Review

Special Issue: Ecological and evolutionary informatics

Data-intensive science applied to broad-scale citizen science

Wesley M. Hochachka[1], Daniel Fink[1], Rebecca A. Hutchinson[2], Daniel Sheldon[2], Weng-Keen Wong[2] and Steve Kelling[1]

[1] Cornell Lab of Ornithology, Ithaca, NY 14850 USA
[2] School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA

Identifying ecological patterns across broad spatial and temporal extents requires novel approaches and methods for acquiring, integrating and modeling massive quantities of diverse data. For example, a growing number of research projects engage continent-wide networks of volunteers ('citizen-scientists') to collect species occurrence data. Although these data are information rich, they present numerous challenges in project design, implementation and analysis, which include: developing data collection tools that maximize data quantity while maintaining high standards of data quality, and applying new analytical and visualization techniques that can accurately reveal patterns in these data. Here, we describe how advances in data-intensive science provide accurate estimates in species distributions at continental scales by identifying complex environmental associations.

The need for citizen-science data in ecology

The conservation of species begins with an understanding of the distribution, abundance, habitat preferences and movements of organisms across wide geographic areas and over long periods of time. Although inroads have been made into automating the collection of information on species occurrence [1–3], typically only human observers can reliably identify organisms to the species level [4]. Intensive surveys made by expert observers can accurately identify patterns of species occurrence within local spatial and short temporal windows. However, the cost and availability of experts to collect data does not scale readily to broad spatial (i.e. continent) or temporal (i.e. year-round) surveys, and essentially all broad-scale ecological data are collected by volunteers, although rare exceptions exist [5].

Such 'citizen-science' projects are research projects in which the public is enlisted in scientific endeavors [6]. Pervasive access to the Internet and information technologies has allowed citizen-science projects, such as Galaxy Zoo, to become global in scale and engage hundreds of thousands of volunteers [7]. In ecology, evolutionary biology and environmental monitoring, citizen-science projects have been designed to provide fine-resolution data with more intensive sampling at a local level (e.g. frequent surveys of water quality at a location [8] and bioblitzes [9]), and have also been deployed across broad extents in space and time (e.g. continental survey of shell polymorphism in Europe [10] or breeding bird surveys [11,12]).

Within citizen-science projects covering large spatial extents and across many years, data may be collected in only some seasons [11,12], or in year-round surveys [13]. The most general collection of citizen-science data (across broad spatial extents, multiple years and across all seasons) presents multiple informatics challenges. First, the acquisition and management of large volumes of data requires planning not only of the data-management infrastructure, but also for appropriate methods for maintaining the motivation of volunteers. Second, quality control and handling of inevitable observational biases are essential owing to variation among observers [14], 'false absences' owing to imperfect detection of organisms [15] and often uneven distributions of data in space and time [16]. Third, appropriate modeling of ecological phenomena over broad spatial and temporal scales is challenging because organisms take part in many kinds of dynamic, scale-dependent processes, including dispersal, invasion, migration, predator–prey interactions, competitive exclusion and mutualistic interactions, any of which can vary through space and time [17]. Addressing challenges such as these has led to the emergence of a data-intensive paradigm in science [18], which is being applied to broad-scale citizen-science projects and species distribution modeling [19].

Here, we review our experience in applying a data-intensive science approach to species distribution modeling that uses observations gathered in eBird [13]. To the best of our knowledge, eBird is currently the largest ecological citizen-science project in existence, currently collecting 2 million to 3 million new species–date–location records monthly across the entire planet, although with most data coming from North America. Working with such a large project has given us insights into many informatics considerations in citizen science. Here, we use examples from our work with eBird to review three general issues in informatics that we believe are general to broad-scale ecological citizen-science projects. We first describe the process of insuring that continual large volumes of data will be acquired from thousands of volunteers, through applied informatics and social networking processes. Second, we describe the use of multiple processes for maintaining data quality, particularly those we believe are relevant to all citizen-science projects. Third, we consider

Corresponding author: Kelling, S. (stk2@cornell.edu)

0169-5347/$ – see front matter © 2011 Elsevier Ltd. All rights reserved. doi:10.1016/j.tree.2011.11.006

the need for the development of novel data analysis techniques, again using examples from eBird to illustrate the diversity of motivations and solutions to analytical problems.

Designing and implementing broad-scale citizen-science projects

Maximizing the information obtained from broad-scale citizen-science projects that gather species occurrence data depends on finding the proper balance between data quantity and quality. Quantity is important, because a sufficiently large volume of data with relatively lower per-datum information content can contain more information for broad-scale species distribution estimates than a smaller amount of higher quality data [20]. High data quality is essential and requires attention to the design and implementation of the project, which should strive to limit incorrect data entry (e.g. spurious species identification and control extraneous sources of variation). These considerations need to feed into the required software, database and hardware infrastructure to support the tens of thousands of contributors [21] that can be needed in broad-scale citizen-science projects.

Data quantity

To obtain large quantities of data, data collection protocols should not be overly complicated, as complex protocols tend to engage fewer participants [6]. For example, eBird collects only basic information to document what was observed (species observed and an estimate of numbers), and where and when the search was conducted (the time, date and location of the search). To account for extraneous sources of variation in the data, eBird also collects basic information to identify the observers and describe how each search was conducted (time spent searching for birds, distance traveled during the search and whether the participant reported all species identified on the search).

Providing appropriate rewards to participants is another way to ensure large quantities of data. Citizen science is a form of crowdsourcing (i.e. an activity that engages and rewards large numbers of people for performing tasks that automated sensors and computers cannot readily accomplish) [22]. Although many of the most successful crowdsourcing projects rely on monetary payment for tasks completed [23], most citizen-science projects must find other rewards for participation. An approach often used attempts to motivate participation by emphasizing the impact that volunteers can have in advancing science. However, our experience with eBird suggests that larger numbers of participants can be recruited and retained if non-altruistic rewards are built into the project. Originally, the main motivation for participation in eBird was that one's data would be used in scientific research, and this approach attracted few project participants. The web site and participant outreach focus of the project was then modified to provide direct rewards to participants, allowing users to: (i) keep track of their bird records; (ii) sort their personal bird lists by date and region; (iii) share their lists with others; and (iv) visualize their observations on maps and graphs. Growth in participation rapidly increased and eBird has gathered more information in 1 month (almost 3 million observations) than it did during the entire first 2 years of the project (2 million observations).

The use of record-keeping facilities as a direct reward for participation may not generalize easily to collection of data from other taxa. However, another aspect of the motivation of eBird is potentially more generalizable: eBird appeals to the competitiveness of participants by providing tools for determining relative status of volunteers (e.g. numbers of species seen) and geographical regions (e.g. checklists submitted per state and province). An appeal to competitiveness has also been successful for other citizen-science projects, such as Folding@home, where team rankings of 'work units' are tracked (http://fah-web.stanford.edu/cgi-bin/main.py?qtype=teamstats).

Data quality

Maintaining high standards in data quality requires appropriate design of data input and management procedures. With eBird, this quality control work has several facets, most or all of which can be applied to any acquisition of observations. First, quality control filters can be implemented in the data entry process to ensure that all required protocol information is accurately entered. Additionally, customized data entry forms containing only those species typically found in a specific geographical region at a specific time of year minimize the possibilities of mistaken data entries. Although participants can report a species that would not normally be expected for a given region and season, they need to take an active additional step to do so. All reports of unusual species, or unusually high numbers of individuals reported, are flagged on entry and subsequently reviewed by a network of volunteers who are regional experts in patterns of species occurrence. Finally, we are developing analytical techniques to identify and quantify high-quality data retrospectively (Boxes 1 and 2).

Extracting ecological and evolutionary insights during analysis

Effective research requires clearly articulated questions. Typically, these questions are formulated prior to data collection, allowing the data collection processes to be tailored to help answer the questions identified [10,24]. Under such circumstances, the types of analysis used will typically be those conventionally used by ecologists (Hochachka, W.M. *et al.*, unpublished data). By contrast, many broad-scale observational data sets, which include citizen-science data, are collected with only general monitoring goals in mind [25], and care must be taken to craft research questions and select analysis methods retrospectively [26]. Often, these analyses will be based on broad-scale hypotheses or be more exploratory in nature. For example, a researcher may want to know if migration timing will change with increasing temperatures, without needing to specify exactly how the timing will change. Our discussion of analysis methods will largely discuss these more exploratory techniques, as their use is less common among ecologists but important for any retrospective analysis of broad-scale citizen-science data (Hochachka, W.M. *et al.*, unpublished data).

Box 1. Glossary of standard statistical models for species distribution modeling

Table I provides terms and notations for single-species distribution models.

Basic species distribution models

In species distribution models using parametric statistical analyses to describe probabilities of presence or absence, the basic model form is a logistic linear model in which the values of regression coefficients β_j are fit for all covariates in the vector x_i (Eqn I):

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\beta_0 + \sum_{k=1}^{d} \beta_k x_{ik}\right) \tag{I}$$

In this parametric model, additive covariate effects are each described by a single coefficient [54]. The function $\text{logit}^{-1}(a) = 1/(1 + e^{-a})$ is the standard inverse link function, which ensures that the resulting value is a valid probability (between zero and one). Non-parametric regression models (e.g. [55]) extend this model by considering more complex functional forms for covariate effects associating predictors and response. Here, we represent the non-parametric function as $f(\boldsymbol{x}_i)$ (Eqn II):

$$\Pr(y_i = 1) = f(\boldsymbol{x}_i) \tag{II}$$

In addition to variables such as climate and land cover, which are typically included in species distribution models, variables such as effort and time of day that capture variability in the detection process are also important. We divide the covariates into two components $\boldsymbol{x}_i = (\boldsymbol{v}_{l(i)}, \boldsymbol{w}_i)$, where $\boldsymbol{v}_{l(i)}$ contains the location-dependent predictors for the observation location $l(i)$ of checklist i, and $\boldsymbol{w}_i$ contains variables describing the observation event.

Occupancy-detection models

Observing a species requires that it is both present and detected by the observer. Occupancy-detection models extend the models above to predict the true occupancy status of a site, represented by the latent variable $z_l \in \{0, 1\}$ with associated function f^{occ}, instead of detection, represented by the variable $y_i \in \{0, 1\}$ with associated function f^{det}. When z_l can be assumed constant (population closure) across a set of repeated data-collection events at location l, then (Eqn III,IV):

$$\Pr(z_l = 1) = f^{occ}(\boldsymbol{v}_l) \tag{III}$$

$$\Pr(y_i = 1) = z_{l(i)} \cdot f^{det}(\boldsymbol{w}_i) \tag{IV}$$

The standard form of occupancy models is a special case of this general form, in which the occupancy and detection functions are logistic linear equations, and $f^{occ}(v_l)$ outputs a probability called ψ and $f^{det}(w_i)$ outputs a probability called p in the notation of [41].

Table I. Terms and notation for single-species distribution models.

Term or notation	Explanation
Observation	The indication of presence or absence of one species at a particular time and place, and an optional count. With eBird, observations of non-zero quantities are recorded by participants and observations of non-detections inferred retrospectively
$y_i \in \{0, 1\}$	The response (detection or non-detection of the organism in observation i)
x_i	Covariates for observation i that are included to explain variation in the response variable. Expressed as the vector $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$
$l(i)$	The location of observation i
v_l	A vector of location-dependent covariates for location l (e.g. elevation or land cover)
w_i	A vector of covariates describing the process of making observation i (e.g. effort and time of day)

Similar to all studies based on observational data, analysis of broad-scale citizen-science data has two basic requirements. First, the observations must adequately represent the inferential population (e.g. the data need to represent the spatial region and time interval of interest or the group of species under consideration). Second, the data must contain sufficient information to estimate the quantity of interest while controlling for important confounding sources of variation, whether these confounding sources arise from other biological processes or from variation in the observation process itself. Because most observational data (i.e. data gathered through sensor networks, remote-sensing techniques, or citizen-science projects) are collected prior to analysis, achieving these goals will often constrain the range of research questions possible. For this reason, extra care is necessary for developing specific research questions and selecting the analytic methods. Similarly, the retrospective nature of this data-driven analysis paradigm places a greater burden on analytic methods to control confounding sources of variation, and many of the existing methods of analysis are not always appropriate [27].

Often, novel analyses of ecological data represent the application of techniques developed in other fields to ecological data, with minor extensions or elaborations. However, at times, entirely novel analysis methods may need to be developed and refined for the specific problems presented by ecological data [28]. We now describe three examples to illustrate motivations for the development and use of novel analysis techniques with ecological citizen-science data. Our goal here is to illustrate that simple application of existing methods may not be sufficient, and that close collaborations between ecologists and statistical and computer scientists can be important for effective use of citizen-science data. Two of our examples fuse traditional parametric statistical techniques with machine-learning analyses [29]. The flexibility of these 'semi-parametric' regression analyses is ideal for exploiting the rich covariate information available from multiple sources, automatically adapting to patterns where too little is known *a priori* to specify a fully parametric statistical analysis while permitting prior knowledge to impose important structure on other parts of the model. Our third example discusses opportunities for entirely new approaches to studying changes in ecological systems, in our case the movement of animal populations during migration.

Broad-scale species distribution modeling in the face of non-stationarity

We believe that one general challenge in analyses of broad-scale citizen-science data is that the relationships between fine-scale environmental predictors and the observed responses of species tend to vary across large spatial and temporal extents [30]. Specifically, the assumption of statistical stationarity, an assumption that the underlying ecological processes that give rise to distributions do not change across time or space, is not valid. Failure to account for non-stationarity can result in potential problems, such as extrapolations of predicted distributions far

Box 2. The occupancy-detection-experience model

We extended the occupancy-detection model (Box 1) with a component modeling the observer experience, producing a model that we refer to as the occupancy-detection-experience (ODE) model [56]. We introduced a new latent variable $e_{b(i)} \in \{novice, expert\}$ that captures the experience of observer $b(i)$ that submitted observation i. The variable $e_{b(i)}$ is a function of experience covariates $u_{b(i)}$, which include attributes of an observer, such as the total number of checklists submitted and the total number of species reported. The observer experience, along with the detection covariates and site occupancy, influences the detection of the species. Thus, when applied to eBird data, occupancy covariates are associated with each site, detection covariates are associated with each checklist and experience covariates are associated with each observer (Eqn I–III).

$$\Pr(z_l = 1) = f^{occ}(\boldsymbol{v}_l) \qquad [I]$$

$$\Pr(e_{b(i)} = 1) = f^{\exp}(\boldsymbol{u}_{b(i)}) \qquad [II]$$

$$\Pr(y_i = 1) = z_{l(i)} \cdot f^{det}(\boldsymbol{w}_i, e_{b(i)}) \qquad [III]$$

The ODE model can also be used for two important tasks. First, it can be used to predict the experience of an observer based on their submission history and attributes about the observer. Second, the model can incorporate observer experience level when predicting the detection of the species and when predicting the true occupancy of a site. Although the importance of the expert–novice distinction varies with the species of bird, our results (Table I) show that incorporating information about the experience level of the observers consistently improves the accuracy of predicting detections.

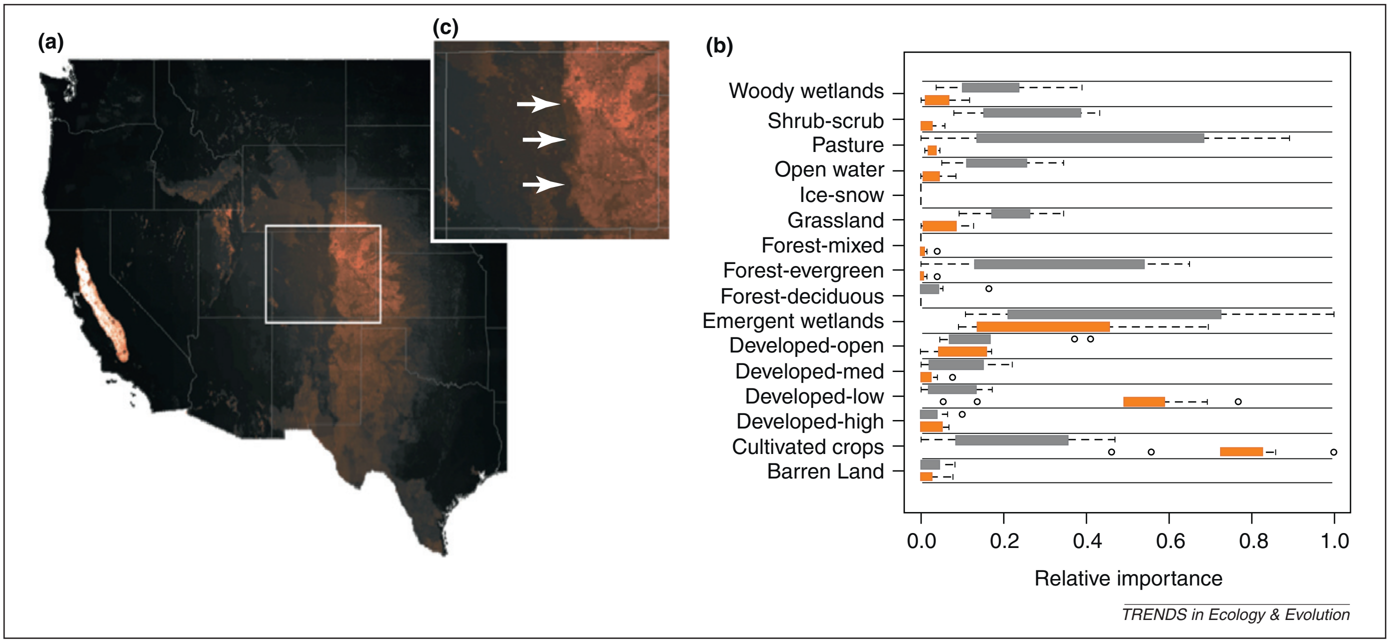
Table I. Differences in average detection probabilities between highly experienced and novice eBird participants for eight bird species[a]

Species	Average $\Delta_{\text{Detection}}$
Blue jay	0.0118
White-breasted nuthatch	0.0077
Northern cardinal	–0.0218
Great blue heron	0.0110
Brown thrasher	0.1659
Blue-headed vireo	0.1158
Northern rough-winged swallow	0.1618
Wood thrush	0.0954

[a] The top four rows correspond to common, easily detected bird species, whereas the bottom four rows correspond to bird species that are harder to detect. Larger positive values indicate higher probabilities for highly experienced participants to detect a species; highly experienced participants for example, have on average, a >16% higher probability of detecting northern rough-winged swallows. These results also indicate opportunities for further training and education of eBird participants.

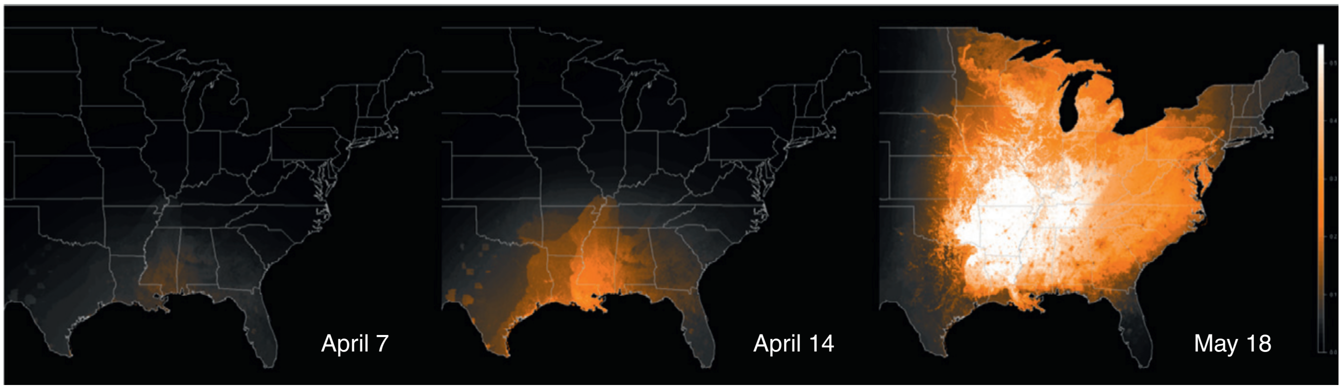
beyond actual distributions indicated by data (e.g. Figure 1 in [31], Figure 3a in [32]).

The most common statistical models used to account for nonstationary spatial and spatiotemporal processes are geographically weighted regression (GWR) [33] and spatially varying coefficients (SVC) [34,35]. GWR uses spatial weights to allow for spatially adaptive coefficients. The SVC approach is based on a hierarchical model that places spatial process priors on those regression coefficients that are thought to vary spatially. Both of these models are fully parametric models and require the analyst to specify fully in advance important predictors and those that are allowed to vary spatially, in addition to other model parameters. Thus, the success of these approaches depends strongly on the analyst's ability to specify the important ecological processes correctly.



TRENDS in Ecology & Evolution

Figure 1. Output from a spatiotemporal exploratory model (STEM) for Swainson's hawks (*Buteo swainsoni*) across the western USA in April 2009 with relative habitat predictor importance in California and Colorado. Swainson's hawk is a common raptor in the Central Valley of California in the arid grasslands of the Great Plains **(a)**. The relative importance of local habitat predictors **(b)** is shown for California (orange) and Colorado (grey) from April 20, 2009 to May 20, 2009. The difference in relative habitat predictor importance between these two regions illustrates how STEM adapts to varying habitat associations in different regions. The population in California is strongly associated with cultivated crops and areas of low human development. These tight associations are reflected in the predicted sharp boundaries of the California population in the Central Valley. The habitat associations of the hawk in Colorado are broader, including areas with pasture, emergent wetlands, evergreen forest and cultivated crops, suggesting a different local ecological niche, and a more diffuse predicted range that lacks sharp boundaries. The insert **(c)** shows that, in Colorado, the most prominent feature is the obvious east–west divide where the Great Plains meet the Rocky Mountains.



TRENDS in Ecology & Evolution

Figure 2. Spring migration of indigo buntings. The predicted probability of occurrence for indigo bunting (*Passerina cyanea*) is shown during the spring migration in 2009. The indigo bunting population crossed the Gulf of Mexico and had begun to make landfall in the south-eastern USA by April 7. By April 14, the population had begun its northward expansion and, by mid-May, it had filled in the majority of its breeding distribution. After this date, birds continued to arrive and filled its southern breeding distribution.

Starting with the observation that full parametric specification of the pattern of nonstationarity is not always possible, we have developed an adaptive framework for handling nonstationarity that can be used with any predictive model used for species distribution modeling. The spatiotemporal exploratory model (STEM) [29] comprises a randomized mixture of overlapping local models, each based on a limited geographic and temporal subset of the data. At a local level, each model automatically identifies and fits important predictor–response relationships. These local patterns are then allowed to 'scale up' as a simple parametric mixture to larger scales. The uniformly distributed mixture of local models can adapt to a wide variety of nonstationary spatiotemporal processes (Figures 1 and 2; [36]) and produces predicted distributions that are judged by expert field ornithologists to be accurate representations of true distributions and seasonal changes in distribution [29,36].

Controlling for sources of variation in citizen-science data

The second species distribution modeling challenge is to control for sources of variation frequently associated with species occurrence observations. Variation associated with many important ecological processes can be modeled using an increasing supply of publicly available environmental data (i.e. land cover, weather, climate, anthropogenic, etc.). These data sets often have large numbers of environmental predictors, which creates a problem for the use of parametric statistical techniques because there is more predictor information than there is prior ecological knowledge to specify the corresponding parametric statistical models based on these predictors. Non-parametric regression and machine-learning techniques [37,38] offer an attractive analytical approach to this challenge and can be viewed as natural and more general extensions of parametric statistical analysis methods (Box 1). These methods are able to identify a small number of important explanatory variables from a larger set of potential predictors, automatically determine the forms of relationship between predictors and responses, and are increasingly being used by ecologists in research fields such as species distribution modeling (e.g. [39,40]).

Not only is the modeling of ecological processes potentially challenging because of the large number of environmental predictors that might be important, but the description of the observation process can also be complex. The same straightforward data collection protocols that make it possible to engage large numbers of citizen-science participants often produce observational data with unwanted sources of variation, many known and some unknown. When variation in the observation process is confounded with ecological processes of interests, it may bias results. To extract as much ecological information as possible, there is the need to identify and account for as many sources of 'nuisance' variation as possible.

Existing automated nonparametric models offer a simple way to do this by including predictors that describe sources of variation in the observation process. For example, by including search duration as a predictor, the model can automatically detect and quantify the association between the probability of detecting a species and the amount of time spent searching for birds. However, this approach for using covariate information to control sources of variation does not explicitly separate the detection process from the species occupancy process and the resultant STEM model can only estimate the relative probability of occurrence [41].

Parametric statistical techniques that explicitly and simultaneously estimate detection and occupancy processes have been developed to estimate the latent, or unobserved, probability of occupancy. These techniques, known as occupancy models [41], are typically used with data collected over relatively small areas, because they require repeated observations at each site and time period, data that are time-consuming to collect. However, as recently demonstrated by Kéry *et al.* [42], data from checklists can be aggregated in a manner that allows them to be analyzed using these methods. This opens the possibility of analyzing citizen-science data with occupancy models. However, as previously discussed, the parametric assumptions of these models can be too rigid for broad-scale citizen-science applications.

Given that neither the fully non-parametric analysis techniques nor fully parametric techniques are completely

Box 3. Semi-parametric occupancy models

Our recent work has extended parametric occupancy-detection models (Box 1), combining traditional site-occupancy models and non-parametric boosted regression trees into a semi-parametric hybrid [57]. Although semi-parametric occupancy models retain the hierarchical structure and assumptions of parametric occupancy models that account for detectability bias, they allow the model probabilities to be non-parametric functions of many inputs. Structurally, the linear terms of the occupancy and detection functions (Box 1) are replaced with ensembles of non-parametric functions, where the j^{th} ensemble member for f^{occ} is referred to as f_j^{occ} (and similarly for f^{det}). The model probabilities f^{occ} and f^{det} are then a logistic transformation of a weighted sum (with weights ρ) of the outputs of J component functions (Eqn I,II):

$$f^{occ}(\boldsymbol{v}_l) = \text{logit}^{-1}\left(\sum_{j=1}^{J} \rho_j^{occ}\ f_j^{occ}(\boldsymbol{v}_l)\right) \qquad [I]$$

$$f^{det}(\boldsymbol{w}_i) = \text{logit}^{-1}\left(\sum_{j=1}^{J} \rho_j^{det}\ f_j^{det}(\boldsymbol{w}_i)\right) \qquad [II]$$

In analyses of both real and synthetic eBird data that compared conventional and semi-parametric site-occupancy models, along with standard methods that do not account for imperfect detection (logistic regression and boosted regression trees), the semi-parametric models performed well [57]. In particular, experiments on synthetic species generated with non-linear occupancy and detection functions showed that semi-parametric occupancy models recovered the true model probabilities better than did other methods. Additionally, they produced estimates of the non-linear relationships between the covariates and the model probabilities that were closer to the truth than were the other models. We emphasize these synthetic data results because they allow the models to be evaluated against the truth, which is unavailable for real data, but we note that the semi-parametric occupancy models also recovered non-linearities in the function estimates for real species.

A major advantage of semi-parametric occupancy models is that they provide great flexibility for exploratory modeling while simultaneously accounting for imperfect detection. Their main disadvantage is that, as with most semi-parametric methods, it is computationally challenging to quantify the uncertainty associated with the model (e.g. confidence limits around predictions). By contrast, conventional parametric occupancy models can be too inflexible for poorly understood systems, but they provide a convenient framework for quantifying uncertainty and testing hypotheses. The use of fully parametric or semi-parametric techniques needs to be decided upon based on study goals and prior knowledge of the system. In some cases, it may be possible to do initial exploration of the data with semi-parametric models to help guide the design of a fully parametric model. For example, semi-parametric occupancy models built with regression trees can automatically discover interactions to be included in a parametric model.

adequate for modeling the process of observation with broad-scale citizen-science data, such as those from eBird, new analytical approaches are required. Box 3 shows how a novel hybrid of parametric and non-parametric methods can be created to meet an identified need, in this case to enable occupancy modeling without the need to specify the structure of a model fully.

Describing broad-scale movements of animals

Our final example of motivation for, and development of, new analysis techniques is based on the desire to use year-round citizen-science data to describe movements of animal populations. Using a temporal sequence of distribution estimates, it is possible to describe patterns of movement in the population of a species during migration (Figure 2). However, for many applications, ecologists need additional information to better understand movement of animals and the factors that drive animal populations. In the case of migration, biologists need estimates of directions and speeds of movement [43]. Although these types of information are more directly gathered by attaching location-recording devices to birds (e.g. [44]), these devices are expensive and can only be attached to a relatively small number of birds originating at a small number of locations.

We are developing novel analysis methods to describe migration using citizen-science observations that are attempting to overcome three significant challenges. First, although migration is a strong factor influencing the spatiotemporal pattern of bird observations, it is a purely latent process and is not measured directly: bird observations record only the occurrence of birds at different instants in space and time, but do not record any direct measurements of migration (in contrast with tracking studies). Second, the occurrence or abundance information we do obtain from citizen-science data is noisy and incomplete. Finally, the migratory behavior of birds is most naturally described at the level of individuals (i.e. tracking studies), but broad-scale citizen-science projects observe patterns at the population level; there is a lack of existing modeling techniques to link these two levels. Similar motivations have been identified and related analytical methods developed for other specific systems [45–48].

In our own work, we address the first two challenges using latent process models [49], a family of stochastic models that explicitly model the relationship between an unobserved process of interest (migration) and the observed variables (occurrence at different locations in space and time) that provide evidence about the hidden process. To address the third challenge, we have developed collective hidden Markov models (CHMMs) [50] (D. Sheldon, PhD thesis, Cornell University, 2010), in which a Markovian model for the movement of individual birds is used to derive a population-level latent process model of migration. Then, given occurrence data, probabilistic inferences are made about the number of birds that make different migratory transitions.

Concluding remarks

In this review, we have illustrated some of the unique challenges inherent in broad-scale citizen-science data sets and how novel data-intensive techniques can overcome these challenges. Although our discussion has focused on one citizen-science project, eBird, the general approaches that we have taken to designing the data collection and analyses processes are more widely applicable. Our experience has shown that an uncomplicated protocol and appropriate rewards for volunteer participation is important for the recruitment and retention of large numbers of volunteers. Then, once the data are collected and passed through quality control processes, we have found that existing methods for analysis may not always be suitable for use with such broad-scale observational data. Ongoing collaboration between ecologists, statisticians and computer scientists enable the development of novel methods for extracting biological insights from these data.

Novel methods for analysis can take multiple forms: (i) adding appropriate structure to existing, less structured analysis methods and models (occupancy-detection-expertise models, Box 2; STEM analysis framework, Figures 1 and 2); (ii) eliminating structural constraints from existing analysis methods (semi-parametric occupancy-detection models, Box 3); or (iii) creating entirely novel methods of analysis designed to estimate specific parameters of interest (i.e. Collective Hidden Markov Models; CHMM). We believe that the future use of broad-scale citizen-science data will involve similar creation of novel extensions to existing analysis methods and the development of new analysis techniques.

From the perspective of ecologists, the modeling of occurrence and habitat associations are increasingly important tools for conservation planning and land management, providing fundamental information for the design of conservation strategies across large landscapes [51]. We have outlined how the careful collection and analysis of species occurrence data from a broad-scale network of citizen scientists can be harnessed to address the need for accurate, year-round predictive models of bird distributions across varied spatial extents. These models provide a framework for range-wide and full life-cycle conservation strategies, necessary to reverse population declines and implement habitat-management objectives for threatened species [21,51–53]. As we continue to collect large volumes of observational data on birds, we can extend these analyses to study year-to-year movement patterns of many North American species and assess the effects of environmental contamination. The application of data-intensive scientific methods to birds in North America, and similar work on other taxa and in other regions, will enable land-managers and conservation biologists to better coordinate national and international conservation efforts with the aid of citizen-science data.

Acknowledgements

We thank R. Bonney, T. Dietterich, B. Sullivan and three anonymous reviewers for improving the clarity of our article. This work was funded by the Leon Levy Foundation, Wolf Creek Foundation and the National Science Foundation (Grant Numbers OCI-0830944, CCF-0832782, ITR-0427914, DBI-1049363, DBI-0542868, DUE-0734857, IIS-0748626, IIS-0844546, IIS-0612031, IIS-1050422, IIS-0905385, IIS-0746500, AGS-0835821, CNS-0751152, CNS-0855167, TG-DEB110008).

References

1 Damoulas, T. *et al.* (2010) Bayesian classification of flight calls with a novel Dynamic Time Warping kernel, In *Proceedings of the 9th IEEE International Conference on Machine Learning and Applications*, DOI: 10.1109/ICMLA.2010.69

2 Stoeckle, M. (2003) Taxonomy, DNA, and the Bar Code of Life. *BioScience* 53, 796–797

3 Turner, W. *et al.* (2003) Remote sensing for biodiversity science and conservation. *Trends Ecol. Evol.* 18, 306–314

4 Hochachka, W.M. *et al.* (2007) Data-mining discovery of pattern and process in ecological systems. *J. Wildl. Manag.* 71, 2427–2437

5 Smith, G.W. (1995) *A Critical Review of the Aerial and Ground Surveys of Breeding Waterfowl in North America*, US Department of the Interior

6 Bonney, R. *et al.* (2009) Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience* 59, 977–984

7 Lintott, C.J. *et al.* (2008) Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices R. Astronomical Soc.* 389, 1179–1189

8 Conrad, C. and Hilchey, K. (2011) A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environ. Monit. Assess.* 176, 273–291

9 Novacek, M.J. (2008) Engaging the public in biodiversity issues. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11571–11578

10 Silvertown, J. *et al.* (2011) Citizen science reveals unexpected continental-scale evolutionary change in a model organism. *PLoS ONE* 6, e18927

11 Freeman, S.N. *et al.* (2007) Modelling population changes using data from different surveys: the Common Birds Census and the Breeding Bird Survey. *Bird Study* 54, 61–72

12 Sauer, J.R. and Link, W.A. (2011) Analysis of the North American Breeding Bird Survey using hierarchical models. *Auk* 128, 87–98

13 Sullivan, B.L. *et al.* (2009) eBird: a citizen-based bird observation network in the biological sciences. *Biol. Conserv.* 142, 2282–2292

14 Cooper, C.B. *et al.* (2007) Citizen science as a tool for conservation in residential ecosystems. *Ecol. Soc.* 12, 11

15 McClintock, B.T. *et al.* (2010) Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology* 91, 2446–2454

16 Boakes, E.H. *et al.* (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol.* 8, e1000385

17 Scott, J.M. *et al.*, eds (2002) *Predicting Species Occurrences: Issues of Accuracy and Scale*, Island Press

18 Hey, T. *et al.*, eds (2009) *The Fourth Paradigm: Data-intensive Scientific Discovery*, Microsoft

19 Kelling, S. *et al.* (2009) Data-intensive science: a new paradigm for biodiversity studies. *BioScience* 59, 613–620

20 Munson, M.A. *et al.* (2010) A method for measuring the relative information content of data from different monitoring protocols. *Methods Ecol. Evol.* 1, 263–273

21 Kelling, S. Using bioinformatics In citizen science. In: *Citizen Science: Public Collaboration in Enviromental Research* (Dickinson, J. and Bonney, R., eds), Cornell University Press (in press)

22 Howe, J. (2008) *Crowdsourcing. Why the Power of the Crowd is Driving the Future of Business*, Crown Business

23 Rogstadiusa, J. *et al.* (2011) *An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets*, Association for the Advancement of Artificial Intelligence

24 Nichols, J.D. and Williams, B.K. (2006) Monitoring for conservation. *Trends Ecol. Evol.* 21, 668–673

25 Silvertown, J. (2009) A new dawn for citizen science. *Trends Ecol. Evol.* 24, 467–471

26 Hargrove, W.W. and Pickering, J. (1992) Pseudoreplication: a *sine qua non* for regional ecology. *Landscape Ecol.* 6, 251–258

27 Kell, D.B. and Oliver, S.G. (2004) Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays* 26, 99–105

28 Phillips, S.J. *et al.* (2009) Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecol. Appl.* 19, 181–197

29 Fink, D. *et al.* (2010) Spatiotemporal exploratory models for large-scale survey data. *Ecol. Appl.* 20, 2131–2147

30 Osborne, P.E. *et al.* (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. *Divers. Distributions* 13, 313–323

31 Phillips, S.J. *et al.* (2004) A maximum entropy approach to species distribution modeling. In *Proceedings of the 21st International Conference on Machine Learning* (Brodley, C.E., ed.), p. 83, ACM International Conference Proceeding Series

32 Pearman, P.B. *et al.* (2010) Within-taxon niche structure: niche conservatism, divergence and predicted effects of climate change. *Ecography* 33, 990–1003

33 Fotheringham, A.S. *et al.* (2002) *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley

34 Finley, A.O. (2011) Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods Ecol. Evol.* 2, 143–154

35 Miller, H.J. and Han, J., eds (2009) *Geographic Data Mining and Knowledge Discovery: An Overview in Geographic Data Mining and Knowledge Discovery* (2nd edn), Champan & Hall/CRC

36 North American Bird Conservation Initiative and U.S. Committee (2011) *The State of the Birds 2011: Report on Public Lands and Waters*, US Department of Interior

37 Elith, J. *et al.* (2008) A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802–813

38 Marmion, M. *et al.* (2009) The performance of state-of-the-art modelling techniques depends on geographical distribution of species. *Ecol. Model.* 220, 3512–3520

39 Buston, P.M. and Elith, J. (2011) Determinants of reproductive success in dominant pairs of clownfish: a boosted regression tree analysis. *J. Anim. Ecol.* 80, 528–538

40 Pan, L.L. *et al.* (2008) A neural network-based method for risk factor analysis of West Nile virus. *Risk Analysis* 28, 487–496

41 MacKenzie, D.I. *et al.* (2006) *Occupancy Estimation and Modeling: Inferring Patterns and dynamics of Species Occurrence*, Academic Press

42 Kéry, M. *et al.* (2010) Predicting species distributions from checklist data using site-occupancy models. *J. Biogeography* 37, 1851–1862

43 Wikle, C.K. (2003) Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology* 84, 1382–1394

44 Gill, R.E., Jr *et al.* (2009) Extreme endurance flights by landbirds crossing the Pacific Ocean: ecological corridor rather than barrier? *Proc. R. Soc. B: Biol. Sci.* 276, 447–458

45 Cook, A. *et al.* (2007) Bayesian inference for the spatio-temporal invasion of alien species. *Bull. Math. Biol.* 69, 2005–2025

46 Walker, D.M. *et al.* (2006) Stochastic modelling of ecological processes using hybrid Gibbs samplers. *Ecol. Model.* 198, 40–52

47 Walker, E. and Bez, N. (2010) A pioneer validation of a state-space model of vessel trajectories (VMS) with observers' data. *Ecol. Model.* 221, 2008–2017

48 Leung, B. *et al.* (2004) Predicting invasions: propagule pressure and the gravity of allee effects. *Ecology* 85, 1651–1660

49 Clark, J.S. (2007) *Models for Ecological Data: An Introduction*, Princeton University Press

50 Sheldon, D. *et al.* (2007) Collective inference on Markov models for modeling bird migration. In *Neural Information Processing Systems conference* (Platt, J.C. *et al.*, eds), pp. 1321–1328, MIT Press

51 Kremen, C. *et al.* (2008) Aligning conservation priorities across taxa in Madagascar with high-resolution planning tools. *Science* 320, 222–226

52 Rich, T.D. *et al.* (2004) *Partners in Flight North American Landbird Conservation Plan*, Cornell Lab of Ornithology

53 Will, T.C. et al. (2005) The Five Elements Process: Designing Optimal Landscapes to Meet Bird Conservation Objectives, Partners in Flight Technical Series No. 1.

54 McCulloch, C.E. *et al.* (2008) *Generalized, Linear, and Mixed Models*, John Wiley & Sons

55 Wood, S.N. (2006) *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC

56 Yu, J. *et al.* (2010) Modeling experts and novices in citizen science data for species distribution modeling. In *Proceedings of the 2010 IEEE International Conference on Data Mining* (Webb, G.I. *et al.*, eds), pp. 1157–1162, IEEE Computer Society

57 Hutchinson, R.A. *et al.* (2011) Incorporating boosted regression trees into ecological latent variable models. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence* (Burgard, W. and Roth, D., eds), pp. 1343–1348, Association for the Advancement of Artificial Intelligence