MSc Project Proposal

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

1 Background

One active area of research in ecological statistics concerns the methodology for combining information from different data sources. There are many different ways that a population of wild animals can be monitored, though these commonly fall into one of three types:

1. mark-recapture-like studies in which animals can be identified, either from man-made tags or natural marks, so that they can be followed over time

2. point counts or censuses in which it is possible to enumerate the number of individuals detected within a single sample but not to associate individuals across occasions,

3. presence/absence (more accurately detect/non-detect) data in which it is possible only to say whether one or more individuals were detected within a single sample but not to enumerate how many.

Mark-recapture data provides the most information of the three but is often costly or difficult to collect. On the other hand, presence/absence data is often relatively cheap and can produce large data sets but provides less information about the behaviour of individuals and the population as a whole. Combining data from multiple sources has the potential to capitalize on these different advantages, but must be done with care.

Blanc et al. (2014) proposed a model for combining data from the latter two categories: i.e., one data set including counts of individuals and one recording only presence absence. As an example, they consider data from a study of Eurasian lynx in the Jura Mountains, eastern France, collected in two different ways. The first involves an "extensive sign sur-vey carried out by a network of observers (state employees, hunt-ers, naturalists, farmers and mountain guides; Duchampet al.(2012)) who collect evidences of presence in the field (i.e. tracks,scat, hair, sightings, livestock and wildlife killed)" (Blanc et al., 2014, pg. 1735). This returns presence/absence information but provides no information about the number of individuals present if signs are found. The second involves an "intensive survey using the non-invasive monitoring technique of camera trapping conducted since 2011 by the French National Wildlife Agency" (Blanc et al., 2014, pg. 1735). The paper indicates that individuals can be distinguished by the camera traps and so this data source produces mark-recapture type data individual capture histories.

2 Models

2.1 Mark-Recapture

The usual goal of mark-recapture modeling (with a closed population) is to estimate the size of a population. I will let N represent the total population size. The capture histories are usually assumed to follow a multinomial distribution conditional on N, but the details are not important. What is important is that (Blanc et al., 2014) assume that N is a realization from a Poisson distribution with rate parameter λ (or mean $1/\lambda$). The idea is that if the study were to be repeated (whatever that means) then the size of the new population would be N is probability $e^{-\lambda}\lambda^N/N!$.

2.2 Occupancy (presence/absence)

The main quantity that can be estimated from presence/absence data is the occupancy probability (i.e., the probability that at least one individual is present at a site at the specified time). Dynamic occupancy models allow for the occupancy status of a site to change over time, but we will consider simple models in which occupancy is static. In (Blanc et al., 2014, pg. 1735), the study area is divided into 14 10×10 km grid cells, each of which is treated as a site. The analysis ignores times (even though it spans 3-years of data) and so the occupancy of a site is assumed to be a constant, denoted by z_i by site i. A site is known to be occupied if signs are observed at that site, so that $z_i = 1$. Otherwise, it is not known whether the site was not occupied, $z_i = 0$, or the site was occupied and signs were simply absent or missed. The probability that site i is occupied is denoted by $\psi_i = P(z_i = 1)$.

2.3 Combined Model

To link these two types of data, (Blanc et al., 2014) note "the fact that when the probability of a site to be occupied is > 0, the abundance at this site is > 0". This in itself is incorrect as it should really say that "the probability of a site to be occupied" (there is no sense in which a site is occupied is > 0) is equal to the *probability* that "the abundance at this site is > 0". However, I think that this is what they meant to say. Putting this mathematically

$$P(z_i = 1) = P(N_i > 0)$$

where N_i is the size of the population (abundance) at site i. If we assume that N_i follows a Poisson distribution with rate λ_i then

$$P(N_i > 0) = 1 - e^{-\lambda_i}$$

and $P(z_i = 1) = \psi_i$ by definition so that

$$\psi_i = 1 - e^{-\lambda_i}$$

or equivalently

$$\lambda_i = -\log(1 - \psi_i).$$

This, in principal, allows the likelihood functions for both the occupancy and mark-recapture data to be written in terms of a single parameter, λ_i, so that both forms of data provide information about the abundance.

3 Concerns

The equations above are perfectly reasonable, but they're not actually what Blanc et al. (2014) present. The problem is that the occupancy and mark-recapture data are modelled on two separate scales, and it doesn't make sense to equate the parameters when this is the case. The equations provided in the paper are (eqn 1)

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

and (eqn 2)

$$\lambda = -\log(1 - \psi).$$

I include these here mainly to emphasize the fact that there are no subscripts to indicate which quantities relate to the overall area and which relate to the individual sites. We know that N is the overall abundance, but ψ, without a subscript, does not exist in the paper. We might interpret this as the probability that the entire area is occupied, but that would surely be equal to 1 since we know that there are some individuals present somewhere. Alternatively, we can add subscripts to all quantities to obtain the equations above, but this also presents a problem because N_i does not exist – not just in the paper, but it doesn't exist biologically either. Blanc et al. (2014) note that individual lynx move between sites and so there is no logical interpretation of the population size at a particular site and summing these values, whatever they are, would not give you the overall population size because there would be overlap between the sites.

Even more confusing is the fact that Blanc et al. (2014) treat this differently (but still incorrectly) in their implementation of the model. Essentially, their JAGS code which is included in the supplementary materials of the paper equates the probability that there are animals in the overall study area and the probability that each site is occupied:

$$P(z_i = 0) = P(N > 0)$$

or equivalently

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

I believe that the effect of this is to force the estimate of the population size is to force N to be as small as possible. The data of Blanc et al. (2014) contain observations of 9 uniquely identified lynx an their estimate of the total population size was 9.96 (95%CI=9.0,13.0), despite the fact that estimate of the abundance from the camera trapping data alone was 14.46 (95%CI=9.0, 35.0) and the estimated detection probability was $.11 \pm .07$, which would suggest that more than one individual eluded detection. In our work with Mehnaz, we applied these models to estimate the abundance of grizzly bears in a region of Alberta and found that this models estimated a population size of 40 with a standard deviation of 1.56 based on observations of 39 unique bears. Other models estimated the size of the population to be 400 or more.

I'm still not exactly clear on why this is happening, but here's my heuristic reasoning. Consider the equation

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

which is what Blanc et al. (2014) actually use in their JAGS code. This shows that λ is an increasing function of ψ_i, which makes sense. Higher abundance means higher occupancy probability and vice versa. However, even if $\psi_i = .99$ then $\lambda = 4.605$ and if the Poisson model is correct then $P(N > 9|\lambda = 4.605) = .02$, which is very unlikely. The data about occupancy pushes ψ_i down, since it's clear that at least some sites are unoccupied and so $\psi_i < 1$. However, the data on abundance push λ, and hence ψ_i, up. The result is that λ increases just enough to make the observed number of individuals possible, i.e. so that the true abundance is greater than the number of observed individuals, and then stops to avoid dragging ψ_i up to high. The resulting estimate is more about the interplay between λ and ψ_i being on different scales and essentially ignores the data on detection.

Going one layer deeper, the essential problem with the method is that you can't draw inference from spatially indexed mark-recapture data without accounting for animal movements. Abundance at a site isn't well defined if animals move because animals may be present at more than one site.

It's not clear then what site abundance would mean because there is overlap between sites and animals would get counted into the abundance of multiple sites.

There are two exceptions to this in which I think the method could work. These occur at the extremes where either site abundance does have a meaningful interpretation, because animals don't move between sites, or because there is effectively only one site so it does make sense to relate the overall abundance and site (now overall) occupancy probability as Blanc et al. (2014) have done. If you know that the sites are far enough apart that animals can't move between them then you can model both abundance and occupancy on a site by site basis and use the relation

$$\lambda_i = -\log(1 - \psi_i).$$

Alternatively, if you know that animals move through the entire study area so that there is really only one site then you can use the relation

$$\lambda = -\log(1 - \psi),$$

although I conjecture that this adds no information because you know that the area is occupied as soon as one animal is observed. In the middle, which is always the case, you need to know how animals move between the sites in order to relate occupancy and abundance, and this requires a model of the animal movements as given in spatially explicit mark-recapture.

4 Proposed Work

The project I envision is to conduct a simulation study to explore this model and show that the method provides erroneous estimates of abundance except in the two extreme cases. This is important because there are lots of biologists with the skills to copy and implement the JAGS code without understanding what it means. In the best case scenario, the researchers will realize that the estimates are ridiculous and abandon the method (as happened to us). However, some will use the results to draw conclusions about the population and to plan management strategies, and this could have serious consequences if the estimates of abundance are not correct.

References

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. *Journal of Applied Ecology*, 51(6):1733–1739.