Spatial genetic mark-recapture methods applied to the Swedish brown bear survey to analyze the problem relating to borders.

Martin Andersson

2022-05-08

Contents

Introduction

The bear survey.

For several years now the population size and trends of brown bears in four regions of Sweden have and continues to be monitored. The total population and how it changes over time is important for conservation efforts and the setting of hunting quotas. If the hunting quotas are too large the survival of the brown bear species could be in jeopardy. On the opposite if it is too low the population could grow large enough to cause problems in the eco-system or for the local human population.

The monitoring and estimation of the bear population is done through the collection of scat samples. The collection is done by volunteers and genotypes are identified through DNA analysis and used to get an estimate of the number of bears from which a scat sample has been obtained. We also expect that for a certain number of bears no samples will be found and as such the population of such bears must be estimated statistically.

The collection of spill samples takes place over 5 years in which spill samples are collected in one region each year in order with the fifth year is an off year when no collection takes place. The samples are collected over 11 weeks in which the volunteers notes down the location of the spill and collects a small piece to send in for DNA analysis.

Region 1 consists of the counties of Gävleborg and Dalarna ,Region 2 is Västerbotten county, Region 3 consists of Västernorrland county and Jämtland, and Region 4 is Norrbotten county.

The total bear population is currently estimated using the Capture-Mark-Recapture method. Traditionally the Capture-Mark-Recapture method involved physically capturing the animals in question but with advancements in DNA analysis it is now possible to perform the analysis by just "capturing" the DNA of the animals instead. This method is sometimes referred to as "Genetic mark-recapture methods." While similar the underlying models at work are quite different. When collecting scat samples for example there is no beforehand known upper limit to the number of "captures" like there might be when using traps.

A thorough explanation of this "genetic mark-recapture" method we use can be found in Miller, Joyce, and Waits (2005).

A full explanation of the methods used in the Swedish survey be found in Kindberg et al. (2011).

A problem with the division of Sweden into regions in this way is that brown bears do not care for these arbitrary borders. A bear could be found on one side of the border one year and the other in another year. Should this happen the bear would be counted twice for the census of the total population in Sweden which introduces bias to the estimation. Another problem with this is deciding which region a bear belongs to. One could assume that each bear has a territory throughout which it wanders regularly. Whichever region contains the largest share of this territory or that contains the territories midpoint could be considered the bears home region. Another method would be counting the bear as the ratio of its territory that lies within each region.

Trying to estimate a bears territory only using the location of scat samples is difficult. For bears which only a single sample has been found you can only get a rough idea of where that bears territory is located. For the bears that no samples were discovered there is no way to assign them a territory.

Another complication with the border problem is how it affects the numbers of samples found for each bear. To estimate the number of bears for which no samples are found the rate at which samples are found needs to be estimated. The rate is assumed to be constant for each bear but due to the way the collection of samples is performed, only samples inside the region currently being sampled can be found. As such a bear that has its territory close to the border is going to have their samples found at a lower rate than one far from the border. This causes bias in the estimation of the rate of samples found and therefore also in the estimation of the number of bears with zero samples found.

The impact of these various sources of bias on the estimate of the bear populations do vary depending on what the actual rate of samples found and the size of the bears territories relative to the regions are. Should

the rate of found samples be very high the number of bears for which no samples are found will be close to zero. If bears have very small territories then the probability that a bears territory will span several regions is also very low.

The method capture-recapture methods that take the positions of animals into account is referred to as spatial capture-recapture and in this case it would specifically be spatial genetic mark-recapture methods. A thorough explanation of why and how spatial methods are useful can be found in Royle, Fuller, and Sutherland (2018).

To analyse how the border problem affects the estimation of the bear population we have performed a simulation study. By simulating a random number of bears over a created region and varying the rate at which samples are found and the size of bears territories the bias can be measured and illustrated. An alternative method for estimating the population by summing the estimated ratio of a bears territory that lies within the region is also analysed to see if the bias can be mitigated.

The simulation values will be chosen based on the values observed for the Swedish survey samples. The population estimates will be performed separately for the male and female bears as we assume the size of their territories to be significantly different. An extra computationally intensive method will also be utilized on the Swedish survey samples as well to get an alternative estimate for the population.

Conclusions

Using the simulation study we observed that the bias in the population estimate scales close to linearly with the size of the bears territories relative to the regions size. The majority of the bias in the estimate was found to be caused by the observation of a large number of bears from outside the region currently being surveyed.

If applied to Swedish samples the relative bias was estimated to lie somewhere between 7% and 17% depending on region and the gender of the bear. However as the estimation of the size of the bears populations and the rate at which samples are discovered were found to be underestimated the actual bias would be assumed to be even larger. The largest region whose population was estimated to be 1033 bears was estimated to on average be overestimated by more than 127 bears using the standard estimation methods.

However as many of the assumptions of the simulation study do not hold for the Swedish regions, the bias is in most likelihood not as severe as the simulation makes it seem. It does however seem likely that the currently employed methods will on average overestimate the population size.

While changing the method for estimation does mitigate the bias somewhat it still causes an on average overestimation of the population size. A better way to fix the bias in the estimation was deemed to instead be changing the method used for the collection of scat samples to include gathering samples a distance outside of the currently surveyed region. That way bears from outside the region could more easily be identified and better estimates of the bears territory sizes and rate of samples observed could be obtained.

Acknowledgments

Method

We will be using R (R Core Team (2021)) as programming language to perform the simulations and analysis. The R package sf (Pebesma (2018)) will be used to create polygonal areas for the spatial part of the analysis. The R package polyCub (Meyer (2019)) will be used to numerically calculate integrals of bivariate normally distributed distribution functions over polygonal areas.

Statistical models and assumptions

For any area S we are performing a survey on, we define the area that lies within distance l from the border of S as O and W as the union of S and O. Both S and O have different bear population sand population densities that are not necessarily the same, however the population density is close to equal along the border between S and O.

In Miller, Joyce, and Waits (2005) a Multinomial distribution is used as a model for the number of samples found for each individual. The idea is that we condition on the total number of samples observed and for each sample assume that each bear has an equal probability of expelling the sample in question. As such the number of samples left by any bear is $Binomial(M, 1/N)$ where M is the total number of samples and N is the total population of bears in the region. Since we condition on the total number of samples found the binomial distribution for each bear is not independent of the other bears. The multinomial distribution is slightly difficult to work with especially as due to spatial reasons the number of samples we expect to observe from each bear is not equal for all bears. Since the observed values for M is relatively large and $1/N$ is relatively small we could approximate the binomial distribution with the poisson distribution $Poisson(M/N) = Poisson(\lambda_0)$. However these Poisson distributions would not be independent of each other as the binomial distributions we are approximating from are not independent. However in McDonald (1980) it is shown that approximating the multinomial distribution with independent Poisson distributions is a reasonable option. As such we choose to use independent Poisson distributions for the number of samples observed for each bear.

We now assume that the number of samples that bear number i leaves is $K_i \sim Poisson(\lambda_0)$. For each bear that has left a sample the j:th sample from the i:th bear has a location that is bivariate normally distributed $N(\mu_i, \sigma I)$ and we let $T_i(\mathbf{x})$ be the probability density function for this distribution. We also use T_i to describe bear i's territory. The amount of time that a bear spends in any area A is $\int_A T_i(\mathbf{x}) \, d\mathbf{x}$. For every sample left by a bear in S it will be observed with probability p. if the sample is not in S then the probability of it being observed is zero. Let $I_i = \int_S T_i d\mathbf{x}$ then the number of samples that bear number i leaves inside S is $K_{Si} \sim Poisson(I_i \lambda_0)$ distributed. Since these samples are only observed with probability p, the distribution for the number of samples observed by bear number i is binomially distributed $Binomial(K_{Si}, p)$, which is a conditional distribution. The binomial distribution where the number of repetitions is conditioned on the outcome of a poisson distributed variable is also poisson distributed. As such $Binomial(K_{Si}, p) \sim Poisson(pI_i \lambda_0)$. As we are not specifically interested in neither p nor λ_0, we can replace $p\lambda_0$ with λ and focus only on the rate at which samples are observed. Therefore the number of observed samples left by bear number i is $O_i \sim Poisson(I_i \lambda)$.

Simulation

For this simulation W is a 9×9 square with centre in origo of the two dimensional Cartesian plane and S is the 2 x 2 square centred around origo. We want S and O to have the same average population density and since S has an area of 4 square units and W has an area of 36 that is 9 times larger we simulate the population of W from a $poisson(9\mu)$ distribution where μ is the mean population size of S.

For the location of bears midpoints in W we choose to use a homogeneous poisson process throughout the plane. As such each simulated bear number i has a territory midpoint $\mu_i = (X_i, Y_i) \sim (Uniform(W))$. Any bear whose midpoint lies within S is marked as a "True Bear" and the rest as "False bears." The total number of "True bears" N is what we are trying to estimate so for each simulation we include a count of the number

of observed "True bears." For each simulated bear we simulate the number of samples they leave from a $Poisson(\lambda)$ distribution. Each simulated sample then has its location simulated from the bivariate normal distribution $N(\mu_i, \sigma I)$. Any sample that is not contained in S is then removed and then any bear with zero remaining samples is also removed and the remaining samples are considered observed. The number of True bears and False Bears for which at least one sample has been observed is then counted.

To see how well we can estimate σ with the current sampling method we will also make an estimation of σ in the simulation. We estimate σ by removing all bears that only have one sample discovered and then using the pooled sample variance method.

λ is estimated using the maximum likelihood estimate for a zero truncated poisson distribution since we cannot observe the bears with zero samples observed.

Let N^{obs} be the number of bears observed, N_T^{obs} be the number of bears observed whose midpoint actually belong to S and N_F^{obs} be the number of bears observed whose midpoint does not belong to S.

Since the probability that an outcome of a $Poisson(\lambda)$ distribution is 1 or larger is $(1 - e^{-\lambda})$ then we have

$$E(\hat{N})(1 - e^{-\lambda}) = E(N^{\text{obs}}).$$

The estimate for the total number of bears in S is then

$$\hat{N} = \frac{N^{\text{obs}}}{1 - e^{-\hat{\lambda}}}.$$

However this is done under the assumption that $N^{\text{obs}} = N_T^{\text{obs}}$. However in our case

$$N^{\text{obs}} = N_T^{\text{obs}} + N_F^{\text{obs}}.$$

Which means that

$$\hat{N} = \frac{N_T^{\text{obs}} + N_F^{\text{obs}}}{1 - e^{-\hat{\lambda}}}.$$

Now the bias in the estimator; $\hat{N} - N$ can be divided into two addends like this.

$$\hat{N} - N = \frac{N_F^{\text{obs}}}{1 - e^{-\hat{\lambda}}} + \left(\frac{N_T^{\text{obs}}}{1 - e^{-\hat{\lambda}}} - N\right).$$

The left addend is the part of the bias introduced by the observation of false bears. The right addend is the bias we would observe if only true bears were observed. With the bias divided like this we can measure how much of the bias is caused by the observation of false bears and how much bias we would have if we could remove all false bears that were observed.

As an alternative method to estimate the population we instead of using N^{obs} in our estimation of N we instead use

$$N_{RatioSum}^{\text{obs}} = \sum_{i=1}^{N^{\text{obs}}} I_i.\text{(Bears indexed in order they were observed)}$$

This being the sum of the estimated ratio of an observed bears territory that lies within the surveyed region. In addition when estimating λ instead of assuming each bear has the same rate of samples found we instead assume each bear has their samples found at a rate of $Poisson(I_i\lambda)$ and use the density function for this distribution when calculating the maximum likelihood estimate for λ.

For the simulation we also need to choose values for μ, λ and σ. Preferably we want to choose these values close to the ones observed in the Swedish samples. A slight complication is that the parameter σ also represents the radius of an area in relation to the area of the region in which it is contained. Therefore we need to normalize σ in some way to make comparisons between regions of different sizes. The solution we choose is to divide σ by the square-root of the area of the region being surveyed. In that way we would for any non-simulation region A theoretically have $\frac{\sigma}{2} = \frac{\sigma_A}{\sqrt{Area(A)}}$. However to not complicate the values used by needing to make transformations of σ anywhere it needs to be used, we instead choose to use "simulation normalized" σ; $\sigma = \frac{2\sigma_A}{\sqrt{Area(A)}}$ when applying the results of the simulation to the real Swedish regions. Figure 1 illustrates the size of the territories in relation to the simulation area for two values of σ.

For λ we choose values 2, 3 and 4 and for σ we use 0.025, 0.05, 0.075, and 0.1. There is little reason to use λ values higher than 4 because in our estimate of the total population size we use the transformation $1/(1 - e^{-\lambda})$ instead of just λ by itself. When λ is larger than 4 the transformations value changes very little as can be seen in Figure 2.

We choose to keep μ at a constant value of 500 as the relative bias in the population estimate does not change with different average populations. A simulation was run to illustrate this and the results can be seen in figure 3.

Application to the Swedish survey

The idea with the alternative model is to take the positions of the bears into account when making the estimations of the underlying parameters. Estimating how much of the territories lies within a certain region requires integrating a bivariate normal distribution function over some irregularly shaped intersections between the territories and the region in question the integrals needs to be performed numerically. These numerical methods will then also be used in combination with an optimizing function to adjust the midpoints of bears since these estimated midpoints are going to be biased to lie inside the surveyed region as we only discover samples there within. The optimizing function causes these calculations to take a very large time when they need to be repeated thousands of times in a simulation to get an estimate for their effectiveness. As such these methods will only be used on the real samples and their effect will need to be trusted based on their soundness.

For the real data we assume that all the assumptions of the simulation are true. Since male bears are known to wander further than female bears and as such would have very different values of σ we choose to estimate their populations separately. σ is estimated again using the pooled variance method. To make comparisons between the real data and the simulated data we also want to get a normalized value for them both. For this we choose to use $\frac{\sigma}{\sqrt{RegionArea}}$. Since the area in the simulation is 4 we can use the value of $\frac{2\sigma}{\sqrt{RegionArea}}$ to make comparisons between the real data and the simulation.

The midpoint coordinates of each bear is estimated as the mean of the samples lateral and longitudal coordinates. The midpoints are going to be biased since all samples are located inside the region so to account for that we will apply an optimizing function to each bear. We find the midpoint μ that optimize the function

$$l(\mu) = \sum_{j=1}^{K_i} log(\phi((x_j - \mu)/\sigma))/ \int_S \phi((x - \mu)/\sigma)\, dx$$

where x_j is the location of the bear in questions sample j and ϕ is the standard bivariate normal distribution. The function applies a penalty to midpoints further from the border which allows midpoints from outside the region in question to be possible outcomes.

After each bears midpoint has been adjusted we can calculate I_i (The integral of the bears territory over the Region) and then get the maximum likelihood estimate for λ where each bears number of samples found is assumed to be a outcome of $Poisson(I_i\lambda)$.
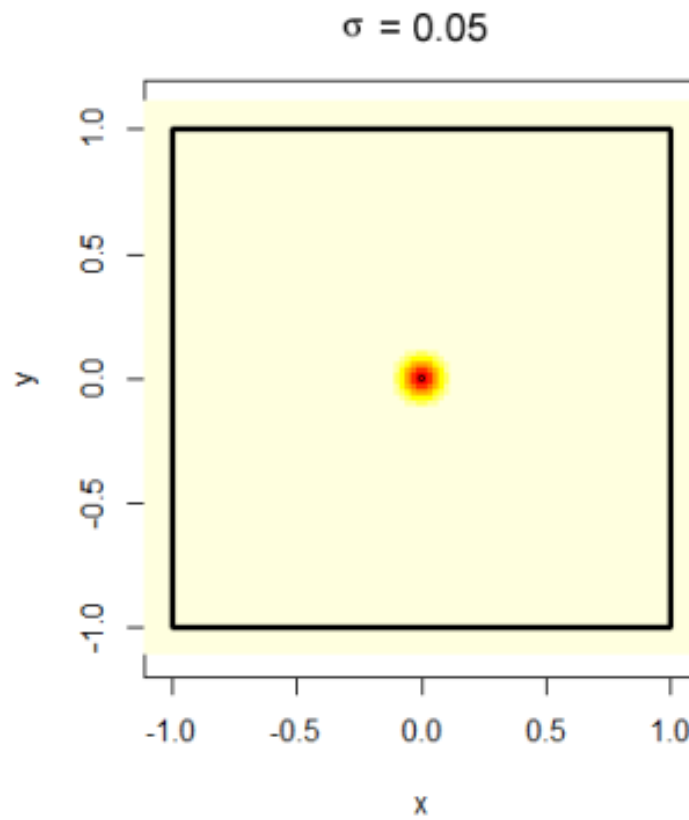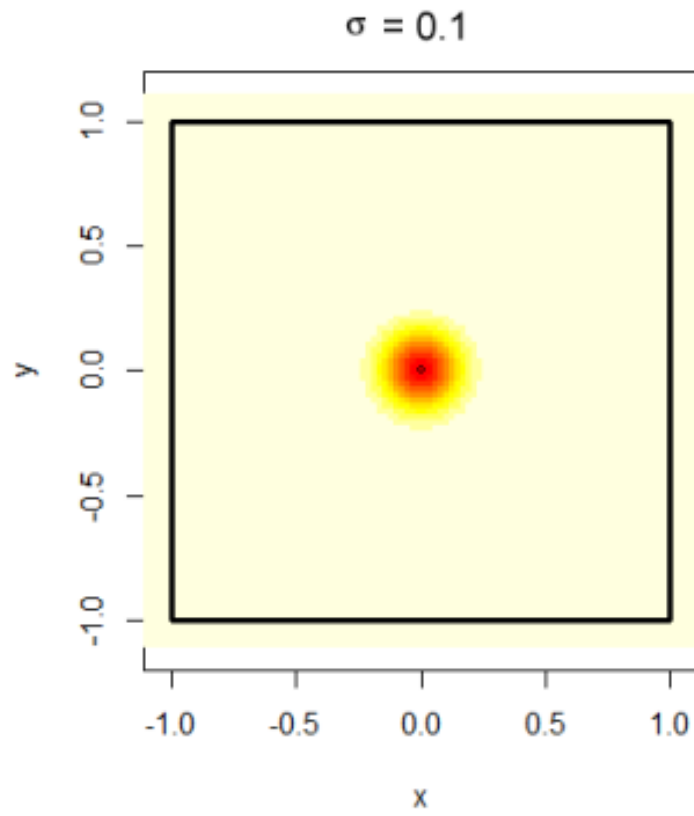
Figure 1: Illustrating the bears territory for two different values of σ. The bear spends more time in the red area than in the yellow.
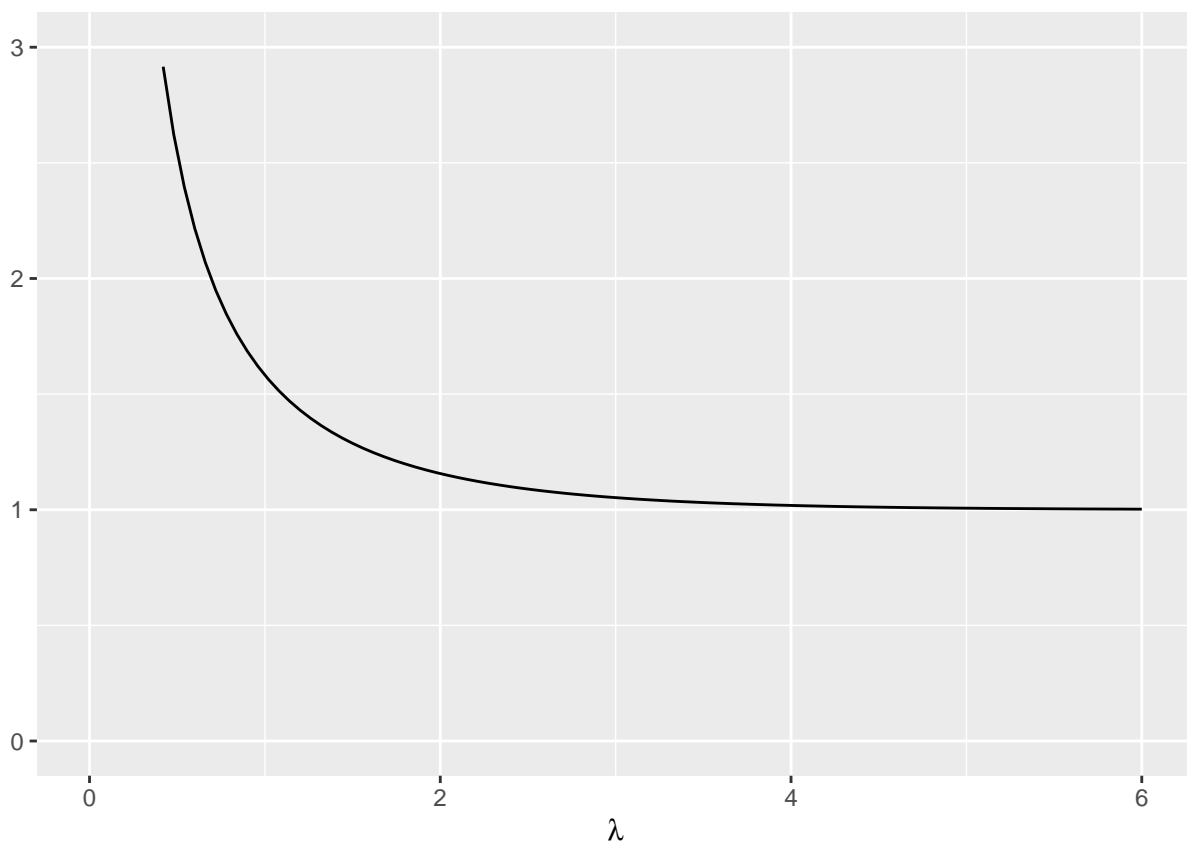
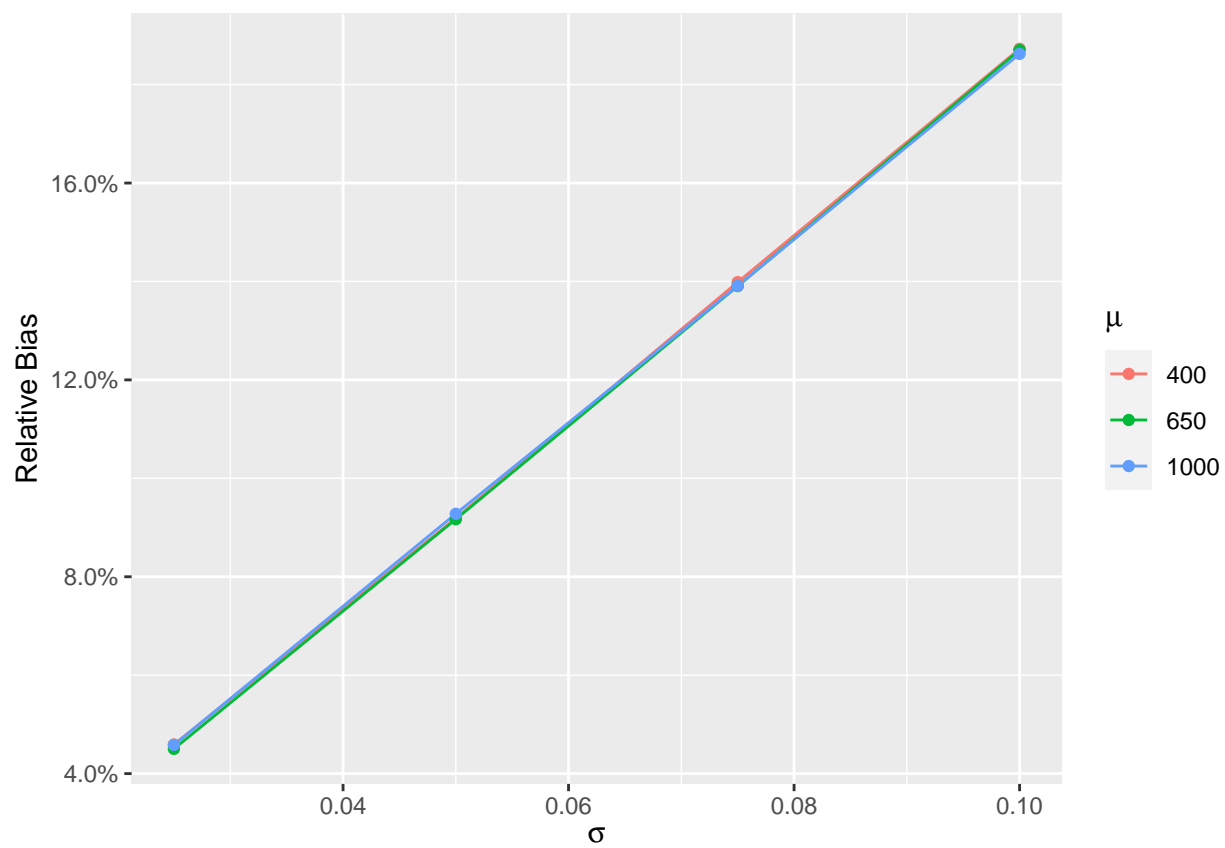Figure 2: Function graph for $\frac{1}{1-e^{-\lambda}}$

Figure 3: Relative bias does not change with total population.

For the estimate of the bear population we use the standard method along with three alternative methods for the estimate. The first is to simply remove any bear whose midpoint was adjusted to lie outside the region and then use $\hat{\lambda}$ to estimate the number of bears for which no samples were discovered. The second is to again using $N_{RatioSum}^{\mathrm{obs}}$ instead of N^{obs}. The third is to subtract the estimated bias from the simulation of the standard method to make a theoretically unbiased estimate of the population.
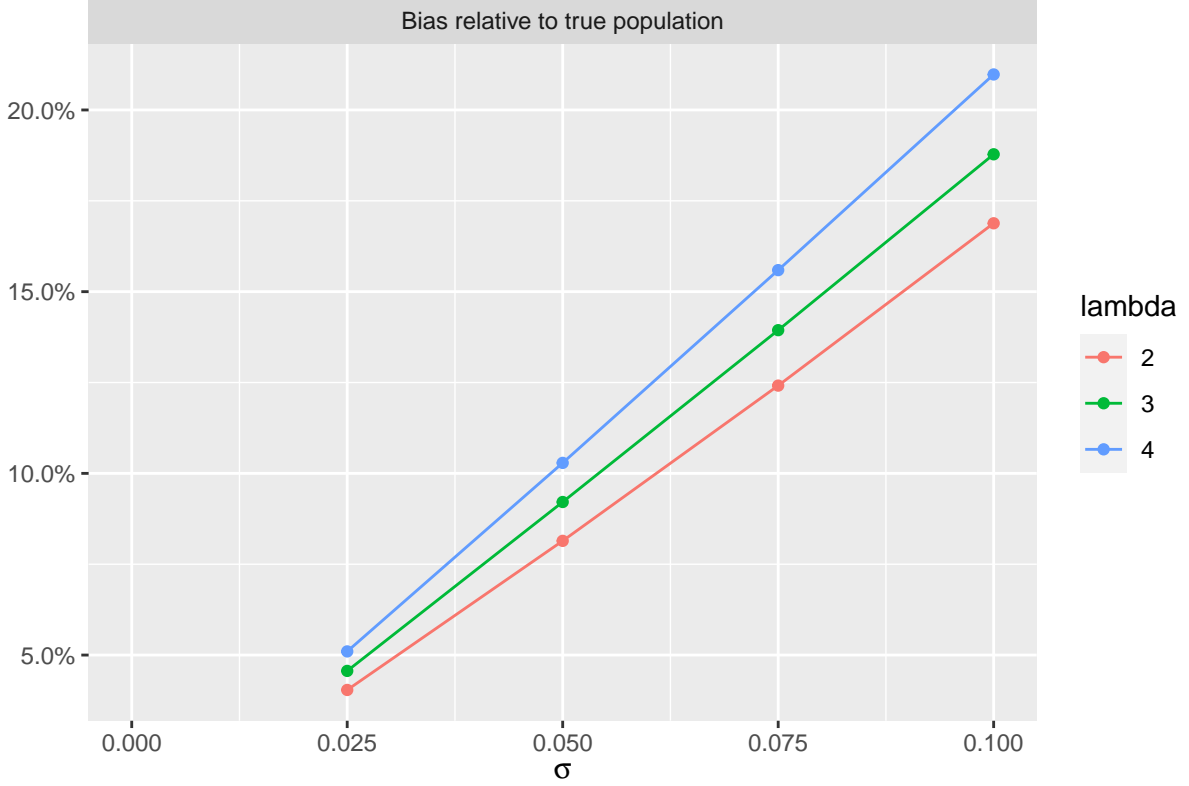
Results

Simulation results

Figure 4: Mean relative bias of population estimate.

Figure 4 shows how the mean bias in $\hat{N}$ grows seemingly linearly as σ (The size of the bears territories) grows. The mean bias grows faster as the value of λ (The rate at which samples are discovered) increases.

$\hat{N} - N = \frac{N_F^{\text{obs}}}{1-e^{-\hat{\lambda}}} + (\frac{N_T^{\text{obs}}}{1-e^{-\hat{\lambda}}} - N).$

Figure 5 shows the bias divided into 2 categories. The lambda bias $\frac{N_T^{\text{obs}}}{1-e^{\hat{\lambda}}} - N$ which is the mean bias in the population estimate we would have if only true bears were observed. The false bear bias is $\frac{N_F^{\text{obs}}}{1-e^{\hat{\lambda}}}$ which is the number of false bears observed in the sample further amplified by the factor we use to estimate the bears for which no samples were discovered.

We can observe how the majority of the bias in population estimate can be attributed to the number of false bears observed and how it grows much faster than the error caused by the bias in the estimate of λ. If all false bears could be eliminated from the sample the remaining relative bias would not exceed 2.5%.

Figure 6 shows the bias when estimating the underlying parameters of the simulation. The methods utilized always underestimate both of the parameters and the bias grows along with σ. As both parameters are biased we cannot directly use the simulated data to get an exact estimation of the mean bias in a sample for which the underlying parameters are unknown.

Figure 7 shows that the ratio sum method for estimating the population is still biased but the rate at which
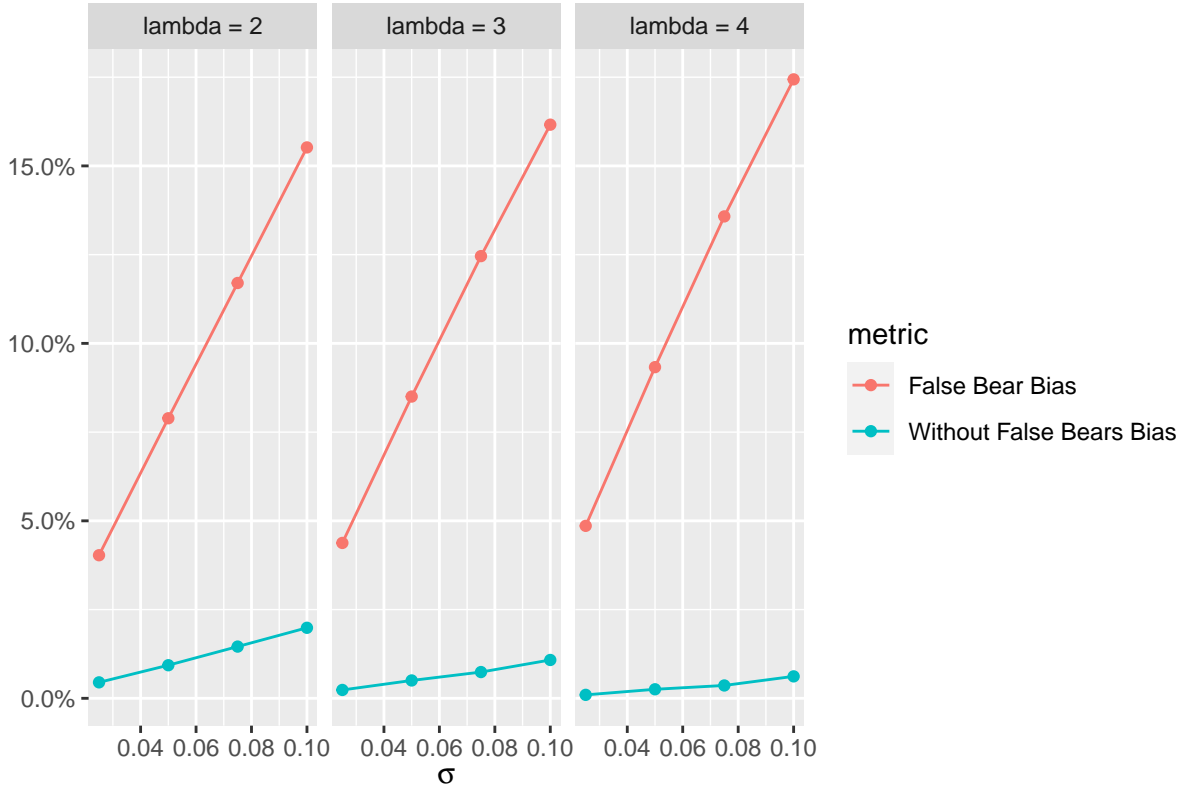
Figure 5: Relative population bias divided into the bias caused by observing false bears; $\frac{N_F^{\mathrm{obs}}}{1-e^{-\hat{\lambda}}}$ and the bias if all the false bears could be removed; $\frac{N_T^{\mathrm{obs}}}{1-e^{-\hat{\lambda}}} - N$.
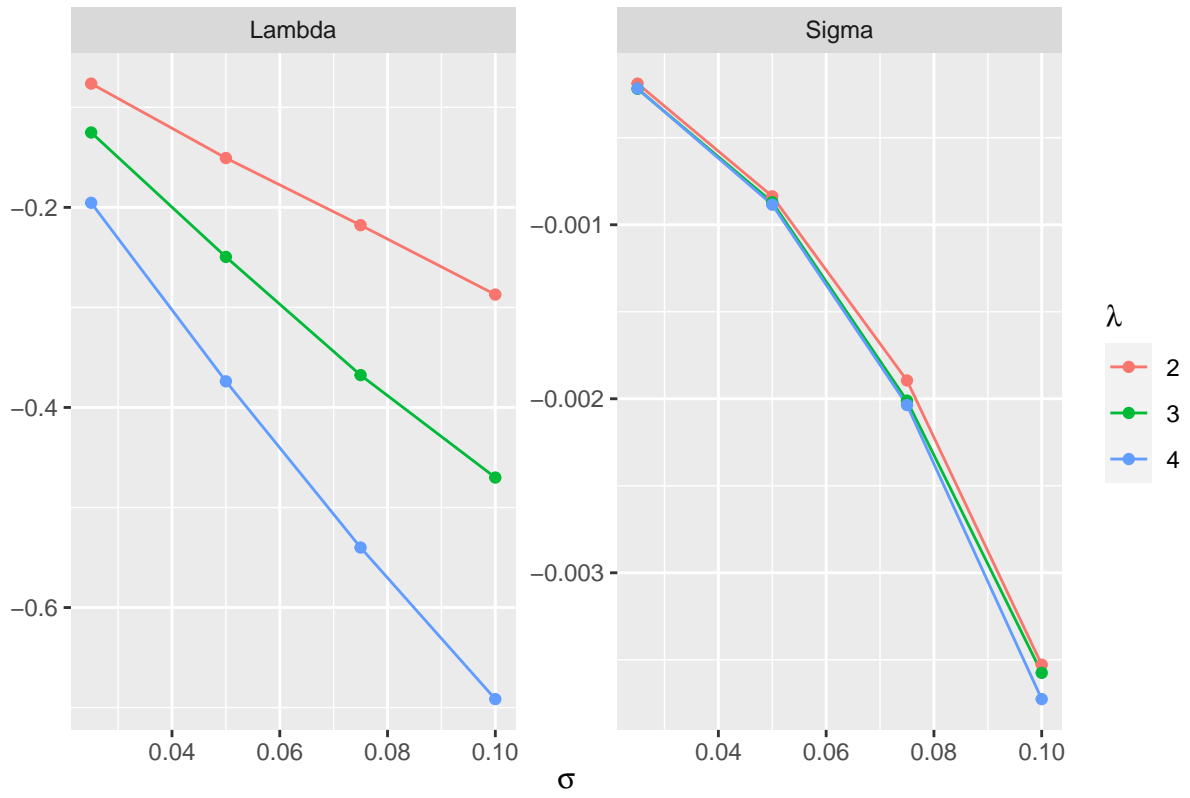
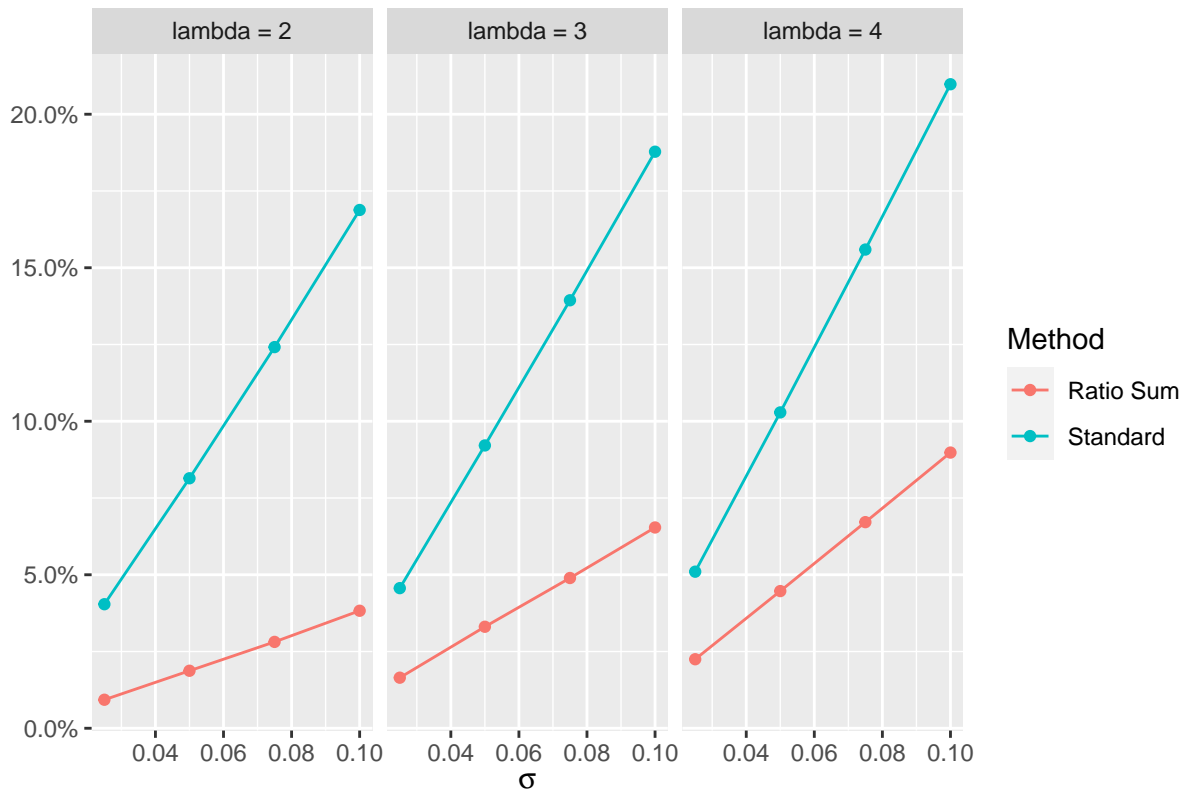Figure 6: Bias in the estimation of the parameters λ and σ.

Figure 7: Comparing relative bias in the population estimate between the standard and the ratio sum method.

it grows is smaller than the standard method for the same parameter values. For small values of σ and λ it is very close to being unbiased.
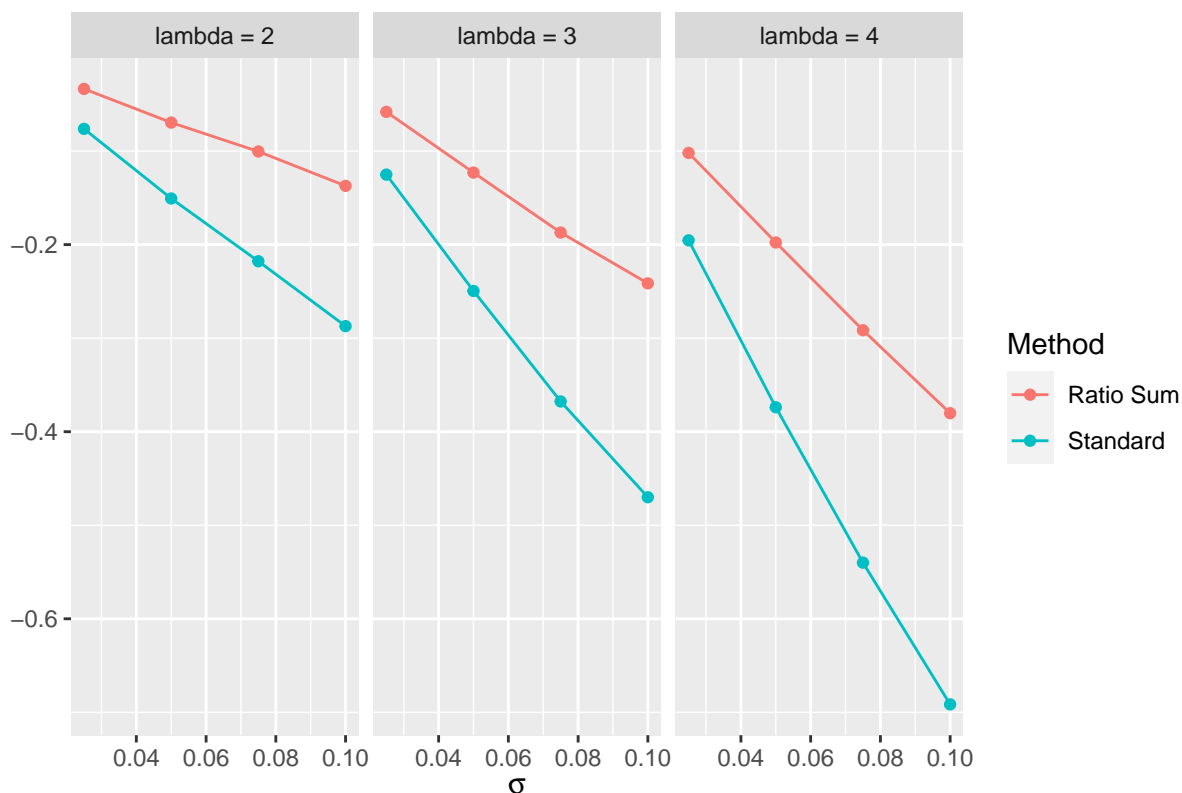

Figure 8: Comparing bias in the estimate of the maximum likelihood estimate of λ between the standard and the ratio sum method.

Figure 8 shows that the ratio sum method is also better for estimating the value of λ than the standard method. Just as in Figure 7 the bias is less than the bias for the standard method for all parameter values.

Swedish survey results

In figure 9 and 10 we can see the estimates of the bear population for 5 years of the Swedish survey. Year 2015 and 2020 were both performed on the same region. The standard estimate is the one that is made without taking the location of the samples into accounts.

The removing outsiders estimate does not differ much from the standard estimate. The ratio sum estimate is noticeably lower than the standard estimate. The simulated unbiased estimate is quite a bit lower than the Ratio Estimate and is much lower than the standard estimate.
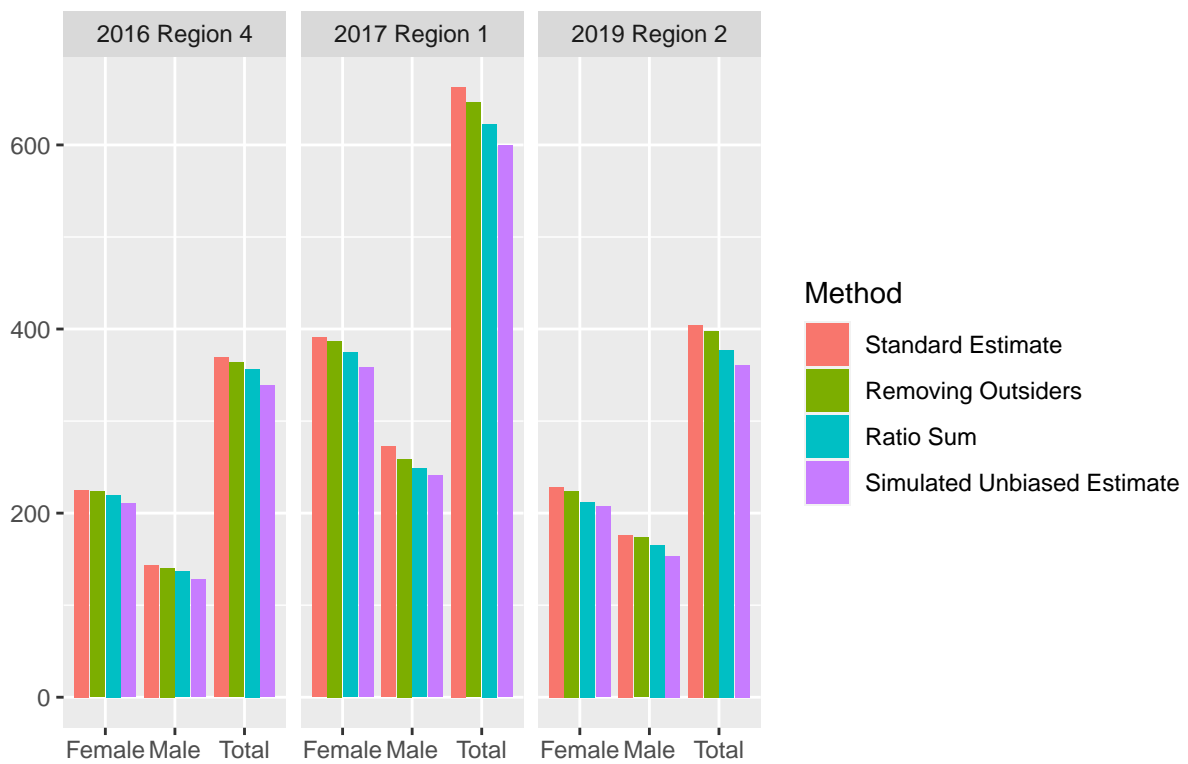
Figure 9: Estimation of the Swedish bear population in region 1, region 2 and region 3 using four different methods.
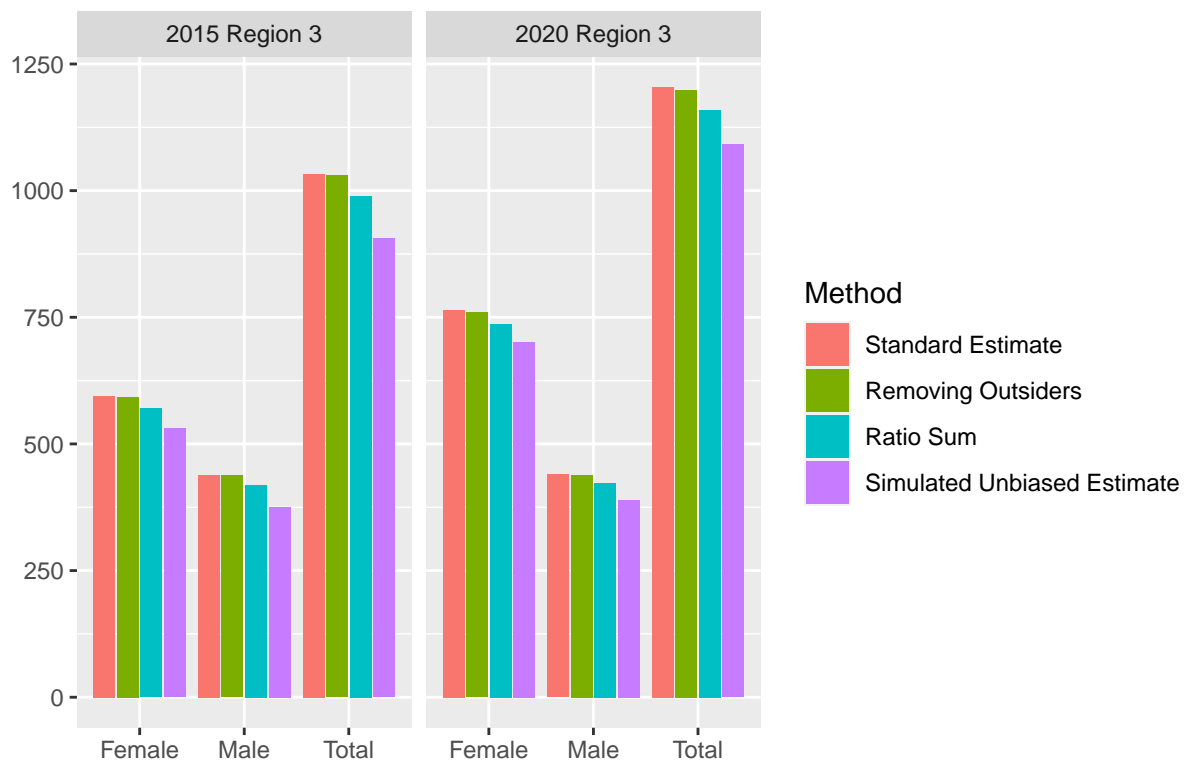
Figure 10: Estimation of the Swedish bear population in Region 3 in the years 2015 and 2016 using four different methods.

Discussion

The bias in the population estimate grows as the parameters of the distribution functions used in the simulation λ (The rate at which samples are found) and σ (The size of the bears territories) grows.

This bias can be attributed to the false bears for which samples have been observed and to the bias in the estimation of λ, However the bias introduced by the false bears vastly exceeds the bias from the estimation of λ.

The mean bias in the population estimate can be seen as a function of the underlying parameters λ and σ that defines the size of a bears territory. As such if we could calculate an expression for this bias function and knew the values of these parameters for a collection of samples we could get an estimate of the bias in that collection. However for a real collection of samples these parameters must also be estimated first which is a problem since the estimates for them are biased and the bias is larger the larger the values for the parameters are. However in addition for this to be a good method the assumptions made for the simulation must hold true.

For each of the 4 regions in Sweden and for each sex separately σ is estimated to lie somewhere in-between 0.043 and 0.092 which from the simulation estimates would suggest that the relative bias lies in-between 7 and 17 percent. However the Swedish regions do differ from the simulation in certain ways. For example part of each Swedish region borders the Baltic Ocean in which brown bears are not known to live. As such the assumption that the population density of the outside region is the same as in the inside region does not hold for any of the Swedish regions. A better approximation of the bias could be achieved by multiplying the bias by the ratio of the border that does not touch the Baltic Sea. Furthermore the bear population is not consistent across the inside of the region either as bears are not known to wander freely throughout larger human settlements.

While the simulation most likely differs quite a lot from the Swedish survey if the spatial part of the collection of samples is ignored then bears from outside the region will be counted twice for the total survey and λ will be underestimated which will cause an on average over-estimation of the regions bear population. By how much is difficult to say and while the ratio sum method could theoretically give better estimations, changing the current surveying methods to better take the border problem into account would be more likely to get better results.

A way to get a better understanding of the border problem would be to extend the search for samples to some distance outside of the region currently being investigated. Not only would we have an easier time identifying bears from outside regions, we could also get a better estimate of the rate at which bear samples are observed by removing bears for which all samples lie very close to the new extended border without losing too much information about the region currently being surveyed.
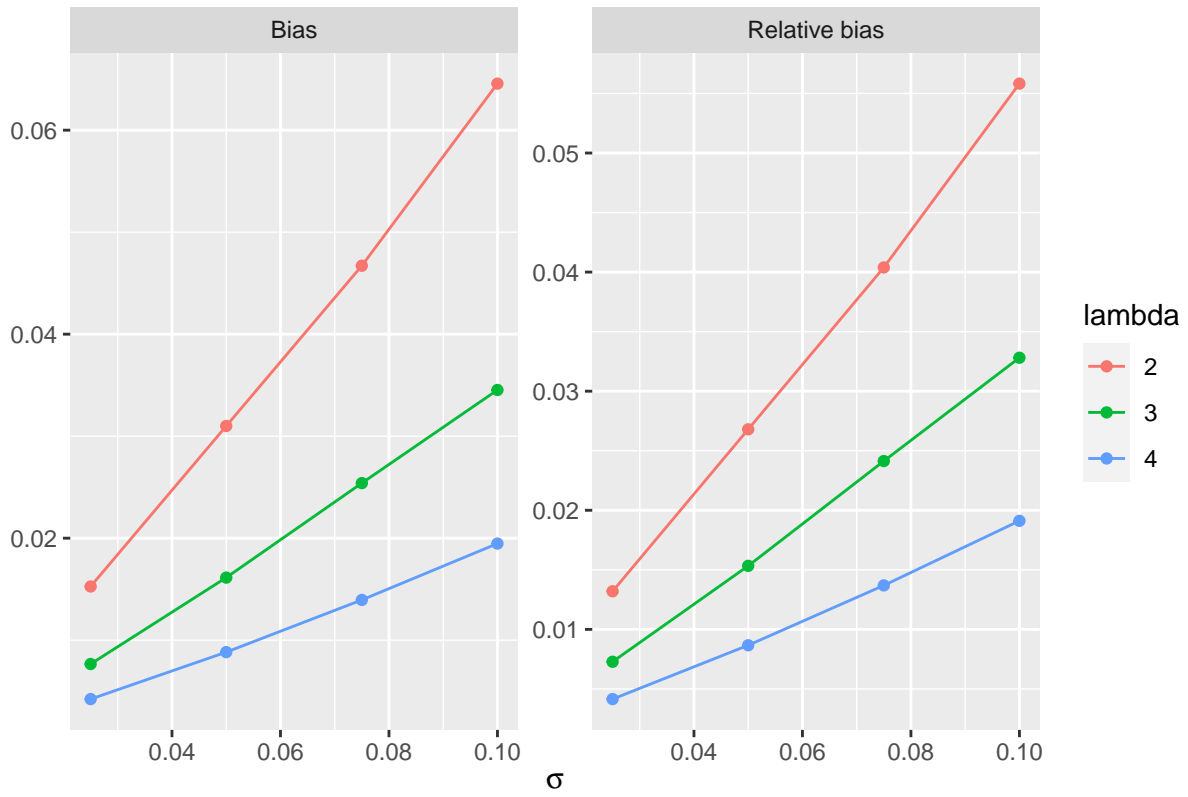
Appendix

Figure 11: Bias in the transformation $\frac{1}{1-e^{-\lambda}}$ of λ which is used in the estimation of the total population $\hat{N} = \frac{N^{\text{obs}}}{1-e^{-\hat{\lambda}}}$.

Table 1: Population and parameter estimates for Sweden.

year	sex	Standard Estimate	Outsiders Removed Estimate	Ratio Sum Estimate	Standard Lambda Estimate	Alternative Lambda Estimate	Simulation Normalized Sigma
2015	Male	438.9666	437.2374	419.2321	4.706508	4.905714	0.0762118
2015	Female	594.2778	592.4432	569.7841	3.801252	3.946863	0.0574793
2015	Total	1033.2444	1029.6805	989.0162			
2016	Male	143.5283	140.1657	136.7642	2.438544	2.529981	0.0654466
2016	Female	225.3331	223.2975	219.6450	2.405327	2.448531	0.0432451
2016	Total	368.8614	363.4631	356.4092			
2017	Male	272.2243	259.0285	248.5888	3.188547	3.356554	0.0669158
2017	Female	390.7397	386.9825	374.3613	3.211857	3.320572	0.0518495
2017	Total	662.9639	646.0110	622.9501			
2019	Male	176.2526	173.7480	165.3290	2.025502	2.125154	0.0915389
2019	Female	228.2444	224.1914	212.1645	1.840133	1.940800	0.0656059
2019	Total	404.4969	397.9395	377.4935			
2020	Male	440.1865	438.4731	422.6845	3.434908	3.559724	0.0657033
2020	Female	764.4728	759.9292	736.3503	3.042619	3.138866	0.0439210
2020	Total	1204.6593	1198.4023	1159.0348			

Table 2: Simulated bias estimate for Sweden

year	sex	Standard Estimate	Simulated Unbiased Estimate	Simulated Relative Bias Estimate	Estimated Simulated Bias
2015	Hane	438.9666	375.1851	0.17	63.78148
2015	Hona	594.2778	530.6052	0.12	63.67262
2015	Total	1033.2444	905.7903		127.45410
2016	Hane	143.5283	128.1502	0.12	15.37803
2016	Hona	225.3331	210.5917	0.07	14.74142
2016	Total	368.8614	338.7420		30.11945
2017	Hane	272.2243	240.9064	0.13	31.31784
2017	Hona	390.7397	358.4768	0.09	32.26291
2017	Total	662.9639	599.3832		63.58075
2019	Hane	176.2526	153.2631	0.15	22.98947
2019	Hona	228.2444	207.4949	0.10	20.74949
2019	Total	404.4969	360.7580		43.73895
2020	Hane	440.1865	389.5456	0.13	50.64092
2020	Hona	764.4728	701.3512	0.09	63.12161
2020	Total	1204.6593	1090.8968		113.76253

Link to github repository containing the code and materials used in this project.

(https://github.com/Martin4188/Bears-Across-Borders)

References

Kindberg, Jonas, Jon E Swenson, Göran Ericsson, Eva Bellemain, Christian Miquel, and Pierre Taberlet. 2011. "Estimating Population Size and Trends of the Swedish Brown Bear Ursus Arctos Population." *Wildlife Biology* 17 (2): 114–23.

McDonald, David R. 1980. "On the Poisson Approximation to the Multinomial Distribution." *Canadian Journal of Statistics* 8 (1): 115–18.

Meyer, Sebastian. 2019. "polyCub: An R Package for Integration over Polygons." *Journal of Open Source Software* 4 (34): 1056. https://doi.org/10.21105/joss.01056.

Miller, Craig R, Paul Joyce, and Lisette P Waits. 2005. "A New Method for Estimating the Size of Small Populations from Genetic Mark–Recapture Data." *Molecular Ecology* 14 (7): 1991–2005.

Pebesma, Edzer. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Royle, J Andrew, Angela K Fuller, and Christopher Sutherland. 2018. "Unifying Population and Landscape Ecology with Spatial Capture–Recapture." *Ecography* 41 (3): 444–56.