Received: 8 November 2023 | Accepted: 14 June 2024

DOI: 10.1111/2041-210X.14386

Methods in Ecology and Evolution

RESEARCH ARTICLE

A Monte Carlo resampling framework for implementing goodness-of-fit tests in spatial capture-recapture models

Yan Ru Choo | Chris Sutherland | Alison Johnston

Centre of Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, UK

Correspondence
Yan Ru Choo
E mail
Email: chooyanru77@gmail.com

Funding information
Engineering and Physical Sciences Research Council

Handling Editor: Andres Lopez-Sepulcre

Abstract

1. Spatial capture-recapture (SCR) models provide estimates of animal density from spatially referenced encounter data and has become the most widely adopted approach for estimating density. Despite the rapid growth in the development and application of spatial capture-recapture methods, approaches for assessing model fit have received very little attention when compared to other classes of hierarchical models in ecology.
2. Here, we develop an approach for testing goodness-of-fit (GoF) for frequentist SCR models using Monte Carlo simulations. We derive probability distributions of activity centres from the fitted model. From these, we calculate the expected encounters in the capture history based on the SCR parameter estimates, propagating the uncertainty of the estimates and the activity centre locations via Monte Carlo simulations. Aggregating these test statistics result in count data, allowing us to test fit with Freeman-Tukey tests. These tests are based on summary statistics of the total encounters of each individual at each trap (FT-ind-trap), total encounters of each individual (FT-individuals) and total encounters at each trap (FT-traps). We assess the ability of these GoF tests to diagnose lack of fit under a range of assumption violating scenarios.
3. FT-traps had the strongest response to unmodelled spatial and trap heterogeneity in detection probability (power = 0.53–0.56), while FT-ind-traps had the strongest responses to random individual variation in detectability (power = 0.88) and non-spatial discrete variation in g_0 (power = 0.35). The tests, designed to diagnose poor fit in the detection parameters, were insensitive to unmodelled heterogeneity in density (power = <0.001). They also demonstrated low false positive rates (<0.001) when the correct models were fitted; therefore, it is very unlikely that they will provide false indications of poor model fit.
4. We demonstrate that these GoF tests are capable of detecting lack-of-fit when unmodelled heterogeneity is present in the detection sub-model. When used jointly, the combinations of test results are also able to infer the type of lack-of-fit in certain cases. Our Monte Carlo sampling methods may be extended to a wider

range of GoF tests, thereby providing a platform for developing more GoF methods for SCR.

KEYWORDS
density estimation, goodness-of-fit, Monte Carlo resampling, spatial capture-recapture

1 | INTRODUCTION

Robust density estimates are important for effective wildlife conservation and management (Morin et al., 2022). One method for estimating animal density is spatial capture-recapture (SCR), a hierarchical model that estimates animal density from spatially referenced encounters of individuals with known identity (Borchers & Efford, 2008; Efford, 2004; Royle et al., 2009; Royle & Young, 2008). SCR uses capture histories that are relatively easy to collect and as such has been adopted widely in many aspects of ecology and conservation, including for monitoring low density, declining species that are of high conservation value (Morin et al., 2022; Tourani, 2022). However, unlike other hierarchical models for analysing ecological data in observational studies, approaches for evaluating goodness-of-fit (GoF) has received relatively little attention, and crucial aspects of these tests such as power and diagnostic ability are not well-understood (Dey et al., 2022; Efford, 2023; Royle et al., 2014b).

Previous assessments of goodness-of-fit with violations of SCR assumptions have focussed on a narrow range of types of lack-of-fit. For example, recent research has focussed on detecting lack-of-fit when SCR models are fit with a misspecified detection function (Dey et al., 2022), with a consensus that existing GoF approaches for SCR appears insensitive to even extreme misspecification (Dey et al., 2022; Efford & Mowat, 2014; Russell et al., 2012). This is perhaps expected given that SCR typically uses sparse data and therefore unlikely to be informative about differences in the specific shape of distance-dependent detection functions. These existing assessments of fit have built a perception of robustness in SCR models, but these assessments focus on detecting lack of fit in aspects of the data and the model that do not relate to core SCR assumptions, in particular that SCR models should not contain any unmodelled heterogeneity (Borchers & Efford, 2008; Royle et al., 2009). The latter should be a greater source of concern to practitioners as it has been widely documented that unmodelled heterogeneity in detectability induces bias in density estimates with SCR models (Dey et al., 2023; Efford & Mowat, 2014; Gardner et al., 2010; Royle et al., 2013; Sutherland et al., 2015; Tobler & Powell, 2013), although this may not be the case with temporal heterogeneity (Sollmann, 2024) or interspersed spatial heterogeneity (Moqanaki et al., 2021). The effectiveness and performance of GoF tests for SCR should thus be assessed against adherence to core SCR assumptions instead.

Spatial capture-recapture is a hierarchical model, consisting of a latent point process model for animal density and an observation model for the observed detections to account for the imperfect detection of animals in the surveyed population. Since SCR estimates are the joint product of two sub-models, of which the density sub-model is latent, it can be challenging to define and measure goodness-of-fit in SCR models (Royle et al., 2014b). Bayesian implementations of SCR models explicitly estimate the unobserved activity centre locations of all individuals (Royle et al., 2009, 2014a; Royle & Young, 2008) and as such is relatively straightforward—and standard practice—to obtain posterior distributions of the activity centres (Royle et al., 2014b). Goodness-of-fit testing in the Bayesian context exploits this feature via posterior predictive checks. Expected encounter rates conditioned on activity centre locations are predicted from the fitted model, against which predicted encounter rates are used to calculate the posterior distributions of fit statistics for the actual data and data simulated from the fitted model. The fit statistics of the actual data are compared to fit statistics of simulated data to calculate a Bayesian *p*-value used to assess model fit. As the expected encounter rates are conditioned on the locations of animals that gave rise to the data, the distributions of the Bayesian GoF test statistics are constrained to the individuals present in the study and hence provide more relevant inferences on model fit to the data at hand, while also accounting for the uncertainty in the activity centre locations.

Frequentist implementations of SCR models generally integrate across all possible activity centres in the region of interest to estimate density without directly estimating their locations (Borchers & Efford, 2008); hence, the Bayesian approaches for GoF testing cannot be directly applied. Existing approaches to GoF testing use Monte Carlo simulations, where estimated SCR parameters are used to generate new capture histories to obtain the empirical distribution of a fit statistic (Efford, 2023). The fit statistic may be a summary statistic, such as the number of individuals captured once, or the model deviance, which would require refitting of the model in every simulation (Efford, 2023). The summary statistics or deviance from the original data are ranked against the corresponding Monte Carlo distributions to measure model fit (Efford, 2023). However, both implementations of summary statistics and deviance have some important disadvantages compared to the Bayesian goodness-of-fit tests. While the deviance is an all-encompassing measure that can pick up overall lack-of-fit, it does not provide a means of identifying specific sources of lack-of-fit and therefore does not offer any insight for improving the model. Conversely, using summary statistics may potentially be more specific, but as they can contain sources of poor fit from both the density and observation sub-models, ad-hoc assumptions

about the goodness-of-fit of either sub-model would have to be made to draw inferences on the source of poor fit. Due to their low specificity, the diagnostic power and utility of these GoF tests may also be limited from a practical perspective.

While there is an obvious need to develop better GoF tests for SCR in general, the existing Bayesian approaches have some attractive diagnostic characteristics that are not yet available for frequentist applications of SCR. Here we present a goodness-of-fit approach for frequentist SCR models which begins to fill this gap. Though not estimated, probability distributions of activity centres can be derived from frequentist SCR models using the capture histories and the estimated SCR parameters (Efford, 2023), allowing conditional (on activity centre) fit statistics to be computed and used in a similar manner to posterior predictive checks in the Bayesian context. However, as the derived location of an activity centre is defined by a probability distribution rather a fixed set of coordinates, uncertainty in activity centre location has to be accounted for.

Here, we describe a framework for incorporating the uncertainty of derived activity centre locations within Monte Carlo simulations from fitted frequentist SCR models. We demonstrate how goodness-of-fit tests can be implemented in frequentist SCR models through these Monte Carlo simulations, using the statistics proposed by Royle et al. (2014b) for Bayesian SCR models, which have to be calculated from information on individual activity centres. We investigate the performance of these goodness-of-fit tests in response to various types of unmodelled heterogeneity, particularly if they offer a means of diagnosing causes of poor fit. In doing so, we aim to close the gap between Bayesian and frequentist SCR models and improve the consistency of GoF assessments across both inference paradigms, thereby enabling practitioners of either camp to draw robust inferences from their SCR models.

2 | METHODS

2.1 | Overview of SCR

Spatial capture-recapture (SCR) is a hierarchical model, comprised of a latent state model for the density of animals and a detection model for the observed encounters conditional on animal density. SCR assumes that each animal i has an activity centre s_i and detection probability of animal i declines with distance from s_i. The captures y_{ijk} represent the encounter of animal i at trap j on occasion k and are assumed to be random variables governed by detection probability p_{ijk}:

$$y_{ijk} \sim \text{Bernoulli}(p_{ijk}).$$

These "traps" may be proximity detectors such as camera traps or acoustic monitors that passively record an observation of an individual at the detector or live traps that physically capture individuals.

The basic SCR model has a detection function that depends on the distance between the animal's activity centre s_i and the trap or location of captures q_j. A common approach is to assume that p_{ijk} follows a half-normal distribution such that p_{ijk} decreases with distance from trap location q_j to activity centre s_i, $\| q_j - s_i \|$ as follows:

$$p_{ijk} = g_{0,ijk} \cdot \exp\left\{ \frac{\| q_j - s_i \|^2}{2\sigma^2} \right\},$$

where $g_{0,ijk}$ represents the detection probability of an individual with an activity centre at s_i and σ is the scale parameter for the half-normal distribution which describes the rate at which p_{ijk} decreases over distance to s_i. The basic SCR model assumes that g_0 and σ are constant, but this assumption may be relaxed by modelling g_0 and σ as a function of relevant covariates:

$$\text{logit}(g_{0,ijk}) = \boldsymbol{X}\boldsymbol{\beta},$$
$$\log(\sigma_{ijk}) = \boldsymbol{X}\boldsymbol{\gamma},$$

where $\boldsymbol{X}$ is a design matrix of the (scaled) predictors, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are respectively vectors of coefficients of the linear predictors of g_0 and σ to be estimated. These covariates may be used to account for individual heterogeneity such as sex differences or environmental heterogeneity (Efford, 2023; Efford & Mowat, 2014). g_0 is modelled with a logit-link function like a binomial generalised linear model to ensure that $\hat{g}_0 = [0, 1]$ and σ—which is a distance—is modelled with a log-link function like a log-linear regression model to ensure that $\hat{\sigma} \geq 0$.

In the simplest case, density is assumed to be uniform, such that the marginal distribution of n_{xy}, the number of animals in any location with location xy in the state space $\mathcal{S}$, follows a homogenous Poisson point process with intensity D:

$$n_{xy} \sim \text{Poisson}(D).$$

Intensity may also be modelled as a function of covariates to account for systematic variation in density:

$$n_{xy} \sim \text{Poisson}(D_{xy}),$$
$$\log(D_{xy}) = \boldsymbol{X}\boldsymbol{\alpha},$$

where D_{xy} is the intensity of activity centres at location S_{xy} modelled by an inhomogenous point process. $\boldsymbol{\alpha}$ is a vector of parameters of the linear predictors of D_{xy} and the relationship is modelled by a log-link function such that $\hat{D}_{xy} \geq 0$.

SCR data may also be modelled as a function of encounter rate λ_{ijk},

$$y_{ijk} \sim \text{Poisson}(\lambda_{ijk}),$$

where the captures y_{ijk} are recorded as independent counts instead of binomial encounters as demonstrated above. We also note that apart from the half-normal detection function used in this section, other detection functions may also be specified for an SCR model. While the count model and alternative detection functions are not considered in this paper, our GoF testing approach should generalise across these variations in model specification in a straightforward way.

2.2 | SCR summary statistics for testing goodness-of-fit

Spatial capture-recapture capture histories $\boldsymbol{y}$ can be represented as three-dimensional arrays, that is the encounter or the lack of an encounter for an individual i (row) in trap j (column) on occasion k (slice). The summary statistics proposed by Royle et al. (2014b) aggregate capture histories in one or more of these dimensions (refer to Table 1 for details on how these dimensions are aggregated). These aggregations result in counts grouped by cells, which lend themselves naturally to testing GoF using well-established approaches such as the Freeman-Tukey test.

The Freeman-Tukey GoF tests from these summary statistics may be characterised as follows:

- *Individual encounters* (*FT-individuals*): This summary statistic aggregates capture histories by individuals. We hypothesise that this data summary should be informative about unmodelled individual heterogeneity, that is it will detect greater variance in each individual's total encounters when these assumptions are violated.
- *Trap encounters* (*FT-traps*): This summary statistic aggregates capture histories by traps. We hypothesise that this data summary should be informative about unmodelled heterogeneity in detectability across traps and/or space.
- *Individual by trap encounters* (*FT-ind-trap*): This summary statistic aggregates capture histories by individuals and traps, that is it sums across sampling occasions. We hypothesise that this data summary should be informative about any non-temporal form of unmodelled heterogeneity in the detection parameters, such as g_0 and σ for the half-normal detection function, including unmodelled heterogeneity that may be detected by FT-individuals and FT-traps.

While not described here, it is also possible to aggregate capture histories by occasions or in conjuction with another dimension to investigate unmodelled temporal heterogeneity in detection.

2.3 | Deriving activity centres in frequentist SCR models

The GoF tests described above require expected encounters $E(y_{ijk})$ of individuals that have been observed, which in turn have to be calculated as a function of distances to their activity centres. Thus, an estimate of each animal's activity centre is required. Bayesian implementations of SCR models explicitly estimate the activity centres as an additional parameter in the model; obtaining a posterior of the expected counts and hence fit statistics is therefore a straightforward task when testing GoF using posterior predictive checks. In contrast, frequentist implementations of SCR models do not immediately provide estimates of the activity centres. A post-hoc approach to derive the activity centres from frequentist SCR models is thus needed to calculate p_{ijk}, the expected probabilities and thereby the expected encounters. This is accomplished by first calculating the conditional probability $\pi(s_i)_{xy}$ that an individual's activity centre lies in any location xy in the state space given the SCR estimates $\hat{\theta}$ and the individual's capture history $\boldsymbol{y}_i$:

$$\boldsymbol{\pi}(\boldsymbol{s}_i) = P\left(\boldsymbol{s}_i | \hat{\theta}, \boldsymbol{y}_i\right).$$

In doing so, we obtain a probability distribution of the latent activity centres of animals in the state space using the maximum likelihood estimates of the SCR model. The R package *secr* (Efford, 2023) provides such probabilistic predictions of an activity centre location within a state space.

2.4 | Propagating uncertainty of $\boldsymbol{\pi}(\boldsymbol{s})$ and $\hat{\theta}$ via Monte Carlo simulations

The expected encounter rates from a fitted model are calculated based on the traps' distances to the individuals' activity centres $\boldsymbol{s}$. As $\boldsymbol{\pi}(\boldsymbol{s})$ is a distribution, the expected encounter rates should account for the uncertainty of the activity centre locations. This requires that we repeatedly sample from $\boldsymbol{\pi}(\boldsymbol{s})$ using Monte Carlo simulations. In this process, we randomly sample a sufficiently large number of

TABLE 1 Summary statistics for testing goodness-of-fit proposed by Royle et al. (2014b), their expected values from the capture histories and the model predictions of detection probability p_{ijk} and the corresponding fit statistics for the Freeman-Tukey test. The summed quantities refer to the cell values after the capture history has been aggregated by the respective dimension(s). The expected encounter rates refer to the number of captures expected in each cell from the SCR model estimates.

	Individual encounters	Trap encounters	Individual by trap encounters
Summed quantities	$y_{i..} = \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk}$	$y_{.j.} = \sum_{i=1}^{n} \sum_{k=1}^{K} y_{ijk}$	$y_{ij.} = \sum_{k=1}^{K} y_{ijk}$
Expected encounter rates	$E(y_{i..}) = \sum_{j=1}^{J} \sum_{k=1}^{K} p_{ijk}$	$E(y_{.j.}) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ijk}$	$E(y_{ij.}) = \sum_{k=1}^{K} p_{ijk}$
Freeman-Tukey test	$\sum_{i=1}^{n} \left(\sqrt{y_{i..}} - \sqrt{E(y_{i..})}\right)^2$	$\sum_{j=1}^{J} \left(\sqrt{y_{.j.}} - \sqrt{E(y_{.j.})}\right)^2$	$\sum_{i=1}^{n} \sum_{j=1}^{J} \left(\sqrt{y_{ij.}} - \sqrt{E(y_{ij.})}\right)^2$
Test name	FT-individuals	FT-traps	FT-ind-trap

activity centres, such that the distribution of the realised activity centres, $\boldsymbol{s}^*$ closely resembles the distribution of $\boldsymbol{\pi}(\boldsymbol{s})$. Using each s^*, we can calculate the distances to the traps which, in combination with the SCR parameters $\hat{\boldsymbol{\theta}}$, gives us the detection probability p_{ijk} and hence expected encounter rates $E(y_{ijk})$.

The distribution of any $\boldsymbol{\pi}(\boldsymbol{s}_i)$ depends on the capture locations of individual i and the values of $\hat{\boldsymbol{\theta}}$. While the capture locations are fixed quantities, $\hat{\boldsymbol{\theta}}$ are estimated and their uncertainty have to be accounted for, which we also achieve using Monte Carlo sampling methods. The joint distribution of the parameters on the link scale can be described by a multivariate normal distribution,

$$\begin{pmatrix} \log(D) \\ \text{logit}(g_0) \\ \log(\sigma) \end{pmatrix} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Here, $\boldsymbol{\mu}$ refers to the maximum likelihood estimates of g_0 on the logit-scale and density and σ on the log-scale, while $\Longrightarrow$ refers to the variance–covariance matrix of the parameters. For each random sample of density, g_0 and σ from the multivariate normal distribution of the fitted SCR model, we derive a distribution of $\boldsymbol{\pi}(\boldsymbol{s}_i)$ for each individual (Figure 1, green to purple boxes). For each $\boldsymbol{\pi}(\boldsymbol{s}_i)$, we obtain one realisation of an individual's activity centre and its associated distances to the traps and calculate expected encounter rates. This results in a distribution of expected capture histories conditional on the derived activity centres, which we can aggregate to obtain the conditional distributions of expected individual by trap encounters, individual encounters and trap enounters, and thereafter, the conditional distributions of FT-ind-trap, FT-individuals and FT-traps for the actual observations.

2.5 | Goodness-of-fit testing with Monte Carlo simulations

When testing the goodness-of-fit of a model, we usually test the null hypothesis that the data are statistically indistinguishable from a typical realisation from the model (Waller et al., 2003). Such tests can be conducted by calculating a fit statistic for the data and comparing the statistic against a reference distribution, to determine if the fit statistic was more extreme than expected under the null hypothesis. With ecological models, the data collected are often sparse; hence, the assumption that the fit statistic approaches an asymptotic reference distribution when the null hypothesis is true may not be valid (Kéry & Royle, 2016). Instead, an empirical sampling distribution of the fit statistic would be needed, which may be obtained by Monte Carlo methods such as parametric or non-parametric bootstrapping, where random samples are drawn from the data or simulated from the model to calculate fit statistics that are expected by the model (Manly, 2007; Waller et al., 2003).

Within the framework of our Freeman–Tukey GoF tests, the empirical distribution of the fit statistic has to be obtained from capture histories simulated from the fitted SCR model. This process is easily accommodated in our Monte Carlo sampling procedure for propagating uncertainty in $\boldsymbol{\pi}(\boldsymbol{s})$ and $\hat{\boldsymbol{\theta}}$. With each Monte Carlo sample of $\boldsymbol{s}^*$, we can simulate a capture history Y_{sim}, which we aggregate by one or more dimensions to obtain the summary statistics needed for FT-ind-trap, FT-individuals and FT-traps. We then calculate the differences between simulated summary statistics and the corresponding expected summary statistics, thereby obtaining a Freeman-Tukey statistic for the simulated data (Figure 1, blue boxes). This is equivalent to the empirical distribution of the Freeman-Tukey statistic for the fitted model, conditional on the locations of $\boldsymbol{s}^*$.

In typical GoF tests for frequentist models, a fixed fit statistic is obtained for the observed data, and its percentile within the theoretical or empirical distribution of the fit statistic is used to calculate the p-value and determine if the model had poor fit. However, the GoF testing framework for SCR here results in a distribution of values for the fit statistics of both the actual and simulated data. To evaluate model fit, we emulate the approach of Bayesian posterior predictive checks, which measures goodness-of-fit by comparing the posterior distribution of fit statistics for the observed data against the posterior distribution of fit statistics for the simulated data. Likewise, here we compare the relative positions of the Monte Carlo distributions of fit statistics for the observed and simulated data (Figure 1, blue to black boxes). Since the Freeman-Tukey statistic measures the differences between the observed and expected values, a larger value of the Freeman-Tukey statistic indicates a larger difference and hence poorer fit. Where the fit statistic for the actual obervations is much larger than the fit statistics for the simulated data, we would conclude that the model fits the actual observations more poorly than we expect if the model was correct. Thus, traditional GoF tests for counts based on the χ^2 distribution, which includes the Freeman-Tukey test, are right-tailed tests, where the fit statistic for the actual observations is considered too extreme if it lies too far to the right-tail of the corresponding χ^2 or empirical distribution. This line of reasoning carries over to Freeman-Tukey tests implemented via Monte Carlo resampling, where goodness-of-fit is evaluated by $1 - P(\boldsymbol{T}_{\text{obs}} > \boldsymbol{T}_{\text{sim}})$, where $\boldsymbol{T}_{\text{obs}}$ and $\boldsymbol{T}_{\text{sim}}$ refer to the fit statistics for the actual and simulated observations respectively. When using the tests described here, a p-value smaller than α (e.g. $\alpha = 0.05$) would indicate that the model has provided a poor fit to the data. The entire GoF-testing procedure is summarised in Figure 1.

2.6 | Simulation design

With the Monte Carlo sampling procedure and corresponding GoF measures defined, we conducted a simulation study to quantify the diagnostic power of our proposed tests to various SCR assumption violations in a range of ecologically realistic scenarios. Unlike previous efforts which test GoF in models with misspecified detection functions (Dey et al., 2022), here we focus specifically on model violations where lack-of-fit has been shown to be a source of bias (Dey et al., 2023; Efford & Mowat, 2014; Gardner et al., 2010; Royle

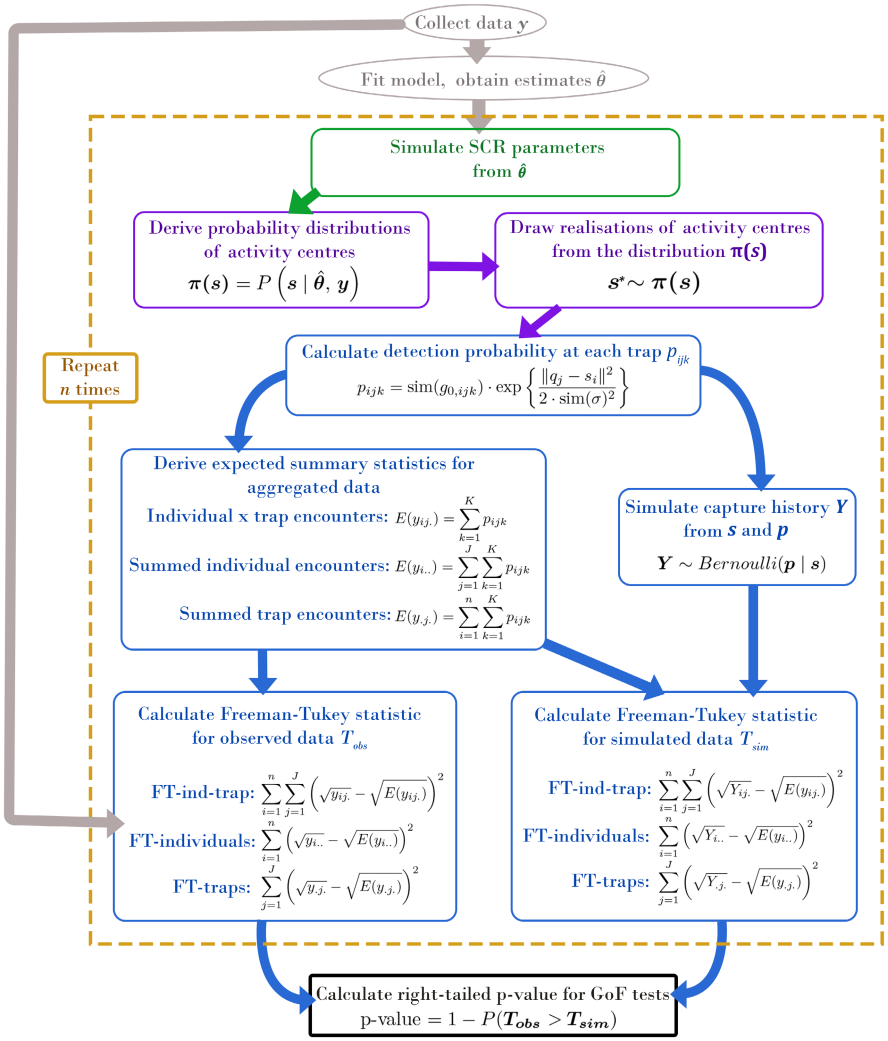

FIGURE 1 Summary of goodness-of-fit testing with Monte Carlo simulations in SCR models incorporating uncertainty of derived activity centres. Solid lines contain single instances of individual processes, while the dotted line encompasses the steps that are repeated *n* times, where *n* is the number of Monte Carlo samples used. Each colour represents a different stage in an SCR analysis. Grey ellipses encircle processes that take place outside of the GoF framework. The green box represents the step where SCR parameter values are sampled from a multivariate normal distribution. The purple boxes cover the steps for deriving the probability distributions of the activity centres. The blue boxes describe the steps for calculating the fit statistics for observed and simulated values, conditioned on the activity centre locations. The *p*-value from the fit statistics is calculated in the black box.

et al., 2013; Sutherland et al., 2015; Tobler & Powell, 2013). In our simulations, we aimed to understand how the GoF tests would respond to correctly fitted models and, more importantly, to models which ignored simulated sources of heterogeneity in detection and, separately, density. We generated data using different combinations of density and detection models (Figure 2). For the detection model, the following data-generating processes were considered (see Table 2 for parameter values):

- *Uniform detection parameters*: g_0 and σ were constant for all individuals in the population (Scenario 1 and 6) (Figure 2: detection function (1)).
- *Random continuous variation in detection parameters*: g_0 and σ varied randomly across individuals, where g_0 was negatively correlated with σ such that each individual would have similar detectability, but distributed differently from their activity centre (Scenario 2) (Figure 2: detection function (2)). We drew values of g_0 and σ on the respective link scales for each individual from a bivariate normal distribution and transformed to real-scale values for simulating capture histories. In our simulations, 25% of the population were expected to be up to 0.24 times as likely as the average animal to be encountered at their activity centres, but with home ranges that were up to 1.8 times larger. Another 25% were expected to be at least 1.33 times as likely as the average animal to be encountered at their activity centre, and with home ranges that were at least 0.55 times smaller. Fitting a basic model here would result in unmodelled individual heterogeneity in g_0 and σ.
- *Two-class variation in individual* g_0: Individuals were split into two classes with different g_0, while σ remained constant (Figure 2: detection function (3)). Here, the intent was to model a system where individual differences within a species resulted in differences in overall activity while home range sizes remained constant. This discrete variation in g_0 was applied to two different scenarios: one where individuals were randomly assigned to classes as might be observed with sex (Scenario 3), and another where individual classes were assigned based on their true activity centre location (Scenario 4), which might resemble behaviour arising from differences in resource availability. In our simulations, animals with higher detection probability (activity centres in the north where variation is spatially structured) were three times as likely to be captured at their activity centres than animals with lower detection probability (activity centres in the south where variation is spatially

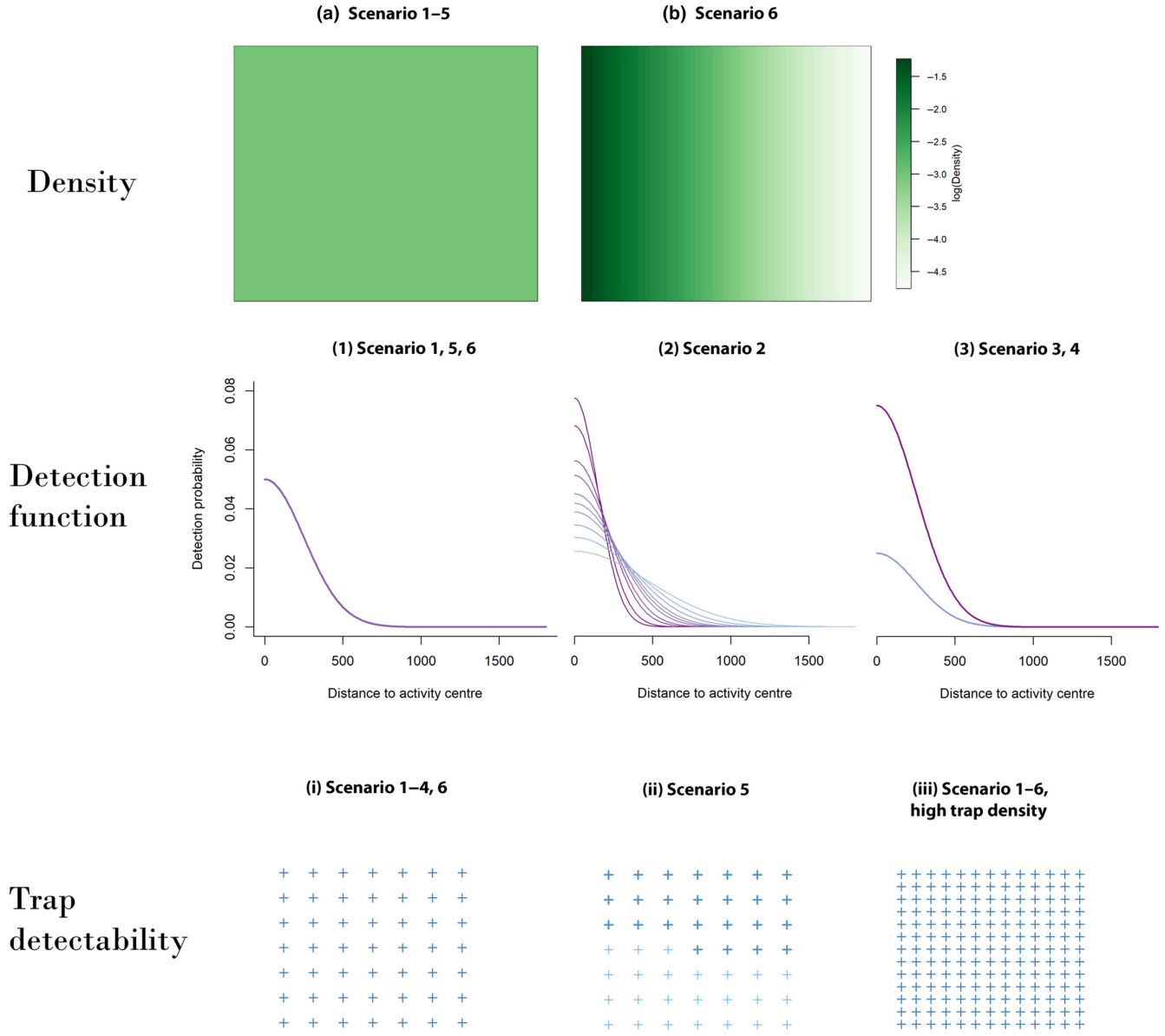

FIGURE 2 Schematic diagram of the density models (top row), detection models (middle row) and traps (bottom row) used to simulate SCR data. The intensity of the shades in the density models represents the gradient in density across the state space. The lines of the detection models represent the detection probability and its rate of decay with distance to the activity centre. Detection model (1) uses a single value of g_0 and σ for the entire population. The detection model (2) randomly samples a value of g_0 and σ for each individual in the population. The detection model (3) uses two paired values of g_0 and σ, which are either randomly allocated to individuals or by their activity centre location. The crosses in the trap array represent the number and positions of traps in the state space. The colours represent the trap efficiency, where a darker colour represents a more efficient trap and hence higher capture probability. An exception is made for traps (iii), where the spatial variation in trap efficiency at high trap density for Scenario 5 is not shown.

structured). Fitting a basic model to both scenarios would ignore individual heterogeneity in g_0. We also conducted more simulations based on non-spatial two-class variation in g_0, where we varied the differences in g_0 to understand how the power of goodness-of-fit tests may vary with the level of heterogeneity (Appendix 1: Figure S5).

- *Two-class variation in g_0 at traps*: Traps were split into two classes with different g_0 based on which half of the state-space they fell within (Scenario 5) (Figure 2: trap detectability (ii)). This might be observed in camera trap studies, where differences in habitat types such as open plains versus closed forests might result in different windows of capture and hence detection probabilities. In our simulations, traps in the north were three times as likely to capture an animal than traps in the south. Fitting a basic model here would ignore trap heterogeneity in g_0.

We considered two state processes for the density model (see Table 2 for parameter values):

TABLE 2 Parameter values of the data-generating processes used to simulate spatial capture-recapture data. cov_j refers to the covariate value specific to trap j. cov_{xy} refers to the covariate value specific to location S_{xy}.

Scenario	Density model	Detection models
Uniform density, constant detection	$N_{xy} \sim \text{Pois}(5)$ $N_{xy} \sim \text{Pois}(2.5)$	$g_0 = 0.05$ $\sigma = 250$
Uniform density[a], random variation in g_0 and σ	$N_{xy} \sim \text{Pois}(5)$ $N_{xy} \sim \text{Pois}(2.5)$	$\begin{pmatrix} \text{logit}(g_0) \\ \log(\sigma) \end{pmatrix} \sim N\left(\begin{pmatrix} \text{logit}(0.048) \\ \log(245) \end{pmatrix}, \begin{pmatrix} 0.15 & -0.25 \\ -0.25 & 0.15 \end{pmatrix}\right)$
Uniform density, non-spatial, discrete variation in g_0	$N_{xy} \sim \text{Pois}(5)$ $N_{xy} \sim \text{Pois}(2.5)$	$g_0 = \{0.025, 0.075\}$ $\sigma = 250$
Uniform density, spatial, discrete variation in g_0	$N_{xy} \sim \text{Pois}(5)$ $N_{xy} \sim \text{Pois}(2.5)$	$g_0 = \{0.025, 0.075\}$ $\sigma = 250$
Uniform density, spatial variation in trap efficiency	$N_{xy} \sim \text{Pois}(5)$ $N_{xy} \sim \text{Pois}(2.5)$	$\text{logit}(g_0) = \text{logit}(0.025) + 1.15\text{cov}_j$ $\sigma = 250$
Spatially varying density, constant detection	$N_{xy} \sim \text{Pois}(\Lambda_{xy})$ $\log(\Lambda_{xy}) = \log(5) + 0.7\text{cov}_{xy}$ $N_{xy} \sim \text{Pois}(\Lambda_{xy})$ $\log(\Lambda_{xy}) = \log(2.5) + 0.7\text{cov}_{xy}$	$g_0 = 0.05$ $\sigma = 250$

[a]Mean g_0 and σ are smaller than corresponding values in other scenarios as the bivariate distribution of g_0 and σ on the real scale is skewed by larger values.

- *Uniform density*: Activity centres were simulated from a homogenous Poisson point process model, where the underlying intensity was constant across the state space.
- *Spatially varying density*: Activity centres were simulated from an inhomogenous Poisson point process model, where the underlying intensity decreased log-linearly across the x-axis of the state space (Scenario 6). Fitting a basic SCR model to capture histories generated from these populations would result in unmodelled heterogeneity in density. The covariate for density was defined such that the expected number of animals would be similar across the uniform density scenarios and the spatially varying density scenarios. In our simulations, the westernmost region of the state-space contained 5 times as many animals as the middle region, and 25 times as many as the easternmost region.

The capture histories for populations with uniform densities were generated from all three detection models, while the capture histories for populations with spatially varying densities were generated only with uniform detection parameters (Table 2).

Our simulations focused on designs using trap layouts of 7×7 traps with 2σ spacing, where $\sigma = 250$ units, for populations with 0.05 individuals per 10,000 square units, corresponding to moderate data quality. However, we also used simulations employing trap designs of 13×13 traps with σ spacing which would result in high data quality, and others with a population density of 0.025 individuals per 10,000 square units, corresponding to low data quality (Table 2). Here, we use data quality to describe the richness of the data and ability to obtain good estimates from an SCR model. Data quality increases with both the number of individuals captured, and the average number of captures per individual. Therefore, increasing individual density or trap density both lead to an increase in data quality. We then defined a state space with a buffer of $3\sigma^*$ between the edges of the state space and the traps for all simulations, σ^* being the 99.7 percentile value of σ from the scenario with random variation in detection parameters (Table 2). This state space was discretised into pixels with dimensions of 125×125 units such that there were at least two pixels between each trap, resulting in a square grid of 53×53 points representing the centre of each pixel. The moderate and high data quality scenarios thus had an expected abundance of 219.5 individuals, while the low data quality scenarios had an expected abundance of 109.7 individuals. More details on the capture histories may be found in Appendix 1 (Figures S1 and S2).

2.7 | Model fitting and goodness-of-fit testing

Simulated data were analysed using a standard SCR model assuming uniform density and detection parameters, which was not the data-generating model whenever the data contained any form of heterogeneity. The correct data-generating model was also fitted to data generated with inhomogenous density and two-class variation in g_0 across individuals and across traps. We did not fit a model with random effects on g_0 and σ as this has yet to be easily implemented in maximum likelihood for SCR. Models were fitted in R v. 4.2.2 (R Core Team, 2023) using the package *secr* (Efford, 2023). The goodness-of-fit for all fitted models were tested using FT-ind-trap, FT-individuals and FT-traps. We simulated 300 datasets under each scenario, within which goodness-of-fit testing was conducted with 1000 Monte Carlo samples. We primarily set the threshold of α at 0.05, that is $1 - P(\boldsymbol{T}_{\text{actual}} > \boldsymbol{T}_{\text{sim}}) < 0.05$, but also across a range at $0.01 \leq \alpha \leq 0.20$. We used the values of α to calculate power and false

positive rates. We measured power as the proportion of iterations with a p-value smaller than α when the wrong model was fitted, and false positive rates as the proportion of iterations with a p-value smaller than α when the correct model was fitted. While we conducted these across simulations of all data qualities, we present only the results for the moderate data quality as they are most likely to mirror data quality from actual studies. The results for low- and high-quality data may be found in Appendix 1 (Figures S3 and S4). We also briefly demonstrate how the power of the GoF tests can change with varying levels in heterogeneity in Appendix 1 (Figure S3).

2.8 | Case study using Fort Drum data

We applied our GoF tests to an SCR capture history of bears from New York, USA. The data was originally published in Gardner et al. (2010). Here, we fit a uniform SCR model, and a SCR model accounting for sex differences in g_0 and σ using maximum likelihood methods via the *secr* R package (Efford, 2023) and in Markov Chain Monte Carlo (MCMC) using JAGS, with R as a console (Kellner, 2024; Plummer, 2003). We subsequently tested the fit of these models using the GoF tests described in this paper. The analysis is described in full in Appendix 2, and the R code for running the goodness-of-fit tests in this case study has been compiled as a package (Choo et al., 2024a) and versioned on Zenodo (Choo et al., 2024b).

3 | RESULTS

Each goodness-of-fit test was sensitive to different sources of unmodelled heterogeneity in the data. All the GoF tests also had low false positive rates when the correct model was fitted to a capture history (Figure 3 column 1). An exception was the scenario with random variation in g_0 and σ as the correct model was not fitted to the capture histories.

We consider first the power of each of the tests to assumption violations, defined as the probability that the test will produce a statistically significant result when the wrong model is fitted. FT-ind-trap, an all-round test for unmodelled non-temporal heterogeneity in detection parameters (Figure 3) was most powerful when both g_0 and σ varied randomly across individuals (power = 0.88), but much lower in the other scenarios where unmodelled heterogeneity occurred only in g_0 (power = 0.34–0.35). FT-individuals, a test of extra variance in summed individual encounters, had low power across all scenarios (power = 0–0.25). FT-traps, a test for extra variance across traps and/or space, had moderate power when g_0 varied across space, either by individuals (power = 0.53) or by traps (power = 0.56) but low power in the remaining scenarios (power = 0.05–0.20).

Where g_0 varied spatially, be it across individuals or traps, FT-traps was the most powerful GoF test, followed by FT-ind-trap then FT-individuals (Figure 3 columns 4 and 5). With individual spatial variation in g_0, 85% of simulations with statistically significant test results contained significant results from FT-traps. In simulations where g_0 varied by traps, this proportion increased to 91%. When heterogeneity in g_0 across individuals was present without spatial constraints, FT-ind-trap was the most powerful test, followed closely by FT-individuals and finally FT-traps. FT-ind-trap was the most powerful test when random continuous variation in g_0 and σ occurred, followed by FT-traps and finally FT-individuals which had practially no power to detect the lack of fit in the model. None of the tests demonstrated any power to detect lack of fit caused by unmodelled heterogeneity in density when the detection model was correctly specified.

Within our tested scenarios, FT-individuals was least likely to be the sole test with a statistically significant result when unmodelled heterogeneity was present (Figure 3 row 4). FT-ind-trap was the most likely to be the only test with a significant result when unmodelled random variation in detectability (71% of all simulations considered) or non-spatial two-class variation in g_0 (17%) was present. FT-traps was the most likely to be the only test when unmodelled spatial variation occurred, be it at the level of individuals (20%) or of traps (28%). FT-ind-trap and FT-traps were almost twice as likely to be the only tests to simultaneously produce significant results with random variation in detectability (18%) compared to spatial variation in individual g_0 (10%), but the probability of their exclusionary co-occurrence in spatial variation in trap efficiency (14%) was broadly similar to both of these scenarios. The simultaneous production of significant results from FT-ind-trap and FT-individuals was most likely in non-spatial two-class variation in g_0 (13%) but $\leq 1\%$ in all other scenarios. Statistically significant results from all tests were most likely when applied to data with spatially varying individual g_0 (13%) followed closely by spatial variation in trap efficiency (10%), and far less likely when applied to non-spatial variation in g_0 (2%).

When α of the goodness-of-fit tests was raised, that is the specified threshold for false positive rates, the power of the GoF tests to detect lack-of-fit increased at a disproportionately higher rate than their false positive rates across all simulation scenarios (Figure 3). The increase in false positive rates with α was non-linear, and the realised false positive rates in the simulations were smaller than the corresponding α. The relative ranks in power of the tests within each scenario was also largely preserved as α increased to 0.20.

3.1 | Case study using Fort Drum bear data

Our goodness-of-fit tests for maximum likelihood SCR models yielded similar results to Bayesian GoF tests. When a uniform model and a model accounting for sex differences were fitted to the data, both estimation approaches resulted in statistically significant test results for FT-ind-trap and FT-traps (Table 3) for a threshold of $\alpha = 0.05$. FT-individuals was not always statistically significant for all models across both estimation approaches, but the differences in p-values between approaches for either model were always ≤ 0.02 (Table 3).

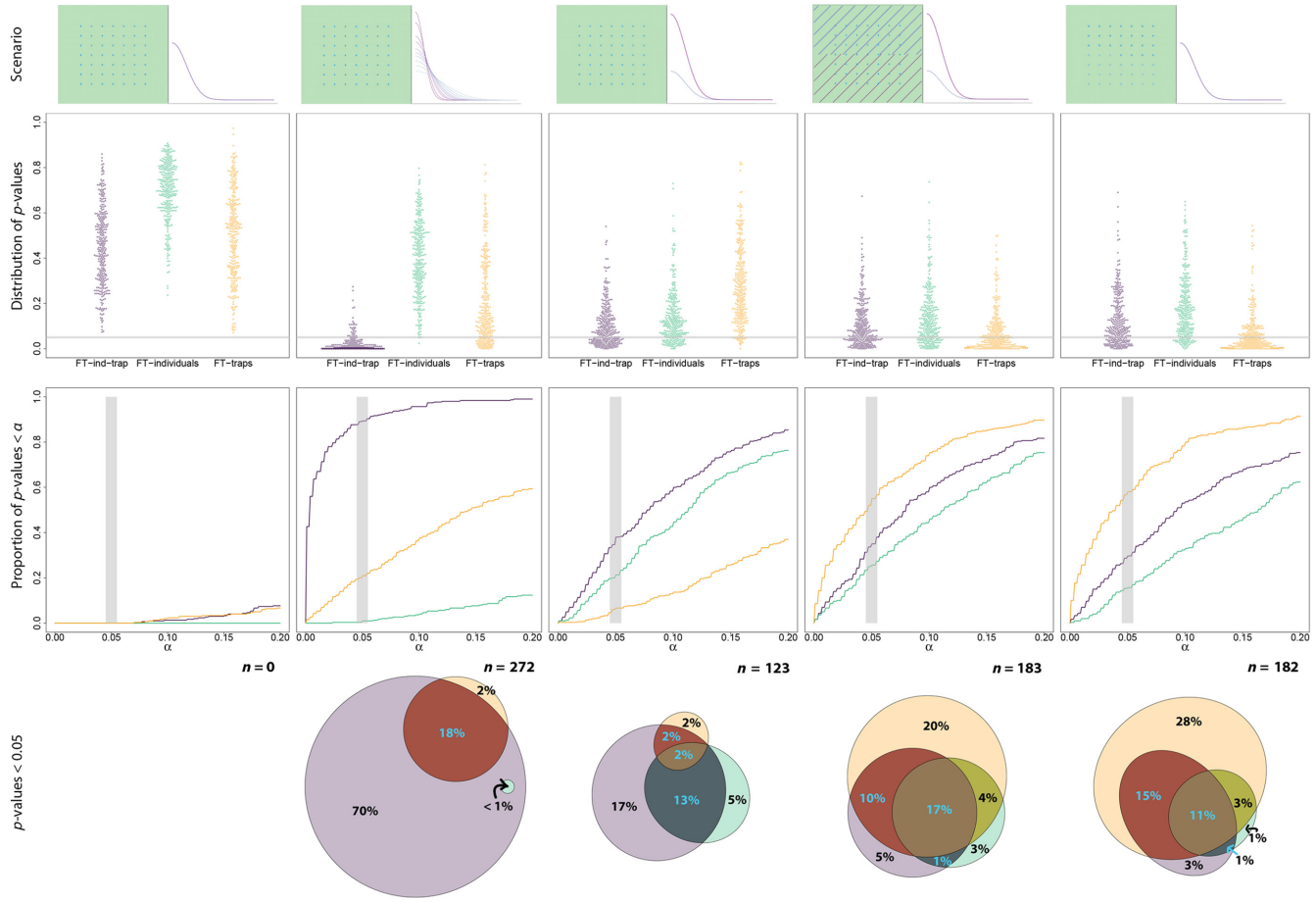

FIGURE 3 Responses of FT-ind-trap, FT-individuals and FT-traps to unmodelled heterogeneity in the detection sub-model. The first column represents the results when the null model is correctly fitted. Top row: The data generating process behind the capture histories. Scenarios from left to right: (1) Homogenous density and detectability, (2) randomly varying detectability, (3) spatially random two-class variation in g_0, (4) spatially varying two-class variation in g_0 and (5) spatially varying trap efficiency. Second row: Distribution of *p*-values from the GoF tests when the corresponding capture histories are analysed using a null SCR model. The horizontal grey band corresponds to $\alpha = 0.05 \pm 0.005$. Third row: Variation in the proportion of statistically significant *p*-values as α increases. The vertical grey band corresponds to $\alpha = 0.05 \pm 0.005$. Bottom row: Overlap in statisically significant test results across the GoF tests at $\alpha = 0.05$. The area of each Euler diagram is proportional to the number of simulations with at least one statistically significant test result, and the labels represent the percentage of simulations in each category.

TABLE 3 Results of goodness-of-fit tests for SCR models fitted to the Fort Drum data. The tests were applied via the Monte Carlo resampling for maximum likelihood models, and from posterior predictive checks for Bayesian models. SCR.0 refers to the null model, and SCR.h2 refers to the model that uses sex as an individual covariate.

Model	Test	Monte Carlo resampling	Bayesian
SCR.0	FT-ind-trap	0	0
	FT-individuals	0.06	0.05
	FT-traps	0	0
SCR.h2	FT-ind-trap	0.01	0
	FT-individuals	0.13	0.11
	FT-traps	0	0

4 | DISCUSSION

Testing goodness-of-fit is a standard procedure for statistical analyses and has become a cornerstone of ecological models, given they feature heavily in wildlife management and conservation decision making (Buckland et al., 2004; Choquet et al., 2009; Kéry & Royle, 2016). Yet, it has been largely neglected in applications of spatial capture-recapture (Tourani, 2022), possibly due to a lack of available tools. Here, we propose a GoF approach based on deriving probabilistic locations of the realised activity centres from fitted SCR models for frequentist SCR models. We use simulations to assess the power and false positive rates of the fit statistics proposed by Royle et al. (2014b). Unlike previous efforts, our simulations indicate that these GoF tests are indeed capable of detecting lack-of-fit, here arising from unmodelled heterogeneity

in the detection parameters g_0 and σ. Our investigations with empirical data also suggest that GoF tests implemented through our Monte Carlo resampling framework may perform similarly to tests implemented with Bayesian methods.

Prior to our study, the diagnostic abilities of available Bayesian fit statistics were not well understood (Royle et al., 2014b). Through our simulations, we found that the power of the GoF tests varied with the type of unmodelled heterogeneity present in the data. Inferring the cause of poor fit is therefore contingent on the combination of tests that were statistically significant. FT-ind-trap, derived from the total encounters of each individual at each trap, appeared to respond to unmodelled heterogeneity in both g_0 and σ and was more powerful when both parameters are affected. In contrast, FT-traps was the most powerful test when the unmodelled heterogeneity in detection probability was spatially structured. This was within expectations as FT-traps was calculated by aggregating encounters at each trap, hence amplifying spatial information in the data. FT-individuals, obtained from the total encounters of each individual, tended to have the lowest power in our simulations. This may be a consequence of SCR essentially being a model for individual heterogeneity in capture rates (Borchers & Efford, 2008), hence stronger effects on individual variance may need to be present to obtain statistically significant results with FT-individuals. The similarities in the performance of our implementation of GoF tests to the examples provided for the Fort Drum bear data in Gardner et al. (2010) suggest that both approaches result in tests with similar properties. Thus, these inferences of the GoF test properties may be transferable across both frequentist and Bayesian paradigms. Given that frequentist SCR models tend to be less computationally intensive to fit than with Bayesian methods, our Monte Carlo resampling framework may also enable other researchers to feasibly conduct larger-scale simulations and investigate more aspects of GoF testing in SCR.

Used jointly, the suite of GoF tests we studied may provide diagnostic insights into certain forms of lack-of-fit. Obtaining statistically significant results from FT-traps alone or from all three tests appears to provide a strong indication that unmodelled spatial heterogeneity is present in the data (Figure 4). The absence of a significant result from FT-traps when other tests behave otherwise may conversely indicate that any unmodelled heterogeneity present is unlikely to have a spatial component (Figure 4). The power of all tests may also vary with the level of heterogeneity present (Appendix 1: Figure S5). In our brief example, all GoF tests had higher power at greater levels of heterogeneity (Appendix 1: Figure S5), while the trends in the overlap of statistically significant results remained consistent with our main results. These results suggest that the tests may be more effective at detecting larger violations which ecologists may be more concerned about, while still allowing the type of violation to be identifiable. Furthermore, the vanishingly low false positive rates when the correct model is fitted (Figure 3) indicate that the multiple comparisons problem typically associated with using multiple hypothesis tests would be of negligible concern here. Employing a full complement of tests may therefore be a feasible means of uncovering different sources of lack-of-fit and inferring their origin, though to do so with greater precision we would likely require a larger suite of tests.

Dey et al. (2022) suggested that the fit statistics we use here are insensitive to misspecification of the detection function. This apparently contradicts our findings as the types of model misspecifications in both studies result in the detection model being misspecified. Unlike distance sampling (Buckland et al., 2004), which also models detection probability as a function of distance, SCR is often used with sparse data to which any single detection function shape is unlikely to provide a significantly better fit and is less reliant on the detection function shape to obtain unbiased estimates of density (Dey et al., 2022; Efford, 2004). An inability to detect lack-of-fit arising from a poorly specified detection function shape may therefore be of little concern to practitioners where density estimates are concerned. In contrast, sensitivity to unmodelled heterogeneity would be of greater practical importance, as these misspecifications can introduce bias in density estimates (Borchers & Efford, 2008; Efford, 2004). This is more likely to result in systematic differences between the expected and actual encounter rates of individuals at each trap, for at least a proportion of the population and/or traps. Such differences may include deviations in the expected and actual number of traps encountered. Thus, the goodness-of-fit tests assessed in our simulations were better able to detect lack-of-fit to unmodelled heterogeneity.

Testing goodness-of-fit in SCR models may appear unnecessary as several studies have demonstrated that density estimates from SCR are often robust to certain violations of its assumptions, particularly those that have garnered more attention from ecologists, such as changes in activity centre locations during surveys (Royle et al., 2016), spatiotemporal correlation in animal detections (Dupont et al., 2022; Gardner et al., 2022; Moqanaki et al., 2021), temporal heterogeneity in detectability (Sollmann, 2024) and heterogeneity from micro-habitats (Theng et al., 2022). Moreover, in the absence of alternatives, a poor-fitting SCR model may be preferable than none (Royle et al., 2014b) and the knowledge of poor fit may seem to be an unnecessary burden. However, we argue that testing model fit here is paramount as it would allow us to gauge if we can place confidence in our estimates or proceed with caution, particularly if a more complex model is needed (Choquet et al., 2005). Furthermore, beyond being tools for estimation, models are ultimately a medium for projecting our perceptions about the systems we study, and goodness-of-fit tests for SCR provide us with means of testing our understanding about the ecological processes that shape our models.

Our study is a start to the discussions around general purpose GoF tests for SCR, and, by providing code for conducting these tests, we offer a development that extends the accessibily of methods for testing model fit. Through our simulations, we quantified the performance of the GoF tests, which will enable practitioners to employ them more effectively. Yet, compared to the suites of GoF tests available for distance sampling and traditional capture-recapture methods that provide a comprehensive assessment of the fitted model (Buckland

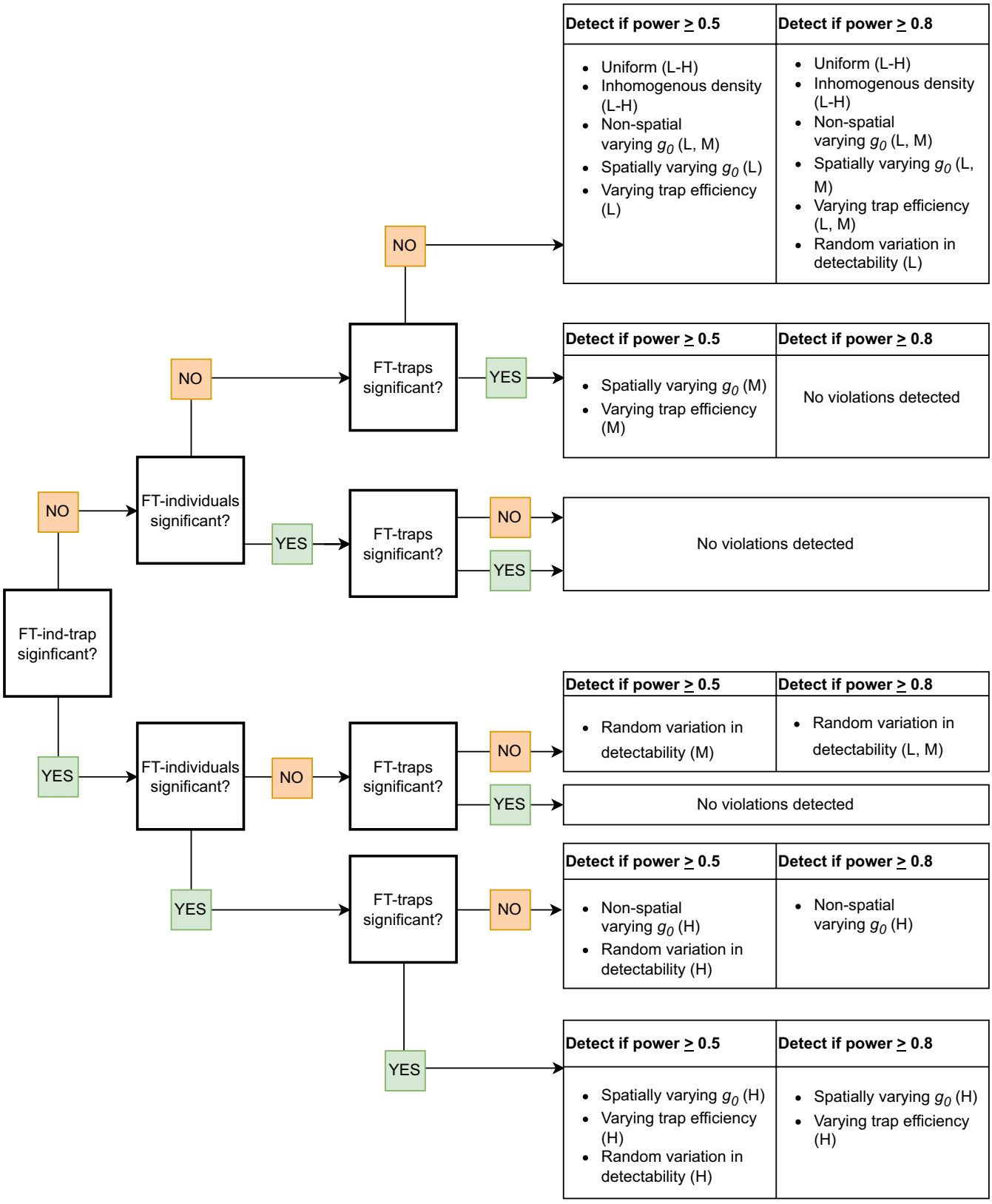

FIGURE 4 Identifying sources of lack-of-fit based on goodness-of-fit (GoF) test results. The thresholds refer to the power of the GoF test at which we arbitrarily expect a significant test result to be obtained, to visualise the combinations of test results that one might get from a (mis)specified model. Power is defined as the probability of obtaining a statistically significant test result when a model is misspecified (see Appendix 1: Table S1 for exact power of tests for each scenario). It is important to note that power is specific to each test, and is not indicative of the probability that any stated combination of tests will have statistically significant results. The letters in brackets refer to the quality of the data that was modelled: L refers to low, M refers to moderate and H refers to high data quality.

et al., 2004; Choquet et al., 2005, 2009; Gimenez et al., 2018; Thomas et al., 2010), the tools available for SCR are still limited in scope and utility. Given that SCR data consists of multiple dimensions over which lack-of-fit can arise, three binary goodness-of-fit tests with overlapping coverage are insufficient to precisely identify the range of potential source(s) of lack-of-fit with great confidence. In addition to these tests, we will need a larger collection of GoF tests which test other aspects of the SCR model with greater specificity.

AUTHOR CONTRIBUTIONS

Yan Ru Choo, Chris Sutherland and Alison Johnston conceived the ideas and simulation scenarios. Yan Ru Choo designed the simulation study and led the analyses. Yan Ru Choo led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Yan Ru Choo is funded by the Engineering and Physical Sciences Research Council.

CONFLICT OF INTEREST STATEMENT

Chris Sutherland is an associate editor with *Methods in Ecology and Evolution*.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14386.

DATA AVAILABILITY STATEMENT

No data were collected for this paper. An R package for running goodness-of-fit tests on *secr* models has been versioned on Zenodo, DOI: https://doi.org/10.5281/zenodo.11283942 (Choo, 2024).

ORCID

Yan Ru Choo https://orcid.org/0000-0002-8852-7178
Chris Sutherland https://orcid.org/0000-0003-2073-1751
Alison Johnston https://orcid.org/0000-0001-8221-013X

REFERENCES

Borchers, D. L., & Efford, M. G. (2008). Spatially explicit maximum likelihood methods for capture-recapture studies. *Biometrics*, *64*(2), 377–385. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2007.00927.x

Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., Thomas, L., Buckland, S. T., Anderson, D. R., Burnham, K. P., Laake, J. L., Borchers, D. L., & Thomas, L. (Eds.). (2004). *Advanced distance sampling: Estimating abundance of biological populations*. Oxford University Press.

Choo, Y. R., Sutherland, C. S., & Johnston, A. (2024a). scrgof: Monte Carlo resampling methods for spatial capture-recapture models fitted in maximum likelihood. https://github.com/chooyr/scrgof

Choo, Y. R., Sutherland, C. S., & Johnston, A. (2024b). scrgof: Monte Carlo resampling methods for spatial capture-recapture models fitted in maximum likelihood. https://doi.org/10.5281/zenodo.11283943

Choquet, R., Lebreton, J.-D., Gimenez, O., Reboulet, A.-M., & Pradel, R. (2009). U-CARE: Utilities for performing goodness of fit tests and manipulating CApture-REcapture data. *Ecography*, *32*(6), 1071–1074. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1600-0587.2009.05968.x

Choquet, R., Reboulet, A.-M., Lebreton, J.-D., Gimenez, O., & Pradel, R. (2005). *U-CARE 2.2 user's manual* (*Utilities-CApture-REcapture*).

Dey, S., Bischof, R., Dupont, P. P. A., & Milleret, C. (2022). Does the punishment fit the crime? Consequences and diagnosis of misspecified detection functions in Bayesian spatial capture-recapture modeling. *Ecology and Evolution*, *12*(2), e8600 https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.8600

Dey, S., Moqanaki, E., Milleret, C., Dupont, P., Tourani, M., & Bischof, R. (2023). Modelling spatially autocorrelated detection probabilities in spatial capture-recapture using random effects. *Ecological Modelling*, *479*, 110324.

Dupont, G., Linden, D. W., & Sutherland, C. (2022). Improved inferences about landscape connectivity from spatial capture-recapture by integration of a movement model. *Ecology*, *103*(10), e3544 https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.3544

Efford, M. (2004). Density estimation in live-trapping studies. *Oikos*, *106*(3), 598–610.

Efford, M. (2023). secr: Spatially explicit capture-recapture.

Efford, M. G., & Mowat, G. (2014). Compensatory heterogeneity in spatially explicit capture recapture data. *Ecology*, *95*(5), 1341–1348. https://onlinelibrary.wiley.com/doi/pdf/10.1890/13-1497.1

Gardner, B., McClintock, B. T., Converse, S. J., & Hostetter, N. J. (2022). Integrated animal movement and spatial capture-recapture models: Simulation, implementation, and inference. *Ecology*, *103*(10), e3771.

Gardner, B., Royle, J. A., Wegan, M. T., Rainbolt, R. E., & Curtis, P. D. (2010). Estimating black bear density using DNA data from hair snares. *The Journal of Wildlife Management*, *74*(2), 318–325. https://onlinelibrary.wiley.com/doi/pdf/10.2193/2009-101

Gimenez, O., Lebreton, J.-D., Choquet, R., & Pradel, R. (2018). R2ucare: An r package to perform goodness-of-fit tests for capture-recapture models. *Methods in Ecology and Evolution*, *9*(7), 1749–1754. https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13014

Kellner, K. (2024). *jagsUI: A wrapper around 'rjags' to streamline 'JAGS' analyses*. R package version 1.6.2.

Kéry, M., & Royle, J. A. (2016). Chapter 2—What are hierarchical models and how do we analyze them? In M. Kéry & J. A. Royle (Eds.), *Applied hierarchical modeling in ecology* (pp. 19–78). Academic Press.

Manly, B. F. J. (2007). Monte Carlo methods. In *Randomization, bootstrap and Monte Carlo methods in biology* (3rd ed., p. 11). Chapman and Hall/CRC.

Moqanaki, E. M., Milleret, C., Tourani, M., Dupont, P., & Bischof, R. (2021). Consequences of ignoring variable and spatially autocorrelated detection probability in spatial capture-recapture. *Landscape Ecology*, *36*(10), 2879–2895.

Morin, D. J., Boulanger, J., Bischof, R., Lee, D. C., Ngoprasert, D., Fuller, A. K., McLellan, B., Steinmetz, R., Sharma, S., Garshelis, D., Gopalaswamy, A., Nawaz, M. A., & Karanth, U. (2022). Comparison of methods for estimating density and population trends for low-density Asian bears. *Global Ecology and Conservation*, *35*, e02058.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *3rd International Workshop on Distributed Statistical Computing (DSC 2003)* (p. 124).

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Royle, J. A., Chandler, R. B., Gazenski, K. D., & Graves, T. A. (2013). Spatial capture-recapture models for jointly estimating population density and landscape connectivity. *Ecology*, *94*(2), 287–294. https://onlinelibrary.wiley.com/doi/pdf/10.1890/12-0413.1

Royle, J. A., Chandler, R. B., Sollmann, R., & Gardner, B. (2014a). Chapter 5—Fully spatial capture-recapture models. In J. A. Royle,

R. B. Chandler, R. Sollmann, & B. Gardner (Eds.), *Spatial Capture-recapture* (pp. 125–170). Academic Press.

Royle, J. A., Chandler, R. B., Sollmann, R., & Gardner, B. (2014b). Chapter 8—Model selection and assessment. In J. A. Royle, R. B. Chandler, R. Sollmann, & B. Gardner (Eds.), *Spatial capture-recapture* (pp. 219–243). Academic Press.

Royle, J. A., Fuller, A. K., & Sutherland, C. (2016). Spatial capture-recapture models allowing Markovian transience or dispersal. *Population Ecology*, *58*(1), 53–62. https://onlinelibrary.wiley.com/doi/pdf/10.1007/s10144-015-0524-z

Royle, J. A., Karanth, K. U., Gopalaswamy, A. M., & Kumar, N. S. (2009). Bayesian inference in camera trapping studies for a class of spatial capture-recapture models. *Ecology*, *90*(11), 3233–3244.

Royle, J. A., & Young, K. V. (2008). A hierarchical model for spatial capture-recapture data. *Ecology*, *89*(8), 2281–2289.

Russell, R. E., Royle, J. A., Desimone, R., Schwartz, M. K., Edwards, V. L., Pilgrim, K. P., & Mckelvey, K. S. (2012). Estimating abundance of mountain lions from unstructured spatial sampling. *The Journal of Wildlife Management*, *76*(8), 1551–1561. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jwmg.412

Sollmann, R. (2024). M_t or not M_t: Temporal variation in detection probability in spatial capture-recapture and occupancy models. *Peer Community Journal*, *4*, e1.

Sutherland, C., Fuller, A. K., & Royle, J. A. (2015). Modelling non-Euclidean movement and landscape connectivity in highly structured ecological networks. *Methods in Ecology and Evolution*, *6*(2), 169–177. https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12316

Theng, M., Milleret, C., Bracis, C., Cassey, P., & Delean, S. (2022). Confronting spatial capture-recapture models with realistic animal movement simulations. *Ecology*, *103*(10), e3676 https://onlinelibrary.wiley.com/doi/pdf/10.1002/ecy.3676

Thomas, L., Buckland, S. T., Rexstad, E. A., Laake, J. L., Strindberg, S., Hedley, S. L., Bishop, J. R., Marques, T. A., & Burnham, K. P. (2010). Distance software: Design and analysis of distance sampling surveys for estimating population size. *Journal of Applied Ecology*, *47*(1), 5–14. https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2664.2009.01737.x

Tobler, M. W., & Powell, G. V. N. (2013). Estimating jaguar densities with camera traps: Problems with current designs and recommendations for future studies. *Biological Conservation*, *159*, 109–118.

Tourani, M. (2022). A review of spatial capture-recapture: Ecological insights, limitations, and prospects. *Ecology and Evolution*, *12*(1), e8468 https://onlinelibrary.wiley.com/doi/pdf/10.1002/ece3.8468

Waller, L. A., Smith, D., Childs, J. E., & Real, L. A. (2003). Monte Carlo assessments of goodness-of-fit for ecological simulation models. *Ecological Modelling*, *164*(1), 49–63.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Figure S1. Number of animals captured across scenarios.

Figure S2. Number of detections across scenarios.

Figure S3. Responses of the GoF tests FT-ind-trap, FT-individuals and FT-traps to unmodelled heterogeneity across 300 simulated capture histories when density = 0.025 animals per km^2 and trap spacing = $2\boldsymbol{\sigma}$.

Figure S4. Responses of the GoF tests FT-ind-trap, FT-individuals and FT-traps to unmodelled heterogeneity across 300 simulated capture histories when density = 0.05 animals per km^2 and trap spacing = $\boldsymbol{\sigma}$.

Figure S5. Power of GoF tests when the tests are applied to capture histories with varying levels of unmodelled heterogeneity.

Table S1. Diagnostic overview of goodness-of-fit tests with respect to the simulation scenarios tested in this study.

How to cite this article: Choo, Y. R., Sutherland, C., & Johnston, A. (2024). A Monte Carlo resampling framework for implementing goodness-of-fit tests in spatial capture-recapture models. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.14386