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 code for the Gantt chart, and a pdf of the top 100 abstracts on citizen science.

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. This involves outlining the key types of bias, the ways in which they occur, the ways in which they are addressed, and the trade-offs inherent in various approaches to address bias.

Some ideas for the title of the thesis are:

- "Crowd-sourced data for citizen social science"
- "Statistical modeling of bias in citizen science data"
- "Hybrid intelligence in the social sciences: applications of machine learning to bias in crowd-sourced data"

I will first provide a non-exhaustive review of the literature on bias in citizen science data, I will then outline three proposals for projects/papers that would constitute the substantive of the PhD thesis. I will discuss the motivation for each of these projects as well as the skills I would need to acquire to pull them off. I will briefly discuss ethics. The appendix contains a provisional Gantt chart for the PhD. A larger version can be found here.

Litterature review:

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

Citizen science is an increasingly popular approach to scientific inquiry, however a number of concerns have been raised about accuracy in citizen science data. For example, Riesch & Potter(2014) conducted qualitative interviews with "scientists who participated in the 'OPAL' portfolio of citizen science projects that has been running in England since 2007", finding that issues around data quality are "almost universally recognized as one of the problems that scientists working in CS need to address" (p 112).

Elliott & Rosenberg (2019), whilst acknowledging the concerns mainly scientists have about citizen science data, notes the existence of a substantive literature in philosophy of science arguing that the quality of data should be evaluated in terms of the purposes for which they are being used, and that "empirical evidence suggests that the quality of citizen science data has often been sufficient for the projects being pursued" (p 3). They further argue that there are a number of ways in which citizen science data-sets can improve their accuracy, such as training, aggregation or statistical modeling (for example, weighting contributions depending on how long the contributor has been active, as volunteers are known to improve accuracy over time).

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

Aceves-Buenoet 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.

Bird, et al (2014) provide a detailed overview of statistical solutions to issues of bias as well as a list of available R packages for their implementation. They provide valuable contextualization of bias in citizen science data, emphasizing three main elements: 1) Types of response data, which refers to the issue of presence only data (for example, FixMyStreet users might use the app after driving over a pothole, but not after a smooth journey to signal the lack of potholes), 2) random error and 3) bias.

They then provide an overview of various broad families of approaches that have been used such as linear and generalized linear models and extensions, mixed-effects models, hierarchical models, machine learning, and species distribution models.

They finish on an optimistic note, anticipating the future "development of novel statistical approaches and survey designs that will break new ground in overcoming some of the problems we have outlined in this paper."

Altwegg & Nichols(2019) look at the use of occupancy models as a means of modeling bias resulting from issues common in app based citizen science data such as "heterogeneous and non-random sampling, false absences, false detections, and spatial correlations in the data."

Similarily, Johnston, et al (2020) use occupancy modeling to spatially biased citizen science data in Great Britain. They noted that whilst the modeling could provide accurate and precise estimates in some areas, there were areas with few observations where this was not the case, with the modeling approach improving accuracy, but not precision. They emphasize that estimation from "spatially biased data should be further validated and tested under a range of different scenarios".

In a pre-print, Johnston, et al (2019) provide a guide to best practices for making inference using citizen science data. They argue that that the collection process must be accounted for explicitly, finding the greatest improvements in accuracy occured from "1) the use of complete checklists rather than presence-only data, and 2) the use of covariates describing variation in effort and detectability for each checklist"

Weiser et al (2020) show that it is possible to obtain unbiased population estimates from citizen science data where the citizen scientists selected the areas they gathered data themselves (non-probaility sites), even in the absence of covariates which explain differences between sites. However they also found that more non-probability sites could require more probability based sites to fully correct for bias. As above, they emphasize the need for further research.

Project proposal:

In this section I outline the key components of my proposed thesis. As each of the three main components are intended to build upon knowledge acquired in previous components, the plans for later components are naturally less detailed at this point.

The first component is a systematic review of citizen science projects which use statistical modeling to address bias. The aim is to gain a deep understanding of which biases are common (and commonly considered a problem), which approaches have been used to mitigate these problems, and to gain a quantitative understanding of the prevalence of various approaches. The understanding gained in this step should help inform the rest of the PhD.

The second component of the thesis is a simulation study evaluate the performance of the various modeling approaches under a number of scenarios using an agent based model to generate observation data. The aim here is to understand the strengths and weaknesses of various approaches in an objective way.

The final component is to propose a united framework for best practice in modeling bias in citizen social science.

Systematic review:

Objective and motivation: 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. It would also allow me to delve deep into the literature and should inform later projects.

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 do not follow a systematic approach. 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.

How I will do it: I will begin by looking at various protocols.

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses framework or PRISMA is often considered the gold standard, however there are concerns that its focus on the synthesis of trial evidence means it can be ill-adapted for reviews of non-interventional research.

An alternative would be the new Non-Interventional, Reproducible, and Open Systematic Reviews framework or NIRO, which emphasizes reproducibility and is designed with non-interventional reviews in mind.

Once I have settled on a protocol to follow, I will perform some initial scoping. Then, and critically before starting the final search for the review, I will pre-register my methodology and intentions on the Open Science Framework.

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 .

Pre-requisites:

- Studying PRISMA and NIRO.
- Registering to the next university two-day workshop on systematic reviews.
- Re-watching Prof Helen Worthington's methods@manchester video 'Intro to systematic reviews'.
- Look into what resources and support I can get from he library, who I see have recently created a new web page dedicated to systematic reviews.

Evaluating solutions:

Objective and motivation: This project provides a large part of the 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 (as identified in the systematic review) under various scenarios.

How I will do it: 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 (both in quantity and quality)? Could bayesian hierarchical models use super-contributors, who are known to be more accurate, to generate informative priors for contributors who are less active/precise?

Pre-requisites:

- Gaining an understanding of the modeling approaches identified in the systematic reviews.
- Learning about what is best practice for running simulation studies.
- Gaining a more in-depth understanding of how to use agent based models to simulate the data collection procedure.
- A good understanding of a (Bayesian?) framework 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).
- 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.

Developing a unified framework for modelling bias in citizen social science:

Objective and motivation: This component seeks to develop a unified framework for addressing bias in citizen science data. The shape of this project will take shape as a result of the previous two projects.

How I will do it: This aspect of the PhD is likely to be be produced in close collaboration with Open Data Manchester, as well as the university library who have expressed interest in best practices for citizen science.

Pre-requisites:

- A good understanding of existing "best practice" frameworks.
- Paying close attention to ongoing debates on bias in citizen science.

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 (Alex is fantastic and most citizen science data has a longitudinal aspect).
- 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 qualitative approaches to social research, no mention of currently active debates such as those surrounding causal inference or reproducibility).
- A good open-sourced modules on geographical data science as most citizen science data has a geospatial component. I signed up to Robin Lovelace's workshop earlier this year but it was canceled due to covid. As I understand the Manchester module is in ArcGis which I'd rather not use. (Perhaps there is a possibility of attending a module in Leeds if they are online?)
- Is there a module on bayesian inference?

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., Wotherspoon, 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).

Johnston, A., W. M. Hochachka, M. E. Strimas-Mackey, V. Ruiz Gutierrez, O. J. Robinson, E. T. Miller, T. Auer, S. T. Kelling, and D. Fink (2019). "Best practices for making reliable inferences from citizen science data: case study using eBird to estimate species distributions." BioRxiv: 574392.

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. Ecological Modelling, 422, 108927.

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

Riesch, H., & Potter, C. (2014). Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. Public understanding of science, 23(1), 107-120.

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.

Weiser, E. L., Diffendorfer, J. E., Lopez-Hoffman, L., Semmens, D., & Thogmartin, W. E. (2020). Challenges for leveraging citizen science to support statistically robust monitoring programs. Biological Conservation, 242, 108411.

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: