16 Jan 2018

Nick P. Gondek

Conservation Metrics Inc.

145 McAllister Way

Santa Cruz, CA 95060

(414) 581-4445

nick@conservationmetrics.com

RH: Gondek et al. • Subsampling for SECR

**Too much DNA? Subsampling strategies for Spatially Explicit Capture-Recapture Estimators**

NICK P. GONDEK,1 *Conservation Metrics, Inc., 145 McAllister Way, Santa Cruz, CA 95060 USA*

DAVID L. GARSHELIS, *Minnesota Department of Natural Resources, 1201 E. Hwy 2, Grand Rapids, MN 55744, USA*

KAREN V. NOYCE, *Minnesota Department of Natural Resources, 1201 E. Hwy 2, Grand Rapids, MN 55744, USA*

JOHN F. FIEBERG, *University of Minnesota, 2003 Upper Buford Cir, St Paul, MN 55108 USA*

**ABSTRACT** Genetic mark-recapture studies estimate animal abundance using non-invasive DNA identification methods to "capture" and subsequently "recapture" individuals that leave genetic material at trap sites. Due to the cost of genotypic analysis, researchers often choose to process only a subsample of this genetic material. Traditional (non-spatial) mark-recapture estimators of abundance have been shown to be biased in this case, especially when individuals display a behavioral trap response following initial capture. Less is known about the impact of subsampling genetic mark-recapture data when using spatially explicit capture-recapture (SECR) models to estimate abundance. We explored the effect of subsampling on SECR estimators using hair-snare data obtained from a 2012 genetic mark-recapture study of black bears (*Ursus americanus*),and also using simulated data with a known population. Similar to effects on non-spatial mark-recapture estimators, subsampling produced density estimates that were lower, on average, than the full data estimate; however non-proportional subsampling (sampling at least one observation from each unique trapping site and sampling period combination) had much less of an effect on estimator performance than simple random sampling, particularly at small sample sizes. Thus, non-proportional subsampling may be preferable to simple random sampling, despite the inherent violations of SECR assumptions that may result.

**KEY WORDS** black bear, genetic mark-recapture, density estimation, Minnesota, subsampling, simulation, spatially-explicit, *Ursus americanus*

Mark-recapture studies are routinely used by wildlife managers to estimate animal abundance; especially in the case of endangered species and game animals, abundance and its associated temporal trends are of critical importance for making informed management decisions (Borchers *et al.* 2002; McCrea & Morgan 2014). In many cases, however, abundance estimates can be difficult to interpret without some understanding of the effective area sampled; extrapolation to a regional scale by way of density is almost always desired, if not necessary (Borchers *et al.* 2002; Royle *et al.* 2013). As such, abundance estimates without associated reliable density estimates may be of limited use to managers.

Density estimates have often been obtained from abundance estimates using ad-hoc methods that range widely in their biological relevance (Royle *et al.* 2013). Spatially-explicit capture recapture methods (SECR), by contrast, link abundance estimators to their associated study area in a statistically rigorous way (Efford *et al.* 2005; Borchers 2012; Royle *et al.* 2013). SECR estimates of abundance scale directly with a given sample area, potentially allowing for more standardization of estimates across space or time. SECR models also make better use of the capture-recapture data by using the information in the spatial capture histories to model detection probabilities as a function of the distance between each animal's (latent) activity center (AC) and the trapping grid. By accounting for each individual's movement tendencies, SECR provides a way to account for an important source of individual heterogeneity in capture probabilities that is unaccounted for in traditional (non-spatial) mark-recapture estimators (e.g., animals with home ranges near the edge of the trapping grid will be less likely to be captured than animals with activity centers near the center of the trapping grid).

SECR models are often fit to capture histories obtained using genetic data (e.g., from hair collected at a trap site), which provides a non-invasive and efficient means of detecting unique individuals (Boulanger *et al.* 2004; Petit & Valiere 2006; Gervasi *et al.* 2008; Buckworth & Territory 2012). Due to the costs of genotypic analysis, however, managers are often forced to subsample clusters of hair left at trapping sites (Boulanger *et al.* 2004; Petit & Valiere 2006; Gervasi *et al.* 2008; Settlage *et al.* 2008).

Subsampling has been shown to be problematic for non-spatial mark-recapture estimators, especially when individuals exhibit a behavioral response to having been previously captured, and this behavioral response is not consistent across individuals (Tredick *et al.* 2007; Ebert *et al.* 2010; Augustine *et al.* 2014). In this case, individuals that leave many hair clusters are likely to be identified in a subsample, whereas individuals that leave few clusters are often excluded. In other words, clusters selected in a subsample are likely to come from individuals that are repeatedly captured. As a result, subsampling tends to result in estimators of capture probability that are biased high and abundance estimators that are biased low; this effect is more pronounced as subsample size decreases (Augustine *et al.* 2014).

A fundamental difference between SECR and non-spatial capture-recapture models relates to how the estimators make use of multiple captures of the same individual during the same time period. Whereas non-spatial models collapse multiple captures at different traps (same time period) into a single capture event, these multiple captures would be used to inform individual movement characteristics in a SECR model (Borchers 2012; Royle et al. 2013). Thus, samples that are redundant in a non-spatial model (captures at trap location in a single period) are of critical importance to SECR models. Importantly, managers often avoid taking multiple samples from a single trap under the assumption that they are likely to be from the same individual, and thus not informative and not worth the cost of genotypic analysis. This strategy is effective for use with non-spatial models; analyzing several samples from the same site often leads to diminishing returns in precision and accuracy (Dreher *et al.* 2009). Instead, preference is given to samples from novel site by session combinations (hereafter referred to as site-sessions), because they are more likely to represent new individuals. Yet, this approach tends to increase inclusion probabilities for samples left at infrequently visited sites and decrease inclusion probabilities for samples left at frequently visited sites. Because SECR models attempt to use information in the spatial capture histories to inform capture probabilities (by modeling movement about activity centers), we hypothesized that the effects of non-proportional subsampling would be problematic for SECR models (more so than for non-spatial capture-recapture models that reduce capture histories to a single binary response for each sampling period). A natural solution to this violation is to avoid preference for novel site-sessions (i.e., use simple random sampling), at the cost of identifying fewer unique individuals and analyzing more redundant data.

Various simulation and empirical studies have addressed the issue of subsampling on non-spatial mark-recapture estimators, with a general conclusion that subsampling genetic mark-recapture data results in estimators that are biased low, but that the magnitude of this bias depends on the actual data at hand and several assumptions such as erroneous genotyping rate and trap spacing (Tredick *et al.* 2007; Dreher *et al.* 2009; Augustine *et al.* 2014, Murphy *et al*. 2016). The tradeoff between potential redundancy (using simple random sampling) and non-representation of movement characteristics (using non-random sampling) has not been investigated in the context of SECR models. Our objective was to use a northern Minnesota genetic mark-recapture dataset in tandem with simulated datasets with known individuals, to compare abundance and density estimates using (1) subsampling strategies commonly utilized by wildlife managers and researchers and (2) various subsampling rates reflective of different budgetary constraints. Using these results, we provide guidance for genetic mark-recapture estimates when budget constraints limit effective sample size.

**STUDY AREA**

**METHODS**

## Genetic mark-recapture data from Minnesota

We considered data from a 2012 genetic mark-recapture study of American black bears (*Ursus americanus*) in northern Minnesota (Garshelis & Noyce 2013). Using stationary hair-snare traps baited with suspended bacon and scent lures, we collected 1642 hair clusters (groups of hair samples obtained from adjacent barbs) from 121 sites over six, 10-day trapping sessions from May through July 2012. Of these 1642 clusters, 1113 were sent to a genetics laboratory for genotypic analysis, and 1019 hair samples from these clusters were successfully linked to individual bears.

Camera traps fitted to a subset of the trap locations indicated that bears occasionally visited the same trap several times in a single session, often using several different entry locations to approach the bait hung at the center of the trap. This individual variation is evidenced in the number of samples left at a given site-session by individual bears; in 46.7% of observations, bears left only a single cluster of hair at a given site-session (Garshelis & Noyce 2013). Conversely, in 25.8% of cases, bears left three or more clusters of hair (up to 11) at a given site-session (Fig. 1).

## Simulated Data

*Capture History Simulation. —* We developed a simulation model to create capture histories allowing for the possibility of among-individual capture heterogeneity, a behavioral response following initial capture at a trap, and the possibility that individuals might leave multiple samples at a site during a single trapping period (e.g., by entering and leaving hair snare traps at different locations or revisiting sites multiple times). To describe the different scenarios, we have compiled a list of terms and definitions used in model notation (Table 1).

We simulated ‘activity centers’, (*Ai*; *i* = 1, 2, …, N), for N individuals using a simple sequential inhibition process with an inhibition distance *ϕ* (Baddeley, 2017). In scenarios with stratified density, the inverse of α represents the proportion of individuals whose activity centers are simulated exclusively in the left half of the trapping grid (Fig. 2B).

A ~ rSSI(N, ϕ, α) (1)

Further, we assigned a normally distributed ‘individual heterogeneity’ parameter, δi, to each individual, which characterized that individual’s heightened or depressed propensity for capture relative to the population:

δi ~ Normal(0, Δ) (2)

We determined the capture probability for individual *i*, at trap *k* during trapping session *t, gi,t,k*, using:

where Ψi,k is the distance between the individual’s activity center and trap k, Ci,k  is 1 if the individual i was previously captured at trap k during any previous trapping period (and 0 otherwise), and δi measures the individual’s propensity for capture (Fig. 2A). Capture histories were then simulated as Bernouli random variables:

c­i,s,t ~Bernouli(gi,t,k) (4)

If captured, we simulated a number of samples left at the trap, Ii,t,k, using a Poisson distribution:

Ii,s,t ~ Poisson(λ­i ci,t,k) with log(λ­i) = γ+δi (5)

By including δi above, we assume that bears that have a higher propensity for being captured are also more likely to leave multiple samples at a trap.

Investigated Scenarios — We simulated eight scenarios with different combinations of behavior, individual heterogeneity, sample redundancy, and uneven density of activity centers to quantify any impacts of subsampling on estimator performance (Fig. 3).

## Subsampling

We chose to explore two alternative subsampling strategies: *simple random sampling* (SRS), and a subsampling method that gives preference to unique site-sessions, which we will refer to as *site-session preferred* (*SSP*) (Figure 2). With simple random sampling, *n* samples were chosen at random from the set of hair clusters pooled across the different sites and trapping periods. Alternatively, with *SSP*, we tried to maximize the number of unique site-sessions represented in the subsample. Let *m* represent the number of unique site-sessions in the full dataset. If *m < n*, we randomly choose *n* unique site-sessions, with 1 sample randomly selected sample from each of these site-sessions (in the survey sampling literature, this would be referred to as a 2-stage cluster sample). When *n* > *m*, we randomly chose one sample at random from each unique site-session and then took a second simple random subsample of size *n – m* from the remaining clusters (from the pooled data) to give a total of *n* samples.

In the context of this study, an observation (i.e. hair cluster) can be considered ‘redundant’ if it does not contribute a unique combination (individual x site x session) to the capture history of interest; i.e., a sample is redundant if there is another observation of the same individual at the particular site-session. To better understand the performance of the estimators under different types of subsampling, we calculated the percentage of each dataset that was not-redundant.

For each dataset (eight simulated scenarios and one empirical dataset), both subsampling strategies were employed, at n=250, n=550, and n=850. Models are then fit to each of these six subsamples, along with the original dataset.

**Model Fitting**

A SECR model is unique from other mark-recapture models in that an animal’s capture probability is derived using the animal’s activity center (AC) (Borchers et al. 2002; Royle et al. 2013). Though many curves are used to characterize how detection probabilities change as a function of distance between an animal's activity center and a trap location, a common and readily understood choice is a half-normal curve, using two parameters: g0 and σ. g0 represents the probability of detecting an animal whose activity center is located exactly at the trap location. σ represents the rate at which this probability decreases as an animal’s homerange center moves further away from the trap. These parameters can be allowed to vary by sex, time, and they may be allowed to change following an initial capture event (i.e., a behavioral trap response (bk)).

For each simulated (and potentially subsampled) data set, we fitted two SECR models to the observed capture histories; a null model (*g0* ~ 1), and model where individuals’ likelihood of capture at a given trap changed after initial capture at that specific trap (*g0* ~ bk). For each empirical subsample, we fit two additional models; a model where likelihood of capture depends on trapping period (*g0* ~ t), and one where both effects influence likelihood of capture (*g0* ~ bk + t). In each case, we assumed was constant for all individuals.

Models were fit using the R programming language (R Core Team 2015), package ‘secr’ for the fitting of the SECR models, and packages ‘foreach’ and ‘doParallel’ for optimization of model fitting and capture history simulation using parallel processing (Analytics & Weston 2014, 2015; Efford 2015). The main function for fitting models to subsampled data is secr.fit in package ‘secr’. This function requires a capture history and a trapping grid to arrive at a derived density estimate and estimated parameters describing the effect of time, sex, and/or trap-specific behavior on capture probabilities (Royle *et al.* 2013).

## Simulation

Note the distinction between a capture history simulation, which refers to the construction of a capture history as described in *Capture History Simulation*, and a subsampling simulation, which refers to the subsampling and subsequent model fitting of an empirical or simulated capture history. We analyzed capture histories of eight unique combinations of individual and population-level variations among simulated bear populations, and one empirical capture history, for a total of nine scenarios (Figure 3). Further, we subsampled each capture history using both SSP and SRS methods, with sample size equal to *n* = 250, *n* = 550, and *n* = 850. We then fit four SECR models as described in *Model Fitting*, and saved resultant model objects for comparison.

# Results

# Discussion

Our simulation results suggest that SECR density estimators, like their non-spatial counterparts, are biased low when applied to subsampled data (Tredick *et al.* 2007; Augustine *et al.* 2014). SSP increased the number of uniquely identified individuals (especially those that left only one or two samples; Figure 4), which outweighed the inherent cost of violating SECR's assumption that the capture history provides an accurate description of individual movement characteristics. These effects were less influential at larger sample sizes (n=550, n=850).

Bears exhibited considerable heterogeneity in their trap-specific behavioral responses (in 47% of cases, bears left only a single sample at a given site-session, but some left as many as 11; Garshelis & Noyce 2013). Thus, subsampling empirical data using SRS was highly likely to select redundant data from bears that tended to leave several samples at a single site-session. Conversely, SSP performed well because much of the data it excluded from the full dataset was redundant (repeated individual by site by session combinations), and thus, increased the likelihood of including novel samples (Figure 4).

Key assumptions in this study include the treatment of the full dataset as a representative sample from the population. Noyce and Garshelis (2013) successfully subsampled 67.7% of samples, so there is likely an unmodelled behavioral effect from the samples that were not originally analyzed when the study took place. Further, these analyzed hair samples were already non-random to some degree (the first 377 clusters were chosen to represent unique site-sessions and 776 clusters were chosen completely at random), so our analysis may be missing novel individuals or known individuals' locations that were similarly not identified in the original genotypic analysis. However, it is also likely that many of the non-sampled clusterswere redundant due to the strong individual heterogeneity noted in this study. Additionally, we analyzed only successful samples, and thus ignored the possibility of failure to identify individuals from a submitted sample; however, because our dataset was not simulated, there may have been false identification of individuals due to allelic dropout/false amplification (Dreher *et al.* 2009), or the 'shadow effect' - erroneously treating a novel individual as a recapture due to similarity in their genotype (Mills *et al.* 2000).

# ACKNOWLEDGEMENTS

[replace with funding sources, Minnesota wildlife society for presentation, fieberg lab, etc]

# LITERATURE CITED

Analytics, R. & Weston, S. (2014). *DoParallel: Foreach parallel adaptor for the parallel package*.

Analytics, R. & Weston, S. (2015). *Foreach: Provides foreach looping construct for r*.

Augustine, B.C., Tredick, C.A. & Bonner, S.J. (2014). Accounting for behavioural response to capture when estimating population size from hair snare studies with missing data. *Methods Ecol Evol*, **5**, 1154–1161.

Baddeley. A. (2017). *spatstat: Spatial Point Pattern Analysis, Model-Fitting, Simulation, Tests.*

Borchers, D. (2012). A non-technical overview of spatially explicit capture–recapture models. *Journal of Ornithology*, **152**, S435–S444.

Borchers, D., Buckland, S. & Zucchini, W. (2002). *Estimating animal abundance: Closed populations*. Springer-Verlag London.

Boulanger, J., Himmer, S. & Swan, C. (2004). Monitoring of grizzly bear population trends and demography using dNA mark-recapture methods in the owikeno lake area of british columbia. *Canadian Journal of Zoology*, **82**, 1267–1277.

Buckworth, R. & Territory, N. (2012). GENETAG: genetic mark-recapture for real-time harvest rate monitoring: Pilot studies in northern australian spanish mackerel fish. *Fisheries Research and Development Corporation*, **78**, 1–17.

Dreher, B., Rosa, G. & Lukacs, P. (2009). Subsampling hair samples affects accuracy and precision of DNA Based population estimates. *The Journal of Wildlife Management*, **73**, 1184–1188.

Ebert, C., Knauer, F., Storch, I. & Hohmann, U. (2010). Individual heterogeneity as a pitfall in population estimates based on non-invasive genetic sampling: A review and recommendations. *Wildlife Biology*, **16**, 225–240.

Efford, M. (2015). *Secr: Spatially explicit capture-recapture models*.

Efford, M., Warburton, B., Coleman, M. & Barker, R. (2005). A field test of two methods for density estimation. *Wildlife Society Bulletin*, **33**, 731–738.

Garshelis, D. & Noyce, K. (2013). Capture heterogeneity in hair-trapping of bears. *Summary of Research Findings*, 71–85.

Gervasi, V., Ciucci, P., Boulanger, J., Posillico, M. & Sulli, C. (2008). A preliminary estimate of the apennine brown bear population size based on hair-snag sampling and multiple data source mark-recapture huggins models. *Ursus*, **19**, 105–121.

McCrea, R.S. & Morgan, B.J.T. (2014). *Analysis of capture-recapture data*. Chapman; Hall, CRC.

Mills, S.L., Citta, J.J., Lair, K.P., Schwartz, M.K. & Tallmon, D.A. (2000). Estimating animal abundance using noninvasive DNA sampling: Promise and pitfalls. *Ecological Applications*, **10**, 283–294.

Petit, E. & Valiere, N. (2006). Estimating population size with noninvasive capture mark recapture data. *Conservation Biology*, **20**, 1062–1073.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Royle, J., Chandler, R., Sollmann, R. & Gardner, B. (2013). *Spatial capture-recapture*. Academic Press.

Settlage, K.E., T, M.F., Clark, J.D. & King, T.L. (2008). Challenges of DNA based mark recapture studies of american black bears. *The Journal of Wildlife Management*, **72**, 1035–1042.

Tredick, C., Vaughan, M., Stauffer, D. & Simek, S. (2007). Sub-sampling genetic data to estimate black bear population size: A case study. *Ursus*, **18**, 179–188.

Figure 1. Histogram of number of samples deposited by an individual at a unique site-session by northern Minnesota black bears (*Ursus americanus)* in a 2014 genetic mark-recapture study.

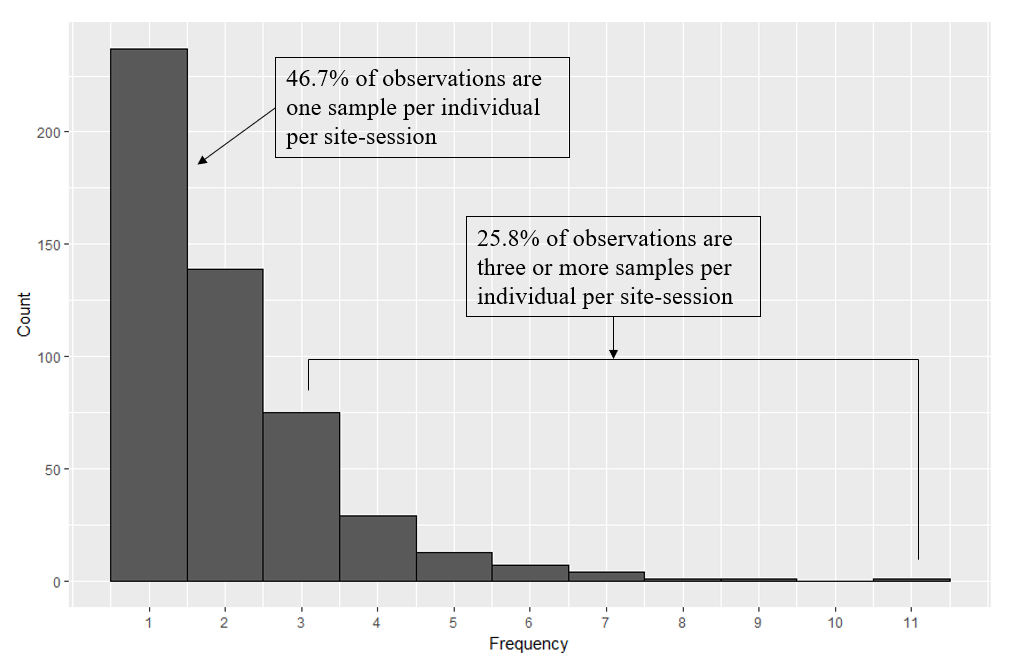
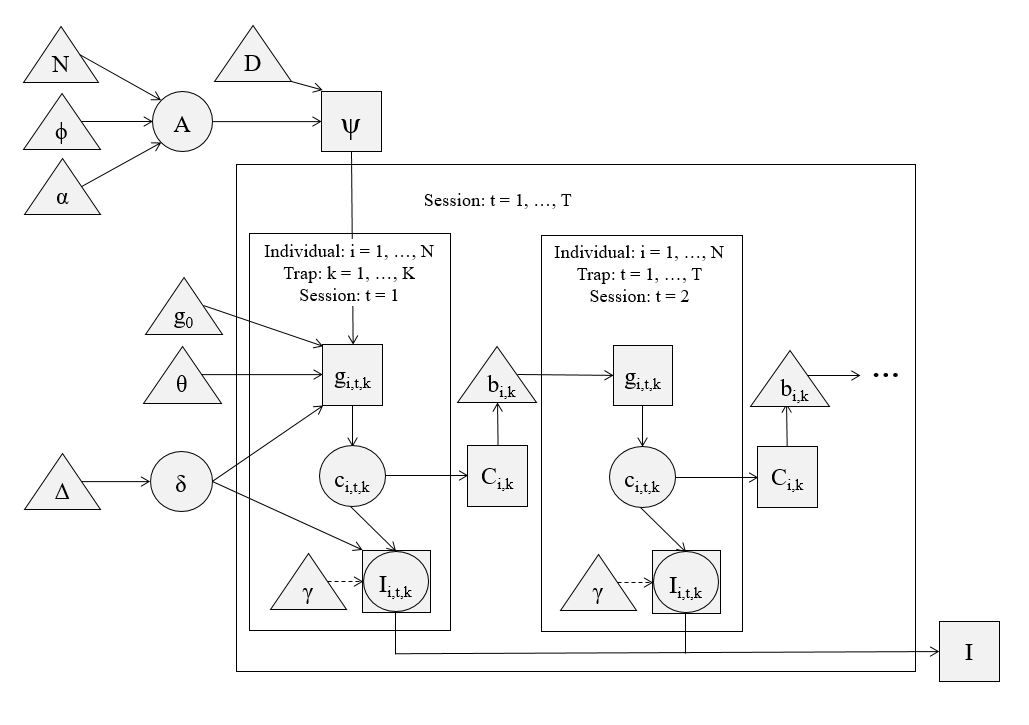
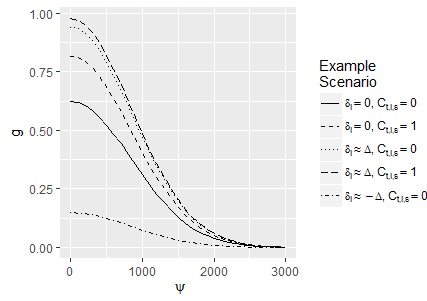
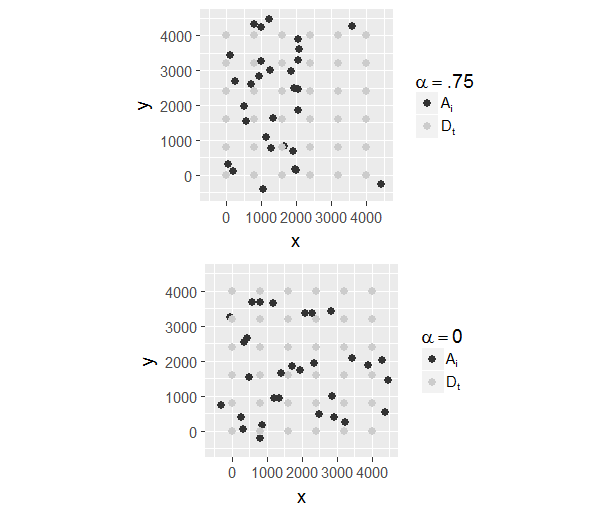


Figure 2. Directed Acrylic Graph of the data simulation process. Triangle nodes represent fixed parameters, circles represent stochastic values, and squares represent deterministic values obtained using stochastic values. All values are defined in Table 1. Ii,t,k is either a deterministic or stochastic node depending on the value of γ.



Figure 3. A) Example half-normal capture probability curves for individuals, dependent on their individual propensity of capture ( and whether the given trap has captured the individual in a previous session ( B) Example “Activity Centers” generated using either a heavily skewed scheme (α = .75) or a normal scheme (α=0).

B

A

Figure 4. Matrix of simulated scenarios investigated in this study. Shaded boxes represent the presence of a given effect on the simulated capture histories of a trial.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Effect** | **Description of effect** | **Relevant terms** | **Effect scale** | **Present in** | | | | | | | |
| **t1** | **t2** | **t3** | **t4** | **t5** | **t6** | **t7** | **t8** |
| trap-specific behavior | Bears are more likely to revisit a given trap after visiting that trap | b | Population |  |  |  |  |  |  |  |  |
| individual heterogeneity of behavior | Individual bears are more or less likely to visit any trap | Δ, δ | Individual |  |  |  |  |  |  |  |  |
| sample redundancy | Bears leave one or more extra samples at a site-session | γ, Δ, δ | Population if Δ = 0; Individual if Δ > 0 |  |  |  |  |  |  |  |  |
| uneven density of activity centers | Bears are distributed disproportionately on trapping grid | α, Φ | Population |  |  |  |  |  |  |  |  |

Figure 5. Proportion of non-redundant samples vs subsampling type for each of the five trials where redundancy is possible. Note that, as redundancy is not introduced for t1, t2, t3, or t8, proportion of non-redundant samples is fixed at 1 for these trials.

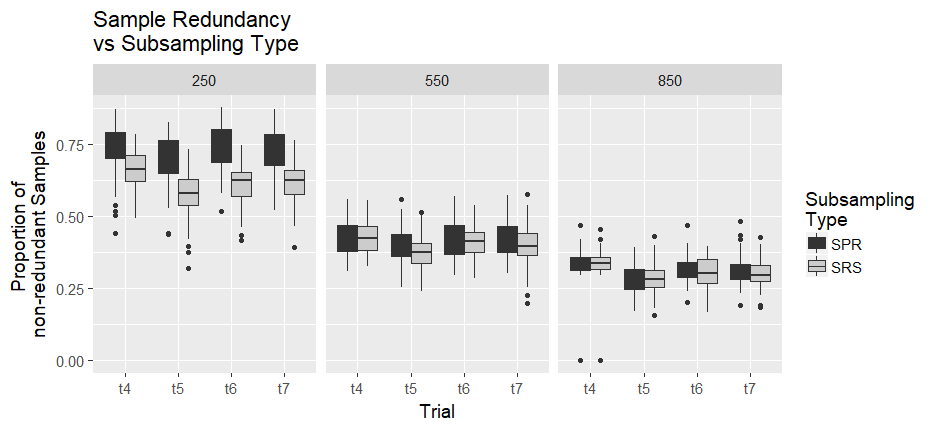


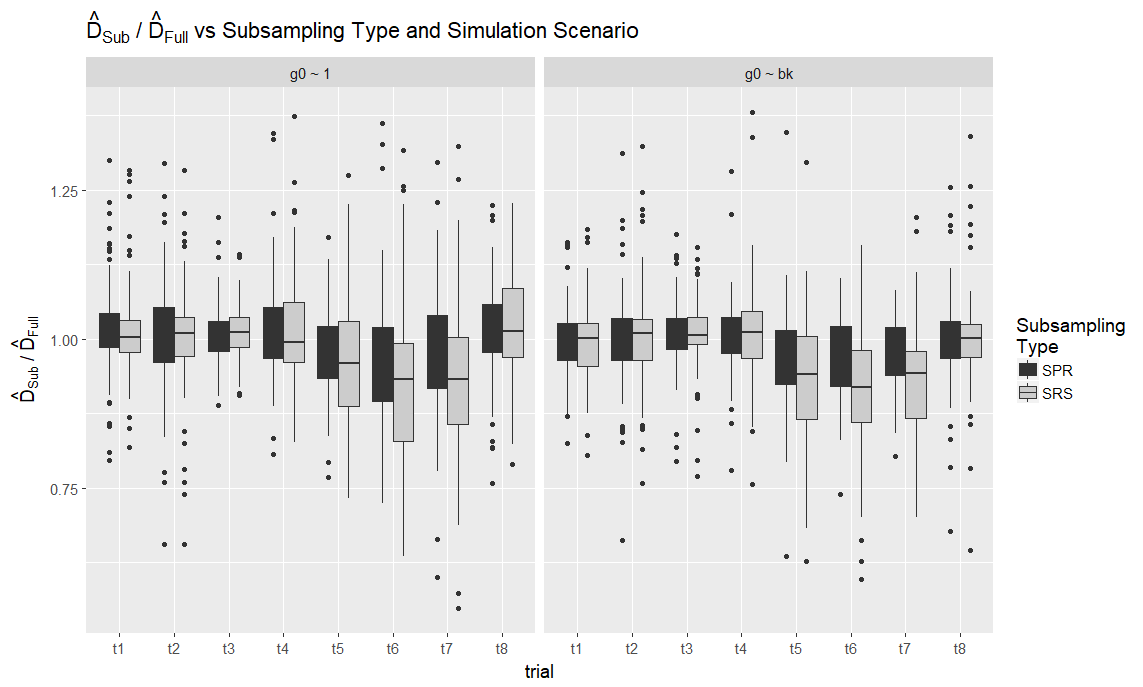
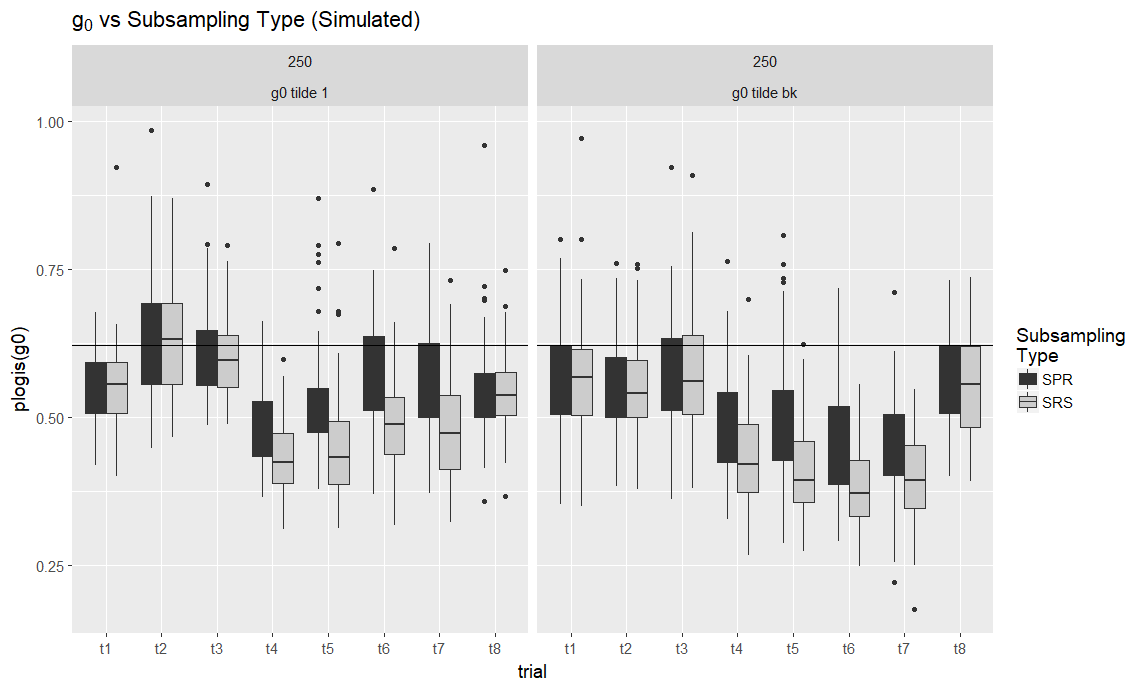
Figure 6A. Ratio of density estimates obtained using subsampled datasets to the estimates obtained using the full dataset for that simulation. 

Figure 6B. Intercept capture probability (g0) obtained from fitted SECR models vs subsampling type and simulation scenario. The horizontal line represents input intercept capture probability.



**Table 1.** Model Notation

|  |  |
| --- | --- |
| Term | Definition |
| *N* | Size of bear population |
| *K* | Number of traps on trapping grid during a single simulation |
| *Dk* | Location of trap on trapping grid during a single simulation{k = 1,...,K} |
| *T* | Number of trapping sessions |
| *Φ* | Inhibition distance between bear activity centers |
| *α* | Parameter describing intensity of stratification of bear activity centers |
| *Ai* | Locations of bear activity centers {i = 1,…,N} |
| *Ψi,k* | Euclidean distances between *Ai* and *Dk* |
| *g0* | Logit capture probability at a given trap for a bear whose activity center is exactly at that trap |
| *bk* | Difference between logit capture and recapture probabilities at a given trap |
| *ϕ* | Inflection point of half-normal distribution which describes capture probability as a function of *Ψ* |
| *δi* | Parameter describing heterogeneity in individual bears’ capture probabilities; this parameter also influences the expected number of redundant samples deposited by an individual. {i = 1,…,N} |
| *Δ* | Standard deviation of normal distribution of *δi* values |
| *γ* | Parameter describing log expected number of samples deposited by an individual bear after being captured. |
| *gi,t,k* | Capture probability for individual i at trap k during time period t{i = 1,…,N},{t = 1,...,T},{k = 1,...,K} |
| *ci,t,k* | Indicator variable equal to 1 if bear i was captured at trap k during time period t, and 0 otherwise. |
| *Ci,t,k* | Indicator variable equal to 1 if bear i was captured at trap k at any time before period t, and 0 otherwise {i = 1,…,N},{t = 1,...,T},{k = 1,...,K} |
| *Ii,t,k* | Number of samples deposited and collected during a simulation {i = 1,…,N},{t = 1,...,T},{k = 1,...,K} |