Joint Nature Conservation Committee

A summary and introduction to “Statistics for citizen science: extracting signals of change from noisy ecological data”

N. Bunch

November 2014

To find out more about JNCC visit http://www.jncc.gov.uk/page-1729

A summary and introduction to “Statistics for citizen science: extracting signals of change from noisy ecological data”

An introduction and summary of the recently published paper:
Isaac, N., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 2014, 5, 1052–1060 doi: 10.1111/2041-210X.12254

This document was subject to peer review by three internal reviewers within JNCC and by Dr Nick Isaac, the primary author of the paper discussed in this summary.

Contents

Acronyms

BRC	Biological Records Centre
CEH	The Centre for Ecology and Hydrology
JNCC	Joint Nature Conservation Committee
NBN	National Biodiversity Network
UKBMS	United Kingdom Butterfly Monitoring Scheme

Background and overview

Most biodiversity data collected is *ad hoc* in nature, and not part of a systematic scheme[1]. This opportunistic data includes the high quantity of records submitted by citizen science programmes and national recording schemes and collated via, for example, the NBN. They have huge value in supplementing long-term, standardized, biodiversity monitoring schemes such as the UKBMS. Opportunistic data have demonstrated the ecological impact of various drivers of biodiversity loss promptly and quantifiably, including climate change (Hickling *et al,* 2006), invasive species (Roy *et al,* 2012) and habitat loss (Warren *et al,* 2001). JNCC and CEH invest resource into the BRC to collate and store these opportunistic records. Having the ability to extract useful signals of biodiversity trend from the large amount of collated data is very helpful to learn more about the natural environment.

Challenges of using opportunistic data – accounting for bias

The variation in sampling methods used to collect opportunistic data can mean that noise may obscure species population changes. Similarly, spatial or temporal variation in recorder effort has the potential to form biases in trend estimates for individual species.
Forms of variation in recorder activity which create biased data include:

- Uneven recording intensity over time, measured as number of visits per year
- Uneven spatial coverage
- Uneven sampling intensity per visit
- Uneven detectability of a target species in question

To deal with these sources of variation in recorder activity, two broad methods can be used:

1. Filtering the data: this involves removing the bias in the data- in theory leaving the signal of change, but reducing the data available with which to find a trend.
2. Statistical Correction Procedures: this seeks to correct for the bias created by variation in sampling in time, space or intensity.

JNCC and CEH have undertaken a study to evaluate the variety of modelling methods that have been used to deal with this variation in recorder activity. The study considers both the models' power to detect changes in species' distribution, as well as how effective they are in providing a reliable indicator of species distribution.

Methods of analysis

A hypothetical ecological community was simulated to trial the variety of models. As the simulated community can be controlled, it is possible to determine whether trends in focal species are correctly identified for a given model under different recording scenarios. Spatially separated sites were generated and randomly populated with species, including one focal species. The simulated sites were subjected to a suite of recording scenarios by virtual observers, as might occur in opportunistic data collection (table 1). These scenarios are based on observed patterns of recording from Great Britain and Netherlands, taking into account the forms of variation in recorder activity stated above.

[1] The individual records are of high quality, and are validated by a record cleaner tool at least once (NBN website). In iRecord- used for online biodiversity record entry which feeds into the NBN- additional verification of records by experts can also occur (iRecord website).

Table 1: *Description of recording scenarios in the simulation*
(From Isaac et al, 2014)

Scenario	*Summary*
Control	Constant recording intensity over years. All species have a fixed probability of being recorded per visit.
MoreVisits	Number of visits per year doubles over the course of the recording period, as would be observed if the number of recorders increased.
MoreVisits+Bias	As *MoreVisits*, but the extra visits are biased toward sites where the focal species is present, as might be observed if the spatial footprint of recording changed over time.
LessEffortPerVisit	Sampling effort per visit declines over time, increasing the proportion of 'short lists' from 60% to 90% of visits, reflecting a shift from systematic to 'incidental' recording.
MoreDetectable	The focal species is 20% easier to detect at the end of the recording period than at the start, for example if a new field guide makes it easier to identify.
NonFocalDecline	50% of nonfocal species are each declining at 30% over the recording period.

Eleven different published methods (table 2), which employ both filtering and statistical correction procedures, were used to identify trends in the distributional changes of the focal species from the simulated datasets. Each method was tested to see how often real trends in the focal species' population were detected. The distribution of the focal species remained unchanged throughout the 10 year simulation, so the test of robustness considered the rate of false positive trends detected (i.e. detecting a trend where none existed). To establish how powerful the methods were at detecting population changes, other simulations were also run in which populations of the focal species declined linearly. Each method's power was assessed in terms of whether they could detect this genuine decline.

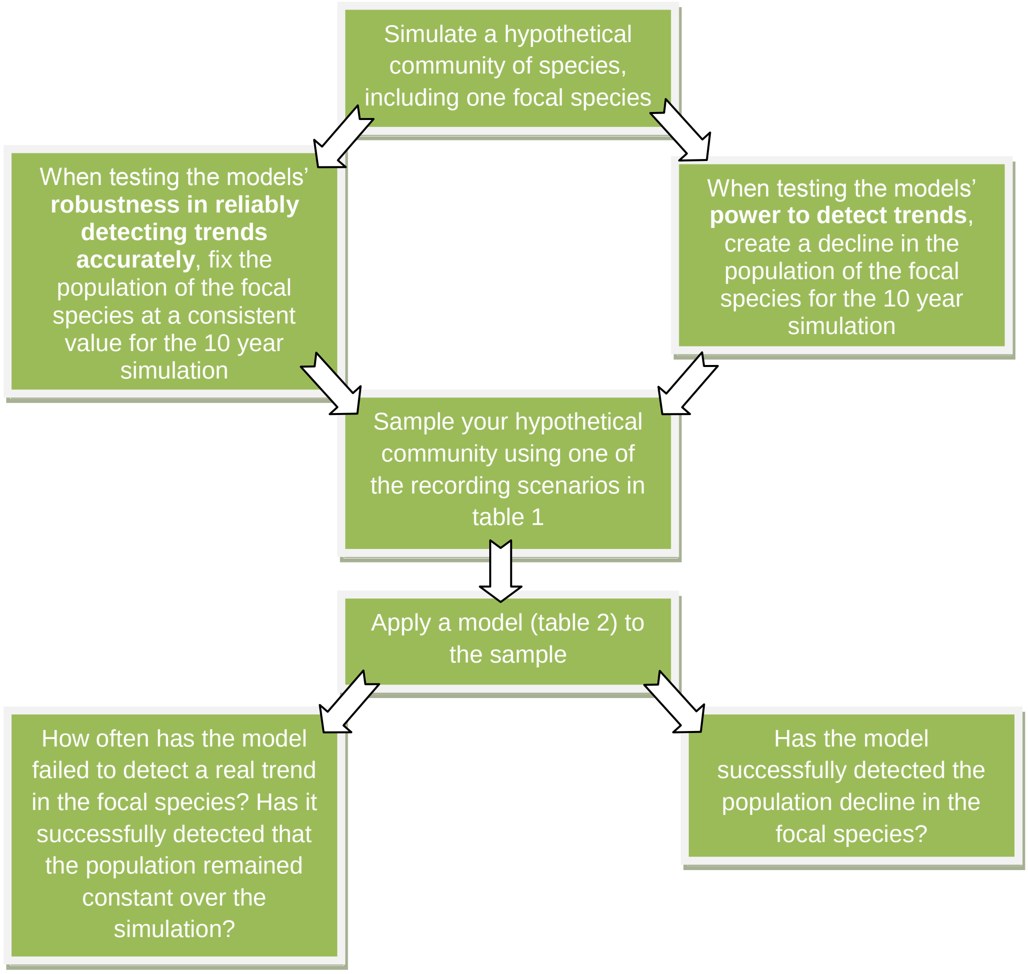
Simulate a hypothetical community of species, including one focal species
When testing the models' **robustness in reliably detecting trends accurately**, fix the population of the focal species at a consistent value for the 10 year simulation
When testing the models' **power to detect trends**, create a decline in the population of the focal species for the 10 year simulation
Sample your hypothetical community using one of the recording scenarios in table 1
Apply a model (table 2) to the sample
How often has the model failed to detect a real trend in the focal species? Has it successfully detected that the population remained constant over the simulation?
Has the model successfully detected the population decline in the focal species?

***Figure 1**: Testing the effectiveness of different published methods to detect trends in biased data sets.*

***Table 2**: Summary of methods and their performance across all tests. Robustness refers to the methods' ability to reliably detect a consistent focal species population over a 10 year period. Power refers to the methods' ability to detect a decline in focal species population. Adapted from Isaac et al. (2014).*

Method	*Method details*	*Summary of method effectiveness*
Naïve	Method simply looks at numbers of the focal species detected. Does not employ any records from other species to control for variation in recorder activity.	A high rate of false positive trends indicated in a majority of scenarios.
Telfer	Measures relative change taking into account variations in the geographical coverage and intensity of recorder effort. Estimated trends in all species together are taken as an indicator of recorder effort.	Robust but least powerful.
Frescalo_P	Uses information about sites' similarity to one another to assign local benchmarks in neighbourhoods, providing site-specific estimates of recording intensity. Data is pooled into two equal time periods.	Occasional false positive trends indicated in two scenarios (*MoreVisits+Bias* & *NonFocalDeclines*) but otherwise robust. Less powerful than 'per-year' methods e.g *Frescalo_Y.*
Frescalo_Y	As *Frescalo_P* but data analysed in ten one-year time periods.	More powerful than *Frescalo_P* but failed to detect trends under 3 scenarios.
ReportingRate	Considers the proportion of visits that record the focal species. Considering focal species as a proportion is thought to make the trend estimate more robust to unevenness in recording over time.	A high rate of false positive trends indicated in a majority of scenarios.
RR + Site Filtering	Uses the *ReportingRate* model but filters the data based on the number of years per site. Only sites with visits in at least 2 of the 10 simulated years were included.	A high rate of false positive trends indicated in a majority of scenarios. Some loss of power due to site filtering and therefore less data used.
RR + List Length	The *ReportingRate* model is sensitive to uneven sampling effort per visit. Adding the *List Length* variable accounts for this as the length of species list collated in a visit is taken to be a proxy for recorder effort.	A high rate of false positive trends indicated in a majority of scenarios.
RR + Site effect	Uses the *ReportingRate* model but incorporates a random effect to take into account the fact that each site has a different identity.	A high rate of false positive trends indicated in 3 scenarios. Most powerful under *Control* scenario.
RR+SF+LL+Site	The *ReportingRate* model, incorporating all three of the above correction factors.	A high rate of false positive trends indicated in 3 scenarios. Some loss of power due to site filtering.
OccDetSimple	The *ReportingRate* model, incorporating an occupancy detection submodel to account for the variation in detectability in species at different visits.	A high rate of false positive trends indicated in *MoreVisits+Bias*. Otherwise robust.
OD+SF+LL+Site	The *ReportingRate* model with all four additional components above included.	Generally robust and powerful.

Results and their implications

The study demonstrates that the more complex methods used to account for recorder variation in opportunistic data sets can be very effective at extracting trends accurately. In particular, the final method (*OD+SF+LL+Site*) performed very well, because the model includes factors to take into account all the four sources of bias created by recorder variation discussed above. Of the other methods, the FRESCALO models were most likely to detect trends. Most methods were robust to variation in the number of visits, but other forms of recorder variation posed problems. Ultimately, the other models demonstrate weakness in their power and/or robustness based on the assumptions they make. For example, many models do not take into account the potential variation in detectability in species over the recording period, and thus fall short in the *MoreDetectable* scenario, where this is the main form of bias. This study has highlighted that in future data collection, it is very useful to capture small amounts of information about sampling intensity, to help to analyse it using a method which will eliminate known bias.

By thoroughly comparing a broad range of methods in a simulated and thus controlled community, this study provides quality evidence that valuable trends in species' distribution can be extracted from opportunistic data if the appropriate analytical method is employed to account for potential biases. The methods analysed are currently used in key areas of ecological monitoring: for example FRESCALO has been used in producing the 2014 Vascular Plant Red Data List for England. This study may help to inform the choice of method used in similar influential publications. It also means it is a real possibility to use the 100 million records from the NBN to confidently establish biodiversity trends. These may be used to create indicators or more broadly to assess biodiversity trends in relation to environmental drivers. The methods outlined here are not guaranteed to generate trends for all individual species- it is exceptionally difficult to do for Great Crested Newts, for example, due to their ecology and extensive data sampling issues. However, this method has huge potential for identifying broader trends such as those affecting multiple taxa.

References

Hickling, R., Roy, D. B., Hill, J. K., Fox, R., & Thomas, C. D. (2006). The distributions of a wide range of taxonomic groups are expanding polewards. *Global Change Biology*, *12*(3), 450–455. doi:10.1111/j.1365-2486.2006.01116.x

iRecord website; "How do I...?"; accessed 19th September; http://www.brc.ac.uk/irecord/how-do-i

Isaac, N., van Strien, A.J., August, T.A., de Zeeuw, M.P., Roy, D.B. (2014). Statistics for citizen science: extracting signals of change from noisy ecological data. Methods in Ecology and Evolution 2014, 5, 1052–1060 doi: 10.1111/2041-210X.12254

National Biodiversity Network website; "NBN Record Cleaner"; accessed 19th September 2014; http://www.nbn.org.uk/Tools-Resources/Recording-Resources/NBN-Record-Cleaner.aspx

Roy, H. E., Adriaens, T., Isaac, N. J. B., Kenis, M., Martin, G. S., Brown, P. M. J., Maes, D. (2012). Invasive alien predator causes rapid declines of native European ladybirds. *Diversity and Distributions*, *18*(7), 717–725. doi:10.1111/j.1472-4642.2012.00883.x

Warren, M. S., Hill, J. K., Thomas, J. A., Asher, J., Fox, R., Huntley, B., Thomas, C. D. (2001). Rapid responses of British butterflies to opposing forces of climate and habitat change. *Nature*, *414*(6859), 65–9. doi:10.1038/35102054