,\nu^2_{\max}]}\left(c_\nu + \frac{m}{2}|\mathcal{C}^G(\boldsymbol{\Lambda})|, d_\nu + \frac{1}{2}\sum_{c\in\mathcal{C}^G(\boldsymbol{\Lambda})} \mathbf{w}_c^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{w}_c\right).\end{aligned}$$

The full conditional distribution of ρ is:

$$\left[\rho \mid \{\boldsymbol{w}_c\}, \nu^2\right] \propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}|\mathcal{C}^G(\boldsymbol{\Lambda})|} \exp\left(-\frac{1}{2\nu^2}\sum_{c\in\mathcal{C}^G(\boldsymbol{\Lambda})} \mathbf{w}_c^\intercal \boldsymbol{\Sigma}^{-1}\mathbf{w}_c\right) \mathbb{I}\{0 < \rho < r\}.$$

B.3 MCMC sampling algorithm for the full dependence joint model

In this section we provide the MCMC sampling algorithm for the full dependence joint model, and we note that the algorithm for the reduced dependence model is quite similar with the appropriate machinery adjusted. The MCMC sampling algorithm for the full dependence joint model is as follows:

1. Define initial values for the parameters of the model.
2. Set $k = 1$.
3. Update $\boldsymbol{V}_{c,m}^{(k)}$ using Gibbs sampling for each $c \in \mathcal{C}^G(\boldsymbol{\Lambda})$ and $m = 1, \ldots, m_c$,

$$\boldsymbol{V}_{c,m}^{(k)} \sim \text{Normal}\left(\mathbf{u}_{c,m}^{(k-1)} + \mathbf{w}_{c,m}^{(k-1)}, \sigma^{2(k-1)}\mathbf{I}\right).$$

4. Update $u_{c,0}^{(k)}$ using Metropolis–Hastings sampling for each $c \in \mathcal{C}^G(\boldsymbol{\Lambda})$ with proposal $u_{c,0}^* \sim \text{Normal}(u_{c,0}^{(k-1)}, \sigma^2_{u_0,\text{tune}})$,

$$\begin{aligned}\log r = &\log P\left(u_{c,0}^* \mid \mathbf{V}_c^{(k-1)}, \mathbf{w}_c^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, \gamma^{(k-1)}, \alpha^{(k-1)}, \sigma^{2(k-1)}\right) \\ &- \log P\left(u_{c,0}^{(k-1)} \mid \mathbf{V}_c^{(k-1)}, \mathbf{w}_c^{(k-1)}, \boldsymbol{\beta}^{(k-1)}, \gamma^{(k-1)}, \alpha^{(k-1)}, \sigma^{2(k-1)}\right)\end{aligned}$$

Draw $u \sim \text{Uniform}(0, 1)$ and if $u < \log(r)$, then set $u_{c,0}^{(k)} = u_{c,0}^*$, otherwise set $u_{c,0}^{(k)} = u_{c,0}^{(k-1)}$.

5. Update $\mathbf{w}_c^{(k)}$ using Gibbs sampling for each $c \in \mathcal{C}^G(\mathbf{\Lambda})$,

$$\mathbf{w}_c^{(k)} \sim \text{Normal}\left(\left[\frac{1}{\tau^{2(k-1)}}\mathbf{I} + \frac{1}{\nu^{2(k-1)}}\mathbf{\Sigma}^{-1}\right]^{-1}\left[\frac{1}{\tau^{2(k-1)}}\left(\mathbf{V}_c^{(k)} - \mathbf{u}_c^{(k)}\right)\right], \left[\frac{1}{\sigma^{2(k-1)}}\mathbf{I} + \frac{1}{\nu^{2(k-1)}}\mathbf{\Sigma}^{-1}\right]^{-1}\right).$$

6. Update $\boldsymbol{\beta}^{(k)}$ using Gibbs sampling,

$$\boldsymbol{\beta}^{(k)} \sim \text{Normal}\left(\left[\frac{1}{\sigma^{2(k-1)}}\sum_{c\in\mathcal{C}^G(\mathbf{\Lambda})}\boldsymbol{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}\mathbf{A}_c\mathbf{X}(\mathbf{s}_c) + \mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1} \left[\frac{1}{\sigma^{2(k-1)}}\sum_{c\in\mathcal{C}^G(\mathbf{\Lambda})}\mathbf{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}(\mathbf{V}_c^{(k)} - \mathbf{w}_c^{(k)}) + \mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right], \left[\frac{1}{\sigma^{2(k-1)}}\sum_{c\in\mathcal{C}^G(\mathbf{\Lambda})}\mathbf{X}(\mathbf{s}_c)^\mathsf{T}\mathbf{A}_c^\mathsf{T}\mathbf{A}_c\mathbf{X}(\mathbf{s}_c) + \mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1}\right]^{-1}\right).$$

7. Update $\gamma^{(k)}$ using Metropolis–Hastings sampling with proposal $\gamma^* \sim \text{Normal}(\gamma^{(k-1)}, \sigma^2_{\gamma,\text{tune}})$,

$$\begin{aligned}\log r = &\log P\left(\gamma^* \mid \mathbf{V}_c^{(k)}, \mathbf{w}_c^{(k)}, \boldsymbol{\beta}^{(k)}, \alpha^{(k)}, \sigma^{2(k-1)}\right) \\ &- \log P\left(\gamma^{(k-1)} \mid \mathbf{V}_c^{(k)}, \mathbf{w}_c^{(k)}, \boldsymbol{\beta}^{(k)}, \alpha^{(k)}, \sigma^{2(k-1)}\right)\end{aligned}$$

   Draw $u \sim \text{Uniform}(0,1)$ and if $u < \log(r)$, then set $\gamma^{(k)} = \gamma^*$, otherwise set $\gamma^{(k)} = \gamma^{(k-1)}$.

8. Update $\alpha^{(k)}$ using Metropolis–Hastings sampling with proposal $\alpha^* \sim \text{Normal}(\alpha^{(k-1)}, \sigma^2_{\alpha,\text{tune}})$,

$$\begin{aligned}\log r = &\log P\left(\alpha^* \mid \mathbf{V}_c^{(k)}, \mathbf{w}_c^{(k)}, \boldsymbol{\beta}^{(k)}, \gamma^{(k)}, \sigma^{2(k-1)}\right) \\ &- \log P\left(\alpha^{(k-1)} \mid \mathbf{V}_c^{(k)}, \mathbf{w}_c^{(k)}, \boldsymbol{\beta}^{(k)}, \gamma^{(k)}, \sigma^{2(k-1)}\right)\end{aligned}$$

   Draw $u \sim \text{Uniform}(0,1)$ and if $u < \log(r)$, then set $\alpha^{(k)} = \alpha^*$, otherwise set $\alpha^{(k)} = \alpha^{(k-1)}$.

9. Update $\tau^{2(k)}$ using Gibbs sampling,

$$\tau^{2(k)} \sim \text{Inverse-Gamma}_{[0,\tau^2_{\max}]} \left(c_\tau + \frac{1}{2}|\mathcal{C}^G(\mathbf{\Lambda})|, \right.$$
$$\left. d_\tau + \frac{1}{2} \sum_{c \in \mathcal{C}^G(\mathbf{\Lambda})} (\mathbf{V}_c^{(k)} - \mathbf{u}_c^{(k)} - \mathbf{w}_c^{(k)})^\mathsf{T} (\mathbf{V}_c^{(k)} - \mathbf{u}_c^{(k)} - \mathbf{w}_c^{(k)}) \right).$$

10. Update $\nu^{2(k)}$ using Gibbs sampling,

$$\nu^{2(k)} \sim \text{Inverse-Gamma}_{[0,\nu^2_{\max}]} \left(c_\nu + \frac{m}{2}|\mathcal{C}^G(\mathbf{\Lambda})|, d_\nu + \frac{1}{2} \sum_{c \in \mathcal{C}^G(\mathbf{\Lambda})} \mathbf{w}_c^{(k)\mathsf{T}} \mathbf{\Sigma}^{-1} \mathbf{w}_c^{(k)} \right).$$

11. Update $\rho^{(k)}$ using Metropolis–Hastings sampling with proposal $\rho^* \sim \text{Normal}(\rho^{(k-1)}, \sigma^2_{\rho,\text{tune}})$,

$$\log r = \log P\left(\rho^* \mid \mathbf{w}_c^{(k)}, \nu^{2(k)}\right) - \log P\left(\rho^{(k-1)} \mid \mathbf{w}_c^{(k)}, \nu^{2(k)}\right)$$

Draw $u \sim \text{Uniform}(0,1)$ and if $u < \log(r)$, then set $\rho^{(k)} = \rho^*$, otherwise set $\rho^{(k)} = \rho^{(k-1)}$.

12. Update $\mathbf{\Lambda}^{(k)}$ using a Gibbs sampling step for each $\lambda_{ij}^{(k)}$, making use of the Gumbel max trick from Gumbel (1954).

If $z_{ij}^{(k-1)} = 0$, then

$$\lambda_{ij}^{(k)} = \ell \quad \text{with probability}$$
$$P\left(\lambda_{ij}^{(k)} = \ell\right) = \frac{1}{\sum_{\ell=1}^N \mathbb{I}\left\{\mathbf{y}_{ij} = \mathbf{s}_\ell^{(k-1)}\right\}},$$

If $z_{ij}^{(k-1)} = 1$, then

$$\eta_{ij\ell}^{(k)} = -\frac{1}{2\sigma^{2(k-1)}} \left(\mathbf{y}_{ij} - \mathbf{s}_\ell^{(k-1)}\right)^\mathsf{T} \left(\mathbf{y}_{ij} - \mathbf{s}_\ell^{(k-1)}\right)$$
$$- \frac{1}{2\tau^2} \left(\boldsymbol{V}_c{}^{(k)} - \boldsymbol{u}_c{}^{(k)} - \boldsymbol{w}_c{}^{(k)}\right)^\mathsf{T} \left(\boldsymbol{V}_c{}^{(k)} - \boldsymbol{u}_c{}^{(k)} - \boldsymbol{w}_c{}^{(k)}\right)$$
$$t_\ell \overset{iid}{\sim} \text{Gumbel}(0,1)$$

$$\lambda_{ij}^{(k)} = \arg \max_{\ell=1,\ldots,N} \eta_{ij\ell}^{(k)} + t_\ell.$$

13. Update $\mathbf{s}^{(k)}$ using Gibbs sampling for each $\mathbf{s}_{j'}^{(k)}$,

$$\mathbf{s}_{j'}^{(k)} \sim \begin{cases} P\left(\mathbf{s}_{j'}^{(k)} = \mathbf{y}_{ij}\right) = \dfrac{1}{n_{j'}^{(k)}} & \text{if } z_{ij}^{(k-1)} = 0 \\ \mathrm{N}_{2,[D^*]}\left(\dfrac{1}{n_{j'}^{(k)}} \displaystyle\sum_{i=1}^{m} \sum_{(j):\lambda_{ij}^{(k)}=j'} \mathbf{y}_{ij}, \dfrac{\sigma^{2(k-1)}}{n_{j'}^{(k)}} \mathbf{I}\right) & \text{if } z_{ij}^{(k-1)} = 1 \text{ and } n_{j'}^{(k)} > 0 \\ \mathrm{Uniform}\left(D^*\right) & \text{if } z_{ij}^{(k-1)} = 1 \text{ and } n_{j'}^{(k)} = 0. \end{cases}$$

14. Update $\mathbf{z}^{(k)}$ using Gibbs Sampling for each $z_{ij}^{(k)}$,

$$z_{ij}^{(k)} \sim \text{Bernoulli}(p_{ij}),$$

where

$$p_{ij}^{(k)} = \begin{cases} 1 & \text{if } \mathbf{y}_{ij} \neq \mathbf{s}_{\lambda_{ij}^{(k)}}^{(k)} \\ \dfrac{\frac{\theta^{(k-1)}}{2\pi\sigma^{2(k-1)}}}{\frac{\theta^{(k-1)}}{2\pi\sigma^{2(k-1)}} + \left(1-\theta^{(k-1)}\right)|B(\mathbf{s}_{\lambda_{ij}}{}^{(k)})|^{-1}} & \text{if } \mathbf{y}_{ij} = \mathbf{s}_{\lambda_{ij}^{(k)}}^{(k)}. \end{cases}$$

15. Update $\theta^{(k)}$ using Gibbs Sampling,

$$\theta^{(k)} \sim \text{Beta}\left(a_\theta + \sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}^{(k)}, b_\theta + \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(1 - z_{ij}^{(k)}\right)\right).$$

16. Update $\sigma^{2(k)}$ using Gibbs sampling,

$$\sigma^{2(k)} \sim \text{Inverse-Gamma}_{[0,b_\sigma]}\left(c_\sigma + \sum_{i=1}^{m}\sum_{j=1}^{n_i} z_{ij}^{(k)}, \right.$$
$$\left. d_\sigma + \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n_i} \mathbb{I}\left\{z_{ij}^{(k)} = 1\right\}\left(\mathbf{y}_{ij} - \mathbf{s}_{\lambda_{ij}^{(k)}}^{(k)}\right)^{\mathsf{T}}\left(\mathbf{y}_{ij} - \mathbf{s}_{\lambda_{ij}^{(k)}}^{(k)}\right)\right).$$

17. Save $\{\mathbf{V}_c^{(k)}, \mathbf{u}_c^{(k)}, \mathbf{w}_c^{(k)}, \boldsymbol{\beta}^{(k)}, \gamma^{(k)}, \alpha^{(k)}, \tau^{2(k)}, \nu^{2(k)}, \rho^{(k)}, \mathbf{s}_{j'}^{(k)}\}_{j'=1}^{N}$, $\sigma^{2(k)}$, $\mathbf{z}^{(k)}$, $\theta^{(k)}$, and $\boldsymbol{\Lambda}^{(k)}$.

18. Set $k = k + 1$ and return to Step 3. Iterate this algorithm through steps 3-17 until the sample size is large enough to adequately approximate the joint posterior distribution.

B.4 Initialization scheme

The initialization scheme for the joint model samplers is as follows. We initialize latent clusters in five main steps such that we match points with identical locations and build out our cluster set by chaining forward through time. The steps are as follows:

1. We partition the full set of tree-location observations by file (time-point) and perform a rapid "exact-match" pass such that any two points within a small tolerance ϵ are given the same cluster ID.
2. Starting from the first file, we "chain-match" forward through time, linking each unused point to its nearest neighbor in the next file provided it lies within a maximum chaining distance.
3. We compute each active cluster's centroid (the average location of its members), then assign any remaining un-clustered points to the nearest centroid if they fall within a maximum assignment distance, or else open a brand-new cluster.
4. We split off any points that lie too far from their assigned centroid, again creating singleton clusters as needed.
5. We pad the latent set by sampling additional jittered centroids up to a user-specified buffer size, and re-assign every observation to its closest centroid to yield the initial $\mathbf{\Lambda}$ labels and padded centroids $\{\boldsymbol{s}_{j'}\}$.

B.5 Point process estimation and mark model training

We utilized the workflow introduced in Chapter 3 to obtain our spatio-temporal point process model. We estimated the point process parameters using the `NLOPT_LN_SBPLX` algorithm in the `estimate_parameters_sc` function. We ran the model for a max evaluation of 350 iterations, with a tolerance of 10^{-4}. We trained the location-dependent mark model using an `XGBoost` model with a tuning grid size of 300 in the `train_mark_model` function.

B.6 Additional simulation results

In this section we provide additional simulation results for the linkage performance across hit-miss distortion levels and signal-to-noise ratio scenarios. The scenarios are described in Section 4.4.2 and a summary is provided in Table B.4.

Table B.4: Overview of the hit-miss distortion probability and signal-to-noise ratio factors.

Factor	Levels / Settings
Hit-Miss Distortion Probability	small: 0.15 medium: 0.25 large: 0.35
Signal-to-Noise Ratio	(1) $\tau^2 = 1.0,\ \nu^2 = 0.5,\ \rho = 0.75$ (2) $\tau^2 = 1.0,\ \nu^2 = 1.0,\ \rho = 0.75$ (3) $\tau^2 = 1.0,\ \nu^2 = 0.5,\ \rho = 0.5$ (4) $\tau^2 = 1.0,\ \nu^2 = 0.5,\ \rho = 1.0$ (5) $\tau^2 = 1.5,\ \nu^2 = 1.0,\ \rho = 1.0$

Figure B.1 provides precision and recall results for the full dependence (FD) and reduced dependence (RD) joint and two-stage model variants across varying hit-miss distortion levels.

Figure B.2 provides precision and recall results for the full dependence (FD) and reduced dependence (RD) joint and two-stage model variants across varying signal-to-noise ratio scenarios.

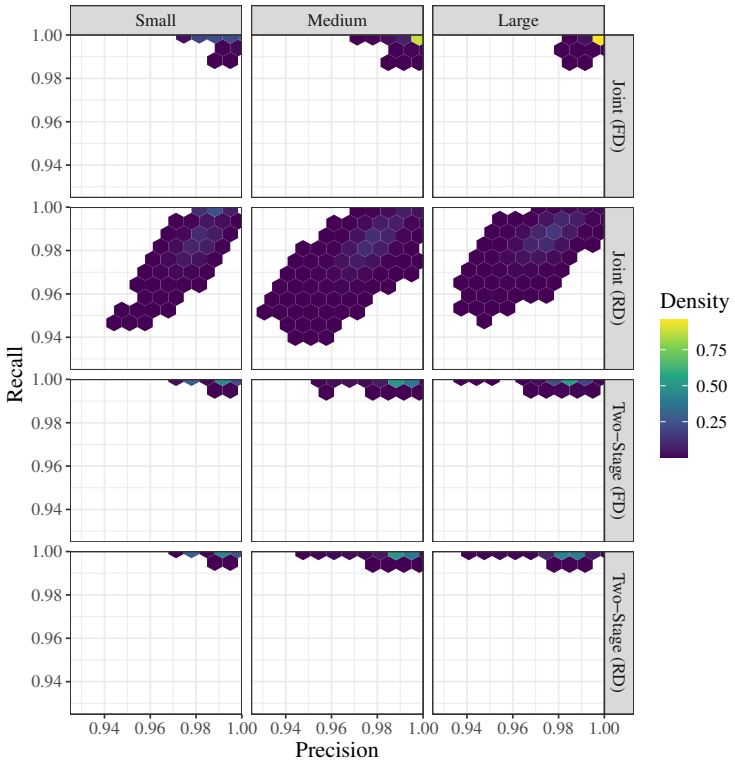

Figure B.1: Precision and recall joint densities for the full dependence (FD) and reduced dependence (RD) joint and two-stage model variants across varying hit-miss distortion levels. The precision and recall values are combined across 100 simulated datasets for each model variant and distortion level.

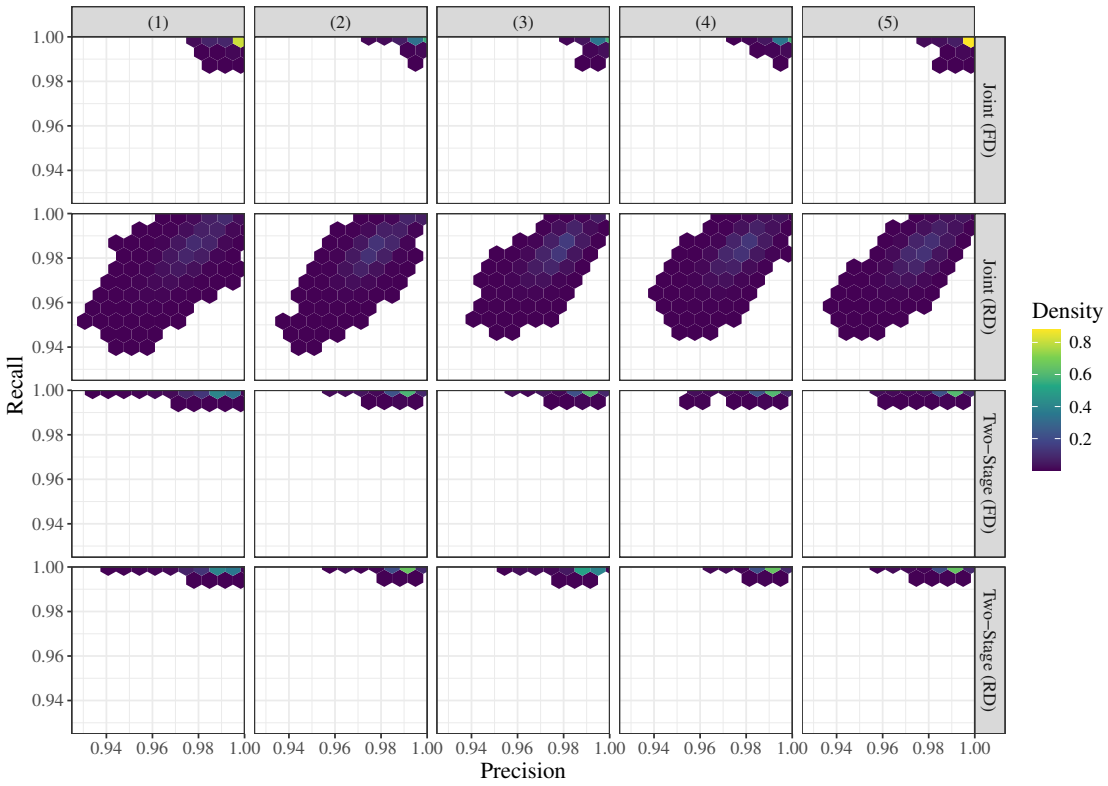

Figure B.2: Precision and recall joint densities for the full dependence (FD) and reduced dependence (RD) joint and two-stage model variants across varying signal-to-noise ratio scenarios. The precision and recall values are combined across 100 simulated datasets for each model variant and signal-to-noise ratio scenario.