

# PhD plan and Research Proposal

Nathan Khadaroo - 8520000

## Introduction / General theme:

This document serves as an outline for the PhD. It is generated using an R Markdown script which can be found here, alongside all previous versions of the document.

The general theme of the PhD is to explore the use of statistical modeling to address bias in citizen science generated data in the social sciences.

I will first provide a non-exhaustive review of the literature, I will then outline some proposals for projects/papers that would constitute the substantive of the PhD thesis.

## Literature review:

In this section I provide a brief non-exhaustive review of the literature on bias in citizen science data sets.

Bird, et al (2014) provide a detailed overview of statistical solutions to issues...

as well as a list of available R packages for their implementation.

Senaratne, et al (2016) provide an overview of quality assessment methods for volunteered geographic information (a type of citizen science data which is highly prevalent in social science applications).

Aceves-Bueno et al (2017) provide a quantitative review of papers comparing citizen science data to some reference data. Specifically they seek to compare these papers own qualitative evaluation of the accuracy of the citizen science data and quantitative assessments, finding that authors can be overly optimistic in their qualitative assessments of their data. Furthermore, they find that what authors consider to be sufficient accuracy varies on a number of factors. Similarly to Senaratne, et al (2016), they provide a list of metrics used to assess accuracy.

## Project proposal ideas:

**Systematic review:** The first project I plan to undertake is a systematic review of citizen science projects which use statistical techniques to address bias in the data collection process. This could potentially also be my masters dissertation.

There is no existing systematic review of the ways statistically modeling has been applied to citizen science data sets across fields. This would therefore constitute a meaningful contribution to the literature.

Prior review of bias in citizen science data so exist, for example Bird, et al (2014). However they are field specific (in this case biology), and are non systematic. This exposes the findings to a higher risk of bias, and prevents the authors from making quantitative statements about the relative prevalence of various approaches.

Key aspects of studies which would have to be recorded would be the type of bias which is being addressed (Measurement error, representativeness, bias, various types of clustering, lack of absence data, etc), what modeling approach is being used (generalized linear models, hierarchical models (bayesian vs classical), additive models, geographically weighted regressing models, occupancy models (technically glm's), etc).

It would also be useful to collect information on openness (whether the data and code are available, if the data was collected using a mobile app, was the software open source and does it have plans for sustainability, etc). A possible template for this would be the framework used in Ostermann and Granell (2017), who review 58 papers on the use of volunteered geographic information in the crisis management field and evaluate .

Before starting the review I will pre-register by plans on the Open Science Framework(OSF).

**Evaluating solutions:** This project provides the core empirical contribution of the thesis. It would necessarily take place after, and build upon, the systematic review, and would involve evaluating the performance of various solutions to bias in citizen science data in various situations.

A potential approach would be a simulation study using an agent based model to evaluate performance of various modeling approaches under different assumptions about underlying true distribution of the data, the clustering of observers, their accuracy, the type of data collection (for example presence only data vs presence/absence data) etc.

This could be used to look at a number of questions such as:

- Do some modeling approaches work better when the target population is generated in certain ways (for example, scarce data vs abundant, etc (obviously yes, but do some model do well over many types. . .)?)
- Are certain modeling approaches useful only when the observation procedure follows different patterns. For example contributions being highly clustered (often as a power law) by contributors, tasks where high accuracy is widespread or not, different distributions of accuracy et cetera.
- How do models account for heterogeneity in contributors? Could bayesian hierarchical models use super-contributors, who are known to be more accurate, to generate informative priors for contributors who are less active?

**Developing a unified framework:**

**Training needs:**

In this section I outline skills which I would need to acquire to accomplish this project.

- A good understanding of a framework (Bayesian?) with which to evaluate and compare various approaches. A highly recommended resource is McElreath (2020). This is in R and Stan, though code for the book is also available in Julia which could be good practice if I choose to code the agent based model in Julia.
- Systematic review training, the university does run some of these (the most recent unfortunately clashed with Sheffield) I think based on Prisma and Cochrane approach. I will also look into NIRO.
- Occupancy modeling *appears* to be the most prominent approach in species monitoring, it could be fruitful to audit a biology module on this, though I can't see any. Additionally MacKenzie et al (2017) looks like a useful overview.
- Small area estimation is quite common with crowd-sourced data (could occupancy be considered a special case of this??). I understand there is a NCRM course on the r sae package but I have been told it might not be worth attending, a colleague who attended previously has kindly sent me the slides.

## Elective module choice:

It is fairly difficult to get detailed information on most courses.

Some potential takes could be:

- Longitudinal data analysis with Alex Cernat.
- Possibly Complex Survey Designs and Analysis though I'm not sure of the software used, and the MARD module by the same lecture was underwhelming and seemed outdated (the focus was mainly on the now quite stale debate between qualitative and quantitative approaches to social research, no mention of currently active debates such as those surrounding causal inference or reproducibility).

## Ethics:

I am committed to making all the output from the PhD freely available and as reproducible as possible. This involves exclusively using scripted open-sourced software for analysis, making data used open whenever possible. I also aim to pre-register any analysis I will be undertaking (including the systematic review) on my Open Science Framework page.

## Bibliography:

Aceves-Bueno, E., Adeleye, A. S., Feraud, M., Huang, Y., Tao, M., Yang, Y., & Anderson, S. E. (2017). The accuracy of citizen science data: a quantitative review. *Bulletin of the Ecological Society of America*, 98(4), 278-290.

Bird, T.J., Bates, A.E., Lefcheck, J.S., Hill, N.A., Thomson, R.J., Edgar, G.J., Stuart-Smith, R.D., Wotherpoon, S., Krkosek, M., Stuart-Smith, J.F. and Pecl, G.T. (2014). Statistical solutions for error and bias in global citizen science datasets. *Biological Conservation*, 173, 144-154.

Elliott, K. C., & Rosenberg, J. (2019). Philosophical foundations for citizen science. *Citizen Science: Theory and Practice*, 4(1).

MacKenzie, D. I., Nichols, J. D., Royle, J. A., Pollock, K. H., Bailey, L., & Hines, J. E. (2017). *Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence*. Elsevier.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Ostermann, F. O., & Granell, C. (2017). Advancing science with VGI: Reproducibility and replicability of recent studies using VGI. *Transactions in GIS*, 21(2), 224-237.

Pickering, J.S., Topor, M., et al (2020). Non-Interventional, Reproducible, and Open(NIRO) Systematic Review Guidelines v0.1

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1), 139-167.

Wijewardhana, U. A., Meyer, D., & Jayawardana, M. (2020). Statistical models for the persistence of threatened birds using citizen science data: A systematic review. *Global Ecology and Conservation*, 21, e00821.

# Appendix:

PhD plan (so far!)

