Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Adaptive Spatio-Temporal Exploratory Models: Hemisphere-Wide Species Distributions from Massively Crowdsourced eBird Data

Daniel Fink
Cornell Lab of Ornithology
Cornell University
Ithaca, NY 14850, U.S.A.
df36@cornell.edu

Theodoros Damoulas
Department of Computer Science
Cornell University
Ithaca, NY 14853, U.S.A.
damoulas@cs.cornell.edu

Jaimin Dave
Department of Computer Science
Cornell University
Ithaca, NY 14853, U.S.A.
jmd447@cornell.edu

Abstract

Broad-scale spatiotemporal processes in conservation and sustainability science, such as continent-wide animal movement, occur across a range of spatial and temporal scales. Understanding these processes at *multiple scales* is crucial for developing and coordinating conservation strategies across national boundaries. In this paper we propose a general class of models we call `AdaSTEM`, for **Ada**ptive **S**patio-**T**emporal **E**xploratory **M**odels, that are able to exploit variation in the density of observations while adapting to multiple scales in space and time. We show that this framework is able to efficiently discover multiscale structure when it is present, while retaining predictive performance when absent. We provide an empirical comparison and analysis, offer theoretical insights from the ensemble loss decomposition, and deploy `AdaSTEM` to estimate the spatiotemporal distribution of Barn Swallow (*Hirundo rustica*) across the Western Hemisphere using massively crowdsourced eBird data.

Introduction

Many environmental and ecological signals arise as the combined effects of simultaneous processes operating across a range of spatiotemporal scales. In his MacArthur Award Lecture, Levin (1992) stated that: "*... there is no single natural scale at which ecological phenomena should be studied... systems generally show characteristic variability on a range of spatial, temporal and organizational scales*". We attempt to automatically discover multiple scales by proposing flexible exploratory models based on adaptive spatiotemporal partitioning using tree data structures. Understanding patterns across different scales is crucial for sustainability, conservation management and decision making under uncertainty as it allows us to appropriately inform policy at different levels of granularity.

The motivation for this work is to estimate the daily distribution of long-distance migrant birds across the Western Hemisphere with the finest spatial resolution possible to date, Fig. 1. Understanding distributional patterns in fine detail across broad extents is a key concern for biodiversity studies. Ecologists and land managers need to know how local-scale patterns of habitat usage vary throughout a species' range to identify critical habitat requirements.

Consider how birds migrating across the Western Hemisphere are affected by processes operating at different scales. At very large spatial and temporal scales, climatic events like the El Niño and The North Atlantic Oscillation (Grosbois et al. 2008) determine when birds begin migrations and the routes they take. During the breeding season, the location of good foraging habitat and nest sites determine bird occurrence at the same location at much smaller scales (Fortin and Dale 2005). The notion of *scale* is formalized here as the *effective spatiotemporal range* ϕ_s, the distance from location[1] s at which the correlation becomes negligibly small (Banerjee, Carlin, and Gelfand 2004). Thus, what is observed at s may be simultaneously affected by processes operating at different ranges, e.g. $\phi_s^{\text{migration}} \gg \phi_s^{\text{foraging}}$.

In addition, the scale of the observation process must also be considered because it determines the smallest scale at which empirical quantities may be estimated. Broad-scale observational data are often irregularly and sparsely distributed. For example, by allowing participants to select the locations where data are collected, the distribution of crowdsourced data tend to follow patterns of human activity, Fig. 2a. In general, as the range of the signal ϕ_s decreases the minimum data density ρ_s necessary to estimate it increases. In the limit $\lim_{\phi_s \to 0} \rho_s = \infty$.

Modeling spatial correlation has been an active research area in statistics and machine learning for the past two decades (Cressie 1993; Rasmussen and Williams 2006). Methodologies such as kriging (Cressie 1986), Gaussian processes (Paciorek and Schervish 2004), Gaussian Markov random fields (Rue and Held 2005), splines (Pintore, Speckman, and Holmes 2006; Kammann and Wand 2003), and autoregressive models (Huang, Cressie, and Gabrosek 2002; Tzeng, Huang, and Cressie 2005) have been proposed to estimate and account for spatial correlation in stationary settings. More recently, research has focused on accounting for nonstationary spatial correlation. Non-stationary covariance functions have been proposed for GPs and kriging models (Stein 2005; Paciorek and Schervish 2004; Jun and Stein 2008; Pintore and Holmes 2004), and spline methods have been developed with spatially varying penalties (Pin-

[1] A location s may be spatial, temporal, or spatiotemporal.

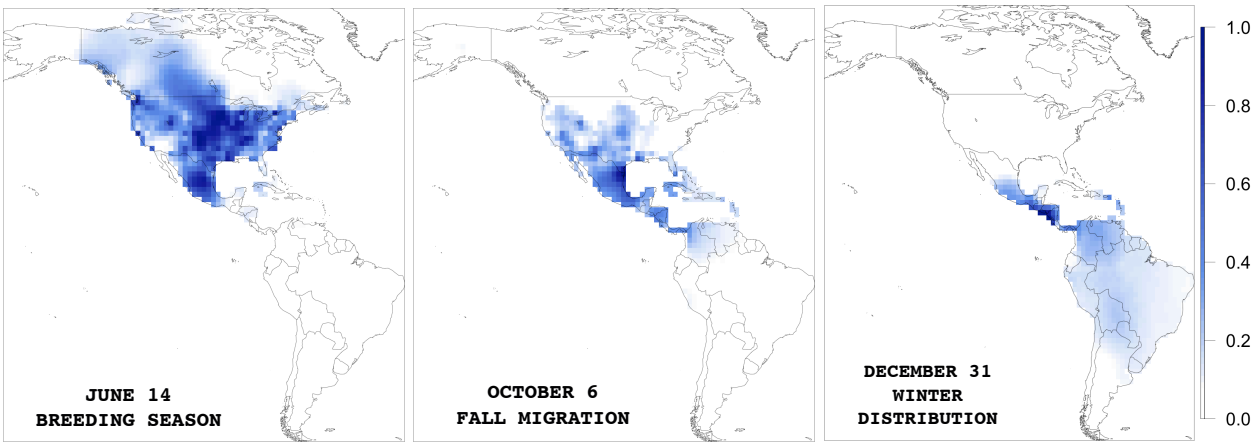


Figure 1: AdaSTEM distribution estimates (Color bar: relative probability of occurrence $\mathcal{P}_s$) for Barn Swallow during the breeding season (June 14, Left), fall migration (October 6, Center), and the winter distribution (December 31, Right). The transitions between the three dates highlights how AdaSTEM captured the southward movement of the population.

tore, Speckman, and Holmes 2006; Crainiceanu et al. 2007). However, the computational complexity of many of these models is high, e.g. GP's and kriging have a dominating term $\mathcal{O}(N^3)$, where N is the number of observations or locations. Thus, many of these methods can be computationally prohibitive for Big Data (Cressie and Johannesson 2008; Gelfand 2012) and crowdsourcing settings where the number of observations and locations is in the millions.

In this paper we present a general class of models designed to discover scale-dependent, non-stationary predictor-response relationships from large numbers of observations with irregular and sparse spatiotemporal density. It is a highly automated ensemble model with a 'pleasingly parallel' implementation. The large-scale experiments were conducted on the *Lonestar* cluster through an allocation on XSEDE (www.xsede.org). Our contributions are:

- An adaptive modeling framework designed to capture multiscale signal from crowdsourcing data.
- Theoretical and empirical analysis of AdaSTEM.
- The first hemisphere-wide population-level spatiotemporal estimate of a long-distance migration Fig. 1.

We start by describing the fixed-scale framework STEM and then the adaptive extension that yields AdaSTEM.

STEM: Spatio-Temporal Exploratory Models

STEM (Fink et al. 2010) is a mixture model designed to adapt to non-stationary, scale dependent processes. This is achieved by creating a dense mixture of local regression models with compact overlapping support. A user-specified regression model, the base model, accounts for variation as a function of predictor values within its support set, which we call a *stixel* for **s**patio**t**emporal p**ixel**. Because the stixels are compact sets the regression model can adapt to local predictor-response associations while limiting long-range extrapolation. Utilizing the fact that stixels overlap, predictions at a specified location, s, are made by taking an average across all base models whose stixels include that location, see Fig. 2b. This combines the bias-reducing properties of local models (e.g. decision trees, (Breiman et al. 1984)) with the variance-reducing properties of randomized ensembles (e.g. bagging, (Breiman 1996)).

The Mixture Model

Let $\{y_n(s), \mathbf{x}_n(s)\}_{n=1}^N$ be the set of observed responses and predictors $\mathbf{x}_i(s) = [x_{i,1}(s), \cdots, x_{i,d}(s)]$ indexed by space-time coordinates $s \in R^k$ within the study area $D \subset R^k$. Formally $y(s)$ is modeled as an ensemble response:

$$\hat{y}_e(s) = \sum_{m=1}^{M} \alpha_m(s) f_m\left(\mathbf{x}(s), D_m, s\right) \qquad (1)$$

with M base models f_m, estimated as $\hat{y}_m(s) = f_m\left(\mathbf{x}(s), D_m, s\right)$, each defined on its own stixel $D_m \subset D$ with mixture weights $\alpha_m(s)$.

The approach described here is based on ensemble modeling (Kuncheva and Whitaker 2003; Hastie, Tibshirani, and Friedman 2009) with a focus on prediction for large data sets. To this end we treat the estimation of the base models independently. The estimated ensemble response $\hat{y}_e(s)$ is computed as the weighted average taken across base models with shared support, Fig. 2b. For simplicity, all supporting base models are weighed equally. The mixture weights at coordinates s are $\alpha_m(s) = n^{-1}(s) I(s \in D_m)$

$$\text{where} \quad I(s \in D_m) = \begin{cases} 1 & \text{if } s \in D_m \\ 0 & \text{if } s \notin D_m \end{cases}. \qquad (2)$$

The *ensemble support* $n(s)$ is the number of base models that support coordinate s: $n(s) = \sum_{m=1}^{M} I(s \in D_m)$. Note that $a_m(s) \geq 0 \, \forall \, m, s \in D$ and $\sum_{m=1}^{M} a_m(s) = 1 \, \forall \, s \in D$.

STEM can be considered as a spatiotemporal wrapper for any base model. Each base model f_m is independently fit within its stixel D_m from N_m, the number of observations falling within D_m. In this paper we consider linear models

fit via least squares for the synthetic experiments, and logistic Generalized Additive Models (GAM) (Wood 2006) for the binary classification of eBird.

STEM uses a simple ensemble design with *fixed size* stixels. First, the study extent D is partitioned into a regular set of M_p square stixels D_m with sides of length λ. Secondly, P such partitions are sampled, by randomizing the position of each left corner π_p, to form an ensemble of overlapping stixels. We require that N_m meet a minimum sample size, γ, for the given class of base models. Stixels where $N_m < \gamma$ are omitted from the ensemble. Thus, every location s has $\max(n(s)) = P$. The algorithm is given in Alg. 1.

Algorithm 1 STEM

```
Set λ by cross-validation
for p = 1 to P do
    Randomize partition corner π_p ~ U(0, λ)
    Partition D into M_p stixels each with length λ
    for m = 1 to M_p do
        if N_m ≥ γ then
            Fit base model f_m in D_m, get estimator ŷ_m
ŷ_e(s) = Σ_{m=1}^{M} a_m(s) ŷ_m(s) (Eq. 1)
```

Ensemble Theory

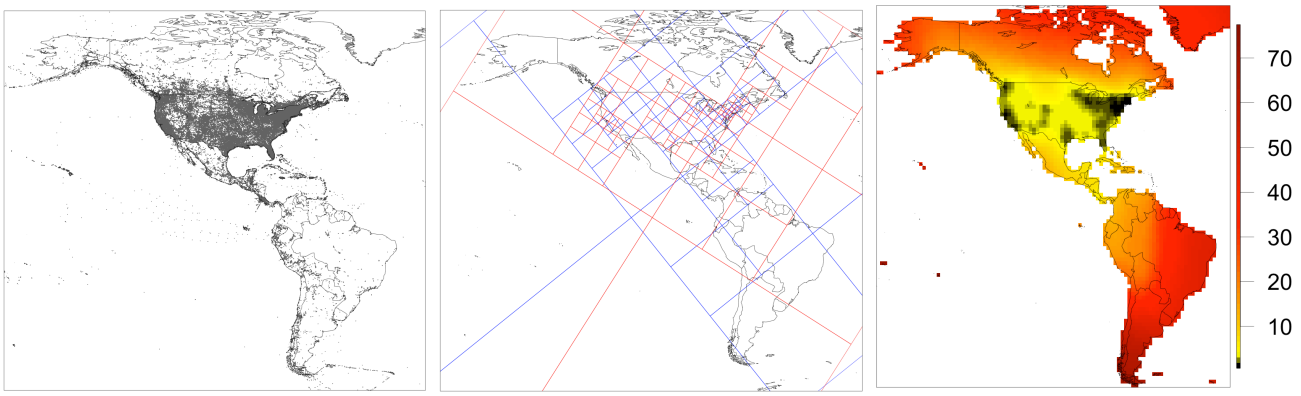
The size and configuration of the stixels $\{D_m\}_{m=1}^{M}$ are important parameters that delineate the neighborhoods where predictor-response relationships are constant and where they may vary. Because the stixels are compact sets, the size of D_m affects the range of spatiotemporal correlation ϕ_s both within and between base models. Theoretical results on ensemble models furnish general guidelines on how to construct D_m to improve predictive performance. We extend the decomposition of the squared error loss (Ueda and Nakano 1996) to the STEM ensemble. Given the ensemble response $\hat{y}_e(s)$ at s, we express Eq. 1 as $\hat{y}_e(s) = n(s)^{-1} \sum_{\{m|s\in D_m|\}} \hat{y}_m(s)$. Intuitively, $\{m|s \in D_m\}$ defines the modeling neighborhood of s as every model m whose stixel D_m contains s. Dropping dependence on s, we decompose the loss as:

$$
\begin{aligned}
&\mathbb{E}\Big\{(\hat{y}_e - y)^{\mathsf{T}}\,((\hat{y}_e - y)\Big\} = \\
&(\mathbb{E}\{\hat{y}_e\} - y)^{\mathsf{T}}(\mathbb{E}\{\hat{y}_e\} - y) + \mathbb{E}\Big\{(\hat{y}_e - \mathbb{E}\{\hat{y}_e\})^{\mathsf{T}}(\hat{y}_e - \mathbb{E}\{\hat{y}_e\})\Big\} \\
&= \underbrace{\bar{n}^{-2} \sum_{\{m|s\in D_m|\}} \sum_{\{m'|s\in D_{m'}\}} (\mathbb{E}\{\hat{y}_m\} - y)^{\mathsf{T}}\,(\mathbb{E}\{\hat{y}_{m'}\} - y)}_{\text{Bias}^{\mathsf{T}}\text{Bias}} \\
&+ \underbrace{\bar{n}^{-2} \sum_{\{m|s\in D_m|\}} \mathbb{E}\Big\{(\hat{y}_m - \mathbb{E}\{\hat{y}_m\})^{\mathsf{T}}\,(\hat{y}_m - \mathbb{E}\{\hat{y}_m\})\Big\}}_{\text{Variance}} \\
&+ \underbrace{\frac{1}{n(n-1)} \sum_{\{m|s\in D_m|\}\{m'\neq m|s\in D_{m'}\}} \sum \mathbb{E}\Big\{(\hat{y}_m - \mathbb{E}\{\hat{y}_m\})^{\mathsf{T}}(\hat{y}_{m'} - \mathbb{E}\{\hat{y}_{m'}\})\Big\}}_{\text{Covariance}}
\end{aligned}
\tag{3}
$$

This is the Bias-Variance-Covariance decomposition of the STEM framework.

Adaptive Multiscale Modeling with AdaSTEM

As we have seen, λ controls the size of the stixels and indirectly the minimum range of spatial correlation that can be modeled, $\hat{\phi}_s$. In STEM λ is a fixed, universal parameter that is estimated via cross-validation in order to indicate the scale of analysis best supported by data. Inspired by the spatiotemporal BVC decomposition, AdaSTEM proposes an adaptive scheme based on tree-data structures (Samet 2006), where $\lambda(s)_{\texttt{AdaSTEM}} = f(\rho(s))$ and f is an isotonically decreasing function of the data density ρ at locality s. In Fig. 2b two resulting partitions from a Quadtree are depicted where the leaf nodes define the corresponding stixels and their density-driven lengths $\lambda(s)$. Fig 2c shows the Quadtree stixel length averaged across all partitions and follows the pattern of data density.



(a) eBird Data Locations (b) Quadtree-generated stixels (c) Average stixel length $\overline{\lambda(s)}_{\texttt{AdaSTEM}}$

Figure 2: Left: (2a) Shows the varying density of observations. (2b) Shows two realizations of QuadTree partions, red and blue. (2c) Shows how the average QuadTree stixel length (in degrees) follows data density.

Letting the stixel size λ vary with data density ρ allows the mixture to better exploit unevenly distributed data in the presence of multiscale signal. Hence, when λ is small in densely sampled extents, the base models can adapt to fine-scale signals producing low bias estimators (term #1 in Eq. 3). The variance (term #2 in Eq. 3) is controlled by the ensemble averaging (Breiman 1996) and lower covariance between the base models in $\{m|s \in D_m\}$ is encouraged by bootstrapping the data and randomizing the angle $\theta_p \sim \mathcal{U}(0, 360]$ and center $c_p \sim \mathcal{U}(D)$ of each tree partition p. Given a base model complexity of $f(N_m)$ and Quadtree complexity $O(N \log_4 T)$, where T is number of nodes, then the complexity for `AdaSTEM` is $O(PN \log_4 T + PM_p f(N_m))$ where P and M_p are the numbers of partitions and stixels. The algorithm is given in Alg. 2.

Algorithm 2 `AdaSTEM`

for $p = 1$ **to** P **do**
  Sample Quadtree center c_p and angle θ_p
  Quadtree on D, get M_p stixels with lengths λ_{m_p}
  for $m = 1$ to M_p **do**
    if $N_m \geq \gamma$ **then**
      Fit base model f_m in D_m, get estimator $\hat{y}_m$
$\hat{y}_e(s) = \sum_{m=1}^{M} a_m(s)\hat{y}_m(s)$ (Eq. 1)

Empirical Analysis

An empirical analysis of `STEM` and `AdaSTEM` on synthetic data illustrates how the quantities $\phi_s, \rho_s, \hat{\phi}_s$ interact to affect predictive performance when the scale of the target function and the observation density varies. For convenience we construct two dimensional spatial regression examples. For this experiment the models are a spatial mixture of linear regression base models. The base model for the m-th stixel, D_m is:

$$z_i = f_m(x, y) = \beta_m + \beta_{x,m} x_i + \beta_{y,m} y_i + \epsilon_i \qquad (4)$$

with independent Gaussian errors $\epsilon_i \sim N(0, \sigma_m^2)$. The parameters of $f_m(x, y)$ are fit by least squares to the observations (z_i, x_i, y_i) falling within the m-th stixel; $(x_i, y_i) \in D_m$, and $i = 1, \ldots, N_m$. Stixels where $N_m < 10$ were omitted from the mixture and $P = 75$. The extent under study is $D = [0, 2]^2$. We perform 100 realizations and set $\lambda_{\texttt{STEM}} = 0.5$ and we let $\lambda_{\texttt{AdaSTEM}}$ vary with a maximum stixel sample size of $\mathcal{S} = 38$. We examine four cases:

- **Case A**. Single-scale ϕ_s with $\rho_s \sim \mathcal{U}$. [Fig. 4: $\mathcal{L}(a)$.]
- **Case B**. Single-scale ϕ_s with $\rho_s \not\sim \mathcal{U}$. [Fig. 4: $\mathcal{L}(b)$.]
- **Case C**. Multiscale ϕ_s with $\rho_s \sim \mathcal{U}$. [Fig. 4: $\mathcal{R}(a)$.]
- **Case D**. Multiscale ϕ_s with $\rho_s \not\sim \mathcal{U}$. [Fig. 4: $\mathcal{R}(b)$.]

Case A: The target function in the single-scale case, see Fig. 4:$\mathcal{L}$, is a surface of undulating, radially symmetric rings with relatively constant curvature across the region being studied. Let z be the height of the surface at a location (x, y) specified by the function:

$$z = f(x, y) = \frac{\sin\left(1.875\pi\sqrt{x^2 + y^2}\right)}{\cos\left(0.375\sqrt{x^2 + y^2}\right)} \qquad (5)$$

The set of observations $N = 300$ used to train the model z_i are i.i.d. and uniformly distributed $\{x_i, y_i\}_{i=1}^{N} \sim \mathcal{U}(0, 2)$. Observations are made with independent Gaussian errors, $z_i = f(x_i, y_i) + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2 = 0.25)$, achieving an SNR ≈ 1.70.

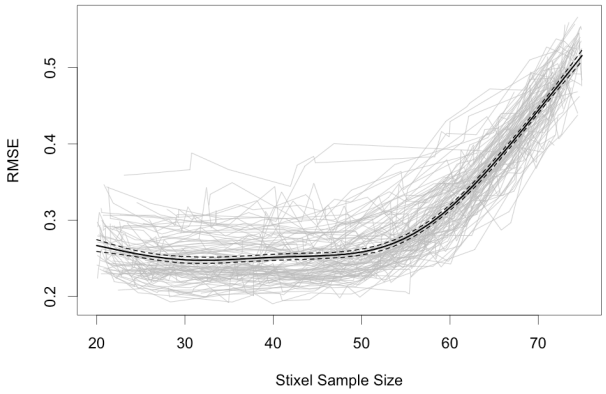


Figure 3: A Bias-Variance trade-off: RMSE vs. $\mathcal{S}$

We evaluate AdaSTEM RMSE as a function of $\mathcal{S}$ with values varying from 20 to 70 for each of the realizations. Then a GAM was fit (dark black line) as a smooth function of $\mathcal{S}$, with 95% confidence bounds (dashed) in Fig. 3. We see that $\hat{\mathcal{S}} = 38$ minimizes the GAM estimate of RMSE. Stixel lengths λ smaller or bigger than $\hat{\mathcal{S}}$ tend to under or over-smooth the surface, respectively. For this case there is a range of λ where $\hat{\phi}_s$ matches ϕ_s the true scale of the signal.

Comparing `AdaSTEM` and `STEM` we find that there is no significant difference in RMSE from a paired t-test (p-value = 1) when the stixels have the same expected size (Fig. 5).

Case B: This case has non-uniform observations with single-scale signal. We add a higher density set of observations in the center of the extent as $x_i, y_i \sim \mathcal{U}(0.625, 1.375)$ to the uniform data for a total of $N \approx 700$, see Fig. 4:$\mathcal{R}$.

The paired t-test shows a small (0.0038) but significant difference in RMSE between `AdaSTEM` and `STEM`, (p-value = 0.004). However, the size of the average difference is very small compared to the amount of variation across replicates, sd=0.0280. In Fig. 5 the range of RMSE values for case B are comparable to those in case A. This demonstrates the ability of the ensemble averaging to control the increased variation as `AdaSTEM` adapts to high density observations.

Case C: The multiscale target function, see Fig. 4:R, is created by adding fine-scale signal to Eq. 5. The fine-scale signal is similar to the broad-scale one in that it too is a smoothly varying, radially symmetric function, but restricted to the center of D. Let $r = \sqrt{((x-1)^2 + (y-1)^2)}$ then the high density signal is $f_h(r) = 1.5\exp(100r^3)\sin(12.5r/\pi)/\cos(2.5r)$. The fine scale signal has an SNR ≈ 1.65 over $[0.75, 1.25]^2$.

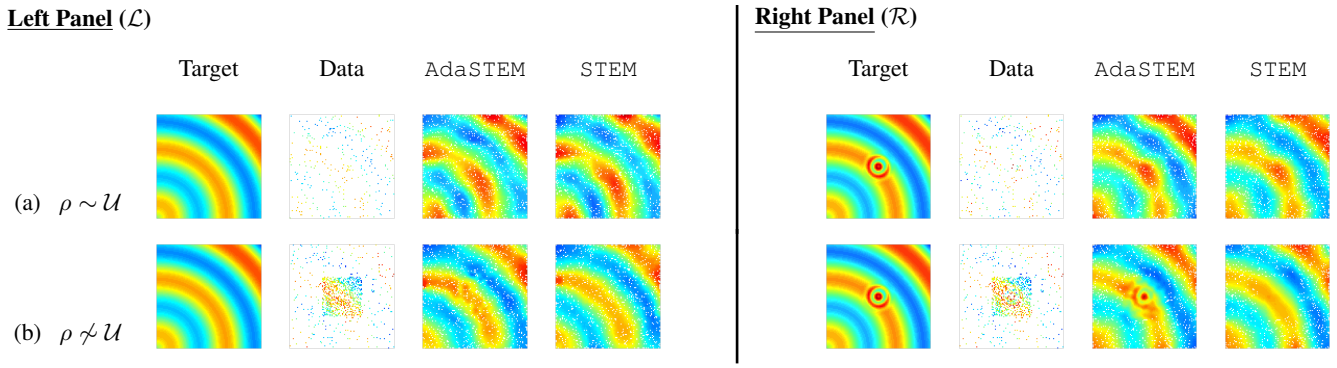


Figure 4: AdaSTEM vs STEM for uniform or non-uniform density of observations and single- or multi-scale signal. **Left Panel** ($\mathcal{L}$): In the presence of single-scale signal (ϕ_s constant throughout the extent) both models perform comparably for both uniform (row (a)) and non-uniform (row (b)) data density. **Right Panel** ($\mathcal{R}$): In the presence of multiscale signal AdaSTEM clearly outperforms STEM when the density of observations is sufficient (row (b)) to contain the small-scale correlation.

In this case, there is too little data to reveal the fine-scale signal of the target function. This results in higher overall RMSE for both models. The paired t-test found no statistical difference between the models (p-value = 0.996).

Case D: In contrast to above there is now sufficient density for AdaSTEM to detect fine-scale signal. This results in lower RMSE values in general and the RMSE for AdaSTEM is significantly less than that for STEM (p-value $<$ 2.2e-16). Note that many problems fall into this category including many species distribution analyses.

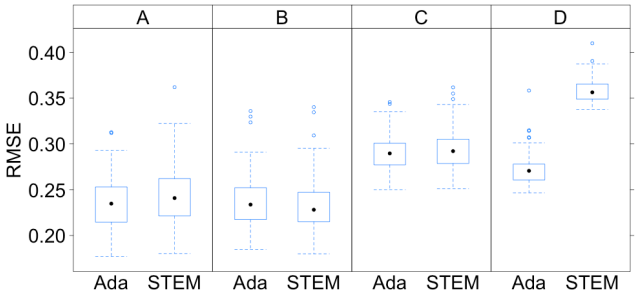


Figure 5: RMSE results on all cases. AdaSTEM outperforms STEM when multiscale structure exists (Case D).

eBird

The ecological goal of this analysis is to estimate the daily distribution of Barn Swallow across the Western Hemisphere, excluding Greenland, throughout its annual cycle. We compare the predictive performance of STEM and AdaSTEM estimating these distributions with eBird data http://www.ebird.org/.

The bird observation data comes from the citizen science project, eBird (Sullivan et al. 2009). eBird is a broad-scale bird monitoring project that collects observations made throughout the year. Participants follow a checklist protocol, in which time, location, search effort, and counts of birds are all reported in a standardized manner. We chose the Barn Swallow for two reasons: 1) it is an easily identified, conspicuous bird commonly seen by birders, and, 2) it has a well known long-distance migration that stretches over most of the Western Hemisphere.

Presence-absence data from complete checklists collected with effort data from January 1, 1900 to December 31, 2011 within the Western Hemisphere were used in this analysis. The data set, Fig. 2a, consists of approximately 2.5M checklists made across 385K unique locations. All models were trained with 2.25M checklists made across 360K unique locations, with the remaining for validation. The observations are most densely distributed in the U.S. where eBird originated, and are sparser in Central and South America. At smaller spatial scales, the data density can be seen to correlate with human population patterns.

The spatiotemporal distribution of Barn Swallow is modeled as a spatial mixture of local temporal trajectories. We estimate the trajectory within each stixel using a binary response GAM. The binary response Y_i indicates the presence or absence of the bird recorded for the i-th search. The logit of the probability of occurrence $\mathcal{P}_i$ varies as an additive function of the day of the year and several other factors describing the effort spent searching for birds. Formally, the i-th search results are Bernoulli distributed $Y_i \sim \text{Bern}(p_i)$ where $\text{logit}(\mathcal{P}_i) = \beta_0 + f(\text{day}_i) + \sum_{j=1}^{4} \beta_j E_{i,j}$. Seasonal variation is captured by the smooth function f of the day of the year covariate day_i and fit with penalized splines. To account for variation in detection rates we include effort covariates for the amount of time spent on a search $E_{1,i}$, the distance traveled while searching $E_{2,i}$, and the number of observers in the search party $E_{3,i}$. The time of the day $E_{4,i}$ is used to account for diurnal variation in behavior, such as participation in the dawn chorus (Diefenbach et al. 2007), that make species more or less conspicuous. The minimum sample size per base model is $\gamma = 500$ and $P = 75$.

Because the objective of this analysis is to study the full species' distribution, we measured the *coverage* of each model as the proportion of locations within the Western

Hemisphere landmass, excluding Greenland, that have at least half the possible ensemble support. Coverage was computed using approximately 3000 locations sampled from a geographically Stratified Random Design (SRD).

Predictive Performance Comparison

The predictive performance of STEM and AdaSTEM were compared for five model coverage levels ranging between 70 and 99%. To assess the quality of the distribution more uniformly across the study area, the test data were subsampled across a grid of one degree cells, with a maximum of 10 observations per cell (Fink et al. 2010). Because of the strong seasonal variation between distributions computed predictive performance metrics were computed independently for each month. AUC and RMSE were used to measure the ability of the model to estimate the expected occurrence rates. Accuracy and the Kappa statistic were calculated to measure the ability of the model to estimate the binary outcome.

AdaSTEM outperforms STEM in terms of AUC, RMSE, Accuracy, and Kappa for all coverage levels tested, Fig. 6. Qualitatively, these results are the same across all 12 months although the greatest differences in performance between STEM and AdaSTEM are achieved during the breeding season when the Barn Swallow population is in North America and observation density is the greatest. Because stixel size decreases with coverage, bias will tend to decrease and variance will tend to increase. Performance in Fig. 6 improves monotonically as coverage decreases, indicating that reducing bias is more significant at these coverage levels, and also that we control for variance through the ensemble averaging.

Fall Migration Estimates

To develop range-wide estimates of the Barn Swallow distribution we selected the smallest stixel size necessary to achieve 90% coverage for AdaSTEM. Then we used this model to estimate one daily distribution surface per week for 52 weeks of the year. The surface is the probability of occurrence on the given day estimated at the SRD locations. All effort predictors were held constant to remove variation in detectability. To control for seasonal variation in detectability when comparing distributions on different days of the year we standardize the predicted probability of occurrence for each day, Fig. 1. The precise quantity estimated is $\mathcal{P}_s$ the relative probability that a typical eBird participant will detect the species on a search from 7-8AM while traveling 1 km on the given day at s.

The northern limit of Barn Swallow's breeding distribution extends from Alaska southeast into the Maritime Provinces of Canada (Brown and Brown 1999). The AdaSTEM estimate provides a good large-scale estimate of the known northern distribution, Fig 1, Left. Looking south, the estimate closely matches the documented range of the species into the conterminous United States and Northern Mexico, where data density is high. Notably, Barn Swallows are known to be absence from the Sierra Nevada, the southwestern portion of California and Arizona, and Florida (AOU 1998). The estimate of the winter distribution also matches known patterns of occurrence. The winter population extends from central Mexico, throughout Central America, in the West Indies mostly in the eastern Lesser Antilles (Bond 1980), and into South America with lower probabilities in eastern Brazil and the southern portion of South America (Ridgely and Guy 1989).

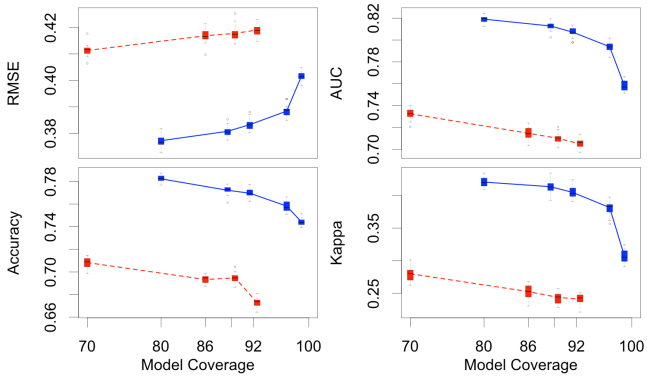


Figure 6: Predictive performance as a function of model coverage for the month of July. AdaSTEM values are shown as blue, solid line and STEM are shown as the red dashed line.

Discussion

We have introduced a spatiotemporal modeling framework based on tree data structures, called AdaSTEM, and demonstrated that is capable of adapting to multiscale signal when there is sufficient data density. Our empirical results suggest that this scheme of density-based adaptation is beneficial for model performance and can be explained with arguments based on the BVC decomposition. Thus, AdaSTEM will also be applicable in other domains where data are irregularly and sparsely distributed, such as applications based on geographic survey and geolocated crowdsourced data. We are currently considering additional adaptation criteria such as base model error residuals.

Until recently, most monitoring and biodiversity collection programs have been national in scope, hindering the ecological study and conservation planning of species with broad distributions. Here we use AdaSTEM to produce the first hemisphere-wide population-level distribution estimates of a long-distance migration using 2.25M eBird checklists. The ability to produce comprehensive year-round distribution estimates that span national borders will make it possible to better understand the ecology of these species, assess their vulnerability under climate change, and coordinate conservation activities.

Acknowledgments

We thank the thousands of eBird participants and K. Webb, W.M. Hochachka, S. Kelling, M.Iliff, C.Wood, B.Sullivan, and the IS eBird team. This work was supported by the Leon Levy Foundation and the National Science Foundation (CCF-0832782, IIS-1017793, CDI-1125098) with computing support from CNS-1059284, OCI-1053575 and DEB-110008.

References

AOU. 1998. Check-list of North American Birds.

Banerjee, S.; Carlin, B.; and Gelfand, A. 2004. *Hierarchical modeling and analysis for spatial data*, volume 101. Chapman & Hall/CRC.

Bond, J. 1980. *Birds of the Western Indies*. Houghton Mifflin Co, Boston.

Breiman, L.; Friedman, J.; Stone, C.; and Olshen, R. 1984. *Classification and regression trees*. Chapman & Hall/CRC.

Breiman, L. 1996. Bagging predictors. *Machine learning* 24(2):123–140.

Brown, C. R., and Brown, M. B. 1999. Barn swallow (hirundo rustica). In Poole, A., ed., *The Birds of North America Online*. Ithaca, NY: Cornell Laboratory of Ornithology. At http://bna.birds.cornell.edu/bna/species/452.

Crainiceanu, C.; Ruppert, D.; Carroll, R.; Joshi, A.; and Goodner, B. 2007. Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics* 16(2):265–288.

Cressie, N., and Johannesson, G. 2008. Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc. B* 70:209–226.

Cressie, N. 1986. Kriging nonstationary data. *Journal of the American Statistical Association* 81(395):625–634.

Cressie, N. 1993. *Statistics for Spatial Data*. New York: John Wiley, 2nd edition.

Diefenbach, D.; Marshall, M.; Mattice, J.; Brauning, D.; and Johnson, D. 2007. Incorporating availability for detection in estimates of bird abundance. *The Auk* 124(1):96–106.

Fink, D.; Hochachka, W.; Zuckerberg, B.; Winkler, D.; Shaby, B.; Munson, M.; Hooker, G.; Riedewald, M.; Sheldon, D.; and Kelling, S. 2010. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications* 20(8):2131–2147.

Fortin, M., and Dale, M. 2005. *Spatial analysis: a guide for ecologists*. Cambridge University Press.

Gelfand, A. 2012. Hierarchical modeling for spatial data problems. *Spatial Statistics* 1:30–39.

Grosbois, V.; Gimenez, O.; Gaillard, J.; Pradel, R.; Barbraud, C.; Clobert, J.; Møller, A.; and Weimerskirch, H. 2008. Assessing the impact of climate variation on survival in vertebrate populations. *Biological Reviews* 83(3):357–399.

Hastie, T. J.; Tibshirani, R.; and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Huang, H.-C.; Cressie, N.; and Gabrosek, J. 2002. Resolution-consistent spatial prediction of global processes from satellite data. *Journal of Computational and Graphical Statistics* 11(1):63–88.

Jun, M., and Stein, M. L. 2008. Nonstationary covariance models for global data. *Annals of Applied Statistics* 2:1271–1289.

Kammann, E., and Wand, M. 2003. Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52(1):1–18.

Kuncheva, L., and Whitaker, C. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning* 51(2):181–207.

Levin, S. A. 1992. The problem of pattern and scale in Ecology. *Ecology* 73(6):1943–1967. The Robert H. MacArthur Award Lecture.

Paciorek, C., and Schervish, M. 2004. Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems* 16:273–280.

Pintore, A., and Holmes, C. C. 2004. Spatially adaptive non-stationary covariance functions via spatially adaptive spectra. Technical report, Department of Statistics, Oxford University. At: http://www.stats.ox.ac.uk/~cholmes/Reports/spectral_tempering.pdf.

Pintore, A.; Speckman, P.; and Holmes, C. C. 2006. Spatially adaptive smoothing splines. *Biometrika* 93(1):113–125.

Rasmussen, C., and Williams, C. 2006. *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Ridgely, R., and Guy, T. 1989. *The Birds of South America*. University of Texas Press.

Rue, H., and Held, L. 2005. *Gaussian Markov Random Fields: Theory and Applications*, volume 104 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Samet, H. 2006. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann.

Stein, M. L. 2005. Space-time covariance functions. *Journal of the American Statistical Association* 100:310–321.

Tzeng, S.; Huang, H.-C.; and Cressie, N. 2005. A fast, optimal spatial-prediction method for massive datasets. *Journal of the American Statistical Association* 100(472):1343–1357.

Ueda, N., and Nakano, R. 1996. Generalization error of ensemble estimators. In *IEEE International Conference on Neural Networks*.

Wood, S. 2006. *Generalized additive models: an introduction with R*, volume 66. Chapman & Hall/CRC.