Methods in Ecology and Evolution
British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 1052–1060 doi: 10.1111/2041-210X.12254

Statistics for citizen science: extracting signals of change from noisy ecological data

Nick J. B. Isaac[1]*, Arco J. van Strien[2], Tom A. August[1], Marnix P. de Zeeuw[2] and David B. Roy[1]

[1]*NERC Centre for Ecology & Hydrology, Crowmarsh Gifford, Maclean Building, Wallingford, OX10 8BB, UK; and* [2]*Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands*

Summary

1. Policy-makers increasingly demand robust measures of biodiversity change over short time periods. Long-term monitoring schemes provide high-quality data, often on an annual basis, but are taxonomically and geographically restricted. By contrast, opportunistic biological records are relatively unstructured but vast in quantity. Recently, these data have been applied to increasingly elaborate science and policy questions, using a range of methods. At present, we lack a firm understanding of which methods, if any, are capable of delivering unbiased trend estimates on policy-relevant time-scales.

2. We identified a set of candidate methods that employ data filtering criteria and/or correction factors to deal with variation in recorder activity. We designed a computer simulation to compare the statistical properties of these methods under a suite of realistic data collection scenarios. We measured the Type I error rates of each method–scenario combination, as well as the power to detect genuine trends.

3. We found that simple methods produce biased trend estimates, and/or had low power. Most methods are robust to variation in sampling effort, but biases in spatial coverage, sampling effort per visit, and detectability, as well as turnover in community composition, all induced some methods to fail. No method was wholly unaffected by all forms of variation in recorder activity, although some performed well enough to be useful.

4. We warn against the use of simple methods. Sophisticated methods that model the data collection process offer the greatest potential to estimate timely trends, notably *Frescalo* and occupancy–detection models.

5. The potential of these methods and the value of opportunistic data would be further enhanced by assessing the validity of model assumptions and by capturing small amounts of information about sampling intensity at the point of data collection.

Key-words: biodiversity, biological records, distribution, *Frescalo*, occupancy modelling, simulations, trends

Introduction

Robust quantitative measures of the stock and rate of change in biodiversity are crucial for assessing species' risk of extinction (Mace & Lande 1991), for measuring progress against international targets (Butchart *et al.* 2010) and testing against predictions about climate change impacts (Maclean & Wilson 2011). The demands for timely information are increasing. For instance, the EU Habitat and Bird directives require changes in species' status to be reported every 6 years, and progress against the Convention of Biological Diversity targets is reported on a decadal basis.

Long-term, standardized, monitoring schemes produce timely and robust estimates of status and trends, often on an annual basis (Gregory *et al.* 2005). Unfortunately, such data are available for only a small number of taxa in a few countries. The next best sources are opportunistic data, such as those available on the Global Biodiversity Information Facility (GBIF), including records submitted by volunteers (Prendergast *et al.* 1993). These data are less structured than monitoring schemes but high in quantity: GBIF comprises >400 million observations of 1·4 million species (http://www.gbif.org). Opportunistic data have delivered substantive insights into the ecological impacts of climate change (Hickling *et al.* 2006), invasive species (Roy *et al.* 2012) and habitat loss (Warren *et al.* 2001).

While opportunistic data have been used to describe coarse-scale changes in biodiversity (Thomas *et al.* 2004; Carvalheiro *et al.* 2013), the absence of standardized protocols presents serious challenges for estimating timely trends in the status of individual species. The noise generated by opportunistic sampling has the potential to swamp any signal of real change, or to produce spurious signals of change where none exists. We use the term 'variation in recorder activity' to refer to the sampling biases inherent in opportunistic data, of which there are four principle forms: (i) uneven recording intensity over time, measured as the number of visits per year (a visit is defined as unique combination of site and date in the records data), (ii) uneven spatial coverage, (iii) uneven sampling effort per visit and (iv) uneven detectability. Each source of variation has the

*Correspondence author. E-mail: njbi@ceh.ac.uk
The copyright line for this article was changed on 6th October 2014 after original online publication

potential to introduce substantial bias in trend estimates for individual species. The growth of citizen science programs (Dickinson *et al*. 2012) is likely to increase data volumes, and affect the nature of recording with potentially far-reaching consequences for how the data may be used to infer biodiversity trends (Tulloch *et al*. 2013).

In the past, opportunistic data were collated into broad time periods, for example published Atlases. This compensates to some degree for variation in recorder activity, allowing changes in species' distributions to be assessed over the years between atlas periods (Thomas *et al*. 2004; Tingley & Beissinger 2009; Botts, Erasmus & Alexander 2012). This approach has limited potential to deliver trends in a timely fashion, because Atlas periods are typically measured in decades. In principle, it should be possible to derive trend estimates on subdecadal time-scales by incorporating information about the data collection process (Szabo *et al*. 2010; Roy *et al*. 2012; van Strien, van Swaay & Termaat 2013). Therefore, a pressing need exists to understand how recorder activity can be treated statistically.

There are numerous methods proposed in the literature for estimating trends in species' distributions from opportunistic data while taking into account recorder activity. A number of authors have proposed methods based on filtering the data (Rich & Woodruff 1996; Maes & Van Dyck 2001; Warren *et al*. 2001; Hickling *et al*. 2006; Kuussaari *et al*. 2007; Van Calster *et al*. 2008; Maes *et al*. 2012; Roy *et al*. 2012), based on the number of years per site and/or the number of species per site (or visit). All are based on the premise that filtering is an effective tool to remove the bias while leaving a signal of biological change.

A second category of methods has a statistical correction procedure to treat recorder activity. These methods are less frequent in the literature than selection methods, but have a greater variety of mechanisms to control for recorder activity. Many authors have sought to correct for uneven sampling intensity over time. Telfer, Preston & Rothery (2002) used the estimated trend in all species together as an indirect measure of how recording intensity differed between two sampling periods. If recorder intensity is higher in the second period, all species are expected to show increases compared with the first period. Any deviation from the overall expected trend is considered as an index of change for the species of interest. Ball *et al*. (2011) proposed that modelling a species' status as the proportion of the total records would be an effective way to control for changes in overall recording intensity over time, under the assumption that the effort per visit does not vary among years. Szabo *et al*. (2010) proposed a modification in which individual visits (or species lists) are the unit of analysis (thereby controlling for variation in the number of lists over time). Their innovation was to treat the number of species on the list (the list length, L) as a proxy for recorder effort per visit. Another innovation is to add the study site (or grid cell) as a random effect (Kuussaari *et al*. 2007; Roy *et al*. 2012), to control for uneven sampling in space. An alternative type of correction factor is to use benchmark species as proxy for recorder activity. Benchmarks are common species whose distribution is assumed to show no overall trend (Hill 2012).

Occupancy–detection models, which are derived from capture–recapture theory (MacKenzie 2006), have recently been successfully applied to large-scale models of distribution (Lahoz-Monfort, Guillera-Arroita & Wintle 2013) and distributional change (van Strien *et al*. 2010; van Strien, van Swaay & Termaat 2013). The key feature of occupancy–detection modelling is the use of replicated visits within a season to estimate the conditional probability that a species is recorded when present. The model consists of two hierarchically coupled submodels: one governing occupancy (presence vs. absence) and the other governing the observations (detection vs. non-detection).

Here, we test the statistical properties of a representative set of methods using computer simulation. We focus on situations where species' occurrences are recorded with high temporal and spatial precision (the site visit), and where an observation of one species can often be used to infer the non-detection of others (i.e. species are typically recorded as an assemblage). Most of the methods we compare have been designed to use such data, but they interpret non-detections in a variety of ways. Specifically, we estimate the Type I error rates of 11 methods under realistic scenarios of recorder activity, and their power to detect genuine trends in species' occupancy. Our aim was to identify methods that produce timely trend estimates and that are robust to multiple forms of variation in recorder activity. Identifying robust and powerful methods would open a vast frontier of previously unexploited data for use in both biodiversity policy and applied ecology.

Materials and methods

SIMULATION OVERVIEW

We constructed a computer simulation to assess the performance of candidate methods under simple deviations from random sampling and changes in community composition. We generated species occurrence matrices using simple rules, which were then subjected to a suite of recording scenarios by virtual observers (Zurell *et al*. 2010) to generate a set of realized data sets. Most recording scenarios simulate temporal trends in recorder activity, generating bias in the pattern of detection and non-detection. Where possible, our scenarios were parameterized using observed patterns of recording in the Great Britain and the Netherlands (see Appendix S1 for details). We ran all simulations over a period of 10 years, but the species occurrence matrix remained unchanged over time for most species under most scenarios.

For each realized data set, we estimated a trend in the distribution of one 'focal' species using 11 candidate methods, which are defined below. The performance of each method–scenario combination was assessed from 1000 simulated data sets. We conducted separate tests of each method's validity and its power to detect change.

All the computer code required to run the simulations is available from a Github repository (https://github.com/BiologicalRecordsCentre/RangeChangeSims). Appendix S2 contains information about how to access and use the code.

SPECIES OCCURRENCE MATRICES

Our system consists of 1000 'sites', which we assert to be separated in space (although for simplicity our simulation is not spatially explicit). Our sites can be thought of as a sample of 1 km^2 grid cells (Roy *et al.* 2012; van Strien, van Swaay & Termaat 2013), but the precise definition is not important. Each test data set consisted of one focal species and 25 non-focal species (preliminary analyses with up to 200 non-focal species produced identical results). Species were distributed randomly among sites: each distribution was determined by drawing 1000 times from a binomial distribution with a species-specific probability of being occupied. For the focal species, we fixed this probability at 50% in all simulations; for non-focal species, we used random numbers from a beta distribution with shape parameters 2 and 2, such that mean species richness among sites was *c.* 13 species, with a variance among sites of *c.* 5.

CONTROL SCENARIO

This section defines the *Control* scenario, which corresponds to random sampling. Each year, a team of virtual observers visited a certain number of sites. In a selection of British and Dutch recording data sets, the distribution of visits among sites each year is characterized by a power law in which the number of sites receiving n visits is 4 times greater than the number receiving $2n$ visits (Fig. S5, Table S1). We therefore defined the probability of any site receiving n visits as $a.n^{-2}$, where a is the probability of being visited once. We selected three levels of overall recording intensity: a = 0·05 (low), 0·07 (medium) and 0·10 (high). This range of values was chosen to generate data sets that superficially resemble the records of British butterflies (high intensity), Odonata (medium) and Hymenoptera (low intensity; Figs S5–S6, Tables S1–S4). The frequency distribution of visits among sites was sampled from the power law function above, truncated so that no site received more than 10 visits in any one year.

Having determined the number of sites to be visited in any year, we then selected the identity of these sites at random, but apportioned visits among them non-randomly. Specifically, visits were allocated to selected sites according to species richness, with the most speciose site receiving most visits. This was carried out to mimic real data sets in which records are clustered around nature reserves and other sites that are known to harbour interesting wildlife.

Site visits under the *Control* scenario have equal survey effort, but do not automatically record all species present. Each species had a fixed probability of being detected if present: the focal species' detection probability was fixed at 0·5 per visit; for non-focal species, the detection probability was drawn at random from the sigmoid curve described in Hill (2012) and varied from 0·16 to 0·88. This species-specific detection probability can be thought of as the product of visual apparency (Dennis *et al.* 2006) and mean abundance. Species' detection probabilities were uncorrelated with occupancy.

BIASED RECORDING SCENARIOS

We devised five biased recording scenarios (Table 1): four capture the major axes of variation in recorder activity identified above and were generated by subsampling records generated by the *Control*. The final scenario simulates changes in community composition.

The first simulates an increase in the number of visits per year (i.e. recording intensity is uneven over time, Fig. S1). In the *MoreVisits* scenario, the expected number of visits per year doubled over the 10-year recording period. We simulated this by subsampling from the *Control* scenario: each year, we sampled (without replacement) a proportion of visits, with the proportion in the final year set equal to 1. Our second scenario, *MoreVisits + Bias,* is a modification in which a trend exists in the ratio of focal:non-focal sites being visited, thus simulating temporal change in the spatial coverage of sites. Specifically, sites containing the focal species are 27% more likely to be visited (than non-focal sites) in year one, but in year 10, the focal and non-focal sites are equally represented.

Uneven sampling per visit is the third major axis of variation in recorder activity (Figs S3–S4). Inter-annual variation in sampling, effort is a potentially serious form of bias for some methods, because it affects species' probabilities of being recorded. We simulated a directional trend towards shorter lists, as might result from changes in recorder behaviour (e.g. a growth in the number of inexperienced recorders with limited identification skills). In the *LessEffortPerVisit* scenario, visits from the *Control* were selected at random and resampled to produce shorter lists. The proportion of visits producing short lists varied from year to year, increasing from 60% to 90% during each simulation. Short lists contained 1, 2 or 3 species, in the ratios 2:1:1, respectively. The total number of records produced by *LessEffortPerVisit* is around half the number produced by the *Control*.

Table 1. Description of recording scenarios in the simulation

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*, the extra visits are biased towards 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
NonFocalDeclines	50% of non-focal species are each declining at 30% over the recording period

We also model situations in which species become more detectable over time, for example through the adoption of new technology or publication of a field guide. In the *MoreDetectable* scenario, we model a gradual increase in the focal species' probability of detection per visit, from 0·4 at the start of the simulation to 0·5 at the end (i.e. a 20% increase over the recording period).

Several of the methods under test measure relative, rather than absolute, change. For this reason, an important consideration is the degree to which these relative trends are impacted by changes in the status of other (non-focal) species. We tested this by simulating a decline of 30% over 10 years in 50% of non-focal species (*NonFocalDeclines*). Declining species were selected at random in each simulation.

RANGE CHANGE METHODS

We compare 11 methods of trend estimation, which we selected to represent the diversity of approaches that have been applied in the literature. Methods differ in the spatial and temporal resolution at which they are applied, but we focus on the underlying assumptions they make (specifically with respect to non-detections). Technical details of all the methods, including mathematical notation, can be found in Appendix S1.

Our simplest measure of change is the linear trend (or difference) in the annual number of sites (or grid cells) on which the focal species was detected [i.e. a Poisson generalized linear model (GLM)]. This model has no mechanism to control for recorder activity, so we refer to it as the *Naïve* method. The *Naïve* method is unique in that it uses only records from the focal species. All others employ records from other species to control for variation in recorder activity, either assuming that a record of one species indicates the absence of others, or as a means of estimating sampling effort.

We included the methods of Telfer, Preston & Rothery (2002) and Hill (2012), both of which are commonly used in the literature (Powney *et al*. 2013; Fox *et al*. 2014). The *Telfer* index for each species is the standardized residual from a linear regression across all species (see Appendix S1 for details) and is a measure of relative change only, because the average real trend across species is obscured. We predict that *Telfer* will be sensitive to scenarios in which recording is biased with respect to the focal species (e.g. spatial bias or changes in detectability). Hill's method, which is known as *Frescalo*, uses information about sites' similarity to one another to assign local benchmarks within neighbourhoods, and provides site-specific estimates of recording intensity. We compare two variants: in *Frescalo_P,* we pooled the data into two equal time periods; in *Frescalo_Y,* the data were analysed in ten time periods (i.e. one per year). *Frescalo* trends are expressed as the reporting rate of focal species relative to that of the benchmarks (Hill 2012; Fox *et al*. 2014). We predict the performance of *Frescalo* will be similar to *Telfer's* method, but more powerful.

Our remaining methods are based on generalizations of a simple model called *ReportingRate,* in which the response variable is the proportion of visits within a given year that produce a record of the focal species (Ball *et al*. 2011), and which we model as a binomial GLM with year as a covariate (Appendix S1). Modelling the focal species as a proportion is expected to make the trend estimate robust to unevenness in recording over time (*MoreVisits* scenario). To this simple model, we can add four components (Table 2, Fig. S8), each of which is designed to address one specific form of variation in recorder activity.

Filtering the data based on number of years per site (+SF) and adding a random effect for site identity (+Site) are both intended to address the problem of uneven spatial coverage over time (MoreVisits + Bias scenario). Our site-filtered (+SF) models contain only those sites that received visits in at least 2 of the 10 years in the simulation. Adding a list length covariate (+LL), as in Szabo *et al*. (2010), provides a means to control for uneven sampling effort per visit (*LessEffortPerVisit*). The fourth component is the addition of a nested submodel for detection, and our *OccDetSimple* model is formulated following MacKenzie *et al*. (2002), but for multiple seasons (one per year: see Appendix S1 for details). We included all four single-component models (Table 2) in our simulation. However, these components can be applied

Table 2. Trend estimation methods and their the key features. The 'trend unit' column refers to the units in which trends are expressed. The 'grain size' defines the basic unit of analysis. The remaining columns indicate which components each method employs for dealing with variation in recorder activity. See Appendix S1 for further details

Method	Trend unit	Grain size	Non-detections	Site filtering	List length covariate	Random site effect	Detection submodel	Other
Naïve	Site	Year						
Telfer	Site	Species	X					X
Frescalo_P	Site	Site : Period	X					X
Frescalo_Y	Site	Site : Year	X					X
ReportingRate	Visit	Year	X					
RR + SF	Visit	Year	X	X				
RR + LL	Visit	Visit	X		X			
RR + Site	Visit	Visit	X			X		
OccDetSimple	Site	Visit	X				X	
RR + SF + LL + Site	Visit	Visit	X	X	X	X		
OD + SF + LL + Site	Site	Visit	X	X	X	X	X	

together; for example, van Strien, van Swaay & Termaat (2013) included a detection submodel with a list length covariate (+LL) and terms to control for phenological variation in detectability (which we do not consider here). We include two multi-component models in our simulation: one with all components except the detection submodel ($RR + SF + LL + Site$) and the model with all four components ($OD + SF + LL + Site$), in which OD refers to the fact that occupancy and detection are modelled separately.

We predict that multi-component models are likely to be more robust than single-component models, which in turn will outperform the simplest models (*Naïve, ReportingRate*). However, we predict that some components are likely to lead to a loss of power. In particular, filtering the data (+SF models) is likely to reduce power because the number of observations is reduced. We also suspect that occupancy–detection models might have low power, due to their greater statistical complexity.

ESTIMATING THE TRENDS AND EVALUATING MODEL PERFORMANCE

For each simulated data set, we tested the null hypothesis of no change in the focal species' distribution using each of the 11 method variants (see Appendix S1 for details). For *Telfer* and *Frescalo_P,* we split the realized data into two 5-year periods. To implement *Frescalo,* we generated a random matrix of neighbourhood weights: randomly generated neighbourhoods would be inappropriate for real data sets where communities show strong evidence of species sorting, but are reasonable for our simulated data in which species were independently distributed. Other parameters of *Frescalo* were set following Hill (2012). We implemented occupancy–detection models in a Bayesian framework using JAGS with three Markov chains, 5000 iterations per chain, a burn-in of 2500 and a thinning rate of three (van Strien, van Swaay & Termaat 2013).

For the test of validity, the distribution of the focal species remained unchanged throughout the simulation: the Type I error rate, α, is the proportion of 1000 simulated data sets in which the null hypothesis was rejected at $P = 0{\cdot}05$. In the test of power, we simulated a linear decline in occupancy: each occupied site had a constant probability of extinction per year, such that occupancy declined on average by 30% over the 10-year period (i.e. the species would qualify as Vulnerable under IUCN Criterion A2). A simple estimate of power would be $1 - \beta$, where β is the rate at which we failed to reject the null hypothesis (i.e. the Type II error rate). However, some scenarios are designed to introduce negative bias in the trend estimates, so β is not comparable across scenarios. Instead, we defined power as $1 - \beta - \alpha$: in cases where $\alpha + \beta > 1$, we set power equal to zero.

Results

About half the methods return appropriate Type I error rates ($\alpha = 0{\cdot}05$) under the control scenario of unbiased even recording, including the *Naïve* model (Fig. 1; Table S5). The *ReportingRate, RR + SF* and *RR + LL* (but not *RR + Site*) methods return significant results around twice as frequently as expected. The methods that split the data into two time periods (*Telfer* and *Frescalo_P*) are both conservative ($\alpha \ll 0{\cdot}05$), as is the most complex model ($OD + SF + LL + Site$).

All methods experience at least one combination of recording scenario and input parameters in which the Type I error rate is inflated by a factor of two compared with the *Control* (Fig. 1, Table S5). Under most scenarios, the failures become

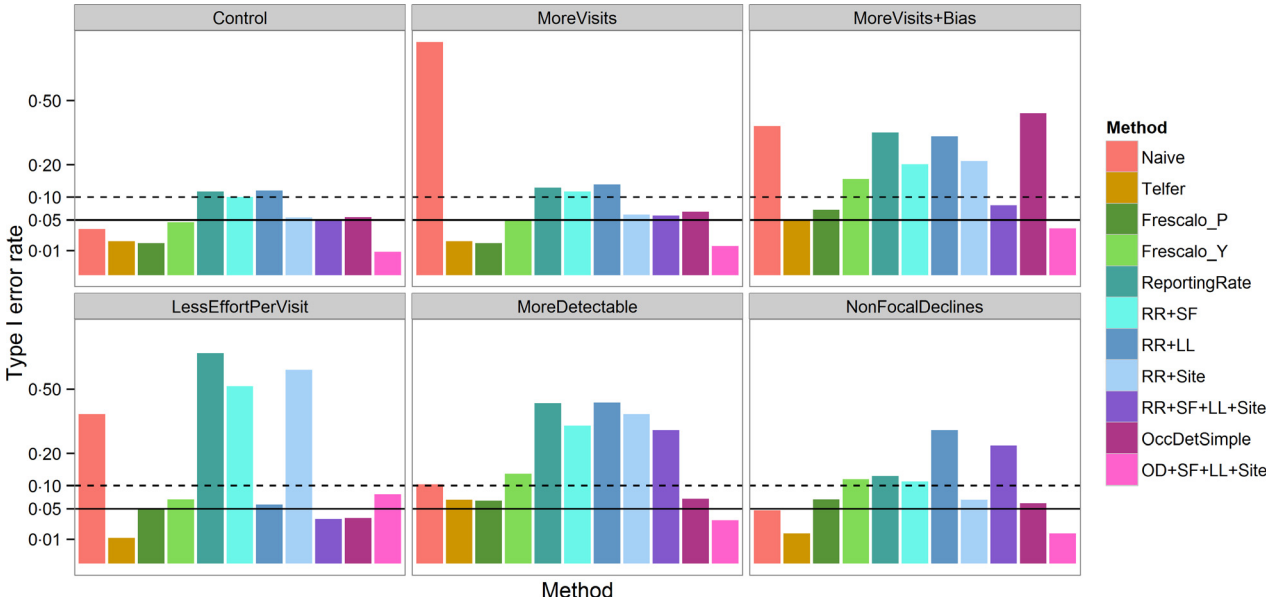

Fig. 1. Type I error rates of all methods under all scenarios (the proportion of simulated data sets in which a significant trend was detected, even though no trend exists). Note the square root scale on *y*-axis. Results are shown for medium levels of recording intensity. The solid and dashed lines indicate $\alpha = 0{\cdot}05$ and $\alpha = 0{\cdot}1$, respectively.

more acute as the recording intensity increases (Fig. S9), reflecting the fact that small data sets contain insufficient data to detect the bias and reject the null hypothesis.

As predicted, the *Naïve* model performs badly under virtually all departures from random sampling. Other methods are robust to growth in the number of visits (*MoreVisits*); that is, the Type I error rate is similar to that observed under the *Control* (Fig. 1).

The performance of most methods deteriorates markedly in our scenario with biased site selection (*MoreVisits* + *Bias*): only *Telfer*, *Frescalo_P* and the two multi-component models (*RR* + *SF* + *LL* + *Site*, *OD* + *SF* + *LL* + *Site*) returned Type I error rates below $\alpha = 0{\cdot}1$ at medium recording intensity (Fig. 1), and only *OD* + *SF* + *LL* + *Site* showed no increase in Type I error rates at high recording intensity (Fig. S9).

When recording visits become progressively more incomplete (*LessEffortPerVisit*), the *ReportingRate*, *RR* + *SF* and *RR* + *Site* all fail, reflecting the fact that it becomes increasingly less likely that the focal species will be recorded on an average visit. Other models, including those with List Length coefficients, performed reasonably well, although *OccDetSimple* returned lower Type I errors than *OD* + *SF* + *LL* + *Site*.

Changes in detectability (*MoreDetectable*) elevate Type I error rates in almost all methods. For *Frescalo_P* and *Telfer*, the elevation is slight ($\alpha < 0{\cdot}1$ under all levels of recording intensity, Fig. S9). Only the two occupancy–detection models are robust.

NonFocalDeclines induce poor performance of *RR* + *LL* models (including *RR* + *SF* + *LL* + *Site*) and *Frescalo_Y*, and moderate elevations for *Frescalo_P* (Fig. 1, Fig. S9).

Not surprisingly, power is strongly affected by overall sampling intensity, with a twofold increase going from low- to high-intensity recording (Fig. 2). Power declines under most deviations from the *Control* (Fig. 3), but the relative power of each method is fairly consistent. Models with site filtering (+*SF*) are less powerful than those without, and two time period models (*Frescalo_P* and *Telfer*) are less powerful than per-year models. All except the two occupancy–detection models lost most of their power under the *MoreDetectable* scenario. Three methods (*Telfer*, *RR* + *LL*, *RR* + *SF* + *LL* + *Site*) completely failed to detect a trend under the *NonFocalDeclines* scenario; the power of *Frescalo* was also much reduced.

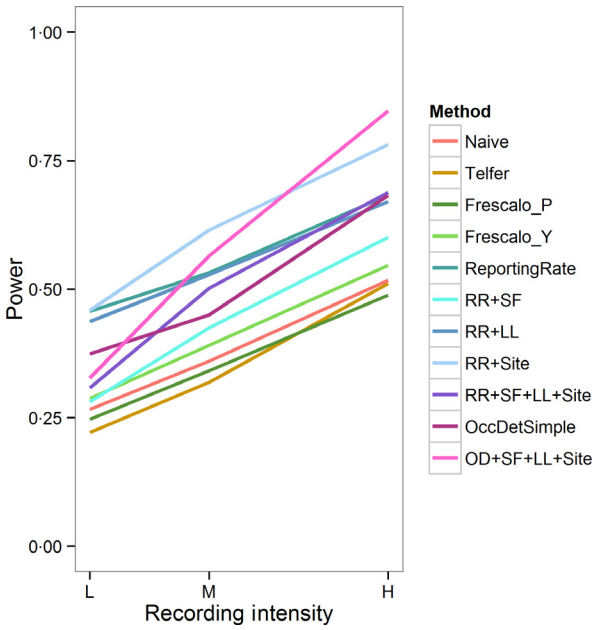

Fig. 2. Power to detect a 30% decline in the focal species under the control scenario, plotted against recording intensity.

Discussion

Our simulations have provided a rigorous test of candidate methods for estimating trends in species' distributions from opportunistic data. Many studies have emphasized the problem that opportunistic data are characterized by uneven sampling effort over time (Prendergast *et al.* 1993; Botts, Erasmus & Alexander 2012; Maes *et al.* 2012), but we observe that most methods are robust to variation in the number of visits (*MoreVisits* scenario). Other forms of variation in recorder activity present serious problems for many methods, yet are rarely discussed. We found that no method is wholly robust under all scenarios, but some perform well enough to be useful, and some general principles have emerged about how to apply them to real-world data sets.

We have clear evidence that simple methods easily fail under realistic scenarios of recording behaviour. The poor performance of the *Naïve* model is not unexpected, but the *ReportingRate* and its 'one component' variants all failed under a majority of scenarios (Table 3). The variants lacking a site effect failed even under the *Control* scenario of random sampling: this occurs because they treat repeat visits to the same site as independent (the proportion of visits to occupied sites varies stochastically from year to year). Our findings draw into question the conclusions of studies that have used such methods (Szabo *et al.* 2011; Breed, Stichter & Crone 2012). *Telfer*'s method, which is also relatively simplistic, was among the most robust methods, but was least powerful, as expected.

Previous studies have compared only simple methods (Botts, Erasmus & Alexander 2012), but our results show that sophisticated methods outperform simple ones. To a large extent, we understand why some methods fail and others perform well. Models with site effects (+*Site*) are more robust than those without, because they control for uneven sampling of sites over time. Models with list length coefficients (+*LL*) are robust to variation in sampling intensity among visits, although this comes at a cost of sensitivity to changes in non-focal species. Occupancy–detection models are robust under *MoreDetectable* because they explicitly model the detection process. As predicted, our multi-component models were robust to more scenarios than single-component models. Our results clearly show that methods which model the data collection process, such as *Frescalo* and occupancy–detection models, have the greatest potential for delivering robust and timely trends from opportunistic data.

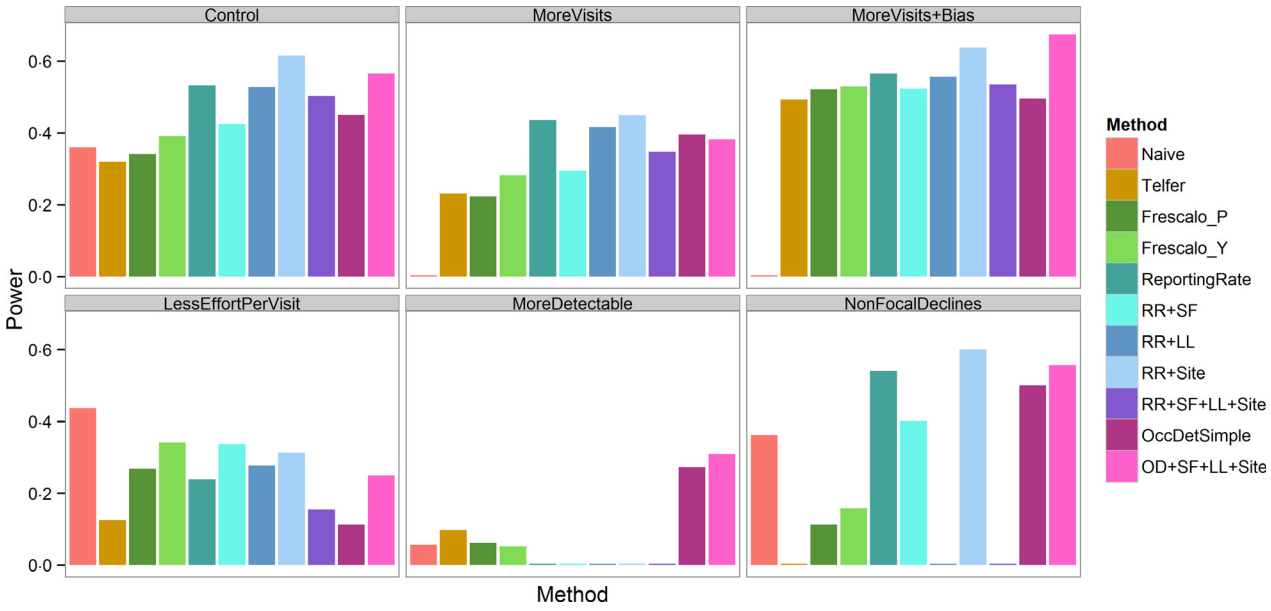

Fig. 3. Power to detect a 30% decline in the focal species under medium recording intensity for all scenarios.

Table 3. Summary of method performance across all tests

Method	Summary of key findings
Naïve	Inflated type I errors under a majority of scenarios
Telfer	Robust but least powerful
Frescalo_P	Mildly inflated under two scenarios (*MoreVisits + Bias* & *NonFocalDeclines*) but otherwise robust. Less powerful than 'per-year' methods
Frescalo_Y	More powerful than *Frescalo_P* but inflated type I errors under 3 scenarios
ReportingRate	Inflated type I errors under a majority of scenarios
RR + Site Filtering	Inflated type I errors under a majority of scenarios. Some loss of power due to site filtering
RR + List Length	Inflated type I errors under a majority of scenarios
RR + Site effect	Inflated type I errors under 3 scenarios. Most powerful under *Control* scenario
RR + SF + LL + Site	Inflated type I errors under 3 scenarios. Some loss of power due to site filtering
OccDetSimple	Inflated type I errors under *MoreVisits + Bias*. Otherwise robust
OD + SF + LL + Site	Generally robust and powerful

We were surprised that *Frescalo_P* (although not *Frescalo_Y*) appears to be moderately robust to scenarios where the focal species undergoes separate treatment (*MoreVisits + Bias, MoreDetectable*). We need a better understanding of *Frescalo*, perhaps using real data incorporating information on neighbourhood weights (which we did not include in our simulation). While *Frescalo_P* performed well in our simulations, we have a number of reservations about its usage. First, using the method requires the user to make a variety of choices, in addition to the number of time periods. The selection of benchmark species and neighbourhoods are defined by input parameters (Hill 2012) which have considerable impact on the trend estimates that are produced (A.J. van Strien, M.P. de Zeeuw and A. Doroszuk, unpublished data). Secondly, our simulations compared all methods at the same spatial scale, but the typical grain size for *Frescalo* is 100-fold larger (100 km^2 vs. 1 km^2) than used by methods which treat the visit as the fundamental unit (Roy *et al.* 2012; van Strien, van Swaay & Termaat 2013), so the number of unique observations (and hence power) is also lower. This coarse-grained approach reflects both computational limitations (neighbourhoods are defined by an *N*N* matrix, where *N* is number of sites), and the need to robustly estimate recording intensity for each neighbourhood. Estimating recording intensity reflects the fact that *Frescalo* was specifically designed for situations in which information from individual visits is unavailable (Hill 2012), which makes *Frescalo_P* the most appropriate method for describing long-term change where the periods are well-defined (e.g. published atlases).

Our multi-component occupancy–detection model (*OD + SF + LL + Site*) is the clear winner in our simulations (Table 3). The Type I error rates were among the lowest of all methods: this is a simple consequence of the fact that it is the most generalized model we tested, with components to deal with multiple forms of bias. It also has considerably more power than we expected. These results clearly validate the use of occupancy–detection models for analysing opportunistic data (van Strien, van Swaay & Termaat 2013), but raise a number of questions about which components should be employed for real data sets.

The worst performance of *OD + SF + LL + Site* came under *LessEffortPerVisit*: the Type I error rates were around eight times higher than under the *Control*, while less sophisticated models (notably *RR + SF + LL + Site* and *OccDetSimple*) were unaffected (Fig. 1). This result implies that the detection submodel may be sufficient to control for uneven sampling

effort per visit, and that the + *LL* component may be superfluous or even counterproductive. We found that filtering the data to include only 'well-sampled sites' (+ *SF* variants) was partially, but not wholly, successful in dealing with uneven sampling of sites (*MoreVisits* + *Bias*). It is conceivable that performance could have been improved by increasing our threshold to 3 years (as in Roy *et al.* 2012), but at the cost of substantially reduced power: under medium recording intensity, about 80% of sites were visited in 1 year, 50% in at least 2 years and just 20% in three or more. Data filtering is fundamentally limited by the assumption that subjective thresholds can separate the signal from the noise, so we need another way to deal with uneven sampling of sites.

Ultimately, the robustness of any model is dependent on its assumptions, and whether those assumptions are valid. We modelled a suite of recording scenarios, but there is a gap between our idealized simulations and the reality of how opportunistic data are collected. There is a clear need to devise diagnostic tests to assess the validity of these assumptions for real data sets, but this is challenging because we lack information about how the records were generated. Most methods we compared assume, at some level, that species are recorded within assemblages during site visits (i.e. failure to record is interpreted as non-detection, as opposed to 'not searched for'). Our + *LL* models relax this assumption using list length as a proxy for sampling effort, which assumes that short lists are the result of incomplete surveys, but this is not universally true (e.g. on sites with few species). The growth of technology in wildlife recording, including smartphone apps, offers great potential to capture metadata about sampling intensity (e.g. start and end times of the survey) with minimal input from the recorder. These data would go a long way to make inferences from opportunistic data more robust in future (Kéry *et al.* 2009; van Strien, van Swaay & Termaat 2013).

Our results provide further confirmation that opportunistically gathered data have enormous potential to make meaningful contributions in biodiversity science and policy-making (Schmeller *et al.* 2009; Tulloch *et al.* 2013). All the variants of our generalized trend model (but not *Frescalo*) can easily incorporate covariates, making them ideal for testing hypotheses about the drivers of biodiversity change (c.f. Roy *et al.* 2012). Our results provide an evidence base for producing quantitative trends from opportunistic data and a benchmark against which future methods can be compared.

Acknowledgements

We are grateful to Gary Powney and three anonymous reviewers for constructive comments on previous versions of this manuscript. We thank Stuart Ball, Mark Hill, Stephen Freeman, Colin Harrower and Thierry Onkelinx for technical advice. This work was funded by JNCC, NERC and the Welsh Government.

Data accessibility

All computer code required to run the simulation and draw the figures is available at https://github.com/BiologicalRecordsCentre/RangeChangeSims. Appendix S2 contains information about how to access and use the code.

References

Ball, S., Morris, R., Rotheray, G. & Watt, K. (2011). *Atlas of the Hoverflies of Great Britain (Diptera, Syrphidae)*. Centre for Ecology and Hydrology, Wallingford.

Botts, E.A., Erasmus, B.F.N. & Alexander, G.J. (2012) Methods to detect species range size change from biological atlas data: a comparison using the South African Frog Atlas Project. *Biological Conservation*, **146**, 72–80.

Breed, G.A., Stichter, S. & Crone, E.E. (2012) Climate-driven changes in northeastern US butterfly communities. *Nature Climate Change*, **3**, 142–145.

Butchart, S.H.M., Walpole, M., Collen, B., van Strien, A., Scharlemann, J.P.W., Almond, R.E.A. *et al.* (2010) Global biodiversity: indicators of recent declines. *Science*, **328**, 1164–1168.

Carvalheiro, L.G., Kunin, W.E., Keil, P., Aguirre-Gutiérrez, J., Ellis, W.N., Fox, R. *et al.* (2013) Species richness declines and biotic homogenisation have slowed down for NW-European pollinators and plants. *Ecology Letters*, **16**, 870–878.

Dennis, R., Shreeve, T., Isaac, N.J.B., Roy, D.B., Hardy, P., Fox, R. & Asher, J. (2006) The effects of visual apparency on bias in butterfly recording and monitoring. *Biological Conservation*, **128**, 486–492.

Dickinson, J.L., Shirk, J., Bonter, D., Bonney, R., Crain, R.L., Martin, J., Phillips, T. & Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Frontiers in Ecology and the Environment*, **10**, 291–297.

Fox, R., Oliver, T.H., Harrower, C., Parsons, M.S., Thomas, C.D. & Roy, D.B. (2014) Long-term changes to the frequency of occurrence of British moths are consistent with opposing and synergistic effects of climate and land-use changes. *Journal of Applied Ecology*, **51**, 949–957.

Gregory, R.D., van Strien, A., Vorisek, P., Gmelig Meyling, A.W., Noble, D.G., Foppen, R.P.B. & Gibbons, D.W. (2005) Developing indicators for European birds. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, **360**, 269–288.

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**, 450–455.

Hill, M.O. (2012) Local frequency as a key to interpreting species occurrence data when recording effort is not known. *Methods in Ecology and Evolution*, **3**, 195–205.

Kéry, M., Dorazio, R.M., Soldaat, L., van Strien, A., Zuiderwijk, A. & Royle, J.A. (2009) Trend estimation in populations with imperfect detection. *Journal of Applied Ecology*, **46**, 1163–1172.

Kuussaari, M., Heliölä, J., Pöyry, J. & Saarinen, K. (2007) Contrasting trends of butterfly species preferring semi-natural grasslands, field margins and forest edges in northern Europe. *Journal of Insect Conservation*, **11**, 351–366.

Lahoz-Monfort, J.J., Guillera-Arroita, G. & Wintle, B.A. (2013) Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, **23**, 504–515.

Mace, G.M. & Lande, R. (1991) Assessing extinction threats: toward a reevaluation of IUCN threatened species categories. *Conservation Biology*, **5**, 148–157.

MacKenzie, D.I. (2006) *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence*, pp. 324. Academic Press, Burlington, Massachusetts, USA.

MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Andrew Royle, J. & Langtimm, C.A. (2002) Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, **83**, 2248–2255.

Maclean, I.M.D. & Wilson, R.J. (2011) Recent ecological responses to climate change support predictions of high extinction risk. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 12337–12342.

Maes, D. & Van Dyck, H. (2001) Butterfly diversity loss in Flanders (north Belgium): Europe's worst case scenario? *Biological Conservation*, **99**, 263–276.

Maes, D., Vanreusel, W., Jacobs, I., Berwaerts, K. & Van Dyck, H. (2012) Applying IUCN Red List criteria at a small regional level: a test case with butterflies in Flanders (north Belgium). *Biological Conservation*, **145**, 258–266.

Powney, G.D., Rapacciuolo, G., Preston, C.D., Purvis, A. & Roy, D.B. (2013) A phylogenetically-informed trait-based analysis of range change in the vascular plant flora of Britain. *Biodiversity and Conservation*, **23**, 171–185.

Prendergast, J., Wood, S., Lawton, J. & Eversham, B. (1993). Correcting for variation in recording effort in analyses of diversity hotspots. *Biodiversity Letters*, **1**, 39–53.

Rich, T.C.G. & Woodruff, E.R. (1996) Changes in the vascular plant floras of England and Scotland between 1930–1960 and 1987–1988: the BSBI Monitoring Scheme. *Biological Conservation*, **75**, 217–229.

Roy, H.E., Adriaens, T., Isaac, N.J.B., Kenis, M., Martin, G.S., Brown, P.M.J. *et al.* (2012) Invasive alien predator causes rapid declines of native European ladybirds. *Diversity and Distributions*, **18**, 717–725.

Schmeller, D.S., Henry, P.-Y., Julliard, R., Gruber, B., Clobert, J., Dziock, F. *et al.* (2009) Advantages of volunteer-based biodiversity monitoring in Europe. *Conservation Biology*, **23**, 307–316.

Szabo, J.K., Vesk, P.A., Baxter, P.W.J. & Possingham, H.P. (2010) Regional avian species declines estimated from volunteer-collected long-term data using List Length Analysis. *Ecological Applications*, **20**, 2157–2169.

Szabo, J.K., Vesk, P.A., Baxter, P.W.J. & Possingham, H.P. (2011) Paying the extinction debt: woodland birds in the Mount Lofty Ranges, South Australia. *Emu*, **111**, 59.

Telfer, M.G., Preston, C.D. & Rothery, P. (2002) A general method for measuring relative change in range size from biological atlas data. *Biological Conservation*, **107**, 99–109.

Thomas, J.A., Telfer, M.G., Roy, D.B., Preston, C.D., Greenwood, J.J.D., Asher, J., Fox, R., Clarke, R.T. & Lawton, J.H. (2004) Comparative losses of British butterflies, birds, and plants and the global extinction crisis. *Science*, **303**, 1879–1881.

Tingley, M.W. & Beissinger, S.R. (2009) Detecting range shifts from historical species occurrences: new perspectives on old data. *Trends in Ecology & Evolution*, **24**, 625–633.

Tulloch, A.I.T., Possingham, H.P., Joseph, L.N., Szabo, J. & Martin, T.G. (2013) Realising the full potential of citizen science monitoring programs. *Biological Conservation*, **165**, 128–138.

Van Calster, H., Vandenberghe, R., Ruysen, M., Verheyen, K., Hermy, M. & Decocq, G. (2008) Unexpectedly high 20th century floristic losses in a rural landscape in northern France. *Journal of Ecology*, **96**, 927–936.

van Strien, A.J., van Swaay, C.A.M. & Termaat, T. (2013) Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. *Journal of Applied Ecology*, **50**, 1450–1458.

van Strien, A.J., Termaat, T., Groenendijk, D., Mensing, V. & Kéry, M. (2010) Site-occupancy models may offer new opportunities for dragonfly monitoring based on daily species lists. *Basic and Applied Ecology*, **11**, 495–503.

Warren, M.S., Hill, J.K., Thomas, J.A., Asher, J., Fox, R., Huntley, B. *et al.* (2001) Rapid responses of British butterflies to opposing forces of climate and habitat change. *Nature*, **414**, 65–69.

Zurell, D., Berger, U., Cabral, J.S., Jeltsch, F., Meynard, C.N., Münkemüller, T. *et al.* (2010) The virtual ecologist approach: simulating data and observers. *Oikos*, **119**, 622–635.

Received 13 May 2014; accepted 4 August 2014
Handling Editor: Barbara Anderson

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Background information on the simulation concept, implementation and results.

Appendix S2. Instructions for reproducing the simulation results