

Response to Improving abundance estimation by combining capture-recapture and occupancy data: example with a large carnivore

Jack H. W. Thomas¹, First2 M2. Last2², etc

¹ Author Affiliation1

² Author Affiliation2

etc

January 12, 2023

SUMMARY. A key area of research in ecological statistics involves combining data sources from multiple streams to get better population estimates. The case study model attempts to integrate mark-recapture and occupancy data to estimate the population size of Eurasian lynx in the Jura Mountains, eastern France. The model has been observed to underestimate population sizes. We propose to conduct a simulation study by generating mark-recapture and occupancy data and analyzing the output of the model when using the generated data. We describe our proposed methods for data generation. We give an outline of our hypothesis on why the model underestimates population sizes. Finally, we analyze the results of the simulation study, which demonstrate that the model is flawed.

Keywords: keyword A, keyword B,...

1 Introduction

One ecological problem we would like to solve is how to estimate the population size of wild animals that live in a measurable area. Ecological statistics is the sub-field of statistics that deals with such problems. The most common methods for monitoring species are censuses, occupancy studies, and mark-recapture-like studies. A census consists of counting all the individuals in the population in a single sample. A census would give the most information about a population if we were interested in how large the population was at a certain time. However, they are often impossible or prohibitively expensive to conduct on a population. Occupancy studies attempt to find signs of presence in the field. These studies are inexpensive and can generate large datasets but can only tell us whether at least one individual

was in a site at a certain time. In mark-recapture-like studies animals are assumed to be identifiable either from unique tags or natural marks. Then the marked animals can be tracked over time. Mark-recapture studies provide a balance in information about a population and the resources required to carry out the study when compared to census and occupancy studies. An active field of research in ecological statistics concerns the methodology for combining information from multiple data sources to get better estimates for population parameters.

Blanc, Marboutin, Gatti, Zimmermann, and Gimenez [1] proposed a statistical model that combined mark-recapture and occupancy data to estimate population sizes. The model attempted to estimate the population size of Eurasian lynx in the Jura Mountains, eastern France. Mark-recapture data of lynx were obtained using camera traps. Lynx were identifiable by natural marks. Observers collected occupancy data by an extensive sign survey for evidence of presence in the field [1]. There are concerns about the model. The method ignores the locations of animals when working with spatial mark-recapture data. There was no assessment of the performance of the method either analytically or via a simulation study. The authors analyzed a dataset but provided no ground truth.

We believe that the method in [1] underestimates population sizes. Jahid et al. [3] used six different models to estimate the population size of grizzly bears in a region of Alberta. The approach of Blanc et al. [1] produced a point estimate of forty bears, despite other methods estimating four hundred bears or more. The main problem with the method is that it fails to account for animal movements when using spatial mark-recapture data. Our proposed work is to conduct a simulation study to explore the model and assess the performance of the resultant population estimates. We hypothesize that the model will produce erroneous estimates in all but two extreme cases: first, when animals do not move between sites during the study, and second when animals move throughout all the sites equally. This work is important because researchers may use this method and plan management strategies without realizing that the estimates are flawed. This could have grave consequences.

2 Literature Review

Several different methods have been proposed to combine data from occupancy and mark-recapture studies. Many different types of occupancy data have been combined with capture-recapture data such as presence-only, absence-only, count, and full occupancy data. All the approaches aim to improve the accuracy of population estimates. We conclude this section with an outline of the method proposed by Blanc et al. [1] that combines full occupancy data and mark-recapture data.

Renner, Louvrier, and Gimenez [6] combined multiple presence-only data into a single likelihood to estimate the abundance of Eurasian lynx in the Jura Mountains, eastern France. The three sources of data consisted of sightings, interference with livestock, and pictures of lynx taken by camera traps. The combined model improved population size estimation by incorporating an area-interaction into the likelihood [6]. Altitude, forest cover, and human population

density were treated as environmental variables. Previous work by [6] showed that there is more domestic interference in areas with higher human population densities. The usual assumption that multiple point locations are independently distributed is violated in real-life data. The new model accounts for this by using spatial dependence. Lasso-type (least absolute shrinkage and selection operator) penalties are also applied to the model to prevent potential over-fitting. A simulation study using generated data confirmed that the model performance was improved when spatial dependence was present in the data. The method takes advantage of prior knowledge in selecting presence-only datasets. Further extensions to the model include adding different types of spatial dependence to the model. The model also assumes a closed population and that sites are always occupied or always unoccupied.

Jiménez, Díaz-Ruiz, Monterroso, Tobajas, and Ferreras [4] combined occupancy data with capture-recapture data to estimate the abundance of stone marten. Camera traps were used to identify some individuals, which constitutes the mark-recapture data. However, many individuals were unidentifiable, leading to the creation of presence-absence data using the camera traps as well. In essence, the camera traps generated both mark-recapture and presence-absence data. The paper found that the integrated model had more accurate and precise estimates by adding presence-absence data. One factor that limits the model is the sparsity of camera traps. Presence-absence data can only be collected by camera traps here. Thus, a full geographic site survey would not be incorporated into the model. To compensate, telemetry data was collected from tagged individuals.

Strebel, Kéry, Guélat, and Sattler [7] combined count, detection/non-detection, presence-only, and absence-only data into a binomial N-mixture model to estimate the abundance of bird species in Switzerland. The paper describes the four types of data as follows: Count data were simply observed counts of animals in a certain site in a certain time range. This is different from mark-recapture data because individuals are not identified or recaptured. Next, detection/non-detection data are described as detecting certain species using a checklist. This is presence-absence data. Presence-only data is a set of sites that animals are known to inhabit, such as breeding sites. The data is a list of species that were present during the breeding season at a certain site [7]. Finally, absence-only data is a set of sites that animals are known to not inhabit. This may include environments that are not suitable for the animal, such as steep elevation gradients. Combining presence-absence data and presence-only data increased the spatial coverage compared to using presence-absence data alone. Absence-only data added more information to the model, which allowed more accurate parameter estimates. The paper cautions against the use of presence-only data as some sites are more likely to be visited than others and this introduces bias [7]. The paper asserts that their approach is valuable when a single source of data is sparse and adding more sources may improve parameter estimates.

Jahid et al. [3] compared integrated multi-sampling models for camera trap and hair trap data. The hair trap data constituted mark-recapture data, while the camera traps constituted the occupancy data. The paper compared six different models that estimated the population size of grizzly bears in the Rocky Mountains, Alberta:



1. Closed capture-recapture model, which is the traditional approach of Otis et al. [5].
2. Combined model of Blanc et al. [1].
3. Spatially explicit capture-recapture (SECR) model from detection data of the hair traps of Efford [2].
4. Combined detection data from hair traps and detection from camera traps model (SECR-O) from Tourani et al. [8].
5. Model SECR with sex as a covariate.
6. Model SECR-O with sex as a covariate.

Jahid et al. [3] found that the precision of the population estimates did not improve when occupancy data was added to the model. One of the assumptions for these integrated models is that the data sources are independent. In the discussion, the paper noted that hair traps and camera traps were not independent. Hair traps and camera traps were often in the same locations. Bears that were captured through hair trapping were likely to be captured by camera traps as well. Jahid et al. [3] found that the method of Blanc et al. [1] had an unusually small posterior standard deviation and that the abundance estimates were almost equal to the number of animals captured. The number of stations with bear detection was 38, which is close to the estimated population of 40 by the model of [1]. The capture-recapture-only model estimated 73.36 bears, and the SECR model estimated 380.28 bears. The authors suggest a simulation study be performed to test the model of [1].

The model by Blanc et al. [1] starts by assuming that population size is a realization of a Poisson distribution with rate parameter λ . Then the probability that a new population N would have size n is calculated using the probability mass function of the Poisson random variable:

$$P(N = n) = \frac{e^{-\lambda} \lambda^n}{n!}. \quad (1)$$

Next, we define presence-absence data to be a Bernoulli random variable. Let $z_i = 1$ and $z_i = 0$ indicate if a site i is occupied or not occupied, respectively. Denote the probability that a site is occupied with:

$$\psi_i = P(z_i = 1). \quad (2)$$

To combine the two data types, the probability that a site is occupied is equated with the probability that there is at least one animal in the overall population:

$$P(z_i = 1) = P(N > 0). \quad (3)$$

Assuming that N follows (1), ~~we have:~~


$$P(N > 0) = 1 - P(N = 0) = 1 - e^{-\lambda}. \quad (4)$$

Using (2), ~~we~~ define:

$$\psi_i = 1 - e^{-\lambda}. \quad (5)$$

Solving for λ , we obtain:

$$\lambda = -\log(1 - \psi_i). \quad (6)$$

This is the key equation of Blanc et al. [1], which links mark-recapture and presence-absence together. In principle, the likelihood functions for both variables inform abundance. 

3 Methods

~~Data will be generated to test the model.~~ Two sets of data are created: capture-recapture data and presence-absence data. **The two sets of data will be generated independently of each other.** This means that there can be different numbers of trapping occasions and detection occasions.

3.1 Capture-Recapture Data Generation

Capture-recapture data will be simulated using the method of camera traps. Individuals will interact with the camera traps and be captured based on probabilities. The first step to generating spatial capture-recapture data is defining home ranges. Animals in the population tend to stay within a certain range. **This is defined to be their home range.** **We can randomly place the centers of these home ranges in the study area.** A Poisson process can be used to generate these home ranges, where the **x and y coordinates are independent uniform random variables.** We will define the study area to be a unit square for simplicity. **Next, we place the camera traps using the same Poisson process into the study area.** We assume that capture probability is related to the distance between a home range center and a camera location and the amount of movement the species exhibits. A half-normal detection function models the capture probability of individual i by trap j :

$$p_{ij} = e^{(-d_{ij}^2/\tau^2)} \quad (7)$$

where τ is the movement parameter for the species and d_{ij} is the distance between the home-range center i and the camera location j . The value τ is defined to be between 0 and 1, where 1 is the most movement, and 0 is no movement. As τ increases the capture probability also increases. As d_{ij} increases the capture probability p_{ij} decreases. Once we place all the camera traps, we can simulate the interactions of individuals with the camera traps. Let the individuals be numbered $1, 2, 3, \dots, i$, the traps be numbered $1, 2, 3, \dots, j$, and the sampling occasions be numbered $1, 2, 3, \dots, k$. Then we obtain a three-dimensional array in which cell i, j, k determines if individual i was captured by trap j at time k . ~~At each time k , we can assign a random number between 0 and 1, and if the random number is less than or equal to the individual capture probability by trap j , we say that individual i was captured at time k by trap j .~~ Collapsing this three-dimensional array along the j axis yields the capture histories for all the individuals. A capture history is a matrix of 0's (not captured) and 1's (captured) where columns indicate times and rows indicate individuals. We now know when individual i was captured by any camera trap at time k .



3.2 Presence-Absence Data Generation

Presence-absence data generation starts with dividing the study area into regions. We define a grid such that each grid cell is equal in size. Each grid cell is henceforth referred to as a site. Now, we assume that an individual's detection or occupancy probability is proportional to the amount of time each individual spends in each site. Each individual has a home range that expands or shrinks based on the movement parameter τ . Let the proportion of time spent by each individual i in each grid cell j equal the probability of occupancy ψ_{ij} . To calculate ψ_{ij} , we perform a double integration of the bivariate normal distribution function with covariance τ^2 . We have:



$$\psi_{ij} = \int_{x_{0j}}^{x_{1j}} \int_{y_{0j}}^{y_{1j}} f(x, y | x_i, y_i, \tau) dx dy \quad (8)$$

where $f(x, y | x_i, y_i, \tau)$ is the bivariate normal distribution function given home-range x_i, y_i , and covariance τ . The bounds of the grid cell j are $x_{0j}, y_{0j}, x_{1j}, y_{1j}$. Each individual i has a two-dimensional probability matrix where each cell represents the detection probability of that site. ~~To simulate presence-absence data, for each time period k we assign a random number for each site and compare it to the detection probability for each individual.~~ This yields our presence-absence matrix which indicates whether at least one animal was detected by each site for each period.

3.3 Simulation

Supplementary code is provided by Blanc et al. [1] that implements their model in R. The R package “rjags” allows for the model to be designed and run in “JAGS” (Just Another Gibbs Sampler). We will use R version 4.1.2 in all of our simulation code. ~~The Digital Research Alliance of Canada provides high performance computing to researchers~~

N	10, 25, 50, 75, 100, 125, 150, 175, 200, 500, 750, 1000
τ	0.15, 0.20, 0.25, 0.30, 0.35, 0.4, 0.45, 0.5, 0.55
nsites	16, 25, 36
nsample_cap	7
nsample_pa	5
ntraps	3, 4, 5, 6

Figure 1: Parameters varied in case study

and students across Canada. We will run our simulations on the Cedar node of the high performance computing network. We can then compare the outputted population estimates with the true population sizes. We will compute the bias and mean square error of the point estimates, the bias of the standard errors, and the coverage and width of the credible intervals. We will then determine if the model outputs reasonable population estimates. The generated capture-recapture and presence-absence data will be used as input for the supplementary code from Blanc et al. [1]. The R package “coda” will generate samples from the JAGS model. Parameters for each simulation will be the actual population size, movement parameter, number of sites in the study area, number of capture-recapture sampling occasions, number of presence-absence sampling occasions, and the number of camera traps in the study. Each simulation will have 50 repetitions. Each simulation constitutes a set of parameters. The parameters will be varied as in Figure 1.

4 Results

Each simulation (1296 total) was repeated 50 times. The total computation time was 14 hours and 7 minutes. The last 9 simulations failed to complete due to running out of time (15 hours for each task). Failed runs were simulations where the actual population was 1000 and the computation time was too long for the HPC network.

All model population estimates were almost one-to-one with the number of animals captured in the simulation (Figure 5). As τ decreased, the presence-absence detections increased (Figure 9). As τ increased, the site occupancy probability decreased (Figure 10). The population estimate converged to the actual population value as τ increased (Figure 6). As τ increased, the standard deviation bias went to zero (Figure 11). Similarly, as the proportion of the actual population captured increased, the standard deviation bias went to zero. The mean ψ value had a small decrease as τ increased, but stayed relatively constant (Figure 10). Varying the number of camera traps, sites, capture-recapture sampling occasions, or presence-absence sampling occasions did not change the results.

5 Discussion

The model by [1] simply outputs the number of individuals captured. All simulations outputted the estimated population close to the number of individuals captured (Figure 5). In fact, the correlation is almost exactly one-to-one ($\rho = 0.999$). The parameter that affected how many individuals were captured was τ , the movement parameter. As Figure 6 shows, we see convergence of the population estimate to the actual population value as τ increases. As τ increases to 0.55, the population estimate converges to the actual population value (Figure 7). We defined standard deviation bias to be the difference between the model output standard deviations for the population size and our own calculated standard deviations of the population estimate. As τ increased, the standard deviation bias went to zero (Figure 12). As before, the model estimates only became accurate when τ was high and we captured almost every animal in the population.

We define the credible proportion to be the percentage of cases where the actual population value was within the 95% credible interval reported by the model output. In all cases, the lower 2.5% credible quantile was the number of individuals captured, which makes sense because that is the smallest possible population. In Figure 2, we see that for most values of τ the 95% credible interval outputted from the model were not capturing the true population value. This is evidence that the Bayesian model is flawed. As τ increases, the number of animals captured also increases and we eventually capture the entire population. Once τ has increased to a sufficient value, we essentially perform a population census and the 95% credible interval outputted from the model will contain the actual population value. This behavior can clearly be seen in Figure 4. As the population proportion captured increases to 1, the 95% credible interval eventually contains the true population value. The amount of animal movement is intrinsically linked to the models performance. The model only succeeds when there is enough movement such that we are capturing almost every animal in the population. As τ increases, the credible interval width decreases, converging to the actual population value (Figure 3).

As the estimated population decreases, the number of presence-absence detections increases. As τ increases, the number of presence-absence detections per site decrease because animals do not stay in one site as often. When the movement parameter decreases, we see a higher number of detections because we have a higher chance of seeing an animal near its home-range. Since τ has a near one-to-one positive correlation with the population estimate, we see the result of Figure 8. There is convergence of the ψ mean as τ increases (Figure 10). The mean ψ_0 parameter was close to one in every simulation. We hypothesize that the sites become saturated in the presence-absence data generation, thus the occupancy probability tends to one. The probability that a site is never occupied is low because the number of individuals tends to overlap the number of sites.

The model of [1] is invalid for all but one extreme case: when animals have a huge movement parameter and there is only one site. Jahid et al. [3] speculated that the model would be valid if animals ranged over the entire study area,

that is when τ is sufficiently high. We have shown this to be true; when the animal movement is high enough to capture all the individuals in the population the model could be considered valid. When there is low movement, the model still outputs population estimates that match the number of individuals captured (Figure 5). Why does the model output these results? In [1], λ is an increasing function of ψ which makes sense because higher abundance means higher occupancy probability and vice versa. But the data on occupancy pushes ψ down, because knowing some sites are unoccupied means that $\psi < 1$. However, the data on abundance pushes λ and ψ up. So λ increases just enough to make the observed number of individuals possible, but then stops to avoid dragging ψ up too high. The resulting population estimate is more about the interplay between λ and ψ being on different scales and essentially ignores the data on detection.

ACKNOWLEDGEMENTS

This research was enabled in part by support provided by (name of the regional partner organization) (Web address) and the Digital Research Alliance of Canada (alliance.can.ca).

References

- [1] Blanc, L., Marboutin, E., Gatti, S., Zimmermann, F., and Gimenez, O. (2014). Improving abundance estimation by combining capture-recapture and occupancy data: Example with a large carnivore. *J. Appl. Ecol.* **51**, 1733–1739.
- [2] Efford, M. (2004). Density estimation in live-trapping studies. *Oikos* **106**, 598–610.
- [3] Jahid, M., Steeves, H. N., Fisher, J. T., Bonner, S. J., Muthukumarana, S., and Cowen, L. L. (2022). Shooting for abundance: Comparing integrated multi-sampling models for camera trap and hair trap data. *Environmetrics* .
- [4] Jiménez, J., Díaz-Ruiz, F., Monterroso, P., Tobajas, J., and Ferreras, P. (2022). Occupancy data improves parameter precision in spatial capture–recapture models. *Ecol. Evol.* **12**, e9250.
- [5] Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildl. Monogr.* pages 3–135.
- [6] Renner, I. W., Louvrier, J., and Gimenez, O. (2019). Combining multiple data sources in species distribution models while accounting for spatial dependence and overfitting with combined penalized likelihood maximization. *Methods Ecol. Evol.* **10**, 2118–2128.
- [7] Strebel, N., Kéry, M., Guélat, J., and Sattler, T. (2022). Spatiotemporal modelling of abundance from multiple data sources in an integrated spatial distribution model. *J. Biogeogr.* **49**, 563–575.

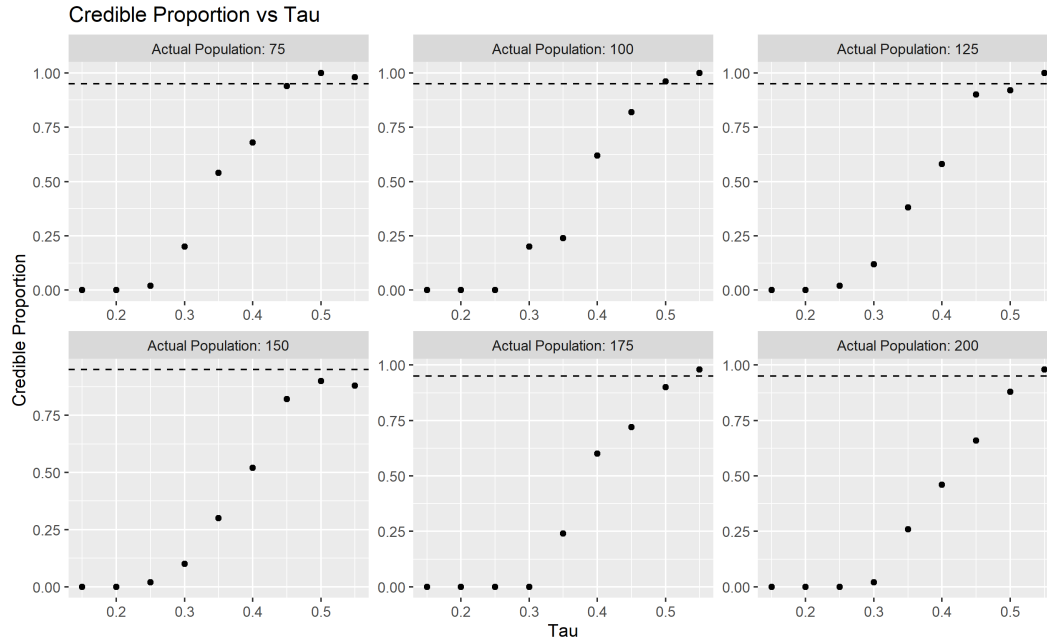


Figure 2: Credible proportion is the percentage of times that the 95% credible interval contained the actual population value. The 95% threshold is displayed as a dotted line.

- [8] Tourani, M., Dupont, P., Nawaz, M. A., and Bischof, R. (2020). Multiple observation processes in spatial capture–recapture models: How much do we gain? *Ecology* **101**, e03030.

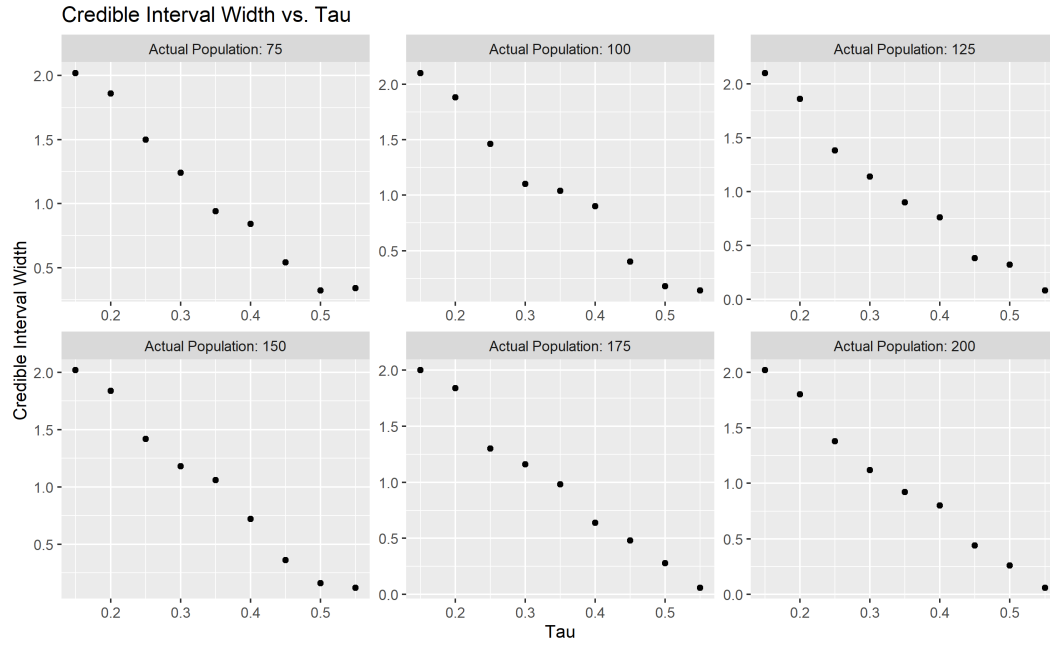


Figure 3: Caption

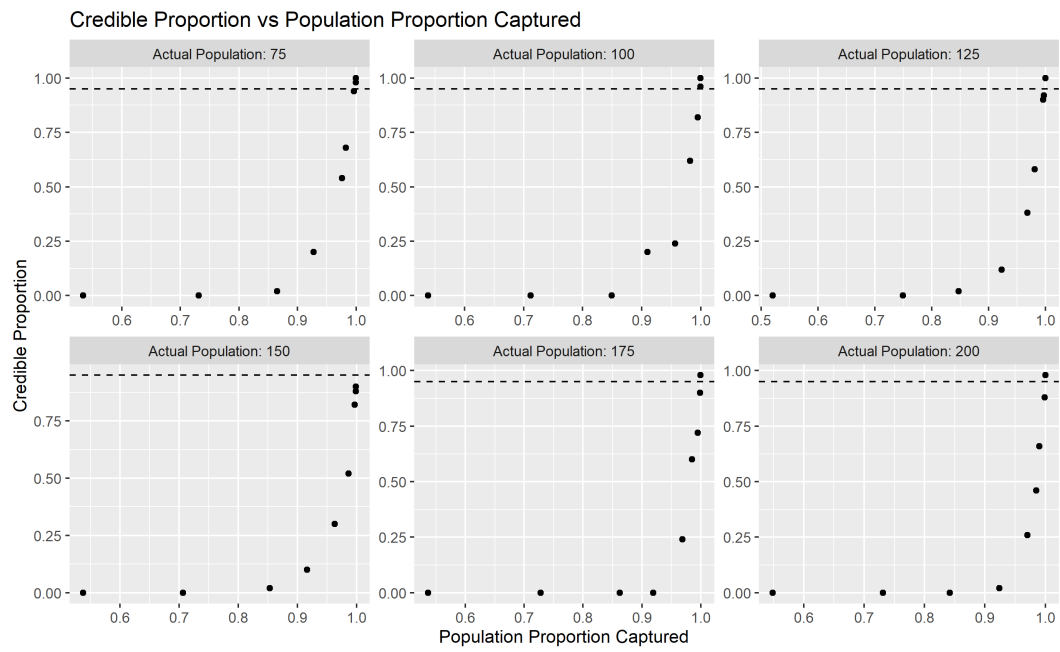


Figure 4: Caption

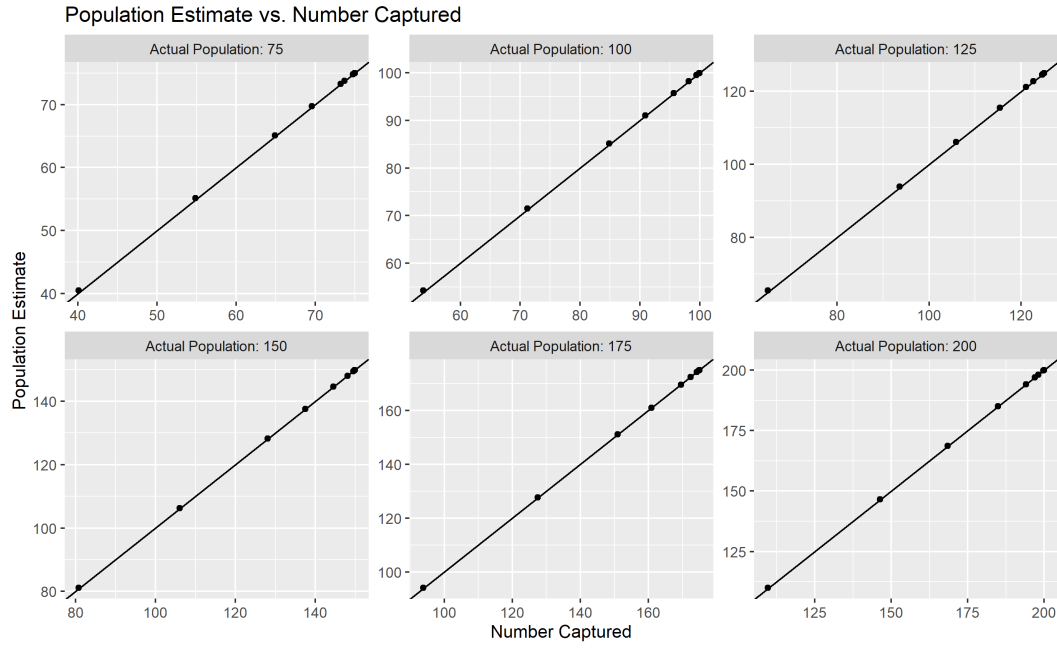


Figure 5: Linear almost perfect correlation ($\rho = 0.999$). Slope=1 line plotted.

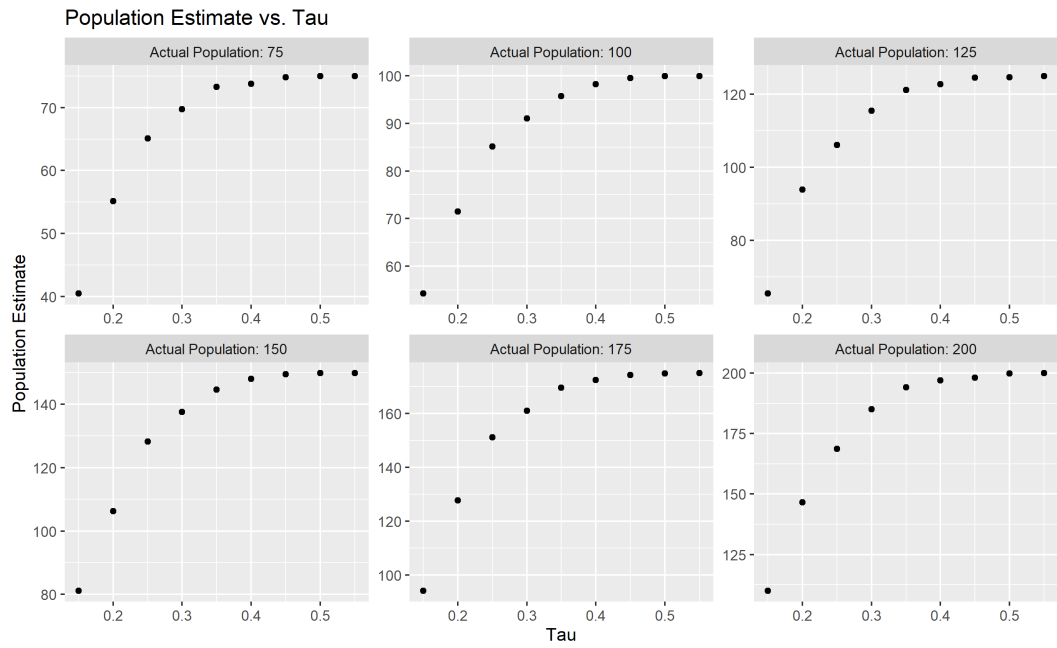


Figure 6: Convergence of population estimate to actual population value.

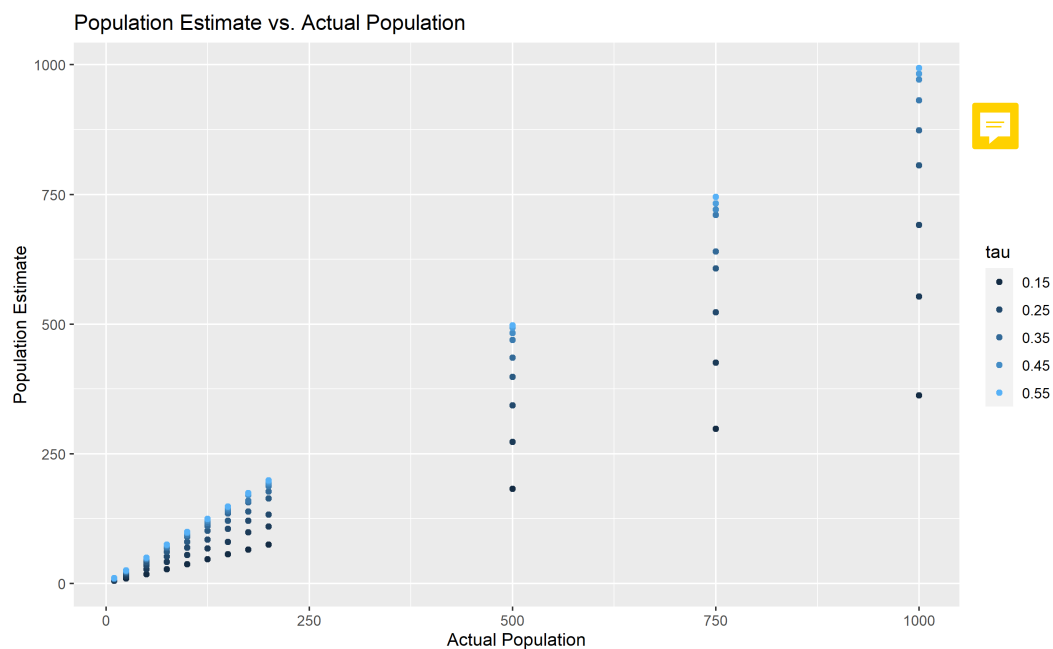


Figure 7: Caption

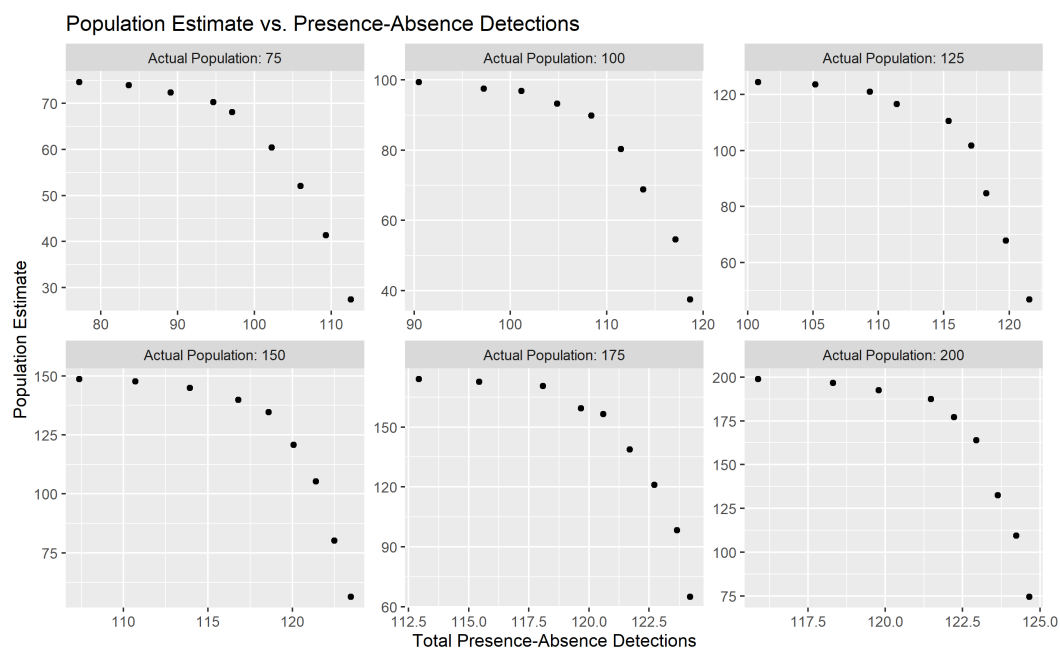


Figure 8: Caption

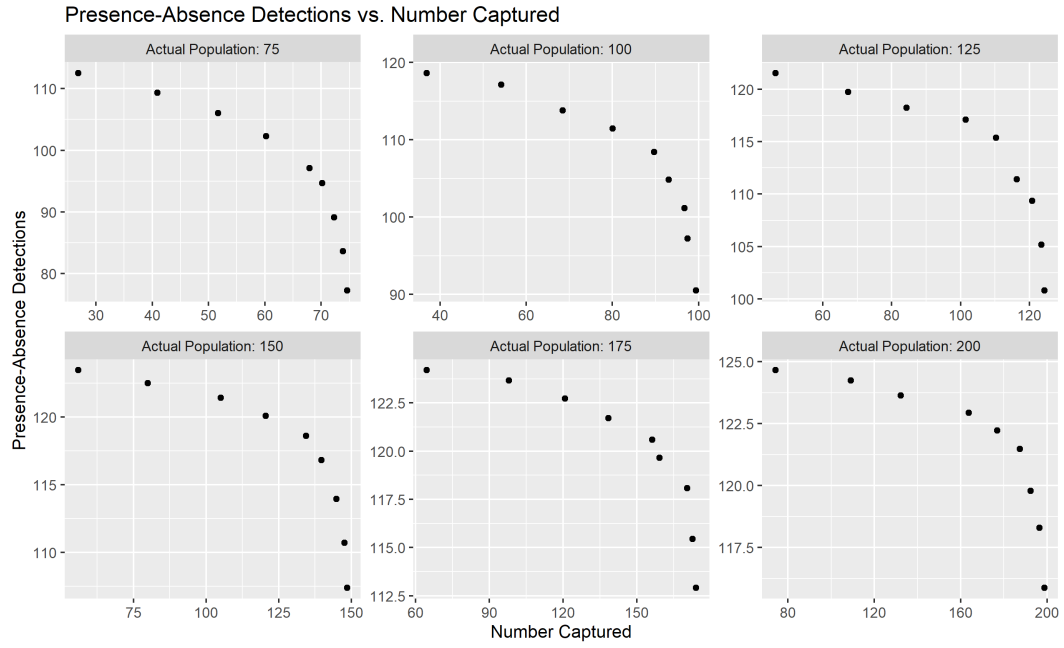


Figure 9: Caption

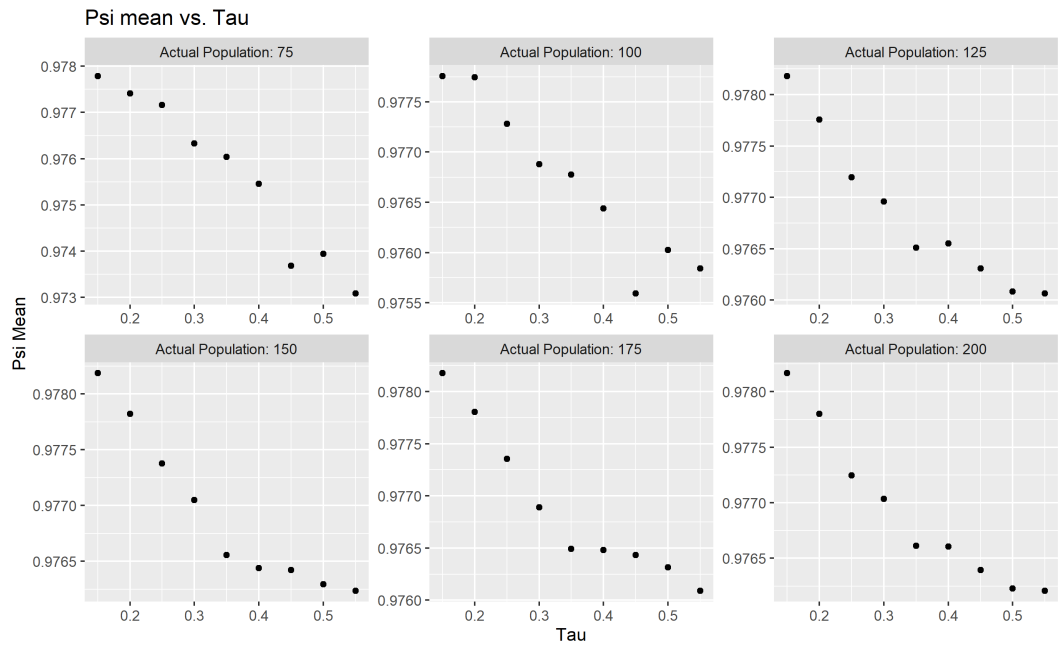


Figure 10: Caption

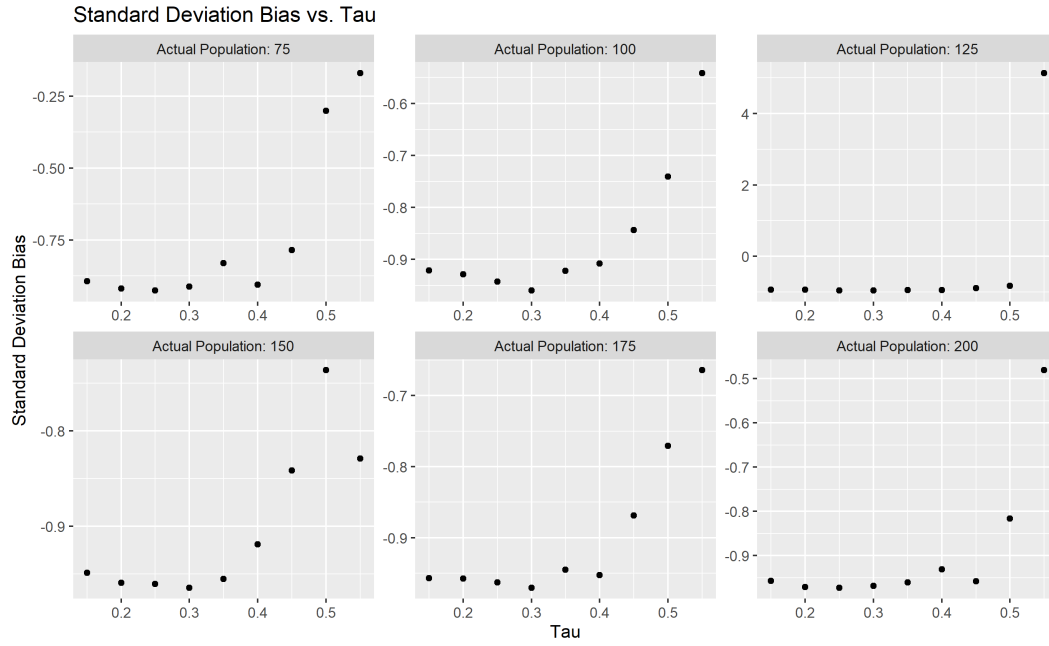


Figure 11: Outlier on N_actual=125

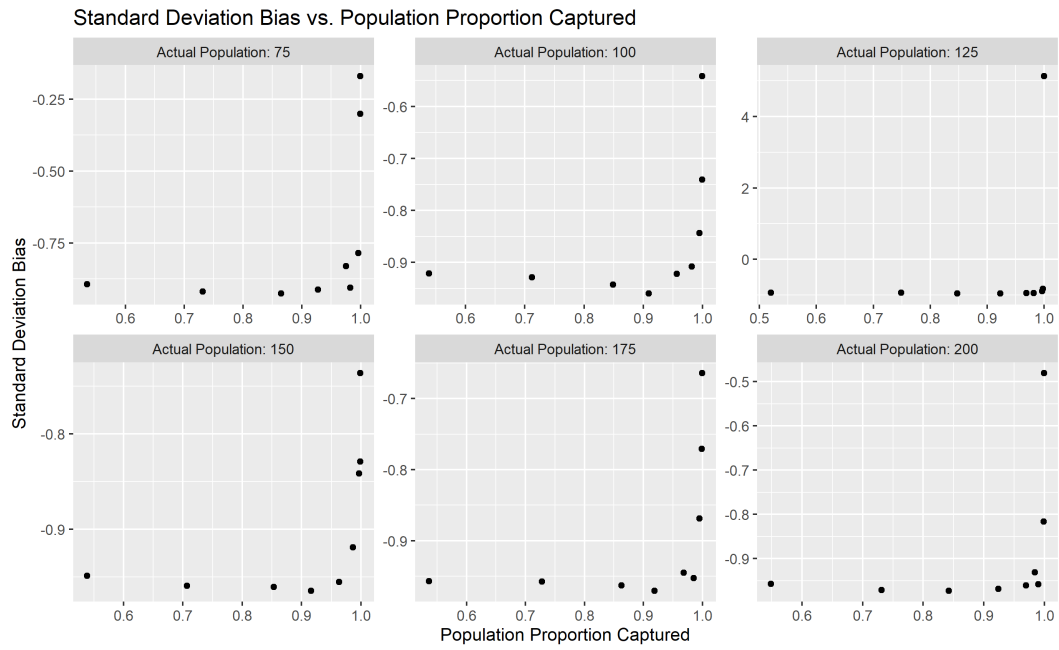


Figure 12: Outlier on N_actual=125