

**Supporting Information.** Daniel Fink, Tom Auer, Alison Johnston, Viviana Ruiz-Gutierrez, Wesley M. Hochachka, and Steve Kelling. 2020. Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*.

## **Appendix S7: Section S1: Trend Simulation Model and Study Design**

A simulation study was used to assess the quality of the trend estimates over the ten-year study period, 2007-2016. The study used spatially explicit simulations to generate data with specified trends while also capturing important aspects of the species' habitat use and the citizen science observation process, both learned from training data. The power, error rate, and bias of the signal filter were assessed along with errors between known and estimated trends. There are three steps in the study:

- 1) Simulate data derived from populations with known trends,
- 2) Using simulated data, estimate trends, and
- 3) Compare known and estimated trends and record statistics to describe errors, power, and bias of estimates.

The remainder of this section describes the simulation model, how the model was used to generate simulated data, the study design, and an evaluation of the breeding and nonbreeding trend estimates for Wood Thrush.

### **S1.1 The Simulation Model**

The simulation model was based on a ZI-BRT, as described above, modified to learn specified trends along with ecological and observational patterns in the training data. Let

$(N, Y, X_e, X_o, year)$  be the set of training data for a given region, season, and species where:

- $N$  is the  $n \times 1$  vector of observed counts on the  $n$  surveys in the training data,
- $Y$  is the  $n \times 1$  vector that indicates the checklists with count greater than zero,
- $X_e$  is the  $n \times k$  matrix of  $k$  predictors that describe the ecological process,
- $X_o$  is the  $n \times j$  matrix of  $j$  predictors that describe the observation process, and
- $year$  is the  $n \times 1$  vector of the year each survey was conducted.

For the given species, the region and season selected for the trend analysis and simulation must be large enough to achieve sufficient sample sizes for good model performance, controlling variance, and small enough to assume stationarity, controlling bias. We conduct two seasonal analyses for Wood Thrush, one across the breeding range from May 30–July 3 and the second across the non-breeding range from Dec 1–Feb 28.

First, we set notation and describe the unmodified ZI-BRT and then we explain the modifications used for the simulation. In the first step of the unmodified ZI-BRT, a Bernoulli response BRT is trained to predict the probability of occurrence:

$$Y \sim \text{Bernoulli}(\pi)$$

$$\text{logit}(\pi) = f(X_e, X_o, year)$$

where  $\pi$  is the probability of occurrence and the function  $f()$  is fit using boosted decision trees. In the second step, the Poisson response BRT,

$$N \sim \text{Poisson}(\mu)$$

$$\text{log}(\mu) = f(X_e, X_o, year)$$


is trained to predict the expected counts  $\mu$ , using the subset of the training data observed and/or predicted to be occupied.

To simulate the data we modify the ZI-BRT as follows. The first modification permutes the *year* predictor variable. This ensures that the ZI-BRT cannot learn year-to-year variation from the training data and effectively removes all trends in the learned ecological and observation processes. The only temporal trend maintained is the increase in the volume of data collected in later years. The second modification trains the ZI-BRT using the year-permuted training data along with an offset constructed with the specified trend. In general terms, the trend offset is  $O = g(\textit{year}^p)$  where  $g()$  is a function of the permuted year value,  $\textit{year}^p$ . The modified fitting procedure begins with the Bernoulli response BRT,

$$Y \sim \textit{Bernoulli}(\pi)$$
$$\textit{logit}(\pi) = f(X_e, X_o, \textit{year}^p),$$

and for the Poisson response BRT is:

$$N \sim \textit{Poisson}(\mu)$$
$$\textit{log}(\mu) + O = f(X_e, X_o, \textit{year}^p).$$



Being on the right side of this equation, the offset can be considered as an adjustment to the observed counts on the log-link scale. Thus, the boosting procedure that adaptively fits  $f()$  has information to estimate  $g(\textit{year}^p)$  from the offset.

## **S1.2 Simulating New Data**

After the modified ZI-BRT is trained, new data are simulated in three steps. First, a new set of eBird observations is generated by sampling checklists with replacement, without regard to the search year, from the training data. Sampling this way replicates the variation observed among participant site selection, search effort, and observer effects. Year-to-year increases in the sample sizes were replicated by repeating this sampling process, independently for each year. In the second step the modified ZI-BRT is used to predict the expected occurrence and abundance,  $\pi^*$  and  $\mu^*$ , for the set of new observations,  $(X_o, X_e, year)^*$ , where the \* denotes the simulated data. Finally, the binary occurrence is simulated  $Y^* \sim Bernoulli(\pi^*)$  and the count, conditional on  $Y^*$  is simulated  $N^* \sim Poisson(\mu^*)$ , generating the simulated data set,  $(N, Y, X_e, X_o, year)^*$ .

## **S1.3 Simulation Study Design**

The simulation study was used to assess the power to detect changes in seasonal population sizes at moderately fine (25.2km x 25.2km) spatial resolution using citizen science data. Qualitatively, we want to understand how performance varies with the strength of the trend and if the method can detect spatial patterns in local trends.

To test how power varied with trend strength, simulations were constructed with increasing and decreasing trends across a range of magnitudes. To test if the method could detect spatial patterns in local trends both spatially constant and spatially varying trends were constructed. Spatially varying trends were constructed so that trend direction and magnitude varied as a function of local population density, giving rise to different

trend directions at the core and edges of population distributions. Flat population trends were also included in the design to assess false positive rates. All together the study consisted of 22 combinations of spatial pattern and magnitude.

The three types of spatial trend offsets constructed were: 1) spatially constant trends, 2) spatially varying trends and 3) no trend. We used the following linear model to construct the trend offsets,  $O = \alpha \textit{year} + \alpha_I \textit{year} X_I$ , where  $\alpha$  controls the strength and direction of the overall year-to-year changes in the expected log count and  $\alpha_I$  controls the strength of the interaction between *year* and  $X_I$ , the interacting variable. Note that because an intercept is fit as part of  $f()$ , we do not include an additional intercept term in the offset.

Spatially uniform trends were generated by setting  $\alpha_I = 0$ . Trends that affect a population uniformly over a region may indicate the indirect effects of broad-spatial scale processes like climate change. Spatially varying trends can be generated by setting  $\alpha = 0$  and specifying a spatially patterned variable  $X_I$  to interact with *year*. To assess if spatial patterns associated with density dependent population processes can be detected, we selected  $X_I$  to be the PLAND cover class predictor with the largest Spearman rank correlation between itself and  $\pi^*$ , used here as an index of population density. Processes like habitat loss, disease, and dispersal can interact with population density to generate spatially varying trend patterns, e.g. Channell and Lomolino (2000) and Massimino, et al. (2015).

Using two parameter sweeps, the spatially constant models were generated with the  $\alpha$  ranging from -0.08 to 0.08 in 11 values spaced 0.016 apart and the spatially varying models were generated with  $\alpha_l$  ranging from -0.40 to 0.40 in 11 values spaced 0.08 apart for a total of 22 simulation treatments. The strongest trends were parameterized to generate relatively large regions within the species' range experiencing changes in population size of at least 6.7% per year over 10 years, one of the IUCN red-list criteria for endangered populations (IUCN 2019).

#### **S1.4 Simulation Evaluations**

Trend estimation proceeds in two steps, as described above, where the signal filter first detects local trends and the trend magnitude is estimated in locations where the direction of trends is consistent. For each simulation we evaluated the power, error rate, and bias of the signal filter along with the correspondence between the magnitude of known and estimated trends. The false detection proportion (FDP) was calculated as the number of locations on the 25km grid where trends were erroneously detected, *as a proportion of the total number of locations where trends were detected*. The power was calculated as the proportion of locations where a trend was correctly identified *out of all locations known to have non-zero trends*. To understand how power varied as a function of the local trend strength, power was also evaluated across all locations with known trends with a minimum magnitude, ranging from 0 to 15% per year. Where the signal filter detected local trends, the coefficient of determination ( $R^2$ ) was computed to describe the proportion of variation in the known magnitudes explained by the estimates.

For each of the breeding and non-breeding seasons, a separate simulation study was conducted for each of the 22 simulation treatments. For each treatment, the training data was spatiotemporally sampled and the year predictor variable was permuted to fit the simulation model. The AdaSTEM base models for each simulation treatment were trained using 100 independent realizations of simulated data.

We measured the performance of the trend estimates averaged across the full suite of simulation treatments to estimate the expected performance across a wide variety of trend scenarios. An important part of this assessment was quantifying directional biases when detecting trends. When biases were found, we adjusted the signal filter to provide robust control against false detection of trends and conservative power estimates. If the FDP was found to exceed a specified error limit (e.g. 5, 10 or 20%) for more than 10% of the of all the locations in all of the simulations, we considered the trend estimator to be biased for that error limit and season. To quantify and adjust for this bias we modified the directional hypotheses used for the signal filter,  $H_0: \delta_s < (0.5 - B^-)$  and  $H_0: \delta_s > (0.5 + B^+)$  where parameters  $B^+$  and  $B^- \in (0,0.5]$  describe the directional biases. As the values of each bias parameter increases, the signal filter requires more consistency in the direction of the trend estimates across the ensemble, thereby reducing the FDP and the subsequent power of the test. Thus, as the values of each bias parameter increases, the fewer locations on the trend map where trends can be identified while guaranteeing the FDR at the specified error limit.

To estimate the directional biases, we performed a parameter sweep evaluating FDP and power across all combinations of values of  $B^+, B^- \in (0.0, 0.01, 0.02, \dots, 0.25)$ . Then we estimated the value of  $B^+, B^-$  that maximized power subject to the constraint that FDP was less than the specified limit (e.g. 5, 10, or 20%) across  $\geq 90\%$  of the simulations.

All of trend estimates reported here, for both breeding and nonbreeding seasons, were made using a FDR limit of 5%. For each of the breeding and nonbreeding simulations we estimated the direction bias parameters ( $B^+, B^-$ ) and used them to estimate trends. Thus, all of the trend maps, power statistics, and  $R^2$  measurements reported in this paper were made using these bias corrections under a 5% FDR limit.

### **S1.5 Wood Thrush Simulations**

Two simulation studies were conducted for the Wood Thrush over the 2007-2016 study period, one for the breeding season (May 30–July 3) across the species' range in the northeastern North America and the second for the non-breeding season (Dec 1–Feb 28) across the species' range in Central America.

The simulations provide qualitative information describing the ability of the method to identify spatially varying trend patterns among locations. Fig.s S1-4 show simulated and estimated trend maps for a sample of simulation treatments across a broad array of spatially constant and spatially varying trends with trend magnitudes that vary in direction and magnitude. The trend magnitudes varied along the rows of each figure with weak (regions with trends  $\sim|1\%/yr|$ ), medium (regions with trends  $\sim|3.5\%/yr|$ ), and



strong (regions with trends  $\sim|6.7\%/yr|$ ) trend magnitudes. This suite of spatial trend patterns is varied enough to begin to assess the method's ability to estimate spatial patterns across locations. The quality of the trend estimates improves from weak to strong trend magnitudes, regardless of spatial pattern or direction. Regional patterns are identified, though with errors. Errors in detecting trends are most frequent when simulated trends are weak, and become less frequent as trends become stronger. The magnitude of the estimates generally varies with simulated trend strength, visible as the correspondence between the darkness of the colors shown for the estimate and simulation trend map pairs (Fig.s S1-4). However, in regions with declining trends the trend magnitude appears to be underestimated in the nonbreeding season and among the spatially varying treatments in the breeding season.

The power curves for 5, 10, and 20% FDR constraints for both seasonal simulation analyses show the expected pattern of increasing power with increasing minimum trend magnitude (Fig. S5). The plots also show the expected tradeoff between FDR and power, with increasing power as the FDR constraint becomes more lenient. We recognize that in some conservation applications the false detection of declining trends, carries a far lower risk for a species than failing to detect a declining trend, and in such circumstances, it may make sense to increase the error limit to 10 or 20% to improve the power to detect trends as can be seen in Fig. S5.

Finally, the correspondence between estimated and simulated known trend magnitudes was stronger in the breeding season ( $R^2=75.6\%$ ) than the nonbreeding season

( $R^2=59.5\%$ ). Overall, these simulation results suggest that breeding season trend estimates will be more accurate, powerful, and less variable than those in the nonbreeding season. In general, this is expected because of the much higher density of data across the breeding range compared to the nonbreeding range.

## **Literature Cited**

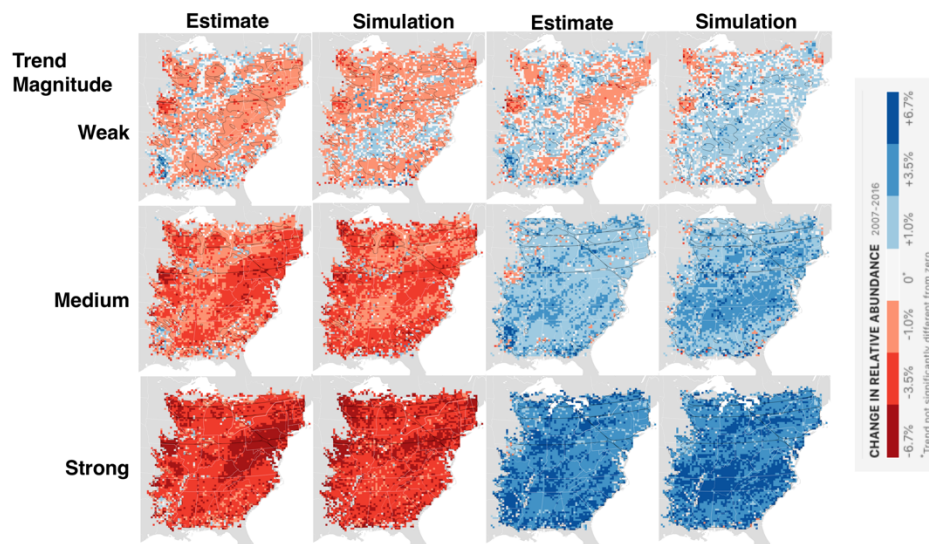
Channell, R., and Lomolino, M. V. (2000). Trajectories to extinction: Spatial dynamics of the contraction of geographical ranges. *Journal of Biogeography* 27:169–179.

IUCN 2019. The IUCN Red List of Threatened Species. Version 2019-1.

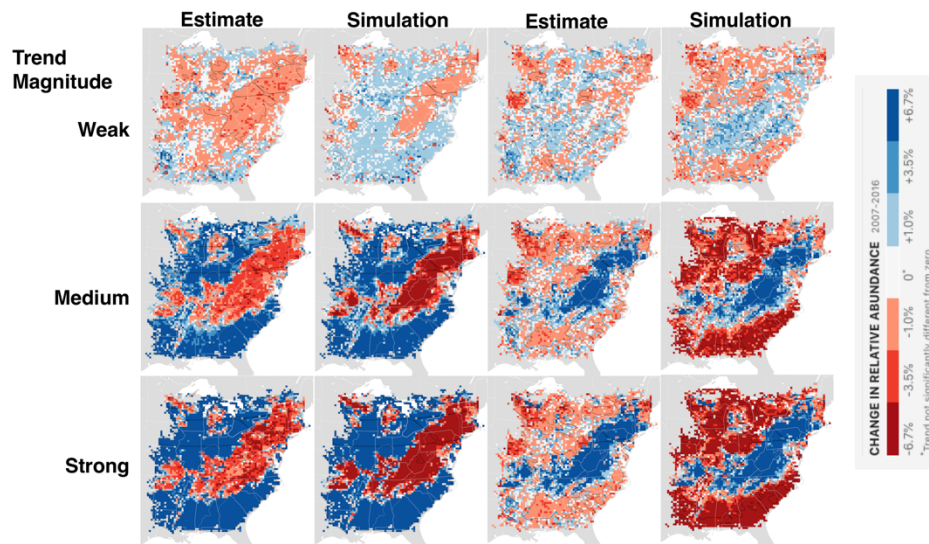
<http://www.iucnredlist.org> . Downloaded on 21 March 2019.

Massimino, D., Johnston, A., Noble, D. G., & Pearce-Higgins, J. W. (2015). Multi-species spatially-explicit indicators reveal spatially structured trends in bird communities. *Ecological indicators*, 58, 277-285.

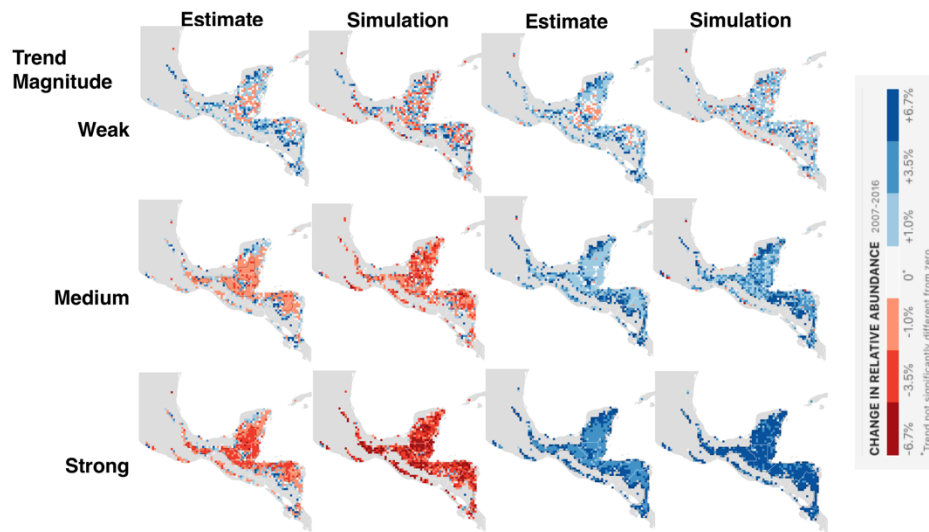
<https://doi.org/10.1016/j.ecolind.2015.06.001>



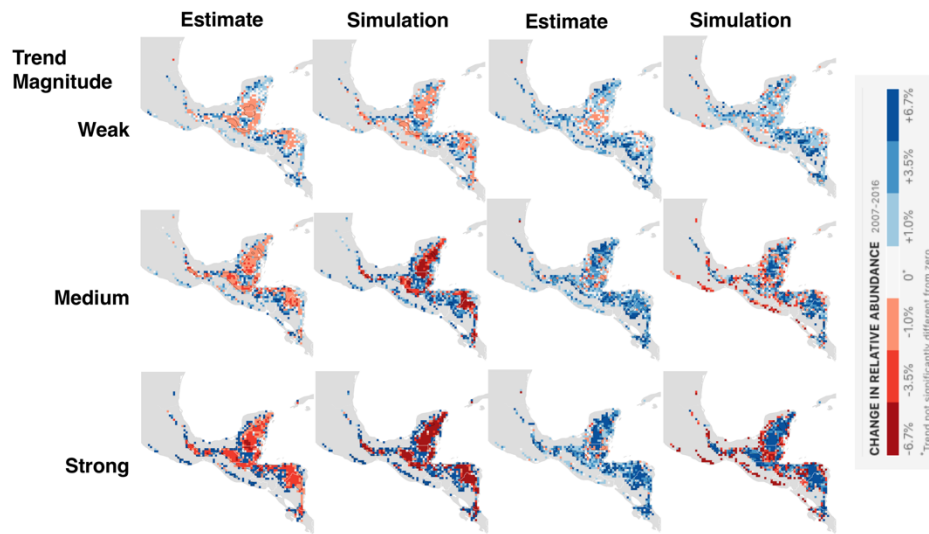
**Figure S1: Wood Thrush breeding season simulated and estimated trend maps for spatially constant treatments.** The trend magnitude varies along the rows with weak (includes regions with trends  $\sim|1\%/yr|$ ), medium (includes regions with trends  $\sim|3.5\%/yr|$ ), and strong (includes regions with trends  $\sim|6.7\%/yr|$ ) trend magnitudes. The first two columns show estimated and simulated trends for decreasing trends. The third and fourth columns show estimated and simulated trends for decreasing trends. The black contours delineate the regions across which the expected False Discovery Rate is at most 5%.



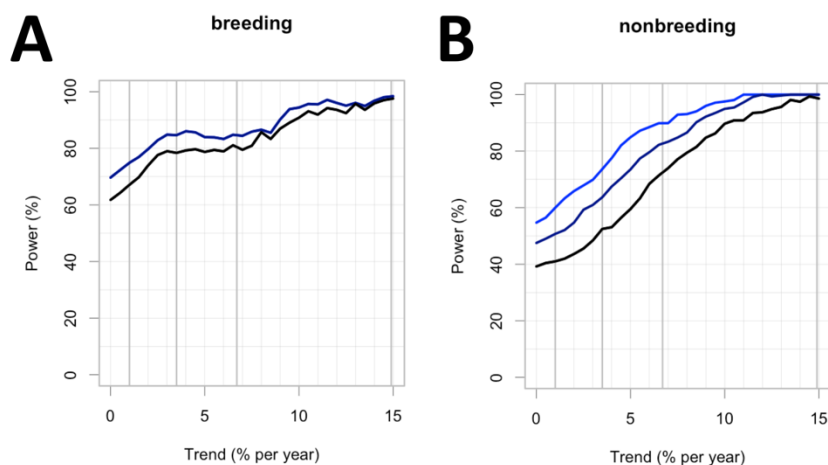
**Figure S2: Wood Thrush breeding season simulated and estimated trend maps for spatially varying treatments.** The trend magnitude varies along the rows with weak (includes regions with trends  $\sim|1\%/yr|$ ), medium (includes regions with trends  $\sim|3.5\%/yr|$ ), and strong (includes regions with trends  $\sim|6.7\%/yr|$ ) trend magnitudes. The first two columns show estimated and simulated trends for decreasing trends. The third and fourth columns show estimated and simulated trends for decreasing trends. The black contours delineate the regions across which the expected False Discovery Rate is at most 5%.



**Figure S3: Wood Thrush nonbreeding season simulated and estimated trend maps for spatially constant treatments.** The trend magnitude varies along the rows with weak (includes regions with trends  $\sim|1\%/yr|$ ), medium (includes regions with trends  $\sim|3.5\%/yr|$ ), and strong (includes regions with trends  $\sim|6.7\%/yr|$ ) trend magnitudes. The first two columns show estimated and simulated trends for decreasing trends. The third and fourth columns show estimated and simulated trends for decreasing trends. The black contours delineate the regions across which the expected False Discovery Rate is at most 5%.



**Figure S4: Wood Thrush nonbreeding season simulated and estimated trend maps for spatially varying treatments.** The trend magnitude varies along the rows with weak (includes regions with trends  $\sim|1\%/yr|$ ), medium (includes regions with trends  $\sim|3.5\%/yr|$ ), and strong (includes regions with trends  $\sim|6.7\%/yr|$ ) trend magnitudes. The first two columns show estimated and simulated trends for decreasing trends. The third and fourth columns show estimated and simulated trends for decreasing trends. The black contours delineate the regions across which the expected False Discovery Rate is at most 5%.



**Figure S5: Wood Thrush seasonal power curves as a function of the minimum simulated trend magnitude.** Power varies as a function of the minimum trend magnitude for the (A) breeding and (B) nonbreeding season analyses. Power is reported as the percentage of all locations in range across the simulated known map that meet the minimum magnitude requirement that were identified with the correct trend direction when FDR was constrained at 5 (black), 10 (dark blue), and 20% (light blue). The maximum false detection proportion was 8% for the breeding season, so no light blue line is shown. The overall power corresponds to a minimum trend of zero, at the leftmost side of the graph. The black line corresponds to the black contour lines in Fig. 4 and 5.