Received: 21 December 2022 | Accepted: 26 June 2023

DOI: 10.1111/2041-210X.14186

Methods in Ecology and Evolution

RESEARCH ARTICLE

A Double machine learning trend model for citizen science data

Daniel Fink[1] | Alison Johnston[2] | Matt Strimas-Mackey[1] | Tom Auer[1] | Wesley M. Hochachka[1] | Shawn Ligocki[1] | Lauren Oldham Jaromczyk[1] | Orin Robinson[1] | Chris Wood[1] | Steve Kelling[1] | Amanda D. Rodewald[1]

[1]Cornell Lab of Ornithology, Cornell University, Ithaca, New York, USA

[2]Centre for Research into Ecological and Environmental Modelling, School of Maths and Statistics, University of St Andrews, St Andrews, UK

Correspondence
Daniel Fink
Email: daniel.fink@cornell.edu

Funding information
The Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, Grant/Award Number: DEB200010; Leon Levy Foundation; National Science Foundation, Grant/Award Number: DBI-1939187, ICER-1927646, 2138259, 2138286, 2138307, 2137603 and 2138296; Wolf Creek Charitable Foundation; Academy of Finland, Grant/Award Number: 326338 and 326327; Swedish Research Council, Grant/Award Number: 2018-02441 and 2018-02440; Research Council of Norway, Grant/Award Number: 295767

Handling Editor: Nicolas Lecomte

Abstract

1. Citizen and community science datasets are typically collected using flexible protocols. These protocols enable large volumes of data to be collected globally every year; however, the consequence is that these protocols typically lack the structure necessary to maintain consistent sampling across years. This can result in complex and pronounced interannual changes in the observation process, which can complicate the estimation of population trends because population changes over time are confounded with changes in the observation process.
2. Here we describe a novel modelling approach designed to estimate spatially explicit species population trends while controlling for the interannual confounding common in citizen science data. The approach is based on Double machine learning, a statistical framework that uses machine learning (ML) methods to estimate population change and the propensity scores used to adjust for confounding discovered in the data. ML makes it possible to use large sets of features to control for confounding and to model spatial heterogeneity in trends. Additionally, we present a simulation method to identify and adjust for residual confounding missed by the propensity scores.
3. To illustrate the approach, we estimated species trends using data from the citizen science project eBird. We used a simulation study to assess the ability of the method to estimate spatially varying trends when faced with realistic confounding and temporal correlation. Results demonstrated the ability to distinguish between spatially constant and spatially varying trends. There were low error rates on the estimated direction of population change (increasing/decreasing) at each location and high correlations on the estimated magnitude of population change.
4. The ability to estimate spatially explicit trends while accounting for confounding inherent in citizen science data has the potential to fill important information gaps, helping to estimate population trends for species and/or regions lacking rigorous monitoring data.

KEYWORDS
Causal Forests, causal inference, citizen science, confounding, Double machine learning, machine learning, propensity score, trends

1 | INTRODUCTION

Information on population trends is essential for conservation monitoring and management (Rosenberg et al., 2019). To date, the estimation of interannual trends has largely been restricted to the analysis of data from structured surveys where, ideally, the same observers follow the same survey protocols at the same locations, dates and times each year. This controlled survey structure is used to minimize the interannual variation in the observation process that can lead to confounding. However, these survey requirements also make it difficult to collect species-observation data at the scales necessary to monitor large groups of species across broad spatial extents, and at arbitrary times of year.

Citizen science projects are collecting increasingly large volumes of data on a variety of taxa (Pocock et al., 2017), however, due to the opportunistic nature of data collection in such projects, these datasets are vulnerable to interannual changes in the observation process. For example, several studies have documented interannual variation in spatial site selection (August et al., 2020; Shirey et al., 2021; Zhang et al., 2021) and its interaction with search effort (Tang et al., 2021). Participant populations change as new participants join projects and continuing participants improve the way they conduct surveys (Johnston et al., 2022). Data collection protocols change, either through deliberate choice or for uncontrollable reasons. Examples of deliberate changes are those caused by the use of short-term incentives or games (Xue et al., 2016), and in the long-term promotion of 'best practice' protocols (e.g. submission of complete checklists; Sullivan et al., 2009). Uncontrollable changes include improvements to equipment, such as binoculars, the development of species identification apps, and external forces shaping observers' behaviour, such as the COVID pandemic (e.g. Hochachka et al., 2021). Studies of citizen science data have also shown how interannual changes in the observation process bias trend estimates. Bowler et al. (2022) showed how changes in the spatial site selection produce species-specific biases in trend estimates. Zhang et al. (2021) showed how bias can even arise despite interannual survey structure, documenting how unexpected changes in survey censoring due to urbanization biased trends in species richness. Controlling for the wide array of potentially confounding factors that give rise to interannual variation constitutes a major challenge when employing citizen science data for trend estimation.

While linear and generalized linear models can account for confounding by incorporating observed confounding features, achieving reliable control of confounding can be challenging (Gelman et al., 2020). This is particularly true when confounding is complex or there are large sets of potentially confounding features (Hernán & Robins, 2020), as is common with many citizen science datasets. Specifying accurate parametric models in these settings requires the selection of relevant features while identifying nonlinear effects and interactions. Machine learning (ML) algorithms are an attractive alternative for this task; however, naïve application of these methods introduce regularization bias when estimating effects (Belloni et al., 2014). Double machine learning (DML) (Chernozhukov et al., 2018) was developed to address this challenge, providing a statistical framework in which generic ML methods (such as penalized regression, lasso, random forests, boosted models and deep neural networks) can be used to control for confounding without the risk of regularization bias. By harnessing large sets of features and ML models, DML enables control for more intricate confounding patterns thereby reducing confounding bias and strengthening inference.

In this paper we consider DML for estimating spatially explicit species population trends from citizen science data. Conceptually, DML divides trend estimation into three separate modelling tasks. The goal of the first task is to predict local population sizes averaged across the study period. To do this, a species distribution model is trained to learn how observations of species vary with a set of features (e.g. climate, landcover, search effort). The goal of the second task is to identify confounding sources of variation in the data. To do this, a propensity score model is trained, which describes how the features vary systematically over time (Ramsey et al., 2019). In the third task, the expected population size and observation year are used as benchmark values to help isolate the trend so it can be estimated without the influence of confounding features.

There are three challenges for adapting the DML framework to the problem of estimating trends with citizen science data. The first challenge is to assess propensity score performance at controlling confounding. Given the critical role played by the propensity score model, it is important to understand how well a given model does controlling confounding. Even with highly automated ML methods, analyst choices about model selection, hyperparameter tuning, feature set selection and engineering can significantly affect performance. However, we are not aware of any standard diagnostics for assessing the effectiveness of the propensity score model at controlling confounding. To meet this need, we propose a novel simulation-based diagnostic measure to identify residual confounding, that is, confounding sources of variation in the available feature data missed by the propensity score model. Information gained from these diagnostics can be used to develop models and adjust the trend estimates.

The second challenge adapting the DML framework to estimate spatially explicit trends is that the original DML framework assumes global linear trend effects, whereas the ability to estimate trends with high spatial resolution (e.g. landscape scale) is valuable for studying the processes affecting populations at these scales (e.g. agriculture, energy development, urbanization) (Rose et al., 2017). For

this reason we chose to use Causal Forests (Athey et al., 2019), an implementation of the DML framework that uses Random Forests (Breiman, 2001) to estimate heterogenous local-linear effects as a function of user-specified features. Thus, spatially explicit trends can be estimated by including spatially varying environmental features that are associated with variation in population trends in the Causal Forests.

The third analytical challenge is created by temporal correlation, which is an important and pervasive characteristic of species abundance data collected over time. The DML framework does not include any structural components to account for correlated observations. Here we assess impact of temporal correlation on model performance through a suite of simulation studies based on temporally correlated data.

We use the DML trend model for a real-world application: estimating population trends with data from eBird, a well-studied citizen science project with a complex observation process and a relatively large set of known potentially confounding features (e.g. Johnston et al., 2019). eBird is a popular citizen science project that has been collecting bird observation data since 2002 (Sullivan et al., 2014). The project engages large numbers of participants who each decide where, when and how to participate. As with many other citizen science projects, the limited structure in eBird has given rise to an evolving, heterogenous observation process where interannual confounding is a central concern when estimating population trends. The goal is to estimate the average annual rate of change in breeding season abundance 2007–2021 at a 27 km resolution across North America for three species of birds with different distributions and habitat preferences: wood thrush *Hylocichla mustelina*, Canada warbler *Cardellina canadensis* and long-billed curlew *Numenius americanus*. Species-specific simulation studies were used to assess confounding control, overall performance and the ability to estimate spatially varying trends when faced with temporally correlated observations.

2 | THE DML TREND MODEL

In this section, we introduce the DML trend model and the simulation-based residual confounding analysis.

2.1 | Double machine learning

To estimate population trends, we begin with the model that describes variation in species abundance Y, the response variable (also called the *outcome* or *label* variable), as

$$Y = \tau T + \mu(X) + \epsilon. \quad (1)$$

The objective is to estimate parameter τ, the rate of change in abundance per unit time T. For convenience we assume Y is a real-valued measure or index of species abundance, but integer counts, or binary indicators of a species' occurrence can be accommodated without loss of generality. We also assume that T measures time in units of years and that τ is the interannual trend, but other units can be accommodated to estimate trends over different time scales without loss of generality. The function μ is a nonparametric function of the vector $X = (X_1, \ldots, X_k)$ consisting of the features (also called *covariates* or *predictor* variables) that capture effects that are constant across years. Features in X can include both ecological process variables (e.g. habitat or climatic conditions) and observation process variables (e.g. search effort or survey time of day). The number of features, k, can be large. The variable ϵ is a stochastic error term.

To understand how confounding can affect trend estimation in (1) it is useful to consider an idealized dataset where the features X include all important sources of variation in abundance and the observations are collected as a random sample across X, independently drawn each year of the study period. Under these conditions, $E[\epsilon \mid X, T] = 0$ and the trend τ can be estimated without bias. In practice, confounding can arise when there are systematic year-to-year changes in the observation process that affect the distribution of X. For example, surveys could be conducted in sites with increasing habitat quality over time.

A common strategy to account for confounding when analysing nonexperimental data uses propensity scores to adjust estimates (Rosenbaum & Rubin, 1983). In this approach a propensity score model is introduced to keep track of confounding, which can be framed here as the dependence of T on features X. The propensity score model is written as,

$$T = s(X) + \delta, \quad (2)$$

where s is a nonparametric function of the features X and δ is a stochastic error term where $E[\delta \mid X] = 0$.

DML solves Equations (1) and (2) using the plug-in estimator of (Robinson, 1988). The plug-in estimator is constructed by substituting the conditional mean response averaged across T,

$$m(X) = E[Y \mid X] = \tau\, s(X) + \mu(X), \quad (3)$$

into Equation (1) to yield the residual-on-residual regression,

$$(Y - m(X)) = \tau\,(T - s(X)) + \epsilon. \quad (4)$$

The residual on the left side of Equation (4) isolates the change in abundance by regressing out year-invariant effects of features X on Y and the residual on the right side removes the effects of confounding by regressing out the effects of features X on T. This formulation motivates the plug-in estimator where $m(X)$ and $s(X)$ are separately predicted and then plugged into Equation (4) to estimate the trend, τ.

The decomposition in Equation (4) was originally designed to use plug-in values from unbiased linear estimators. In practice, linear models can be restrictive because they can be hard to specify accurately and are difficult to scale to large numbers of features, making it difficult to generate accurate plug-in predictions. Chernozhukov et al. (2018) showed how the plug-in estimator can used to accurately estimate τ even when the predictions of $m(X)$ and $s(X)$ are noisy and suffer from regularization bias. This makes

it possible to take advantage of large feature sets X using generic statistical and ML methods (e.g. penalized regressions, lasso, random forests, boosted models, deep neural networks and ensembles of these methods). The ability to accurately predict propensity scores from large, complex sets of features is important because it can improve inference by strengthening the control of confounding.

The goal in this paper is to use DML to strengthen inference about population trends by controlling for interannual confounding, but with additional assumptions DML can also be used for causal inference. The model in Equation (1) is closely connected to the potential outcomes framework (Rubin, 1974) that describe the conditions necessary for casual inference. Within this framework, DML estimators are unbiased and normally distributed (Chernozhukov et al., 2018). Thus, unlike many ML methods, DML can be used for statistical and causal inference. (See SI Section S2 for more information about the potential outcomes framework.)

2.2 | Residual confounding

Propensity scores provide a theoretically sound strategy to fully control for confounding in X *provided* that the propensity score model estimated from the data $\hat{s}(X)$ can fully and accurately capture the confounding. If $\hat{s}(X)$ fails to fully capture the confounding, then trend estimates will be biased. We refer to this situation as *residual confounding* with respect to X.

Here we propose a fully data-driven simulation-based diagnostic for residual confounding that can be computed for any DML model. Recognizing that the conditional mean model estimated in Equation (3) $\hat{m}$ is a zero or null-trend model, we use $\hat{m}$ to simulate datasets $\{X^*, W^*, T^*, Y^*\}$ with a known population trend $\tau^{\text{sim}} = 0$ while maintaining the interannual confounding in the original features. Then to diagnose residual confounding, we look for systematic differences between τ^{sim} and τ^*, the DML trend estimate based on the simulated data. The differences $\left(\tau^{\text{sim}} - \tau^*\right)$ can be used to compare models and adjust τ^*. The algorithm is,

1. Generate a synthetic feature set $\{X^*, W^*, T^*\}$ by resampling with replacement from $\{X, W, T\}$ stratified by T, and, then,
2. Compute synthetic responses $Y^* = \hat{m}(X^*) + \epsilon^*$, where ϵ^* is generated by bootstrapping the residuals.

The first step generates realistic feature sets that maintain the joint distribution of the features including the interannual variation in X, while avoiding extrapolation in the feature space. Using the conditional mean model estimated in Equation (3) in the second step ensures that the synthetic data have zero trend while maintaining year-invariant patterns of variation in abundance associated with X. Moreover, because these synthetic data are based on a zero or null-trend model, additional assumptions (beyond those of the DML) about unknown trends are avoided. (See SI Section S3 for an illustration of the residual confounding diagnostic and adjustment using a toy example based on synthetic data.)

2.3 | Spatial heterogeneity and Causal Forests

Causal Forests (Athey et al., 2019) is an implementation of the DML framework that uses Random Forests (Breiman, 2001) for the conditional mean and propensity score models. It also uses a modified random forest (Wager & Athey, 2018) to estimate the trend τ as a nonparametric function of the feature vector $W = (W_1, \ldots, W_m)$, where the number of features m can be large. This extends the global linear trend in Equation (1) to

$$Y = \tau(W)T + \mu(X) + \epsilon. \quad (5)$$

The ability to estimate trends conditional on a set of features W can be used to identify and study heterogeneity in population change. For example, by including spatial features in W, spatially explicit trends can be estimated. In the causal inference literature, Equation (5) is known as a *heterogenous treatment effect* or *conditional average treatment effect* estimator. In statistics it is equivalent to a varying coefficient model for τ (Hastie & Tibshirani, 1993). (See SI Section S1 for a brief review of DML literature.)

3 | TREND ANALYSIS OF NORTH AMERICAN BREEDING BIRDS

We estimate trends based on data from eBird, a popular citizen science project that has been collecting bird observation data since 2002 (Sullivan et al., 2014). This application presents the challenges of estimating spatially explicit trends in abundance across large geographic extents in the face of confounding and temporally correlated observations. The eBird dataset also includes a relatively large number of potentially confounding features.

In this section we describe the eBird data, the species abundance model underlying trend estimation, the Causal Forest implementation, the residual confounding analysis and the species-specific simulation study used to assess the performance of the method. All computing was done in the R statistical computing language (R Core Team, 2019).

3.1 | Data

eBird is a semistructured survey (Kelling et al., 2019) because its flexibility allows participants to collect data in the ways they choose, but with ancillary data collected to describe the variation in the observation process. To help control for variation in observation process, we analysed the subset of the data for which participants report all bird species detected and identified during the survey period, resulting in *complete checklists* of bird species. This limits variation in preferential reporting rates across species and provides a basis to infer species

nondetections. We also required all checklists to include key ancillary variables describing characteristics of each birdwatching event, for example the time of day and distance travelled. These variables and others can be used to adjust for variation in detection rates (Johnston et al., 2019).

We calculated trends for three species that represent a range of different breeding niches, observation processes and processes driving population change. Wood thrush is a commonly reported bird of the deciduous forest in eastern North America. Canada warbler is a less commonly reported forest bird that breeds in the boreal forests of North America. Long-billed curlew is an infrequently reported shorebird that breeds in the grasslands of the arid interior of North America.

We analysed eBird data from 2007 to 2021 within each species' previously identified breeding range and season (Fink, Auer, Johnston, Strimas-Mackey, et al., 2020). To prepare the data for the trend analysis we aggregated data within cells of a (27 km × 27 km × 1 week) grid based on checklist latitudes, longitudes and dates. We computed grid cell averages for four classes of information: (1) The number of individuals of the given species reported in each grid cell was used as the response variable (*Y*); (2) Five observation-effort features describing how participants conducted surveys were used as features to account for variation in detection rates (Johnston et al., 2019); (3) 14 features describing short-term temporal variation—date, time of day and hourly weather—were used as features to account for variation in availability for detection; and (4) A suite of 57 spatially varying features describing the composition and configuration of landscapes in each grid cell were used to capture associations between species and elevation, topography, land and water cover, land use, hydrology and road density. Please see the SI Section S4 for details about data and data processing.

3.2 | Species abundance model

A species' expected abundance can be defined as the product of the species' occurrence rate and the expected count of the species given occurrence, within a given area and time window (Zuur et al., 2009). Based on this definition and the chain rule, the rate of change in species abundance is the sum of two terms: (1) the rate of change in the species occurrence, and (2) the rate of change in species count given occurrence. Intuitively, trends in species abundance can arise from trends in the occurrence rate (e.g. as a function of whether the habitat is even suitable for a species) and/or trends in expected counts given occurrence.

We estimated each trend component with its own DML model. To estimate the interannual rate of change in occurrence rate a Causal Forest was trained based on the binary response variable indicating the detection/nondetection of the species and the features. Then to estimate the interannual rate of change in the log-transformed species count, a separate Causal Forest was trained based on the continuous response variable (log-transformed count) and features, using the subset of surveys where the species was detected (i.e. all counts were positive).

To quantify the sampling variation in abundance trend estimates arising jointly from the estimated trend in species occurrence rates and the estimated trend in species counts, given occurrence, we adopted a data resampling approach and computed an ensemble of 100 estimates. We calculated 80% confidence intervals using the lower 10th and upper 90th percentiles across the ensemble. Additionally, averaging estimates across the ensemble provides a simple way to control for overfitting (Efron, 2014). Please see the SI Section S5 for additional information about the resampling approach used to construct the ensemble.

3.3 | Causal Forest implementation

Causal Forests were fit using the GRF package (Tibshirani et al., 2020) and were grown with 2000 trees using automatic parameter tuning for all parameters. The feature sets for the conditional mean and propensity score models included (1) observation effort, (2) short-term temporal and (3) spatial features. We also included latitude and longitude as features in the conditional mean model to account for residual spatial patterns of abundance. To account for spatial variation in trends we included all spatial features in **W** along with latitude and longitude to account for residual patterns.

3.4 | Residual confounding

For the eBird analysis we extended the residual confounding simulations that were introduced in Section 2.2 to assess confounding bias under a range of nonzero trend scenarios. (See Section 3.5 for a description of the scenarios.) It was important to evaluate the method across a broader set of plausible trends to provide a basis for generalizing results when applied to real data where the trend is unknown. We implemented the residual confounding diagnostics and adjustments at the species level because interannual variation in how participants conduct surveys can generate distinct biases for each species. The goal was to find a set of parameters to describe and correct for residual confounding that would generalize well for all locations in the species range, regardless of the direction, magnitude or spatial pattern of the unknown trend. To do this we fit a linear regression model for each species, estimated using all locations within the species' range for all simulation scenarios. Predictions from this regression model were used to make residual confounding adjustments (see Figure 1).

3.5 | Simulation study

The species-specific simulations were constructed to create data meeting four conditions for each species: (1) realistic patterns of year-invariant patterns of occurrence and counts on eBird checklists;

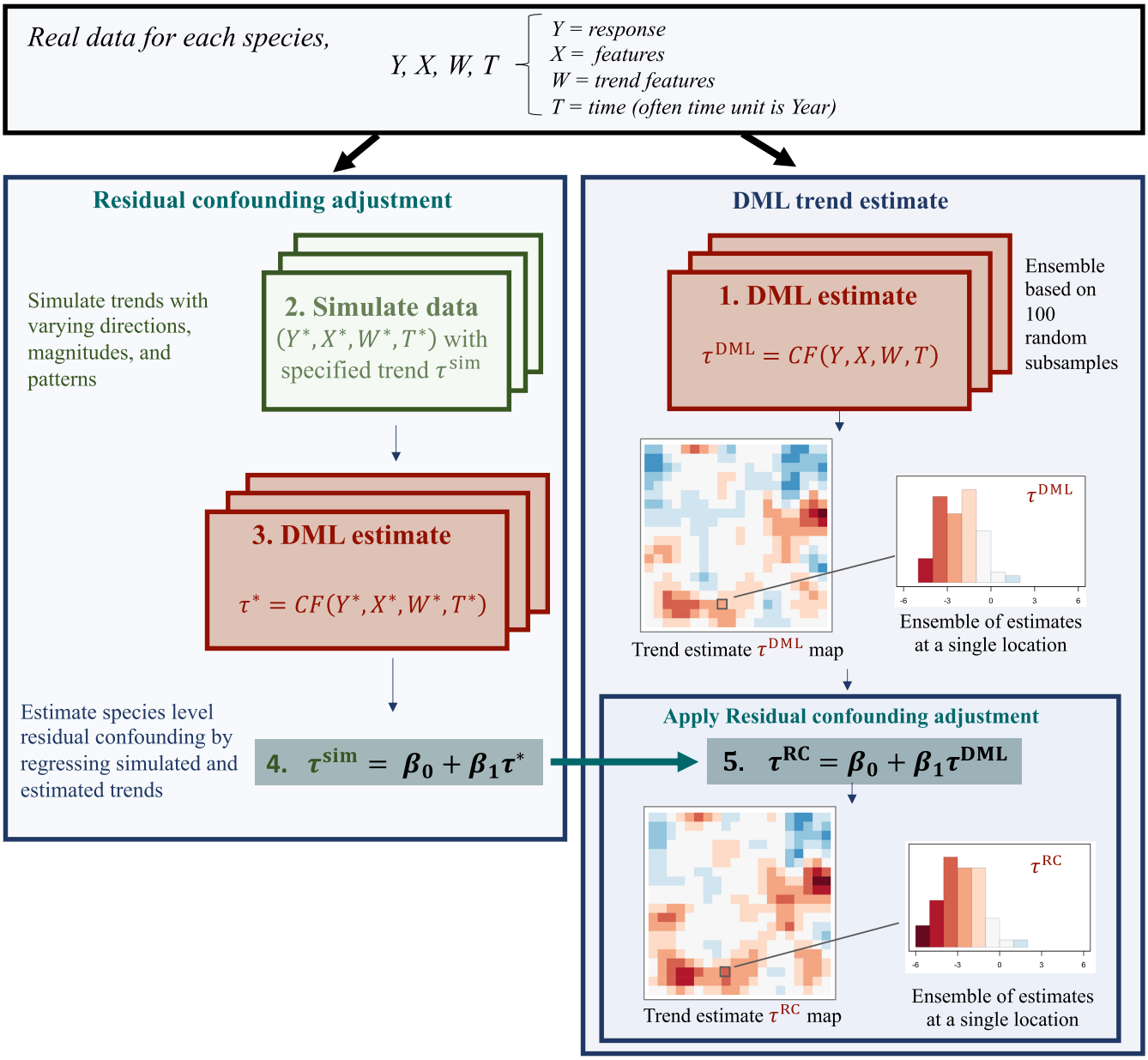

FIGURE 1 A schematic workflow for the DML abundance trend model with residual confounding adjustment. This schematic is based on the eBird analysis with numbers corresponding to the steps in Section 3.5. The residual confounding adjustment (LEFT) is based on the systematic differences between simulated trends (Step 2; expanded in Figure SI-1) and the DML estimates of these trends (Step 3, red boxes). Step 4 estimates the residual confounding coefficients by formally linking Step 2 and Step 3. The real data analysis (RIGHT) begins by calculating the DML trend estimate (Step 1) and then adjusts for residual confounding (Step 5). An ensemble of trend estimates is generated at each location, from which the point estimates and confidence intervals are calculated.

(2) with specified trends in abundance; (3) including temporal correlation generated from environmental stochasticity in population growth rates; while (4) maintaining the interannual confounding in the original eBird data. (See the SI Section S6 for additional details about the simulations.)

To assess the overall performance in detecting and describing spatial trend patterns, trends were simulated at a 27 km×27 km spatial scale across each species range and across 10 scenarios with zero and nonzero trends varying in direction, magnitude and spatial pattern. Magnitudes were set to <1% (weak), 3.3% (moderate) and 6.7% per year (strong) based on IUCN Red List criteria (IUCN, 2019). The spatially varying trends were constructed to vary in direction and magnitude along a gradient from the core to the edge of the species' population (Figure 3; Figures SI-3 and SI-4). (See the SI Section S6.3 for details about the simulation scenarios.) The population dynamics at each location were generated using a discrete-time stochastic exponential growth rate model. The deterministic growth rate component varied by scenario and location. The stochastic growth rate realizations were independent among years and locations with a standard deviation of approximately 6.7% per year. The compounded effects of this

stochasticity across years generated temporal correlation as would be expected from the landscape-scale effects of processes like extreme weather or reproductive success. (See the SI Section S6.4 for details about the discrete-time stochastic exponential growth rate model.)

The simulations were used for two tasks: (1) The training task to compute the residual confounding estimates, and (2) the testing task to evaluate the trend estimates after accounting for residual confounding. To maintain independence among these tasks, we independently generated two sets of simulated data for training and testing. Each simulated training dataset contained data from one of five different simulation scenarios (one null and four with varying magnitude, spatially constant and variable trends). Ten datasets were independently generated for each of the five scenarios. Thus, the simulated training data included 50 datasets generated under five different trend scenarios. Fifty simulated test datasets were generated independently of the training datasets, from five comparable but different scenarios.

Analysis workflow: To compute each species' DML trend estimate with the residual confounding adjustment we followed the steps in the schematic workflow shown in Figure 1:

1. Estimate τ^{DML} using Causal Forests with propensity score adjustments based on the original data $\{X, W, T, Y\}$ (Sections 3.2 and 3.3),
2. Simulate data $\{X^*, W^*, T^*, Y^*\}$ with specified trends τ^{sim} (Section 3.5),
3. Estimate τ^* using Causal Forests with propensity score adjustment based on the simulated training data $\{X^*, W^*, T^*, Y^*\}$,
4. Fit $\tau^{\text{sim}} = \beta_0 + \beta_1 \tau^*$, the residual confounding regression (Section 3.4),
5. Apply the residual confounding adjustment to the original DML estimate: $\tau^{\text{RC}} = \beta_0 + \beta_1 \tau^{\text{DML}}$ using simulation-based parameters (β_0, β_1) to adjust τ^{DML} based on the original data.

3.5.1 | Confounding bias in eBird

To measure the strength of the confounding bias and the performance of the propensity score and residual confounding adjustments we performed a simulation analysis and ablation test comparing the trend estimates, τ^1, the Causal Forest estimates *without* adjusting for the propensity scores to τ^{DML} and τ^{RC}, with both trends computed using the workflow above.

To measure the performance of the trend estimates we used the simulated test data with the specified trends $\tau^{\text{sim}}_{\text{test}}$ as the original data. This allowed us to assess the species-level confounding bias by fitting the regression $\tau^{\text{sim}}_{\text{test}} = \alpha_0 + \alpha_1 \tau$ separately for each of the estimates $\tau = (\tau^1, \tau^{\text{DML}}$ and $\tau^{\text{RC}})$. The intercept parameter α_0 measured the distance between a zero trend estimate and the corresponding expected value of the simulated trend. Thus, the intercept described the bias when estimating the trend direction (increasing or decreasing), with a value of zero indicating no *directional bias*. The slope parameter α_1 measured how simulated trends scaled with the direction and magnitude of the estimated trends, with a value of 1 indicating no *scaling bias*.

The impact of the propensity score on directional bias is assessed by comparing bias coefficients between τ^1 and τ^{DML}, where an improvement is α_0 moving towards 0 and α_1 moving towards 1. Estimates show that the propensity score adjustments reduced directional bias for all three species (Table 1). The propensity score adjustment also reduced the scaling bias for Wood thrush, with a smaller reduction for Canada warbler and a small increase for long-billed curlew (Table 1).

The impact of the residual confounding adjustment on directional bias is assessed by comparing bias coefficients between τ^{DML} and τ^{RC}. The residual confounding adjustment strongly reduced both directional and scaling bias for long-billed curlew but had smaller effects on the other two species (Table 1). For Canada warbler the residual confounding adjustment increased the magnitude of the directional bias and slightly decreased the magnitude of the scaling bias. For wood thrush the residual confounding adjustment led to a slight increase in the magnitude of the directional bias and it decreased the magnitude of the scaling bias.

3.5.2 | Trend performance

Next we assessed the performance of τ^{RC}, the DML trend estimates computed with both the propensity score and the residual

TABLE 1 Species-level estimates of confounding bias. Slope and intercept estimates and standard errors (SEs) are presented for each species for trend estimates τ^1 without any correction for confounding, trend estimates τ^{DML} with the propensity score (PS) adjustment, and trend estimates τ^{RC} with the PS and the residual confounding (RC) adjustment.

Species	Estimator	PS	RC	Intercept α_0	Intercept SE	Slope α_1	Slope SE
Wood thrush	τ^{RC}	Yes	Yes	0.295	0.005	1.045	0.001
	τ^{DML}	Yes	No	0.254	0.005	1.342	0.001
	τ^1	No	No	−0.610	0.005	1.480	0.001
Canada warbler	τ^{RC}	Yes	Yes	0.641	0.011	0.931	0.002
	τ^{DML}	Yes	No	−0.212	0.010	1.159	0.002
	τ^1	No	No	−1.589	0.011	1.171	0.002
Long-billed curlew	τ^{RC}	Yes	Yes	0.119	0.011	0.998	0.002
	τ^{DML}	Yes	No	−3.663	0.010	1.153	0.002
	τ^1	No	No	−4.767	0.012	1.143	0.002

confounding adjustments, using the simulated test data as the original data in the workflow and then comparing estimates with the specified trends τ^{sim}_{test}. We evaluated the quality of the estimated trend magnitude (the average per cent-per-year [PPY] rate of change in abundance 2007–21) and the trend direction (increasing/decreasing), two important inferential objectives for population monitoring. Directional errors were defined to occur when trends were estimated to be significantly different from zero but were in the opposite direction to the simulated trend. We considered estimates to be nonzero if the 80% confidence interval did not contain zero. Because directional errors varied strongly with trend magnitude (Figure SI-2), we reported the mean directional error rate, binned into categories of trend magnitude (see SI Section S7 for more details about the directional error). We also computed Pearson's correlation between simulated and estimated trends for nonzero trend estimates. Finally, we assessed the coverage of the resampling-based uncertainty estimates as the percentage of all 27 km locations where the estimated intervals contained the simulated trend value. All assessments were based on independent test set data.

The mean directional error rate among nonzero trends was low for all species (Table 2). The correlations among nonzero estimates and simulated true values τ^{sim}_{test} were high. As expected, the directional error rates increased and the correlations decreased with the volume of species' data that were nonzero counts; from wood thrush (a commonly reported species in a region with high data density), to Canada warbler (less commonly reported in regions with lower data density), to long-billed curlew (infrequently reported compared to wood thrush and Canada warbler within a relatively low data-density region of the continent). Interval coverage increased with decreasing amounts of species data (note, both sample sizes and detection rates decrease among species), though it was markedly less than the nominal confidence 80% level for all species.

3.5.3 | Identifying spatial trends

Finally, we assessed model performance for describing spatial heterogeneity in trends. We compared model performance between spatially constant and spatially varying scenarios for each species (Table 2). The similarity in performance between constant and varying trends highlights the ability of the model to adapt to heterogenous trends. Figure 2 shows trend maps for wood thrush for a single realization of each simulation scenario (see SI Figures SI-4 & SI-5 for Canada warbler and long-billed curlew). These maps show how the model adapted to simulations with different directions, magnitudes and spatial patterns.

3.6 | Species trend estimates

Figure 3 shows maps of the estimated average annual per cent-per-year change in abundance from 2007 to 2021 for all three species based on the real data. The wood thrush population shows steep declines in the northeast and increases in the southwest of its breeding season population, a pattern similar to other published studies (e.g. Fink, Auer, Johnston, Ruiz-Gutierrez, et al., 2020) The estimated population change for Canada warbler also shows spatial patterning, though the uncertainty is relatively high. Long-billed curlew shows strong, significant range-wide declines consistent with previous analysis (Rosenberg et al., 2019).

4 | DISCUSSION

Our work shows that the DML trend model is a promising method for estimating spatially explicit interannual trends based on citizen science data. Simulation results demonstrated the ability to control confounding in realistic settings based on temporally correlated eBird observations with relatively large sets of confounding features. Model estimates accurately estimated trend direction and magnitude and were sufficiently accurate to distinguish between spatially constant and spatially varying patterns at a 27 km × 27 km resolution.

This study highlights the importance of considering interannual confounding when estimating interannual trends with citizen science data. It also highlights several challenges implementing the DLM trend model including the tasks of modelling propensity scores, estimating uncertainty, and accounting for spatial and temporal correlation. In this

Species	Trend scenarios	Directional error	Correlation	CI coverage
Wood thrush	All	2.4%	96%	47%
	Constant	1.9%	97%	44%
	Varying	1.9%	92%	48%
Canada warbler	All	6.8%	88%	49%
	Constant	5.1%	92%	45%
	Varying	5.5%	79%	51%
Long-billed curlew	All	3.4%	88%	61%
	Constant	2.5%	91%	56%
	Varying	2.6%	86%	62%

TABLE 2 Trend estimate performance. The directional error, correlation and interval coverage are presented for each species, averaged across all evaluation scenarios (All) the spatially constant evaluation scenarios (Constant), across the spatially varying scenarios (Varying).

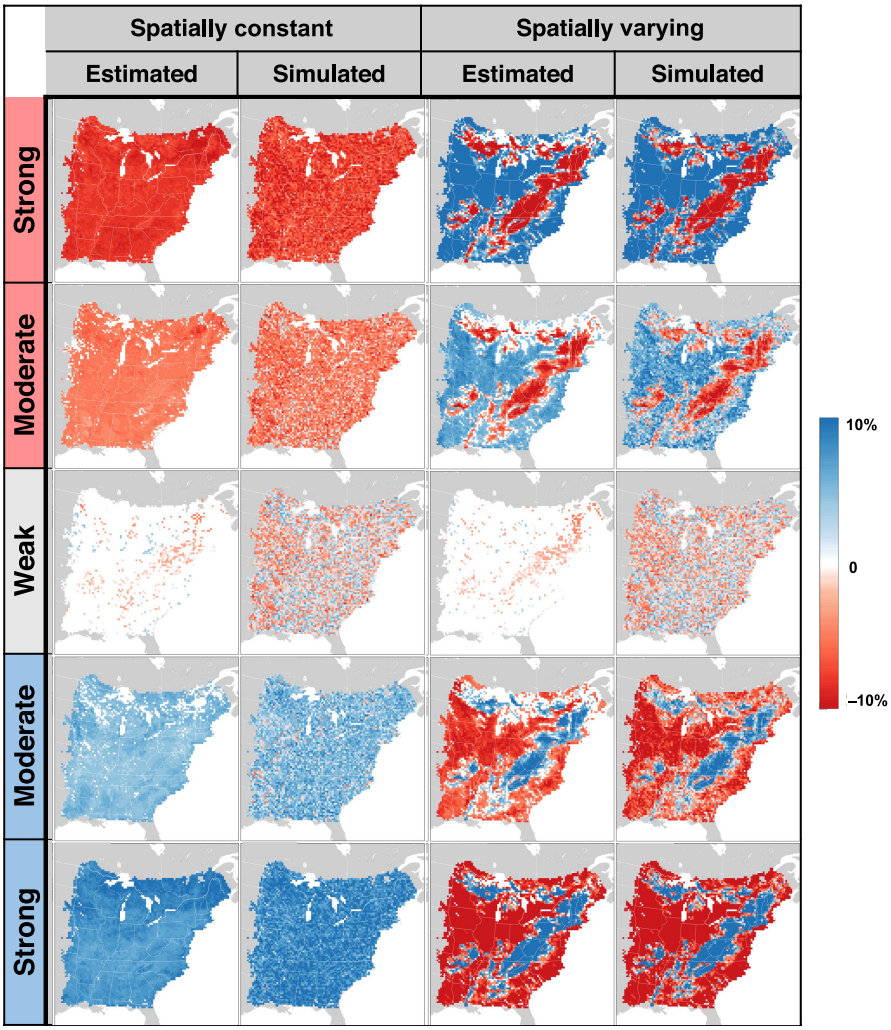

FIGURE 2 Wood thrush trend simulations. All trend maps show the average annual per cent-per-year change in abundance from 2007 to 2021 within 27 km pixels (red = decline, blue = increase, white = 80% confidence interval contained zero), intensity (darker colours indicate stronger trends). Simulated trends show scenarios varying by direction and magnitude along rows: weak (includes trends ~| 1% / year |), moderate (includes regions with trends ~| 3.5% / year |) and strong trends (includes regions with trends ~| 6.7% / year |). The columns show simulated and estimated trends for spatially constant and varying simulation scenarios.

section we discuss these challenges and how DML trend models may be used for other applications and citizen science datasets.

4.1 | Confounding in citizen science data

Without any correction for confounding, the simulation results showed that confounding bias can be strong (−4.77% per year directional bias for long-billed curlew), though it varied among species (−0.61% per year directional bias for wood thrush). These results align with other studies highlighting the risk of interannual confounding bias when analysing opportunistically collected data (Bowler et al., 2022; Zhang et al., 2021). However, to date, most trend analyses based on opportunistically collected citizen science data have largely ignored the issue of interannual confounding, implicitly assuming its absence (e.g. Bianchini & Tozer, 2023; Boersch-Supan et al., 2019; Horns et al., 2018; Walker & Taylor, 2017). Comparing control for confounding with other, common trend models (e.g. generalized linear models or occupancy model) are an important direction for future research.

Confounding bias could help explain the species-level variation and relatively low degree of overall alignment between trend estimates based on structured survey and citizen science datasets (e.g. Boersch-Supan et al., 2019; Walker & Taylor, 2017). Accounting for confounding bias will also be important when integrating citizen science data with data from other surveys. Failure to do so would likely result in variable performance among species and decreased precision.

We presented a two-stage approach to control confounding. The first stage uses propensity score adjustments made as part of the DML model. The ablation study in Section 3.5.1 showed that the propensity scores largely reduced the confounding bias for wood thrush and Canada warbler; reducing the directional bias to approximately |0.25|% per year. For long-billed curlew, the propensity score reduced but did not eliminate the confounding bias (directional bias was reduced from −4.77% to −3.67% per year), despite using fully tuned Casual Forest models with relatively large sample sizes and a relatively well-studied citizen science dataset with known confounding features (e.g. Johnston et al., 2019). This result highlights the challenging, species-specific nature of the confounding problem and it motivated our development of the residual confounding diagnostic and adjustment.

In Section 3.5.1 the second stage residual confounding adjustment largely eliminated the residual bias for long-billed curlew

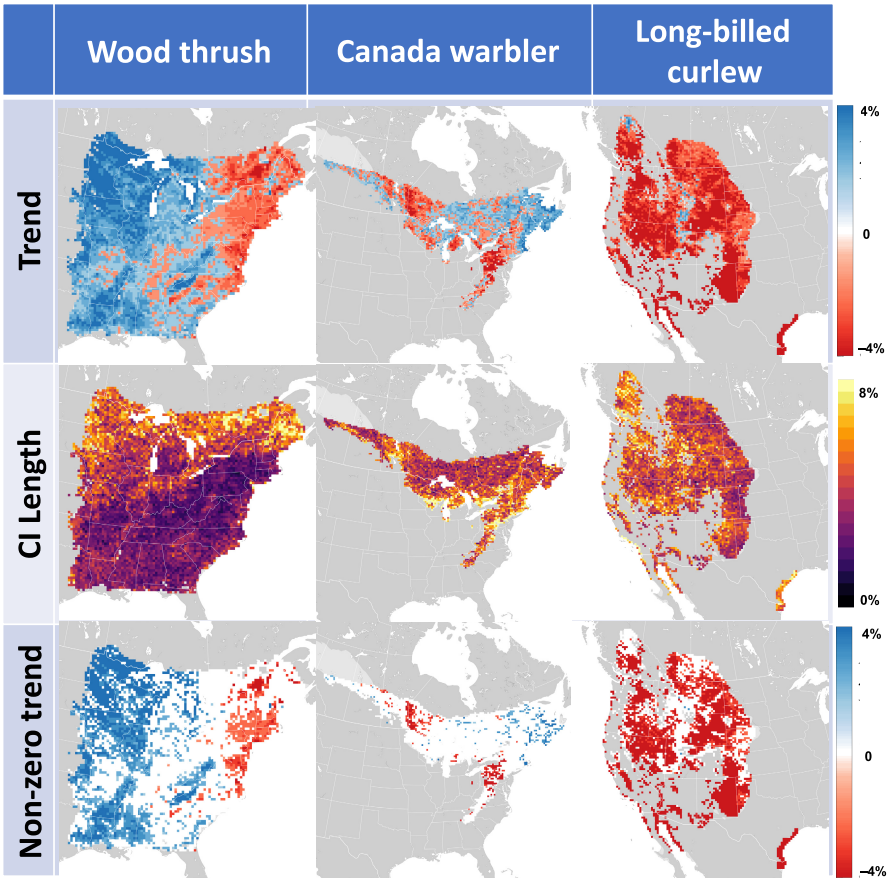

FIGURE 3 Trend estimates wood thrush, canada warbler and long-billed curlew. All trend maps show the average annual per cent-per-year change in abundance from 2007 to 2021 within 27 km pixels (red = decline, blue = increase), intensity (darker colours indicate stronger trends). The top row shows the estimated trends, middle row shows confidence interval length and the bottom panel shows the nonzero trends in red and blue with white in locations where 80% confidence interval contained zero.

reducing the directional bias from −3.67 to 0.12% per year and reducing the scaling bias from 1.15 to 1.00. The results of the residual confounding adjustment were mixed for wood thrush and Canada warbler, producing increases directional bias (from 0.25% to 0.30% per year and from −0.21% to 0.64% per year respectively) while decreases in scaling bias (from 1.34 to 1.05 and from 1.16 to 0.93 respectively). These results suggest a strategy where the residual confounding adjustment is applied selectively when it is most needed. This could be achieved by implementing a test or shrinkage estimator based on residual confounding diagnostics to determine when and how the adjustment is applied. This is an area for future research.

One of the most important steps in specifying ML propensity score models is feature selection. The goal of the propensity score model is to capture sources of interannual variation in the observation process that also impact the reported abundance of the species. This suggests a feature selection strategy that includes all features from the conditional mean model in the propensity score model. This strategy ensures that all important sources of variation in year-invariant abundance are included as the set of potential confounders. This is the strategy implicit in our presentation of the DML trend model. Nonetheless, the propensity score model is not limited to this set of features, and we expect there are situations where it may be advantageous to consider additional features. However, we caution against indiscriminately including features in the propensity score models to avoid introducing bias (Hünermund et al., 2023). For example, bias can be introduced in the DML trend model by including propensity score features that are themselves affected by changes in species abundance and are also changing from year to year. Intuitively, including such features would make it difficult to isolate population change from other sources of interannual variation. The challenge of feature selection in causal inference has received considerable attention and several good references are available (e.g. Cinelli et al., 2020; Hernán & Robins, 2020). This challenge highlights the importance of future work to improve our understanding of citizen science observation processes and how they evolve over time. Incorporating instrumental variables within DML analyses (Chernozhukov et al., 2018) could also help strengthen confounding control and is another area for future research.

The estimation of residual confounding provides a practical way to assess confounding control. This can be used to guide model development (e.g. feature selection and engineering) when implementing DML. Residual confounding can also be used to adjust estimates to reduce bias when propensity score deficiencies are found. In general, we expect residual confounding will be most valuable when the data generating processes responsible for confounding are complex and not well understood, the set of available features is not well suited to model interannual variation in the observation process, or signal to noise ratios are low.

Critical to the success of any residual confounding analysis is the construction of the underlying simulated data. In general, the simulated data need to have a known population trend while maintaining

all the interannual confounding in the original features. The zero or null-trend simulation in Section 2.2 is a convenient, general purpose and completely data-driven analysis based on the estimated conditional mean model. Thus, the applicability and performance of this approach will also depend on the quality of the conditional mean model.

For the eBird data analysis, it was important to assess performance estimating zero, nonzero and spatially structured trends based on temporally correlated observations. To do this we extended the residual confounding simulations to include trends that varied in direction, magnitude and spatial patterns. Our strategy was to inform the simulation data generating process by using real data as much as possible while reducing the synthetic components (and assumptions therein) to a minimum (see Knaus et al. (2021) for other examples of empirically driven simulations). The synthetic components in our simulation were based on two key assumptions, (1) that populations change at the same rate across the study period, and (2) that spatial patterning aligned with edge-core population structure. Expanding the simulations to include nonconstant temporal dynamics and additional spatial patterns is an area for further research.

4.2 | Inferential scope

An important goal of this paper was to investigate the use of DML to control for interannual confounding when estimating trends based on citizen science data. The results show that DML can be used to reduce confounding bias leading to more accurate estimates and stronger associative inferences. These results are in line with other research that has sought to improve associative inferences by harnessing approaches and ideas originally developed for causal inference (Bühlmann, 2020; Cui & Athey, 2022).

With additional assumptions DML can also be used for causal inference. The key assumption to make causal inference is to assert the absence of confounders that are missing from the analysis and independent of the original features (sometimes called *missing, hidden or unmeasured confounders*). Practically, asserting the absence of missing confounders requires assumptions that go beyond the data in hand (Hernán & Robins, 2020). Neither the propensity score model nor the simulation-based residual confounding analysis can detect or control for missing confounders. Thus, end-users need to carefully consider the strength of their domain knowledge and the limits of inference.

4.3 | Uncertainty, temporal correlation and spatially structure

The simulation results showed that confidence interval coverage for the eBird analysis was below the nominal 80% level. We believe this is caused, at least partially, by the outcome model not accounting for the temporal correlation. Nevertheless, these same simulation results demonstrated that there was strong control of directional error when we used the interval estimates to identify nonzero trends, that is, trend estimates whose intervals did not contain zero. We interpret these two results to indicate that the uncertainty may not be scaling appropriately with the magnitude of the trend. Accounting for temporal correlation in the outcome model is an interesting direction for further research into use of the DML framework. Incorporating more powerful rules to identify nonzero trends that control for multiple comparisons (e.g. false detection rate thresholding) and spatial correlation could also serve to improve the power of the approach and the scope of inference for species with weaker trend signals like Canada warbler.

In this study we showed how spatial features can be used to estimate spatial patterns of variation in the trends (Figure 3; Figures SI-3 and SI-4). Residual spatial structure is another common feature of large-scale geographic studies that is absent from the DML. Given that the processes driving population change are more likely to vary locally when data come from large geographic extents (Rose et al., 2017), an important avenue of additional research is into accounting for spatial nonstationarity of the drivers and confounders of trends.

4.4 | DML trend applications

The DML trend model and the simulation-based adjustment presented here can be used with other applications and data types. The Causal Forest implementation (Tibshirani et al., 2020) can accommodate binary and real-valued outcome variables, making it possible to estimate trends in species occurrence rates, expected trends and other indices of abundance. For example, using the DML trend model to estimate trends with higher temporal resolution by generating estimates over shorter time study periods could be useful for studying short-term population fluctuations like those generated by weather, demographics or population cycles. Investigating the bias-variance trade-off associated with trend estimation across temporal resolutions is another area for further research.

The ability to associate features and trends can also be used to study a variety of questions. For example, including indicators in the trend feature set can be used to estimate the effects of different survey protocols, management actions or policies on population change. This could be useful for conservation planning, assessment, to inform integrated analysis or for future survey design. Moreover, by including other features in the trend model that capture other potential trend effects (e.g. changes in landcover) it is possible to study systematic management differences after accounting for changes in landcover. This may be useful for Before-After-Control-Impact studies with citizen science data where accounting for the simultaneous impacts of other management and environmental changes is a challenge (Kerr et al., 2019). Finally, learned associations between trends and features can also be used to forecast expected population changes, conditional on a given set of features values and the assumption that the underlying processes driving population change during the study period will persist into the future.

For the eBird analysis presented here we used the fact that each species' counts were collected as parts of complete checklists of birds, which allowed us to infer the zero counts associated with nondetection. The same approach can be used with other checklist-based citizen science projects to analyse counts as well as binary response presence–absence data (e.g. birds (Johnston et al., 2020) and butterflies (van Swaay et al., 2008)). Even when observations are not collected in the form of complete lists, observations from some taxa can be assembled into pseudo-checklists (Henckel et al., 2020; van Strien et al., 2013) making them amenable to DML trend analysis. It may also be possible to analyse presence-only data (e.g. iNaturalist.org) by carefully selecting (Valavi et al., 2021) or weighting (Fithian & Hastie, 2013) background data. However, more research will be needed to carefully consider biases and confounding associated with presence-only data (e.g. Stoudt et al., 2022). The DML trend model may even be useful for the analysis of data collected from structured surveys where confounding can arise despite survey structure (e.g. Zhang et al., 2021).

5 | CONCLUSIONS

The volume of citizen science data is rapidly growing, but the lack of structured protocols has rendered most of these data unsuitable for estimating population trends without requiring strong assumptions about the absence of confounding. Bias can be introduced by changes over time in how people participate. The DLM trend model can account for these confounding changes over time. When used appropriately, including assessments of the propensity score model used to account for sources of confounding variation, this approach has the potential to increase the biodiversity monitoring value that we can obtain from citizen data. This could enable us to better track population changes in areas of the world with fewer structured monitoring programmes.

AUTHOR CONTRIBUTIONS

Daniel Fink and Alison Johnston conceived the ideas and designed methodology; Tom Auer, Matt Strimas-Mackey, Shawn Ligocki, Chris Wood, Steve Kelling and Amanda D. Rodewald collected the data; Daniel Fink, Alison Johnston, Tom Auer, Matt Strimas-Mackey, Wesley M. Hochachka, Shawn Ligocki, Lauren Oldham Jaromczyk and Orin Robinson analysed the data; Daniel Fink and Alison Johnston led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

The eBird and the eBird Status and Trends projects rely on the time, and dedication and support from countless individuals and organizations. We thank the many thousands of eBird participants for their contributions and the eBird team for their support. We also thank three anonymous reviewers for their thoughtful comments.

FUNDING INFORMATION

This work was funded by The Leon Levy Foundation, The Wolf Creek Foundation and the National Science Foundation (ABI sustaining: DBI-1939187). This work used Bridges2 at Pittsburgh Supercomputing Center and Anvil at Rosen Center for Advanced Computing at Purdue University through allocation DEB200010 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603 and #2138296. Our research was also funded through the 2017–2018 Belmont Forum and BiodivERsA joint call for research proposals, under the BiodivScen ERA-Net COFUND program, with financial support from the Academy of Finland (AKA, Univ. Turku: 326327, Univ. Helsinki: 326338), the Swedish Research Council (Formas, SLU: 2018-02440, Lund Univ.: 2018-02441), the Research Council of Norway (Forskningsrådet, NINA: 295767) and the U.S. National Science Foundation (NSF, Cornell Univ.: ICER-1927646).

CONFLICT OF INTEREST STATEMENT

The authors do not have any conflict of interests to report.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14186.

DATA AVAILABILITY STATEMENT

The R language source code and data for the eBird study and the toy example vignette presented in the Supplemental information are available here https://doi.org/10.5281/zenodo.8092408 (Fink et al., 2023). All bird observation data used to conduct the eBird study are freely available on the eBird website https://ebird.org/science/use-ebird-data. All environmental data used to conduct the eBird study are publicly available. See Supplemental Information and Table SI-1 for references and links.

ORCID

Daniel Fink https://orcid.org/0000-0002-8368-1248
Alison Johnston https://orcid.org/0000-0001-8221-013X
Matt Strimas-Mackey https://orcid.org/0000-0001-8929-7776
Tom Auer https://orcid.org/0000-0001-8619-7147
Wesley M. Hochachka https://orcid.org/0000-0002-0595-7827
Shawn Ligocki https://orcid.org/0000-0001-7163-5570
Orin Robinson https://orcid.org/0000-0001-8935-1242
Amanda D. Rodewald https://orcid.org/0000-0002-6719-6306

REFERENCES

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2), 1148–1178.

August, T., Fox, R., Roy, D. B., & Pocock, M. J. O. (2020). Data-derived metrics describing the behaviour of field-based citizen scientists provide insights for project design and modelling bias. *Scientific Reports*, *10*(1), 11009. https://doi.org/10.1038/s41598-020-67658-3

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Bianchini, K., & Tozer, D. C. (2023). Using breeding bird survey and eBird data to improve marsh bird monitoring abundance indices and trends. *Avian Conservation and Ecology*, *18*(1). https://doi.org/10.5751/ACE-02357-180104

Boersch-Supan, P. H., Trask, A. E., & Baillie, S. R. (2019). Robustness of simple avian population trend models for semi-structured citizen science data is species-dependent. *Biological Conservation*, *240*, 108286. https://doi.org/10.1016/j.biocon.2019.108286

Bowler, D. E., Callaghan, C. T., Bhandari, N., Henle, K., Benjamin Barth, M., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H., & Bonn, A. (2022). Temporal trends in the spatial bias of species occurrence records. *Ecography*, *2022*(8), e06219. https://doi.org/10.1111/ecog.06219

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bühlmann, P. (2020). Invariance, causality and robustness. *Statistical Science*, *35*(3). https://doi.org/10.1214/19-STS721

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097

Cinelli, C., Forney, A., & Pearl, J. (2020). *A crash course in good and bad controls* (SSRN Scholarly Paper No. 3689437). https://doi.org/10.2139/ssrn.3689437

Cui, P., & Athey, S. (2022). Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, *4*(2), 110–115. https://doi.org/10.1038/s42256-022-00445-z

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, *109*(507), 991–1007. https://doi.org/10.1080/01621459.2013.823775

Fink, D., Auer, T., Johnston, A., Ruiz-Gutierrez, V., Hochachka, W. M., & Kelling, S. (2020). Modeling avian full annual cycle distribution and population trends with citizen science data. *Ecological Applications*, *30*(3), e02056.

Fink, D., Auer, T., Johnston, A., Strimas-Mackey, M., Robinson, O., Ligocki, S., Hochachka, W., Wood, C., Davies, I., & Iliff, M. (2020). *EBird status and trends, data version: 2019; released: 2020*. Cornell Lab of Ornithology.

Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W., Ligocki, S., Oldham Jaromczyk, L., Robinson, O., Wood, C., Kelling, S., & Rodewald, A. D. (2023). A double machine learning trend model for citizen science data. *Zenodo*, https://doi.org/10.5281/zenodo.8092408

Fithian, W., & Hastie, T. (2013). Finite-sample equivalence in statistical models for presence-only data. *The Annals of Applied Statistics*, *7*(4), 1917–1939. https://doi.org/10.1214/13-AOAS667

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Hastie, T., & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x

Henckel, L., Bradter, U., Jönsson, M., Isaac, N. J. B., & Snäll, T. (2020). Assessing the usefulness of citizen science data for habitat suitability modelling: Opportunistic reporting versus sampling based on a systematic protocol. *Diversity and Distributions*, *26*(10), 1276–1290. https://doi.org/10.1111/ddi.13128

Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If* (Revised 2022). Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Hochachka, W. M., Alonso, H., Gutiérrez-Expósito, C., Miller, E., & Johnston, A. (2021). Regional variation in the impacts of the COVID-19 pandemic on the quantity and quality of data collected by the project eBird. *Biological Conservation*, *254*, 108974. https://doi.org/10.1016/j.biocon.2021.108974

Horns, J. J., Adler, F. R., & Şekercioğlu, Ç. H. (2018). Using opportunistic citizen science data to estimate avian population trends. *Biological Conservation*, *221*, 151–159. https://doi.org/10.1016/j.biocon.2018.02.027

Hünermund, P., Louw, B., & Caspi, I. (2023). Double machine learning and automated confounder selection: A cautionary tale. *Journal of Causal Inference*, *11*(1). https://doi.org/10.1515/jci-2022-0078

IUCN. (2019). *The IUCN Red List of Threatened Species. Version 2019*. https://www.iucnredlist.org

Johnston, A., Hochachka, W., Strimas-Mackey, M., Ruiz Gutierrez, V., Robinson, O., Miller, E., Auer, T., Kelling, S., & Fink, D. (2019). Analytical guidelines to increase the value of citizen science data: Using eBird data to estimate species occurrence [preprint]. *Ecology*. https://doi.org/10.1101/574392

Johnston, A., Matechou, E., & Dennis, E. B. (2022). Outstanding challenges and future directions for biodiversity monitoring using citizen science data. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13834

Johnston, A., Moran, N., Musgrove, A., Fink, D., & Baillie, S. R. (2020). Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, *422*, 108927. https://doi.org/10.1016/j.ecolmodel.2019.108927

Kelling, S., Johnston, A., Bonn, A., Fink, D., Ruiz-Gutierrez, V., Bonney, R., Fernandez, M., Hochachka, W. M., Julliard, R., Kraemer, R., & Guralnick, R. (2019). Using semistructured surveys to improve citizen science data for monitoring biodiversity. *BioScience*, *69*(3), 170–179. https://doi.org/10.1093/biosci/biz010

Kerr, L. A., Kritzer, J. P., & Cadrin, S. X. (2019). Strengths and limitations of before–after–control–impact analysis for testing the effects of marine protected areas on managed populations. *ICES Journal of Marine Science*, *76*(4), 1039–1051. https://doi.org/10.1093/icesjms/fsz014

Knaus, M. C., Lechner, M., & Strittmatter, A. (2021). Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, *24*(1), 134–161. https://doi.org/10.1093/ectj/utaa014

Pocock, M. J. O., Tweddle, J. C., Savage, J., Robinson, L. D., & Roy, H. E. (2017). The diversity and evolution of ecological and environmental citizen science. *PLoS ONE*, *12*(4), e0172579. https://doi.org/10.1371/journal.pone.0172579

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramsey, D. S. L., Forsyth, D. M., Wright, E., McKay, M., & Westbrooke, I. (2019). Using propensity scores for causal inference in ecology: Options, considerations, and a case study. *Methods in Ecology and Evolution*, *10*(3), 320–331. https://doi.org/10.1111/2041-210X.13111

Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, *56*, 931–954.

Rose, K. C., Graves, R. A., Hansen, W. D., Harvey, B. J., Qiu, J., Wood, S. A., Ziter, C., & Turner, M. G. (2017). Historical foundations and future directions in macrosystems ecology. *Ecology Letters*, *20*(2), 147–157.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., & Parr, M. (2019). Decline of the North American avifauna. *Science*, *366*(6461), 120–124.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701. https://doi.org/10.1037/h0037350

Shirey, V., Belitz, M. W., Barve, V., & Guralnick, R. (2021). A complete inventory of North American butterfly occurrence data: Narrowing data gaps, but increasing bias. *Ecography*, *44*(4), 537–547. https://doi.org/10.1111/ecog.05396

Stoudt, S., Goldstein, B. R., & de Valpine, P. (2022). Identifying engaging bird species and traits with community science observations. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(16), e2110156119.

Sullivan, B. L., Aycrigg, J. L., Barry, J. H., Bonney, R. E., Bruns, N., Cooper, C. B., Damoulas, T., Dhondt, A. A., Dietterich, T., & Farnsworth, A. (2014). The eBird enterprise: An integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31–40.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292.

Tang, B., Clark, J. S., & Gelfand, A. E. (2021). Modeling spatially biased citizen science effort through the eBird database. *Environmental and Ecological Statistics*, *28*(3), 609–630. https://doi.org/10.1007/s10651-021-00508-1

Tibshirani, J., Athey, S., & Wager, S. (2020). *Grf: Generalized random forests*. R package version 120.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2021). Modelling species presence-only data with random forests. *Ecography*, *44*(12), 1731–1742.

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*(6), 1450–1458. https://doi.org/10.1111/1365-2664.12158

van Swaay, C. A. M., Nowicki, P., Settele, J., & van Strien, A. J. (2008). Butterfly monitoring in Europe: Methods, applications and perspectives. *Biodiversity and Conservation*, *17*(14), 3455–3469. https://doi.org/10.1007/s10531-008-9491-4

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Walker, J., & Taylor, P. (2017). Using eBird data to model population change of migratory bird species. *Avian Conservation and Ecology*, *12*(1). https://doi.org/10.5751/ACE-00960-120104

Xue, Y., Davies, I., Fink, D., Wood, C., & Gomes, C. P. (2016). Behavior identification in two-stage games for incentivizing citizen science exploration. *Principles and Practice of Constraint Programming*, 701–717. https://doi.org/10.1007/978-3-319-44953-1_44

Zhang, W., Sheldon, B. C., Grenyer, R., & Gaston, K. J. (2021). Habitat change and biased sampling influence estimation of diversity trends. *Current Biology*, *31*(16), 3656–3662.e3. https://doi.org/10.1016/j.cub.2021.05.066

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). Zero-truncated and zero-inflated models for count data. In A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, & G. M. Smith (Eds.), *Mixed effects models and extensions in ecology with R* (pp. 261–293). Springer. https://doi.org/10.1007/978-0-387-87458-6_11

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Data S1. The R language source code and data for the eBird study and the toy example vignette (presented in the Supplemental Information).

How to cite this article: Fink, D., Johnston, A., Strimas-Mackey, M., Auer, T., Hochachka, W. M., Ligocki, S., Oldham Jaromczyk, L., Robinson, O., Wood, C., Kelling, S., & Rodewald, A. D. (2023). A Double machine learning trend model for citizen science data. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.14186