Design and inference for line transect surveys in occupancy studies

Milly Jones

School of Engineering, Mathematics and Physics, University of Kent, Canterbury, UK
email: mlj23@kent.ac.uk

and

Eleni Matechou

School of Mathematical Sciences, Queen Mary University of London, UK

email: e.matechou@qmul.ac.uk

and

Diana Cole

School of Engineering, Mathematics and Physics, University of Kent, Canterbury, UK

email: d.j.cole@kent.ac.uk

and

Robin Lines

Wildlife Conservation Society, Cambodia

email: rlines@wcs.org

and

BIOMETRICS 64, 1–25 DOI: 10.1111/j.1541-0420.2005.00454.x
December 2008

Nicolas J. Deere

Durrell Institute of Conservation and Ecology (DICE), School of Natural Sciences,

University of Kent, Canterbury, UK

email: n.j.deere@kent.ac.uk

SUMMARY: Line transects, fixed paths along which observers record detections of animals or signs of their presence, are a core tool in ecological monitoring. Occupancy models commonly utilise line transect surveys to infer species' presence/absence in surveyed areas whilst accounting for imperfect detection and varying availability for detection along the transect. However, which ecological parameters can be inferred using such surveys depends on how detections are recorded, ranging from detection/non-detection in transect segments to continuous-time detections. We present the first comprehensive comparison of common line transect occupancy models within a unified framework, structured around data collection design, model formulation, parameter identifiability, and practical performance. We show that differing detection protocols give rise to structurally distinct statistical models, with varying identifiability properties and inferential performance. Using algebraic analysis and data cloning, we show that widely used models are often not practically identifiable. We develop a novel, data-based diagnostic to detect identifiability issues in applied settings and perform an extensive simulation study that illustrates the practical implications of data-collection approaches, model choice, and corresponding inference. We make recommendations regarding choice of detection protocol and line transect survey design and demonstrate the real-world consequences of these findings using a case study on leopards in the Zambesi-Transfrontier Conservation Area. Our results offer methodological innovation and actionable guidance for ecologists and statisticians.

KEY WORDS: Bayesian; Data cloning; Identifiability; Line transect studies; Occupancy models.

1. Introduction

Globally, biodiversity is declining at an unprecedented rate, reinforcing the need for wildlife monitoring methods that support reliable inference to guide conservation action (Gonzalez et al., 2023). Line transect surveys, fixed paths along which observers record detections of animals or signs of their presence, have become a cornerstone of such monitoring. Their appeal lies in logistical simplicity, scalability and ability to generate spatially structured data across large study regions (Hines et al., 2010; Whittington et al., 2015). They are routinely employed for large carnivore monitoring programmes (Hines et al., 2010; Lines et al., 2018; Sharma et al., 2024), for detecting signs of species poorly captured by emerging technologies (Wagner et al., 2025) and can be paired with new monitoring techniques such as telemetry and environmental DNA data (Kendall et al., 2023; Katayama et al., 2024) to validate ecological inference. Consequently, line transect data continue to be collected at scale and are increasingly used to inform high-stakes management decisions.

However, the spatial structure that makes transect surveys efficient also introduces statistical challenges. Line transect data are often analysed using occupancy models, a broad class of frameworks that infer species occupancy probabilities at surveyed sites whilst accounting for imperfect detection from detection/non-detection data (MacKenzie et al., 2002). Such models typically assume observations are independent, but detections collected along transects are rarely independent. Line transects often follow habitat features, such as rivers or roads, that are convenient for sampling whilst also functioning as low-resistance travel routes for wildlife. Consequently, detections recorded along transects are often spatially correlated. Ignoring this can result in bias and misleading inference (Hines et al., 2010), a pronounced problem when inference from transect surveys is expected to support management decisions and biased or unstable estimates can translate into misplaced conservation effort or delayed intervention.

The spatial autocorrelation in line transect data can be addressed using hierarchical

occupancy models that separate biological processes governing space use from the observation process that generates detections. During sampling, line transects are typically divided into segments of equal (and often arbitrary) length. Observers travel along the transects at a standardised pace and record, on each segment, direct encounters with animals or indirect signs of their presence such as dung or vocalizations (Hedges et al., 2013; Gestich et al., 2016); referred to as detections. A site is occupied if it intersects all or part of a species' home range during the sampling period. Within an occupied site, an individual transect segment is occupied if it intersects the home range. In this way, not all segments at occupied sites are used by the species, such as segments with high human footfall, which are routinely avoided by cryptic or sensitive species. An observer can only make detections on occupied segments (assuming no false positive observation error). Furthermore, detection probabilities on occupied segments are typically less than 1, so non-detection does not imply segment non-occupancy (false negative observation error). The detection process is therefore shaped by the spatial relationship between the species' home range, the sites and the transects, as illustrated schematically in Figure 1 a)i.

[Figure 1 about here.]

Building on this segment-level formulation, occupancy models for line transect data (also termed multi-scale single-season occupancy models, Nichols et al. (2008)) comprise three levels: site level occupancy, segment occupancy conditional on site occupancy, and detections conditional on segment occupancy. Spatial dependence in transect segments is most commonly modelled using discrete Markov processes (MP) (Hines et al., 2010; Whittington et al., 2015), with alternatives including intrinsic conditional autoregressive models (ICAR) (Crosby and Porter, 2018) and continuous-time Markov chain (CTMC) formulations (Guillera-Arroita et al., 2011). The segment-level detection data for these models depend critically on the detection protocol used, differing in their observation method and the num-

ber of observations made on each segment (segment-level replication). Common observation methods include binary detection/non-detection protocols (DND - record a 1 if a detection is made on a segment and a 0 otherwise) and counts (C - count of all detections on segment, Emmet et al., 2021) which can be implemented without complex equipment. Distance to first detection (DTD - distance between start of segment and first detection Henry et al., 2020) observations are increasingly taken following technological improvements (for example using camera time-stamps or GPS locations). Finally, independent replicate observations per transect segment can be taken by multiple devices or observers (Nichols et al., 2008; Whittington et al., 2015).

The choice of detection protocol is typically governed by cost, feasibility, and species detectability. Critically, an overlooked criterion is their effect on model identifiability. While the aforementioned occupancy models were introduced to account for spatial autocorrelation in line transect data, they also introduce additional latent structure and associated parameters. In the context of conservation monitoring, non-identifiability is particularly problematic because parameters that cannot be uniquely identified from the data cannot be estimated reliably, and non-identifiability can result in biased parameter estimates or large variability (Cole, 2020). Identifiability issues can arise from inherent model structure (structural non-identifiability) or from the underlying data set (practical non-identifiability). In multi-scale occupancy models, confounding between occupancy and detection processes has been identified previously and addressed through constraints or replicated observations (Nichols et al., 2008; Kendall et al., 2013). Hines et al. (2010) reported convergence problems in the absence of segment-level correlation, recommending aggregation of detection histories and modelling segments as independent. These studies demonstrate that identifiability in line transect occupancy models is highly sensitive to assumptions about spatial autocorrelation, their explicit representation in the model, and the detection protocol used. Identifiability

issues caused by model formulation will cause inferential problems whether using frequentist or Bayesian approaches. In the former these will often appear as convergence issues and in Bayesian approaches will often result in poor mixing in chains. In this paper, we employ Bayesian inference but employ uniform priors on all model probabilities so that results regarding identifiability are applicable within a frequentist setting as well.

While recent work has explored how detection protocols affect site-level occupancy probabilities (Pautrel et al., 2024), much less attention has been paid to segment-level occupancy probabilities and their identifiability in the presence of spatial autocorrelation. Furthermore, few guidelines exist for the effect of line transect design (such as length or number of segments) despite the influence these have on parameter interpretations and resulting model identifiability. This gap is particularly concerning given the potential of line transect surveys to infer fine-scale space-use patterns. Increasingly, conservation questions extend beyond simple presence or absence at the site level to understanding space use within home ranges, such as episodic use of habitat elements to satisfy ecological demands (Wright and Hooten, 2025). Species' habitat preferences can be modelled as functions of environmental covariates to quantify hierarchical habitat selection (McGarigal et al., 2016), providing information that is critical for targeted management. Importantly, negative impacts on species often manifest first at fine spatial resolutions through subtle changes in abundance or range contractions. Such changes can serve as early warning indicators of population decline before site-level extirpation becomes detectable through traditional occupancy analyses.

This paper places identifiability at the centre of line transect occupancy modelling. In Section 2, we develop a unifying framework that classifies line transect occupancy models according to their latent segment occupancy process (L), detection protocol (O), and number of independent replicate observations per segment (J), denoted L/O/J. In Section 3, we conduct an extensive simulation study to examine how these modelling and data-

collection choices affect identifiability and inference under ecologically realistic scenarios, with particular emphasis on spatial autocorrelation. Section 4 investigates identifiability and study-design considerations in detail and introduces a novel, data-based diagnostic for detecting identifiability problems using simple, cheap summaries of observed data, supported by extensive and computationally expensive data-cloning simulations. Finally, in Section 5, we apply our framework to leopard data from the Zambezi Trans-frontier Conservation Area, illustrating how detection protocol and transect design govern the reliability of fine-scale occupancy inference across the landscape and, ultimately, the value of line transect surveys for robust, timely conservation decision-making.

2. The Models

In this section, we develop the modelling framework introduced in Section 1. We consider S sites, where each site $i = 1, \cdots, S$ is occupied independently with probability ψ. The occupancy status of site i, denoted $w_i \in \{0, 1\}$, follows $w_i \sim \text{Bern}(\psi)$. As is standard in occupancy models, site occupancy is assumed to remain unchanged over the sampling period (MacKenzie et al., 2002).

At site i, M_i transects of length L_{im} are surveyed, with total site effort $L_i = \sum_{m=1}^{M_i} L_{im}$. Each transect is divided into segments of constant length R, resulting in $K_{im} = L_{im}/R$ segments for transect m at site i. Conditional on site occupancy, transect m may intersect wholly, partially, or not at all with the home range of the species (see Figure 1). A segment is occupied if it intersects with the species' home range. The latent segment occupancy status is denoted $z_{i,m,k} \in \{0, 1\}$ for segments $k = 1, \cdots, K_{im}$ at site i and transect m.

This defines a general multi-scale occupancy model (Kery and Royle, 2015) comprising three linked processes: site-level occupancy, segment-level occupancy conditional on site

occupancy, and detections conditional on segment occupancy:

$$
\begin{aligned}
w_i &\sim \text{Bernoulli}(\psi) && \text{(site-level occupancy)} \\
z_{i,m,k} \mid w_i &\sim \text{Bernoulli}(w_i \theta_{i,m,k}) && \text{(segment-level occupancy)} \\
y_{i,m,k,j} \mid z_{i,m,k} &\sim f(z_{i,m,k}, \beta) && \text{(detection process)}
\end{aligned}
$$

where $\theta_{i,m,k}$ is the probability that segment k at site i and transect m is occupied (given site occupancy). We denote by $y_{i,m,k,j}$ the j-th replicate observation on that segment, where observations arise from a distribution with density function f and detection parameters β. Assuming no false positive detections can be made, when $z_{i,m,k} = 0$ the probability of making a detection under density function f across any replicate on that segment is zero.

As discussed in Section 1, it is common to model the segment occupancy process using a first order spatial MP. In this framework, if a segment is occupied, the next segment is occupied with probability θ_{11}, and if a segment is unoccupied, the next is occupied with probability θ_{01}. The first segment at occupied sites is occupied with probability η. It is common to assume stationarity with $\eta = \theta_{01}/[\theta_{01} + 1 - \theta_{11}]$ or to set $\eta = \theta_{01}$ (Hines et al., 2010); alternatively η can be estimated as an additional model parameter. Stationarity may be applicable if the transect begins in the middle of a trail, and $\eta = \theta_{01}$ may be applicable if the survey is started from a point where individuals are unlikely to occupy the preceding space (such as a forest boundary).

The models in this paper share the site occupancy and segment occupancy process described above but differ by the detection protocols, which we describe in Section 2.1. The different models are outlined in Figure 1 a), showing the latent segment occupancy, the observation methods, and replication in the observation process.

2.1 *Detection process and model likelihoods*

The general framework for likelihood functions for line transect occupancy models is:

$$\mathcal{L} = \prod_{i=1}^{S} \left\{ \psi \prod_{m=1}^{M_i} H_{i,m} + (1-\psi)\mathbb{I}(\delta_i = 0) \right\}, \tag{1}$$

where δ_i is the number of species detections at site i and $H_{i,m}$ is the likelihood function for the observations at site i and transect m conditional on the site being occupied. The form of $H_{i,m}$ varies both by the segment occupancy model and the detection protocol. When modelling the latent segment occupancy process with a MP, $H_{i,m}$ takes the following form:

$$H_{i,m} = [1-\eta, \eta] \left\{ \prod_{k=1}^{K_{im}-1} \text{diag}(P_{i,m,k})\Theta \right\} P_{i,m,K_{iM_i}},$$

where

$$\Theta = \begin{bmatrix} 1-\theta_{01} & \theta_{01} \\ 1-\theta_{11} & \theta_{11} \end{bmatrix},$$

and $P_{i,m,k} = [P^0_{i,m,k}, P^1_{i,m,k}]$. $P^0_{i,m,k}$ and $P^1_{i,m,k}$ are the probabilities of observations on segment k at site i and transect m conditional on $z_{i,m,k} = 0$ and $z_{i,m,k} = 1$ respectively.

In each of the following models, the latent segment occupancy process is modelled using a MP, and along occupied segments detections are modelled as a Poisson point process with rate λ. The probability of a single observer making at least one detection on an occupied segment of length R is $p = 1 - \exp(-\lambda R)$.

2.1.1 *MP/DND/J.*

At site i, transect m, and segment k, $j = 1, \cdots, J$ independent detection/non-detections observations, $h_{i,m,k,j} \in \{0, 1\}$, are made. Let $\delta_{i,m,k} = \sum_{j=1}^{J} h_{i,m,k,j}$ denote the number of detections per segment k. Then:

$$P_{i,m,k} = [\mathbb{1}(\delta_{i,m,k} = 0), \prod_{j=1}^{J} p^{h_{i,m,k,j}}(1-p)^{1-h_{i,m,k,j}}]^T$$

2.1.2 *MP/DTD/J.*

At site i, transect m, and segment k, $j = 1, \cdots, J$ independent measurements, $l_{i,m,k,j}$, are made between the start of the segment and the first detection. If

no detection is made, we let $l_{i,m,k,j} = R$. Then:

$$P_{i,m,k} = [\mathbb{1}(\delta_{i,m,k} = 0), \prod_{j=1}^{J} \lambda^{\mathbb{1}(l_{i,m,k,j} \neq R)} \exp\big(-\lambda l_{i,m,k,j}\big)]^T$$

2.1.3 *MP/C/J.* At site i, transect m, and segment k, $j = 1, \cdots, J$ independent counts $c_{i,m,k,j}$ are made of the number of detections. Then:

$$P_{i,m,k} = [\mathbb{1}(\delta_{i,m,k} = 0), \prod_{j=1}^{J} \frac{(\lambda R)^{c_{i,m,k,j}} \exp(-\lambda R)}{c_{i,m,k,j}!}]^T$$

2.2 *Assumptions*

The modelling framework above makes the following assumptions, which are common to occupancy models. The occupancy status of a site or segment is unchanged throughout the survey period (otherwise known as closure). Relaxations to this assumption include multi-season occupancy models (MacKenzie et al., 2003). Replicate detections on a segment are independent given segment occupancy, and there are no false positives in the detection process. Occupancy models can be relaxed to allow for false positives (Royle and Link, 2006). In this paper we also assume that site occupancy probabilities, segment occupancy transition probabilities, and detection probabilities are constant. However, the models readily extend to parameters that vary as a function of site or segment level covariates, a point to which we return in Section 6.

3. Simulations

In this section, we compare the mean square error (MSE) of posterior distributions of models MP/DND/1, MP/DND/2, MP/DTD/1, MP/DTD/2, and MP/C/1 (see Web Appendix 1 for a short description of each model) under a range of different parameter settings. In Section 1, the absence of correlation in segment occupancy in the MP/DND/1 model is identified as a cause of convergence problems (Hines et al., 2010). We therefore simulate data varying the amount of correlation in segment occupancy (as $\theta_{01} \to \theta_{11}$) under high and low detectability

p. We denote by β the true value and by $\hat{\beta}_q$ the q-th posterior sample of a parameter. For Q samples, we let the MSE for this parameter be $\frac{1}{Q}\sum_{q=1}^{Q}(\hat{\beta}_q - \beta)^2$. All models are implemented in NIMBLE (de Valpine et al., 2017) in R (R Core Team, 2025).

We simulate $n = 10$ sites, $M = 10$ transects, and $K = 20$ segments of length 500. All simulations fix $\psi = 0.75$. Data are simulated using the discrete latent MP process with $\theta_{11} = 0.7$ and $\theta_{01} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. The first segment on each transect is drawn from stationarity. We consider two levels of detectability, one high with $\lambda = 1/300$ ($p \approx 0.81$), and one low with $\lambda = 1/1200$ ($p \approx 0.34$). For each parameter set, we simulate $N = 100$ data sets where detections on occupied segments follow a Poisson process with rate λ; detections are summarised according to the corresponding model. For example, detections per segment are binarised for MP/DND/1 or only the distance to first detection is kept for MP/DTD/1. Details on prior distributions and MCMC runs are in Web Appendix 2.

Figure 2 shows, for each model, boxplots of log(MSE) for parameters ψ, p, θ_{01} and θ_{11} across $\theta_{01} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. The distribution of MSE for the site occupancy parameter, ψ, is comparable across values of θ_{01} and all models. However, MP/DND/1 generally has much higher MSE for parameters p, θ_{01} and θ_{11} compared to the other models. It is also the only model where MSE increases rapidly for all parameters as $\theta_{01} \to \theta_{11}$. Overall, Figure 2 indicates that MSE is reduced by using continuous observations (MP/DTD/1 versus MP/DND/1) and by increasing replication at the segment level (MP/DTD/2 versus MP/DTD/1). Web Figure 1 shows results for $p \approx 0.34$; as expected, the MSE increases for parameters p, θ_{01} and θ_{11} for all models as detectability decreases. Results are comparable for ψ between low and high detectability. As in the high detectability case, MP/DTD/1 continues to have on average the second highest MSE behind MP/DND/1 for parameters p, θ_{01} and θ_{11}.

Even under ideal conditions, when p is high, as $\theta_{01} \to \theta_{11}$ the MSE for MP/DND/1 increases

drastically across parameters p, θ_{01}, and θ_{11}. Since $\theta_{01} \to \theta_{11}$ corresponds to decreasing spatial auto-correlation in segment occupancy, this leads to practical non-identifiability, which is investigated in the next section.

[Figure 2 about here.]

4. Identifiability

This section explores the identifiability of the models described in Section 2, focusing on the latent segment process and its implications for inference and survey design. Section 4.1 investigates structural identifiability using symbolic differentiation (Cole, 2020), and shows identifiability issues occur when correlation in segment occupancy is zero. This leads to practical non-identifiability when the correlation is close to zero. A method for diagnosing this is data cloning (Ponciano et al., 2012). Data cloning assesses parameter identifiability by replicating the observed data set, with the variances of practically identifiable parameters scaling inversely to the number of data copies. This method is computationally very demanding and requires expertise to fit. Section 4.2 introduces a novel data-based metric, Δ, as an indicator of correlation in the latent segment process. Crucially, we show that the data-based metric, which is simple and computationally cheap, can diagnose when this problem occurs. We explore the minimum number of segments required for identifiability using symbolic differentiation (Section 4.3), and extend this through data cloning to reveal weak identifiability at low segment numbers. Finally, in Section 4.4 we discuss the implications of segment length for survey design.

4.1 *Independence of spatial replicates*

An identifiable model can have certain points in the parameter space that are non-identifiable. These points can be identified algebraically using the methods outlined in Cole (2020) and Web Appendix 3. We find that MP/DND/1 is non-identifiable when

$\theta_{01} = \theta_{11} = \theta$. At this point, whether a segment is occupied is independent of adjacent segments. As only a single replicate observation is taken ($J = 1$), this model reduces to the restricted three-level occupancy model Kery and Royle (2015). The observation process becomes $y_{i,m,k,1}|z_{i,m,k} \sim \text{Bernoulli}(z_{i,m,k} \times \theta \times p)$, and is known to be non-identifiable (Nichols et al., 2008) as it is only possible to estimate θp jointly. In other words, a known non-identifiable model is nested within the MP/DND/1 design. We find the other models remain identifiable when $\theta_{01} = \theta_{11} = \theta$. Although other points in the parameter space lead to non-identifiable models, these are linked with boundary conditions (at points $\psi = 0, p = \lambda = 0, \theta_{01} = 0, (\eta, \theta_{01}) = (0, 0)$, and $\theta_{11} = 1$ when $\eta = \theta_{01}/(\theta_{01} + 1 - \theta_{11})$), full details of which can be found in Web Appendix 3.

4.2 *Metric* Δ *- the data diagnostic*

Model identifiability issues occur for MP/DND/1 when $\theta_{01} = \theta_{11} = \theta$. This condition cannot be checked prior to analysis as the segment occupancy process is unobserved. However, the detection process is observed; in this section we consider the link between correlation in the latent segment occupancy process and the observed detection process.

Conditional on site occupancy, $\mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 1, w_i = 1)$ denotes the probability a detection is made on segment k given a detection was made on segment $k - 1$. Conversely, $\mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 0, w_i = 1)$ denotes the probability a detection is made on segment k given no detection was made on segment $k - 1$. We then have the following result.

THEOREM 1: *Let site i be occupied and the latent segment occupancy process be modelled as a MP. Let η be the stationary probability for segment occupancy, and $h_{i,m,k} \in \{0, 1\}$ denote detection or non-detection on segment k of transect m. Then the following holds:*

$$\mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 1, w_i = 1) = \theta_{11}p, \tag{2}$$

$$\mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 0, w_i = 1) = \frac{p\theta_{01}(1 - \theta_{11}p)}{1 - \theta_{11} + \theta_{01}(1 - p)}. \tag{3}$$

253 *Proof.* See Web Appendix 4.1

254 Equality between Equations (2) and (3) implies that $\theta_{01} = \theta_{11}$, $p = 0$, or $\theta_{11} = 1$. These are
255 all points within the parameter space previously identified as leading to a non-identifiable
256 model under MP/DND/1. The point $p = 0$ is a boundary point occurring when there are
257 no detections. When η is modelled as the stationary probability and $\theta_{11} = 1$, then the first
258 segment on all transects is occupied, and all subsequent segments must remain occupied,
259 so that the MP is modelling a process in which no state switches occur. Finally, $\theta_{11} = \theta_{01}$
260 means there is no correlation in the latent segment occupancy process.

261 We let $\Delta = \mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 1, w_i = 1) - \mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 0, w_i = 1)$ denote
262 the magnitude of the difference in the conditional probabilities of observing a detection.
263 This is a statistic that is easily calculated from observed detection/non-detection data,
264 conditioning on site occupancy, and is a function of the detection probabilities p and segment
265 occupancy parameters θ_{01}, θ_{11}. We have shown that $\Delta = 0$ leads to identifiability issues under
266 MP/DND/1, and now further explore how Δ can indicate weakly-estimable parameters under
267 different detection protocols.

268 Theoretical identifiability does not ensure practical identifiability, that is model parameters
269 may be identifiability but weakly-estimable (Cole, 2020). We investigate this issue (across
270 all models in the paper) as $\Delta \to 0$ using data cloning. Data cloning assesses parameter
271 identifiability by replicating the observed data set D times and examining how parameter
272 variances scale as the amount of data increases. If a parameter is practically identifiable,
273 then its variance scales with $1/D$ (Ponciano et al., 2012). Fixing $\psi = 0.75$, $n = 10$ sites,
274 $M = 10$ transects per site, and $K = 20$ segments per transect, 5 data sets are generated
275 for each of 12 parameter settings (see Web Appendix 4.2 for full details). Each data set is
276 cloned $D \in \{1, 2, 5, 10, 15, 20\}$ times (by copying the detection histories resulting in a data

set of nD sites). The scaled variances, $\sigma^2_D(.)/\sigma^2_1(.)$, are computed from posterior variances $\sigma^2_D(.)$ for each parameter.

We compute the MSE of scaled variances with respect to expected scaled variances using $\sum_D \left(\frac{\sigma^2_D(.)}{\sigma^2_1(.)} - \frac{1}{D} \right)^2$ over $D \in \{1, 2, 5, 10, 15, 20\}$. Figure 3 shows log(MSE) for scaled variances of each parameter and data set against Δ for each model. Ten data sets for MP/DND/1 were removed where $\Delta < 0.1$ as the chains did not converge due to identifiability problems. As Δ decreases, the MSE for MP/DND/1 increases - indicating that parameters become increasingly non-identifiable even when $\Delta > 0$. The MSE for MP/DND/1 is comparable to other models (across all parameters) when $\Delta \approx 0.5$. This corresponds to simulations with highly ideal model parameters: high detectability, and strong spatial auto-correlation in segment occupancy. The MSE for MP/DTD/1 increases at small values of Δ in all parameters, but is comparable to other models when $\Delta \geq 0.2$. This suggests that MP/DTD/1 is more robust to low detectability and lack of auto-correlation in detections than MP/DND/1, but less so than the other models considered. MP/C/1 shows increases in MSE for θ_{01} and θ_{11} when $\Delta \approx 0$. These five points correspond to simulation parameters with the lowest detectability rather than lack of spatial auto-correlation in segment occupancy. MP/DND/2 and MP/DTD/2 are stable across Δ, even when detectability is low, showing that increased segment-level replication improves inference.

Crucially, we show that MP/DND/1 requires not just the existence of, but strong correlation in segment occupancy and high detectability in order to be fully identifiable. Whilst data cloning can be used to investigate identifiability, this is difficult for non-experts and is time consuming. Figure The metric, Δ, offers a practical and easy to use alternative that picks up when there is an issue of practical identifiability. Future data sets under this modelling framework can be quickly assessed for identifiability concerns using Δ, and it can serve as a tool to guide study design (more in Section 4.4).

[Figure 3 about here.]

4.3 *Study Design*

We investigate, under a range of different modelling assumptions, the number of segments required per transect to fully identify model parameters using symbolic differentiation (Cole, 2020) (full details in Web Appendix 5.1). We consider $K_{im} = 2, 3$, and 4 segments per transect, where 2 segments is the minimum to justify a MP to model segment occupancy. We look at whether ψ is to be estimated or assumed known and equal to 1 (for example where occupancy is known from other data in integrated models), and common forms for η. Results are shown in Web Tables 4 - 8 (in Web Appendix 5.1); table entries denote the deficiency in model identifiability, with 0 showing that the model is identifiable, or the level of deficiency otherwise. Web Table 4, for MP/DND/1, shows that $K = 4$ segments are required to guarantee model identifiability for all modelling constraints considered. Although fewer segments could be used, this would prevent analysis of the data under different constraints if required. Results for other models show that $K = 3$ segments are required across constraints in the same way. Practitioners should ensure that these conditions are met independently of the Δ value for the data set.

Since theoretical identifiability does not ensure practical estimability, we use data cloning to assess how estimability varies with the number of segments K. We take $n = 10$ sites, $M = 10$ transects per site, and vary $K \in \{4, 6, 8, 10, 15, 20\}$. We fix $\psi = 0.75$, $\theta_{11} = 0.7$, and $\theta_{01} = 0.1$ and let the first segment on each transect be drawn from stationarity. The detection rate is $\lambda = 1/300$ ($p \approx 0.81$). These parameter values have $\mathbb{E}(\Delta) = 0.46$ which is sufficiently great to distinguish the effects of K from the effect of spatial auto-correlation in segment occupancy. We generate $N = 10$ data sets for each K, and clone the data $D \in \{1, 2, 5, 10, 15, 20\}$ times. For full details see Web Appendix 5.2.

Figure 4 shows the box plot of log(MSE) of scaled variances across K for parameters p,

θ_{01}, and θ_{11} and each of the models. For MP/DND/1, the MSE increases as the number of segments decreases, and is greater than for other models when K is small. Whilst $K = 4$ ensures identifiability, model parameters for MP/DND/1 are weakly estimable and require far more segments to be strongly estimable. Results for models MP/DND/2, MP/DTD/1, MP/DTD/2, and MP/C/1 show that the distribution of MSE is relatively constant across K, and so model parameters remain estimable even at low number of segments.

[Figure 4 about here.]

4.4 *Choosing segment lengths*

The segment length R and number of segments K per transect are closely related in survey design. In this section, we show that the choice of R has a considerable impact on Δ, and thus may determine the identifiability of the resulting survey data. The true latent process is likely to be continuous and we consider modelling occupancy along the transect using a continuous time Markov chain (CTMC). Whilst the transect is occupied, detections are again made according to a Poisson process with rate λ. This formulation describes an interrupted Poisson process (IPP), full details of which are in Web Appendix 6.1. Under this modelling framework, we find $\mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 1, w_i = 1)$ and $\mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 0, w_i = 1)$, and derive the following result:

THEOREM 2: *Let the site be occupied, the latent segment occupancy process be modelled using a CTMC, η be the stationary probability of occupancy at the start of the transect, and $h_{i,m,k} \in \{0, 1\}$ denote detection or non-detection on segment k of transect m. Let segments have length R. Then the following holds:*

$$\lim_{R\to\infty} \Delta = \lim_{R\to\infty} \left\{ \mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 1, w_i = 1) - \mathbb{P}(h_{i,m,k} = 1|h_{i,m,k-1} = 0, w_i = 1) \right\} = 0. \tag{4}$$

Proof. See Web Appendix 6

Equation (4) confirms the known pattern that longer segments reduce correlations in detections—consistent with Hines et al. (2010), who recommend aggregating detection histories to weaken spatial dependence and enable independent segment modelling. Web Figure 2 shows how Δ varies with segment length R (for fixed transect length) across simulated detection rates (low-high) and switching rates between occupied and unoccupied states (low-high). As R increases, the mean Δ approaches zero in all scenarios, while variability rises because fewer segments, K, are observed. Across simulations, there is a segment length R that maximises the expected Δ, though under low detection and high switching rates, Δ remains flat. The maximal Δ also varies considerably between simulations. The maximal Δ increases with detection rate and lower switching (i.e. larger home ranges). Therefore, there may exist no R that guarantees sufficient identifiability for a chosen detection protocol. Theorem 2 and Web Figure 2 show that segment length as well as species characteristics (such as home range sizes and detectability) substantially impact Δ. Selecting an appropriate R thus determines which detection protocols are expected to lead to identifiable models in practice.

5. Case Study

We consider a case study of apex predators in the Kavango-Zambezi Transfrontier Conservation Area (KAZA TFCA) focussing here on the leopard (*Panthera pardus*) (Lines et al., 2018). From May to October 2015, a systematic randomized spoor and sightings survey was conducted on foot over 102 4-km transects mapped onto road networks (Lines et al., 2018). Each transect was visited $J = 3$ times over 10 days, and detection/non-detection data recorded on each segment of length 1km ($K = 4$). The surveyed region comprised $n = 10$ sites, with between $M = 8$ and 19 transects per site (see Figure 6 (a) for the distribution of transects over the KAZA TFCA). There were a total of 279 leopard sign detections over the 1224 segments, made at 61 of the 102 transects and 7 of the 10 sites.

The data are analysed using MP/DND/1, MP/DND/2, and MP/DND/3, and we assume

that the segment occupancy process starts at stationarity. See Web Appendix 7 for details on prior distributions and for posterior convergence summaries for each parameter. For $J = 1$ we have $\Delta = 0.303$, $J = 2$ gives $\Delta = 0.382$, and $J = 3$ gives $\Delta = 0.382$. Figure 3 indicates that for this value of Δ, MP/DND/1 is likely to experience identifiability problems for parameters p and θ_{11}. Furthermore, Figure 4 shows that identifiability issues are likely for these parameters when $K = 4$ segments are used. Figure 5(a) shows posterior means (circles) and 95% posterior credible intervals (PCIs -bars) for each model and model parameter. MP/DND/2 and MP/DND/3 show narrower PCIs in model parameters p and θ_{11} than MP/DND/1. Posterior distributions are comparable for ψ. Figure 5(b) shows the scaled variances for $D \in \{1, 2, 3, 4, 5, 10, 15, 20\}$ data clones. For each model parameter, MP/DND/2 and MP/DND/3 show a scaled variance that decreases in line with $1/D$. However, MP/DND/1 shows slower decrease for p and θ_{11}, and faster for θ_{01}. These results are consistent with those from Section 4.

Posterior means for θ_{11} for MP/DND/2 and MP/DND/3 are high, indicating a high probability that a segment following an occupied one is also occupied. This suggests the species uses the road network to facilitate long distance travel. Conversely, results for θ_{01} indicate a low probability of the species occupying the next segment if the current one is unoccupied. This inference is supported by Figure 6(a), which shows the placement of the transects within each of the 10 named sites, and Figures 6 (b), (c), and (d) which show the posterior probabilities of segment occupancy for models MP/DND/1, MP/DND/2, and MP/DND/3 respectively. Figure 6 shows that leopards widely use transect segments in the north-eastern region of the study area, and are largely absent from the south-western region. Crucially, this fine-scale map of leopards' use of the transects show that site level occupancy probabilities would miss details such as the spatial heterogeneity of segment occupancy at sites 'Moo' and 'Kalo'. Both these sites lie on the boundary between occupied

and non-occupied sites, highlighting how spatial use of the site becomes more fragmentary towards boundaries of the species' home range. Finally, we can see that posterior probabilities for model MP/DND/1 are more often close to 0.5 than they are for MP/DND/2 and MP/DND/3. This indicates greater uncertainty about latent segment occupancy states when only $J = 1$ replicate is used. This follows naturally from Figure 5(a), where the PCIs for θ_{11} were much wider when $J = 1$ than when $J = 2$ or $J = 3$. Issues with model parameter estimability therefore lead to considerably poorer inference about segment occupancy and hinder the ability to create reliable fine-scale maps of occupancy.

[Figure 5 about here.]

[Figure 6 about here.]

6. Discussion

Occupancy models are widely used in species monitoring programmes. In multi-scale designs where spatial autocorrelation among segments must be accounted for, Markov process models for segment occupancy are commonly applied. Using algebraic results and extensive simulation, this paper demonstrates that, despite being widely used, the single-replicate detection/non-detection per segment model with a Markovian latent process is particularly susceptible to identifiability issues. In contrast, continuous detection protocols (such as distance-to-first-detection and count data) and replicated segment-level designs are considerably more robust. We present a unified framework for analysing line transect data within occupancy models, enabling direct comparison of inferential performance across detection protocols. Simulation results show that the data-based diagnostic, Δ, derived from detection correlations, effectively identifies identifiability issues, with both segment number and length strongly influencing parameter identifiability. Together, these results provide practical guidance for designing more reliable surveys.

We make the following recommendations for survey design. Single detection/non-detection observations per segment are the most susceptible to identifiability problems. Continuous detection protocols, such as distance-to-first-detection or count data, generally produce lower bias, reduced uncertainty, and fewer identifiability issues, consistent with previous findings (Henry et al., 2020). However, gains from distance-to-first-detection protocols are limited under low detectability, and identifiability issues persist when data are sparse. Increasing segment-level observations substantially improves robustness, reducing sensitivity to low detectability and weak spatial dependence. In some cases replicated detection/non-detection designs outperform continuous protocols. Careful attention must be paid to the closure assumption when increasing replication; further investigation into fractional replication (where only a subset of sites are revisited; Doser and Stoudt, 2023) may show improved robustness while remaining feasible. Accordingly, when data are collected continuously, we strongly recommend against aggregating them into single-replicate detection/non-detection observations; aggregating to segment counts offers greater robustness than distance-to-first-detection. Overall, replicated segment-level observations are most robust, with both detection/non-detection and distance-to-first-detection performing comparably.

Choice of segment length and number is a critical but often overlooked element of survey design. While four segments per transect are theoretically sufficient for identifiability in single-replicate detection/non-detection models, substantially more are needed for reliable estimation. Similarly, four segments per transect are generally sufficient for continuous detection protocols and replicated segment-level observations when detectability is high. Practically, the primary consideration should be segment length rather than segment number, though these are closely linked. Segment length influences spatial dependence in detections across segments and, in turn, identifiability under single-replicate detection/non-detection and distance-to-first-detection protocols. As line transect surveys typically collect data con-

tinuously, we propose a post-hoc optimisation to guide model selection and data processing. Segment lengths should vary from near zero to one-quarter of the transect length, discretising the data at each step. The optimal segment length maximises the metric Δ. If $\Delta < 0.5$ at the optimum, we recommend against single-replicate detection/non-detection models; if $\Delta < 0.2$, we recommend against single-replicate distance-to-first-detection models. When continuous data collection is infeasible at large scale, a pilot study can optimise segment length and detection protocol for future surveys by recording detection locations and applying the same optimisation method. These recommendations extend to legacy line transect data, allowing re-analysis with more robust models that reduce bias and uncertainty. For historic data sets restricted to single-replicate detection/non-detection observations, the diagnostic Δ can assess the reliability and identifiability of parameter estimates.

Across all simulation scenarios, estimates of site-level occupancy ψ were relatively robust, consistent with previous findings (Pautrel et al., 2024), and comparable across detection protocols. However, when the inferential focus shifts to detection or segment-level occupancy, parameters critical for understanding fine-scale space use and hierarchical habitat selection, survey design choices become more influential. This was particularly evident in the leopard case study in the KAZA-TFCA, where segment-level occupancy revealed detailed spatial variation toward home range edges, patterns missed at the site level. The single-replicate detection/non-detection protocol, however, showed substantially greater uncertainty than multi-replicate designs.

We excluded covariates to establish the baseline identifiability of each model, following recommendations that covariates should not be used to resolve non-identifiability (Kery and Royle, 2015). Extending the framework to incorporate heterogeneous observation protocols (for example, mixing discrete and continuous detections) and sign types (spoor vs. scat) is feasible, if independence assumptions are maintained, to alleviate time and resource constraints

in field surveys. While our recommendations focus on single-species frameworks, further work should extend these results to multi-species and multi-season models, particularly where species behaviour or detectability differences influence design. Although we focus on Markov process latent dynamics and often assume stationarity, alternative latent processes and non-stationary initial conditions warrant investigation, especially on short transects, where misspecification may bias inference (Crosby and Porter, 2018).

This study provides the first comprehensive exploration of identifiability under multi-scale occupancy models with Markovian latent processes. By linking identifiability to directly observable data, we present an accessible approach to diagnosing and optimising survey design. While specific recommendations depend on species traits such as detectability and home-range size, our results establish a mechanism by which an optimal model can be selected. Broadly, species with low detectability and small home ranges require continuous detection protocols or segment-level replication to ensure robust inference. More generally, this optimisation-based approach provides a transparent, data-driven foundation for designing and evaluating species monitoring programmes, allowing practitioners to tailor strategies that maximise identifiability and minimise uncertainty in ecological inference.

Funding

The PhD project of the corresponding author, MJ, is funded by the Engineering and Physical Sciences Research Council. NJD was supported by a Leverhulme Trust Research Leader Award to Dr Matthew J. Struebig.

Supplementary Materials

Supplementary material is available at Biometrics online. Web Appendices, Tables, and Figures referenced in Sections 4-5 are available with this paper at the Biometrics website on Oxford Academic.

Data Availability

The R code and case study data used in this paper to support our findings are available at `https://github.com/millyljones/Line-Transect-Occupancy`

References

Cole, D. J. (2020). *Parameter Redundancy and Identifiability.* CRC Press.

Crosby, A. D. and Porter, W. F. (2018). A spatially explicit, multi-scale occupancy model for large-scale population monitoring. *The Journal of Wildlife Management* **82,** 1300–1310.

de Valpine, P., Turek, D., Paciorek, C., Anderson-Bergman, C., Temple Lang, D., and Bodik, R. (2017). Programming with models: writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics* **26,** 403–413.

Doser, J. W. and Stoudt, S. (2023). "Fractional replication" in single-visit multi-season occupancy models: Impacts of spatiotemporal autocorrelation on identifiability. *Methods in Ecology and Evolution* .

Emmet, R. L., Long, R. A., and Gardner, B. (2021). Modeling multi-scale occupancy for monitoring rare and highly mobile species. *Ecosphere* **12,**.

Gestich, C. C., Caselli, C. B., Nagy-Reis, M. B., Setz, E. Z., and da Cunha, R. G. (2016). Estimating primate population densities: the systematic use of playbacks along transects in population surveys. *American Journal of Primatology* **79,** e22586.

Gonzalez, A., Vihervaara, P., Balvanera, P., Bates, A. E., Bayraktarov, E., Bellingham, P. J., Bruder, A., Campbell, J., Catchen, M. D., Cavender-Bares, J., et al. (2023). A global biodiversity observing system to unite monitoring and guide action. *Nature Ecology & Evolution* **7,** 1947–1952.

Guillera-Arroita, G., Morgan, B. J. T., Ridout, M. S., and Linkie, M. (2011). Species

occupancy modeling for detection data collected along a transect. *Journal of Agricultural, Biological, and Environmental Statistics* **16,** 301–317.

Hedges, S., Johnson, A., Ahlering, M., Tyson, M., and Eggert, L. S. (2013). Accuracy, precision, and cost-effectiveness of conventional dung density and fecal DNA based survey methods to estimate Asian elephant (Elephas maximus) population size and structure. *Biological Conservation* **159,** 101–108.

Henry, D. A. W., Lee, A. T. K., and Altwegg, R. (2020). Can time-to-detection models with fewer survey replicates provide a robust alternative to traditional site-occupancy models? *Methods in Ecology and Evolution* **11,** 643–655.

Hines, J. E., Nichols, J. D., Royle, J. A., MacKenzie, D. I., Gopalaswamy, A. M., Kumar, N. S., and Karanth, K. U. (2010). Tigers on trails: occupancy modeling for cluster sampling. *Ecological Applications* **20,** 1456–1466.

Katayama, N., Yamamoto, S., Baba, Y. G., Ito, K., and Yamasako, J. (2024). Complementary role of environmental DNA for line-transect bird surveys: A field test in a Japanese rice landscape. *Ecological Indicators* **166,** 112442.

Kendall, C. J., Bracebridge, C., Lynch, E. C., Mgumba, M., Monadjem, A., Nicholas, A., and Kane, A. (2023). Value of combining transect counts and telemetry data to determine short-term population trends in a globally threatened species. *Conservation Biology* **37,**.

Kendall, W. L., Hines, J. E., Nichols, J. D., and Grant, E. H. C. (2013). Relaxing the closure assumption in occupancy models: staggered arrival and departure times. *Ecology* **94,** 610–617.

Kery, M. and Royle, J. A. (2015). *Applied Hierarchical Modeling in Ecology Analysis of Distribution, Abundance and Species Richness in R and Bugs - Prelude and Static Models*. Elsevier Science & Technology Books.

Lines, R., Tzanopoulos, J., and MacMillan, D. (2018). Status of terrestrial mammals at

the Kafue–Zambezi interface: implications for transboundary connectivity. *Oryx* **53,** 764–773.

MacKenzie, D. I., Nichols, J. D., Hines, J. E., Knutson, M. G., and Franklin, A. B. (2003). Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. *Ecology* **84,** 2200–2207.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J., and Lagtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology* **83,** 2069–2360.

McGarigal, K., Wan, H. Y., Zeller, K. A., Timm, B. C., and Cushman, S. A. (2016). Multi-scale habitat selection modeling: a review and outlook. *Landscape Ecology* **31,** 1161–1175.

Nichols, J. D., Bailey, L. L., O'Connell Jr., A. F., Talancy, N. W., Campbell Grant, E. H., Gilbert, A. T., Annand, E. M., Husband, T. P., and Hines, J. E. (2008). Multi-scale occupancy estimation and modelling using multiple detection methods. *Journal of Applied Ecology* **45,** 1321–1329.

Pautrel, L., Moulherat, S., Gimenez, O., and Etienne, M. (2024). Analysing biodiversity observation data collected in continuous time: Should we use discrete- or continuous-time occupancy models? *Methods in Ecology and Evolution* .

Ponciano, J. M., Burleigh, J. G., Braun, E. L., and Taper, M. L. (2012). Assessing Parameter Identifiability in Phylogenetic Models Using Data Cloning. *Systematic Biology* **61,** 955–972.

R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Royle, J. A. and Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology* **87,** 835–841.

Sharma, K., Alexander, J. S., Durbach, I., Kodi, A. R., Mishra, C., Nichols, J., MacKenzie,

D., Ale, S., Lovari, S., Modaqiq, A. W., Zhi, L., Sutherland, C., Khan, A. A., McCarthy, T., and Borchers, D. (2024). Chapter 34 - PAWS: Population Assessment of the World's Snow leopards. In Mallon, D. and McCarthy, T., editors, *Snow Leopards (Second Edition)*, Biodiversity of World: Conservation from Genes to Landscapes, pages 437–447. Academic Press, Second edition.

Wagner, B., Garnick, S. W., Ryan, M. F., Isaac, J. L., Begg, A., and Nitschke, C. R. (2025). Thermal drone surveys to detect arboreal fauna: Improving population estimates and threatened species monitoring. *Ecological Applications* **35,**.

Whittington, J., Heuer, K., Hunt, B., Hebblewhite, M., and Lukacs, P. M. (2015). Estimating occupancy using spatially and temporally replicated snow surveys. *Animal Conservation* **18,** 92–101.

Wright, W. J. and Hooten, M. B. (2025). Continuous-space occupancy models. *Biometrics* **81,**.

Received October 2007. *Revised February* 2008. *Accepted March* 2008.

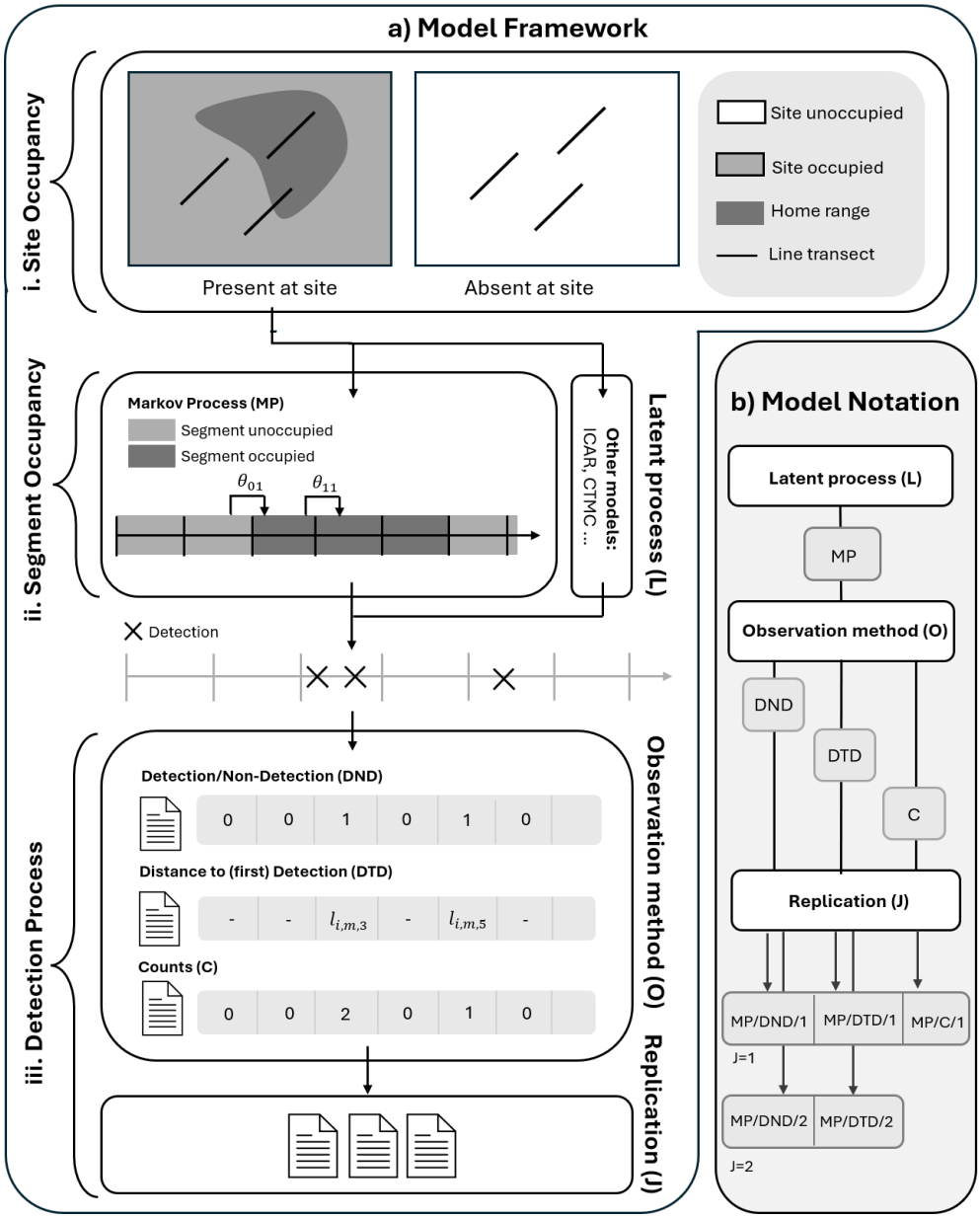
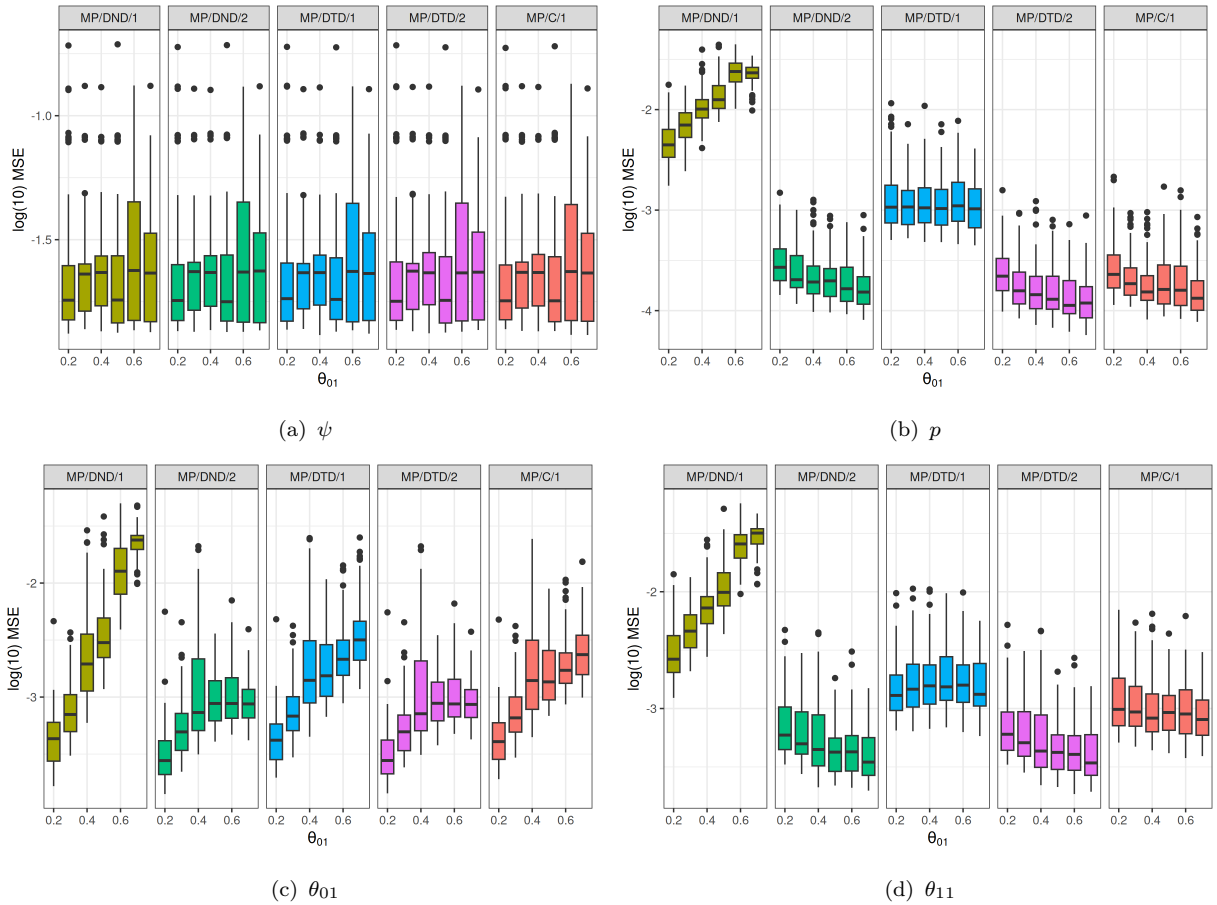


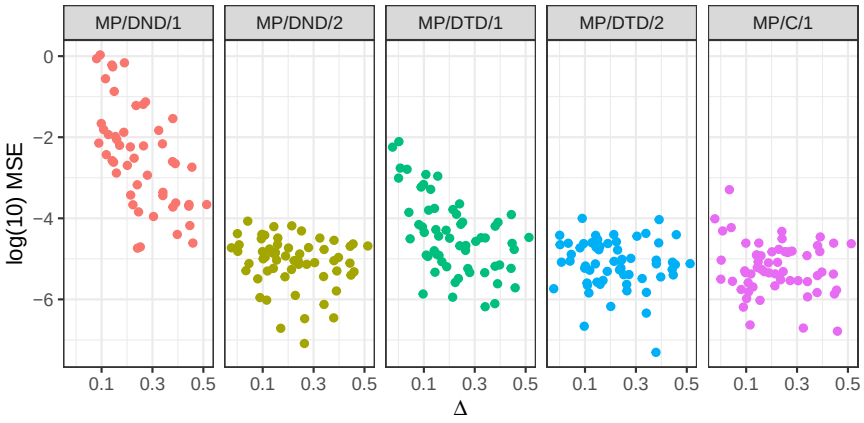
Figure 1. a) Modelling framework including site occupancy, segment occupancy and the detection process. i. A site may be unoccupied or occupied, with transects wholly, partially, or not intersecting the home range of species' at occupied sites. ii. The latent segment occupancy process models the intersection of transects with the home range. iii. The detection process comprises the observation method and replication of observations. b) Models are labelled L/O/J according to their latent segment occupancy process (L), observation method (O), and amount of replication (J)



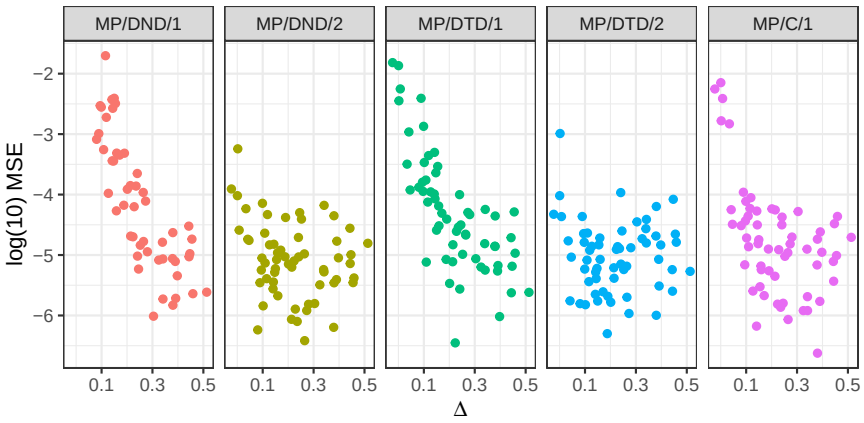
(a) ψ (b) p

(c) θ_{01} (d) θ_{11}

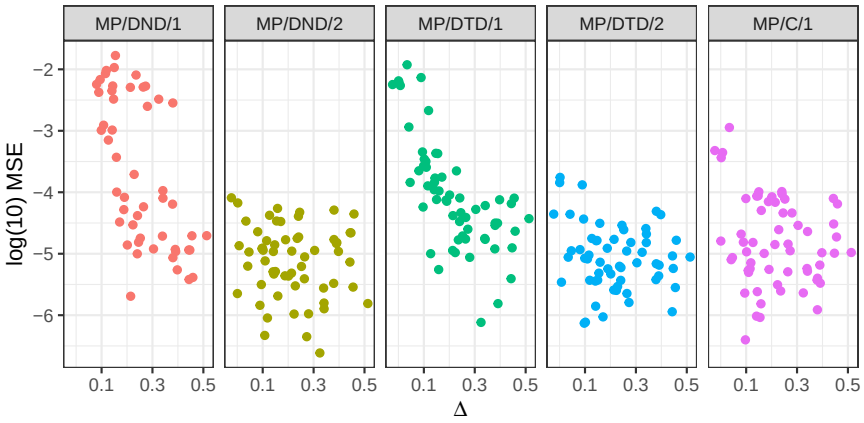
Figure 2. Boxplot of log(MSE) for parameters ψ, p, θ_{01} and θ_{11} across $\theta_{01} \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ when $\theta_{11} = 0.7$ and $p \approx 0.81$, for models MP/DND/1, MP/DND/2, MP/DTD/1, MP/DTD/2, MP/C/1.



(a) p

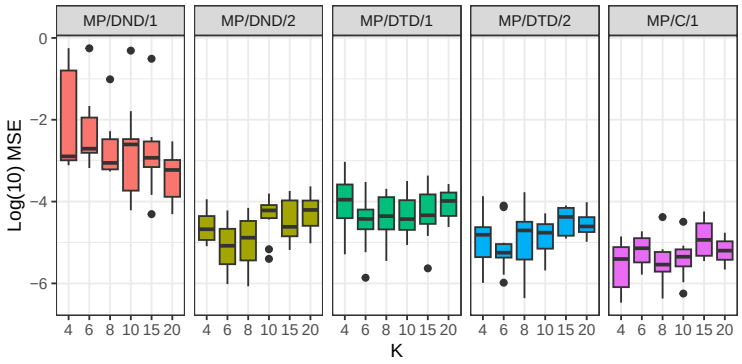


(b) θ_{01}

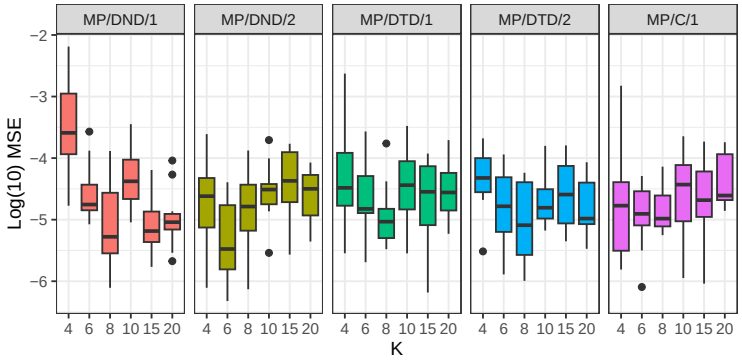


(c) θ_{11}

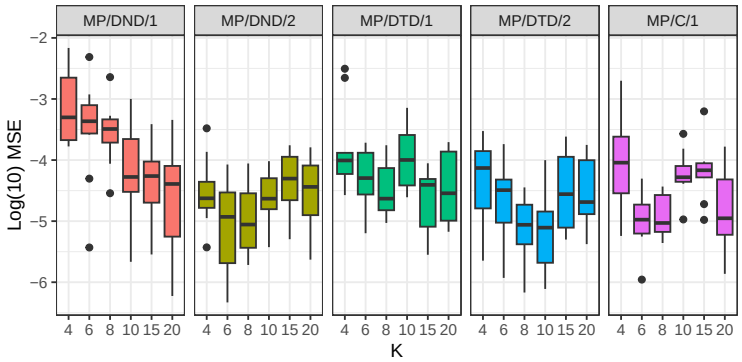
Figure 3. Log(10) MSE for parameters p, θ_{01} and θ_{11} across metric values Δ and models MP/DND/1, MP/DND/2, MP/DTD/1, MP/DTD/2, MP/C/1.



(a) p

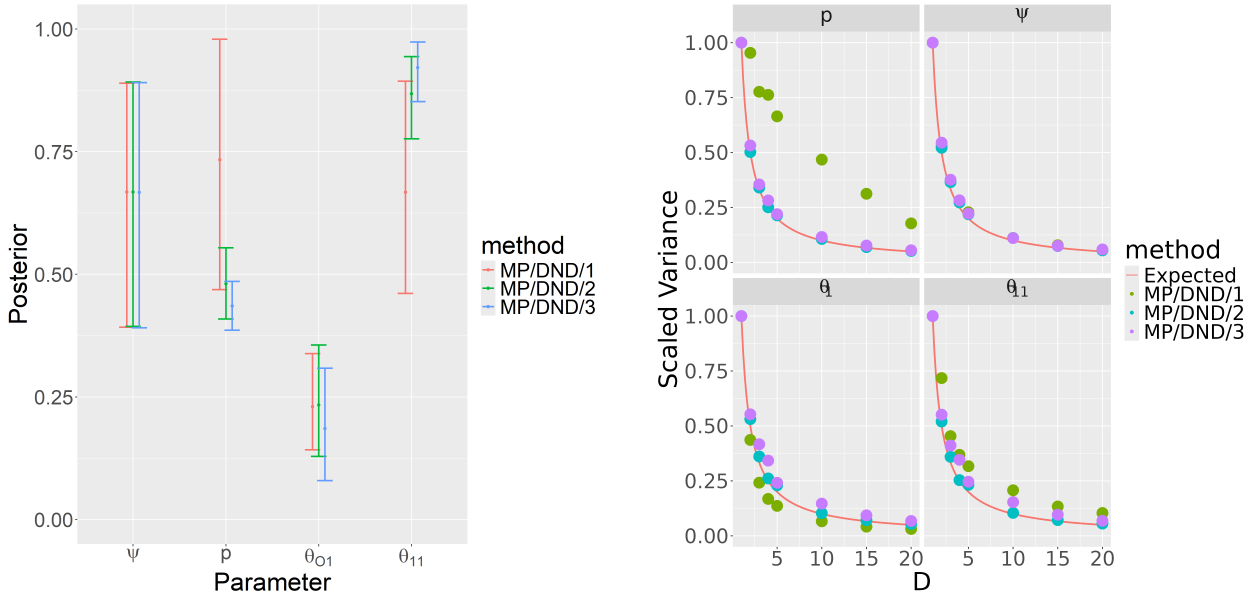


(b) θ_{01}



(c) θ_{11}

Figure 4. Log(10) MSE for parameters p, θ_{01} and θ_{11} across number of segments $K \in \{4, 6, 8, 10, 15, 20\}$ and models MP/DND/1, MP/DND/2, MP/DTD/1, MP/DTD/2, MP/C/1.



(a) (b)

Figure 5. (a) Posterior means (circles) and 95% PCIs (bars) for MP/DND/1, MP/DND/2, and MP/DND/3. (b) Scaled variances (points) for models MP/DND/1, MP/DND/2, and MP/DND/3 for $D \in \{1, 2, 3, 4, 5, 10, 15, 20\}$. The expected scaled variance $1/D$ is the plotted line.

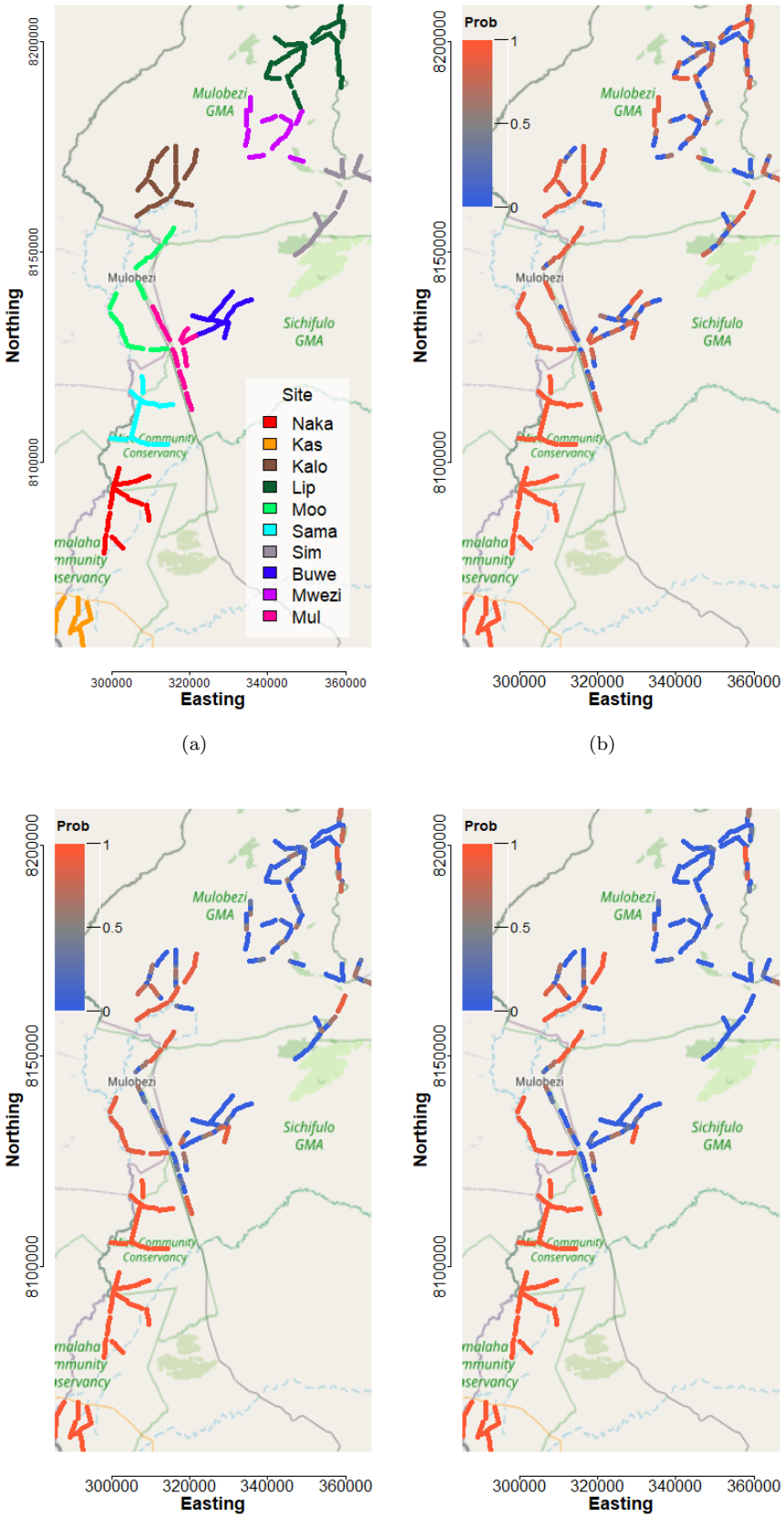


Figure 6. Map of the study area with transect segments coloured according to the (a) site in which they belong, or the posterior probability of segment occupancy from model (b) MP/DND/1, (c) MP/DND/2, (d) MP/DND/3