Response to review by Philip Verhagen

Many thanks for your thoughtful feedback. We tried to address the things you mentioned as best we could without making the paper too long, and our response follows the text of your comments.

I will start my review with some feedback on the paper's structure: because of its substantial introduction to the SPACE project in lines 1-68, it is not immediately clear to the reader that the paper in fact mainly discusses a single module that is still under development - the title also seems to imply that the paper reports on the full project. It is of course necessary to provide some background, but the two flowcharts in Figure 1 and 2 make it seem as if there already is a fully developed system, whereas in practice there is a blueprint, and a first module. I would therefore suggest to discuss the outlook of the project in more detail in a discussion section at the end, and to clearly state the aims of the paper in the introduction.

We actually had an earlier version that discussed the whole project in more detail but, in light of comments we had, decided we should focus this paper on the module that was functional, with just more general discussion of the parts that are in development (some of them not very far along at all). Figs. 1 and 2 were our attempt to give just a taste of how we see the whole product functioning eventually.

In the introduction, the term 'survey' is introduced without proper definition. I assume the SPACE project is supposed to cover all types of archaeological survey (field walking, core sampling, test pitting, perhaps even trial trenching?), but the paper is only concerned with field walking, so make sure to keep this consistent in the text.

Although it is true that we use "survey" vaguely in the introduction, we do clarify farther down that we anticipate that the SPACE will eventually deal with a number of kinds of survey, but that for now we have concentrated on fieldwalking aspects.

The motivation for the project is provided by stating that existing knowledge is not used to advantage because it is 'too difficult, makes too little difference, or requires math'. I would love to see more evidence for this, also because the 'too difficult' and 'requiring math' are clearly related. However, this may not be the only issue involved, since archaeologists are more than happy to use other 'difficult' techniques and tools that they don't understand in full detail when they perceive them to be useful…. In that sense, I am curious to know what will be the strategy of the project of achieve its aims - you mention two major target groups in line 203-210, but history shows that it needs a lot of effort to convince people to actually use these tools, so how are you going to approach this?

We're not sure how to respond to this. We certainly haven't done a study of why none of these things have typically been implemented in surveys, so our comments are really only speculation (thus, we use "perhaps"). We agree that it will take a lot of effort to convince people but we also hope that making such tools available in an easy format will help to break down the barriers.

The topic is introduced with a discussion of the mock surveys before explaining the exponential model discussed in lines 110-146. Please add a few lines after line 76 to explain why such an empirical approach is needed.

Although full explanation would take a lot more space, we added a few lines to indicate the importance of knowing the actual locations of artifacts for more accurate assessments of effectiveness.

In lines 89-92 you mention a number of field data collection apps, it would be good to provide links to these.

We added links to these apps in a footnote.

Then, crucially, you fail to clearly explain the concept of the elipses (lines 96-100), while this is central to the approach. Please make sure that this is explained in more detail (why not circles, what is the size of the elipses and why?).

We haven't made much change here, as a full explanation would likely be a distraction through review of psycological literature. However, in this iteration of the software we have actually gone back to using circles, instead of ellipses, as the latter were giving us glitches in the code. So we're using circles until we can fix that. The short answer is that the "target" zones are elliptical because psychological studies have shown that there are differences in people's perception of distance versus angle of observation. We do cite the psychological literature briefly.

The example given in Figure 3 is based on a forensic search case study - is this one that you carried out yourselves? While the principles are of course the same, it would be nicer to show the results for 'real' archaeological artifacts.

We have done this with both archaeological and forensic test fields, and I hope we have now made that clearer. In both cases, we have done the work ourselves (well, different groups within our larger group). We added a sentence to clarify that we have tried this out on both archaeological and forensic artifacts. The later examples (fig. 6) use actual artifacts.

Finally, it was not completely clear to me if the exponential function was chosen on the basis of the data collection results or not. The detection function itself uses the k parameter as a catch-all for effects influencing detection detection probability - I assume that this is calibrated on the basis of the results of the mock surveys in Jordan and Cyprus, and not calculated on the basis of each of the contributing factors?

Our favouring of the exponential curve has been the result both of practical results and theoretical expectations (and common sense). Even before examining the data from the calibrations, it is clear that detection, even right on the line, is usually less than 100% and it makes sense that detection probability should decline with distance. The real question is whether that decline should be as in fig. 4b or fig. 4c, and fig. 4c is both more realistic (in that it doesn't assume 100% detection close to the path) and resembles the actual data better.

The idea to set up an Open Access database for further reference data collection is really good. However, how are you going to convince colleagues to set up these experiments all over the world? Can you be a bit more specific about what such a mock survey set up would entail in terms of effort? Otherwise, you might run into more complaints about the approach not being realistic in practice.

Thanks. We agree that this might be a hard sell, but we hope that by making it relatively easy to do we can gradually get some buy-in. We did add a few lines on how much time it took for us to do the calibration runs in Jordan to show that it was not unreasonably costly, even though more calibrations would have been better in terms of sample size.

Finally, you provide a strategy for sustainability of the tools by relying on GitHub facilities, but long-term Open Access is not necessarily the same as long-term maintenance. In this context it may also be useful to refer to the Dig It, Design It and Dig It, Check It tools developed Amy Mosig Way, since only five years after publication these already seem to be offline (10.1016/j.jasrep.2018.06.034, 10.1016/j.jasrep.2018.07.007, 10.1016/j.dib.2018.08.131). I would appreciate some more of your thoughts on this aspect.

You bring up a really important point here, but we're not really prepared to answer it yet.

All in all, I found this an interesting paper, generally well written, but I think it will profit from extending the discussion on the necessity and implementation of the tools in practice, as well as on the long-term strategy for maintenance and further development. Also, the explanation of the mock survey analysis needs a bit more detail. And, as indicated above, I recommend to adapt the paper's structure somewhat.

In light of this and the other reviewer's comments, we did make some very minor structural change (mostly just moving up the "why do it" section to the end of the introductory sections).

Response to review by Tymon de Haas

Again, many thanks for the thoughtful suggestions and corrections.

We give our responses next to your text.

1) in the introduction it is stated that archaeologists have for various reasons not used the available statistical theory to aid designing surveys. it would be useful to explain/argue why this is problematic (financially, logistically, scientifically, ...). perhaps by reference to some practical examples of high cost surveys, cases where aims could not be fully realised, ...? I note that the very short Why do this section could actually be embedded in the introduction to strengthen the rationale there.

   As for the previous review, we agree that this is an issue, but we aren't really prepared at the moment to deal with it in any detail. We can only speculate. However, we did move the "why do it" section up near the end of the introductory sections and modified it slightly.

2) where it comes to quality control over survey data and more realistic approximation of artefact density estimations, the discussion of the Sweep width module is convincing. It does not yet convincingly argue the suggested use of the module for adapting the general survey strategy: isn't the kind of archaeology one will map, especially in surveys that are primarily interested in ceramic sites (ADABS, POSIs, ...) more dependent on the interwalker distance than on the sweep width?

   Actually, one of the purposes of the sweep width estimate is to provide a rationale for using a particular transect spacing, while another is to assess the thoroughness of a suvey done with some transect spacing. We did mention this (currently lines 183-84).

   And if the module will propose a sweep width rounded of to Meters (if I read lines 136/137 correctly), the question also is whether the variations in sweep width are such as to really significantly deviate from current practices (which assume a sweep width of, say, 2 meters).

   Sorry. We did not mean to give that impression. We only stated that we measure the sweep width in meters, not that we round it to the nearest meter. Often, we have sweep widths of 0.8 or 1.5 or 3.5 m. As mentioned above, we would argue that this supplements current practices by providing a basis for deciding what interval to use instead of arbitrarily using 2 m or 10 m, for example.

   In order to clarify these points and/or to better 'sell' the benefits I would suggest to elaborate further on these points and perhaps illustrate the use of the module/through discussion of one or two examples from the author's work test on Cyprus and in Jordan: how have the actual survey practices (sweep widths/walker spacings) been modified based on the calibrations?

   I hope we have now clarified some of this, although we did not want to include a detailed example from Cyprus or Jordan as we are using those in another paper on the Bayesian allocation algorithm.

3) a practical point: for surveyors to adopt the calibration procedure much will depend on the necessary time investment in relation to the time available for the survey as a whole. How much time did the actual calibration tests take in the field?

   This is an excellent point. Without going into great detail, we added a few lines to indicate that our time investment in this was not that onerous (e.g., six 2 to 3-hour sessions spread over a field season). These allowed multiple "runs" so that we could get a decent sample size for the curve-fitting.

4) The suggestion to have a database of calibration data (lines 170-176) available will probably for many be the preferred (less time-consuming) option. Do you foresee that with more of such data you could provide a set of more general recommendations regarding Sweep width in relation to typical survey conditions?

It is actually not ideal to use other people's calibrations, but we mention that option because we recognize that some projects will be reluctant to invest the time to do their own. It is not technically possible to generalize very much because the myriad factors that affect detectability, and visibility and crew composition, alone, will cause a lot of variation. However, at least with our own crews, we have found that differences in sweep widths for different kinds of fields or artifacts pretty much match our expectations, and one of us has actually applied our sweep widths, provisionally, to a different project in Jordan where the visibility was somewhat similar (although the crews were different).

a few other points:

-should the calibration besides accounting for variations in land use/visibility not also use artefact density as a variable, as overall density will variably affect the capacity of wlakers to obsereve different kinds of artefacts?

We're glad you noticed that. And you're quite right. However, so far we have not had time to test for the effect of artifact clustering on the shapes of the detection functions. Intuitively, we think that more clustered artifacts will be found more often than dispersed ones. We plan to investigate this in future.

-The SPACE project needs a slightly more elaborate introduction: explain in the text what the acronym stands for, who are participating, for how long does the project run, ...

We added a few lines toward the end of the introductory sections on the location, participants and funding of the project, although we are unsure that it belongs where we put it. Possibly it should go into an acknowledgements or a footnote.

-fig 3 heading (2. locations....) should be removed and caption reproduces sections of text; fig 6 caption refers to "c" that is not included.

Thanks for catching that. We reduced the redunancy in fig. 3's caption and made the corrections to fig. 6 (originally we had included 3 graphs in that figure but reduced to 2 to make it less cluttered).