**Associate Editor**  
Comments to the Author:  
The authors have provided a new method to deal with an estimation of population size by digital aerial surveys. Please find enclosed the  three reviewers' comments which are generally positive.  
They do provide some suggestions to improve the readibility of the paper to the audience.   
There are some comments on the application to the real example too. I hope you can address them adequately.  it is also important to provide simulation results which can cover a wide spectrum of the situation and it would be even better if they can compare with some existing one in some extreme situation.  
  
**Referee: 1**  
  
Referees comments to the Author. These may be passed without any edits.  
This is a very well-written and interesting article. I too believe that the future of mark-recapture, line transect surveys, distance sampling, etc. on wildlife population will be more or less replaced by digital aerial surveys, drones and other modern technologies. So it makes sense to develop new robust and computationally efficient statistical methods that can handle these types of platforms.  
  
My main concern is that the real data example doesn't completely align with the proposed methodology in its full entirety. Instead, a special case of the model is used. As an applied statistician, I was hoping to see the entire model fitted on some real data. However, I think it's good that the author's have at least acknowledged this limitation at the start of page 15.  
  
Also, I understand that the proposed method is "new" because it uses a full likelihood which leads to some nice modelling advantages; however it seems that the methods of Stevenson et al. (2019, <i>Biometrics</i>) give similar performances (as mentioned on page 20: "The performance of the LCE and CCR estimators is very similar, with the LCE estimator making a slight gain in precision as sample size increases"), and the CCR method can handle larger data sets (as mentioned on page 21: "The CCR method, by contrast, scales well and is able to deal with much larger numbers of detections.").  
  
Minor comments:  
  
- Page 1, first sentence: Could you add a reference for this opening claim?

Done  
  
- Page 5, 3rd line: The Zhang et al. citation is missing "in press".

Now published; new citation inserted.  
  
- Page 5, line 13: I've never heard of a Palm likelihood, is there a reference or some further theory for this?

Added a reference to a paper by Tanaka that contains details of the likelihood. Palm distributions are covered in most books on spatial statistics.  
  
- Page 5, line 16: What do you mean by "true likelihood"?

Changed wording to say “the CCR ``likelihood'' is an approximation to the actual likelihood”, which we hope clarifies the point.  
  
- Page 6, equation (1): I think the support, that is, t > 0 should also be included here?

Included it.  
  
- Page 6, second paragraph of Section 2.2: I feel like a diagram could be added here? It may help readers visualize the movement and availability on the line transect, and to see where the introduced notation fits in.

Added a figure ~~and moved the table with all the simulation results into Supplementary Material~~  
  
- Page 8, line 5: Could the authors expand more on the assumption that the up/down state is independent of the in/out state? Not sure if this was explored in the simulations. Would this really hold in practice?  
  
- Page 10, line 6: I couldn't quite understand what is meant by "there are additional observations on the time delay t between detection by the first and second observers"?  
  
Page 10, line 15: Theta is introduced here but not fully defined. You could add: "see end of Section 4.2" after \theta in introduced?  
  
- Section 4.1: Is a seperate section (Section 4.1) really needed for the homogeneous density? Could this be just be mentioned as a special case after equation (11) is introduced?  
  
- Page 17, line 20: It wasn't clear to me why 100 knots is added here? How is this simulated and why was it set to 100?  
  
- Page 17, last line: L=1100 km seems like a very large distance to travel ... doubling and tripling this amount, even more so. Does this mean the plane or drone will continuously fly for 1100 km taking photos or information? Is this realistic?  
  
- Table 1: There is no CI coverage probability % for CCR? Were there these worse, better or similar cf with LCE?  
  
**Referee: 2**  
  
Referees comments to the Author. These may be passed without any edits.  
This paper presents a new method to estimate density (and other relevant disturbance parameters) when considering feeds of two cameras that have a time lag between them. The methods are illustrated with harbor porpoise derived data and the performance evaluated under simulated data. I suspect these and similar methods will have demand soon as surveying with drones becomes more and more common. The method is contrasted with an earlier method to deal with the same problem by Stephenson et al in Biometrics.  
  
I present below some comments and suggestions that might help to improve the manuscript. I really don’t have any “Fundamental issues”, so there’s only a few (2) “General comments”, which are issues that are rather general in nature (a few of them might mostly reflect me thinking about this for the first time and possibly getting the wrong end of the stick!) and (3) “Specific comments”, which are much more self-contained and mostly only editorial in nature.  
  
General comments  
  
When one reads the abstract one is left wondering…”Because detection of animals from the air is imperfect” – the way I see it, this method will help with availability bias, but not perception bias – or, in other words, what is perception bias in an image? What is the problem you are proposing to solve?  
  
I am a bit confused, as you state “Animal movement” as being the issue in page 1, but surely, if the two cameras are mounted on the same airplane, the speed of the animals would have to be ridiculously high for it to cause issues? What am I missing? You then later state that “In practice, the time delay will need to be sufficiently long relative to the duration of the diving cycle to ensure that the data are adequate to fit the availability model.” – I find it hard to believe that for most whales the time delay on a rear-facing and a forward-facing camera is enough to be on the order of the say tens of seconds successive breaches might take to occur. Ah.. but then you state “Mounting both cameras on the same UAV has the advantage of creating a different viewing  
aspect for the two cameras: an animal that is obscured from one camera by a bush or shadow might be detectable from the other camera.” –I like that idea. Gives one the opportunity to turn some otherwise availability bias into effectively perception bias. So maybe you chose your example wrongly, whales would not have been ideal, but say deer on a low canopy forest perhaps?  
  
You state “Ensuring that the narrow search strips of two UAVs overlap adequately can be difficult in some environments and a cheaper alternative is to mount two cameras on a single aircraft:”. While difficult, it does not seem impossible, and hence this might open the question to what might be the optimal strategy, depending on how likely this is. Two have two cameras on 1 UAV, with the small-time delay that this necessarily leads to, versus two UAV’s traveling with a delay, or even the same UAV getting back on itself every now and then, say. Have you thought about it? You later state “because animal movement in and out of the field of view of the cameras is itself an availability process.” Which seems to imply that the methods could deal also with only partly overlapping strips?  
  
In page 5 you state that “we use a hidden Markov model formulation of the likelihood” – but it seems rather obscure what this means at this point since you’ve not defined what the states might be? It will become obvious later, but still…  
  
I was confused when reading that “Previous literature has devoted substantial attention to each of the problems of availability and uncertain capture histories, but rarely together”, since there’s no point to think about capture histories, either certain or uncertain, if you are not interested in estimating availability. Can you perhaps reword?  
  
In page 6 you say “They did not allow animal movement in the direction of aircraft travel” – given airplane vs animal speed, would that be a big problem for them?   
  
In page 8, The matrix derived from equation 5, U(t)=exp(Qt) does not seem to have rows adding up to 1 (which should be the case for a state transition matrix, as say equation 4)? Is there some normalizing constant missing, or am I missing something?

The exponential in this equation is a matrix exponential, not a matric consisting of each element of Q exponentiated. We have added the words “, where here exp() indicates a matrix exponential function”  
  
Might section 5.2 be an appendix? In some sense, this is just a description of the computational algorithm you used, so it becomes quite distracting while there’s no need to read it to understand the methods/message of the paper?

We have moved this to an appendix.  
  
Figure 1 implies that correlation is higher when animals are moving. I would have assumed intuitively that if animals are moving, hence changing position, that would increase independence, not correlation? While not moving then there’s no possibility of going from in to out or out to in, and hence there are only two relevant states (up or down), and the correlation would not tend to increase – am I being deceived by intuition?

Yes, that intuition is misleading. Suppose animals did not dive at all  
  
A correlation of 0.75 between two estimators means that almost (r^2=0.56) 50% of the variation remains unexplained. That seems a quite large value to me, assuming that both estimators are unbiased… I would assume a larger correlation across estimators – what are the sources for between estimator variability? The fact that you get differences as large as those observed in practice not more than 20% of the times does not fill me with confidence either. What does this mean… one can easily expect to be 20 % off the truth when one chooses one vs. the other when one does not know which is closer to the truth?  
  
I find it a bit confusing in figure 3 (which I find hard to read), but it seems like the mean bias increases for larger samples sizes, in particular for the crosses (CCR)?  
  
In the discussion, you state: “For UAV surveys of non-diving animals, including land surveys, the parameters \tau and  \gamma are not needed,”, but one could come up with an availability process for terrestrial animals too… even without animal movement, a forward-facing camera might not see an animal obstructed by a bush that is visible to a backward-facing camera… will this be able to be accounted for?  
  
Specific comments:  
  
In page 2, I suggest “aircraft's path projected on the ground” would be more accurate than just “aircraft's path”  
  
In page 3, while the example “and birds or amphibians may be available only when vocalizing” is correct, it does not apply in the case of cameras, so since this is the 3rd example, I’d remove it?  
  
In page 4, should “in whether animals detected in similar locations by the two cameras is the same animal” be “in whether animals detected in similar locations by the  
two cameras are the same animal”? Difficult one, the correct would depend on whether there are 2 animals or 1, which we don’t know! If you replace “animals” by “detections” my suggestion works though.  
  
Page 6, something is wrong with “Zhang et al., ress”  
  
In page 9, perhaps better to be explicit and replace “so that only animals in the `up' state can be detected” by “so that only animals in the `up' and `in' state can be detected”  
  
Re notation, choosing \theta for the spatial density parameter is misleading since the entire unknown parameter vector what bold \theta – I think choosing a different letter helps to avoid confusion since if I got it right from equation 12 the bold \theta will have several non \theta parameters within it too.   
  
In page 12 out of 31 you have “For the rest of this paper, we focus on the constant density model with identical detectors and no covariates, which has five parameters” – I’d put the \bold theta there, before the 5 parameter set. Presumably, that’s a good way of reminding the reader what it is, the set of parameters.  
  
I’d had a “typically” or some similar qualifier to the statement “The field of view of a digital camera is such that objects towards the periphery of the image are as easily detected”, as if the cameras were really oblique to increase the timing between detections one could observe considerable distortion at the edges of the image?  
  
The sentence “A CSP is a triple P = (X,D,C)” seems meaningless to me? And what is a simple P?  
  
Equation 16 – what is P\_m?  
  
Figure 2 – I find it awkward that one can’t know which scenario is which? Or is that considered irrelevant?  
  
Reference typos:  
  
In 3 Borchers references, twice “markov” instead of “markov”, and also “poisson” vs “poisson”  
  
Typo in Hamilton et al 2018 – “andgroup size” and “354?362”  
  
In link et al 2010 – “multinomial”, not “multi-nomial”  
  
Pike & … - “sightingsfrom” and “arctic” and “resreport”???  
  
There’s a doi for Zhang et al, aleready more than in press at online early – use it ? <https://doi.org/10.1111/biom.13030>  
  
  
Referee: 3  
  
Referees comments to the Author. These may be passed without any edits.  
The manuscript addresses the important and challenging problem of animal population size estimation using digital aerial surveys. The bottom-up approach for the construction of the state-space models is very interesting, and the proposed estimation methods seems effective and was able to produce consistent estimates of model parameters with those by a different approach reported in the authors' recent work. The presentation of the work is also of high quality. I have only a few comments for the authors to consider in a revision:  
  
1. The model for in-out availability of the  animals was based on a 2-D Brownian motion model for individual animal movement, but the up-down availability is simply assumed to follow a Markov process with up/down times following exponential distributions, in contrast to the inverse Gaussian distributions for the in/out times. This apparent inconsistency of the models for horizontal and vertical movements of the animals seems a bit odd. I understand that it might lead to technical difficulties to use a 3-d Brownian motion (with truncation/reflection at the surface) for animal movement, but is it possible to give some sort of justification to the modelling choice in this paper, or show that the choice does not have material influence on the final estimate of the abundance parameter?  
  
2. An interesting point of the reported approach is that the capture history of the detected animals is not assumed to be known with certainty. However, this makes the form of the data a little ambiguous. In the classical capture-mark-release-recapture experiment, the data effectively consists of 3 numbers: the number of animals captures in the first capture occasion only, that in the second occasion only, and that in both occasions. It seems that the data to be modeled in the current paper consists of the numbers of animals detected by the first and second observers respectively, and the coordinates of these animals. For the sake of clarity, it is perhaps worthwhile to specify the form of the data available somewhere in the paper.