Machine Learning for Computational Sustainability

Tom Dietterich, Ethan Dereszynski, Rebecca Hutchinson, Dan Sheldon
School of Electrical Engineering and Computer Science
Oregon State University
Corvallis, OR, 97331 USA
tgd@eecs.oregonstate.edu

***Abstract*—To avoid ecological collapse, we must manage Earth's ecosystems sustainably. Viewed as a control problem, the two central challenges of ecosystem management are to acquire a model of the system that is sufficient to guide good decision making and then optimize the control policy against that model. This paper describes three efforts aimed at addressing the first of these challenges—machine learning methods for modeling ecosystems. The first effort focuses on automated quality control of environmental sensor data. Next, we consider the problem of learning species distribution models from citizen science observational data. Finally, we describe a novel approach to modeling the migration of birds. A major challenge for all of these methods is to scale up to large, spatially-distributed systems.**

Keywords-computational sustainability; species distribution models; dynamical ecosystem models; hidden Markov models

I. Introduction

The world-wide spread and rapid growth of human populations and the associated modification of the earth’s ecosystems has resulted in large changes in the functioning of these ecosystems. There has been a huge conversion of land to agricultural production and, consequently, large decreases in the range and size of the populations of many species. Paradoxically, some invasive pest species have greatly increased their range and population sizes so that they are interfering with ecosystem services upon which humans and other species rely.

From a control systems perspective, we do not know what these large-scale changes imply for the future trajectory of the Earth system. Are we headed toward a world-wide ecosystem collapse accompanied by the extinction of most or all humans? Or will the Earth shift to a new quasi-stable operating point that can sustainably support 9-10 billion people? What controls should we apply to the system to achieve desirable outcomes?

The central problem is that we lack an accurate model of the dynamics of the earth’s ecosystems. Doak et al., [4] document a long list of “ecological surprises”—cases where either an ecosystem behaved in a way that is still not understood or where an attempted intervention had major unforeseen consequences.

The same lack of models at global scales is also seen at smaller scales. For example, the habitat requirements and population dynamics of most wild bird species are not well-understood. While the populations of many species are declining, some are increasing.

Fortunately, there are two trends that are helping address the lack of models. First, we are in the midst of multiple revolutions in sensing. One revolution is driven by sensing technology: the number, diversity, and capability of sensors is rapidly increasing. Another revolution is driven by the development of citizen science and crowd sourcing where people (and often their smart phones and laptops) collect observational data at a scale that dwarfs what professional scientists could ever collect. The second trend is the rapid development of machine learning algorithms that can fit complex models to the massive amounts of data that are becoming available.

At the moment, the machine learning techniques lag behind the data collection. The new data sources raise challenges—both old and new—for machine learning, and this paper describes research on three such challenges:

- **Automated data quality control (QC).** In the past, human data technicians have manually inspected sensor data streams to identify data quality problems (e.g., sensor and communications failures, configuration errors, etc.). However, the number of deployed sensors is rapidly outstripping the ability of people to QC the data. Methods are needed to automate the quality control process.
- **Fitting models to citizen science data.** Citizen observers vary tremendously in their expertise. For example, bird watchers may fail to detect a bird species even though it is present at a site. In addition, citizen scientists choose when and where to make their observations—they do not follow a carefully-designed statistical sampling plan. Consequently, their data may exhibit a wide range of sampling biases and errors. Machine learning methods are needed that can compensate for all of these data quality issues.
- **Fitting models of population dynamics to count data.** To create models of population dynamics, the ideal form of data collection would be to track each individual in the population so that all interactions (e.g., predation, mating, reproduction, mortality) could be directly observed. However, in virtually all ecosystems, we lack this kind of data. Instead, we often have only (noisy) counts of the number of individuals observed at selected times and places. Can we develop machine learning methods that can fit dynamical models to such data?

This work was supported by NSF grants 0832804 and 1125228.

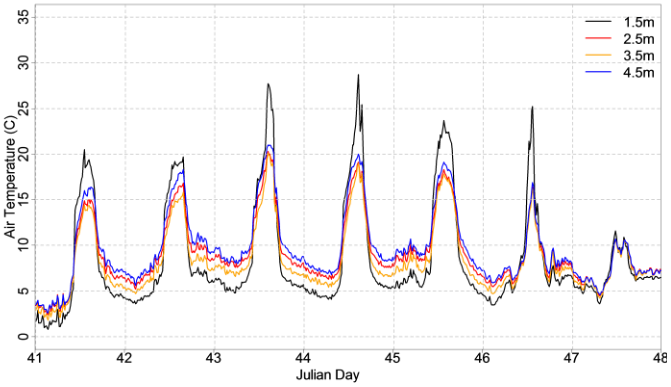

Figure 1. Temperature readings from four thermometers. The 1.5m thermometer reads incorrectly high because of a broken sun shield.

The remainder of this paper describes research in progress that addresses each of these challenges. We hope to give the reader a sense of the computer science questions underlying this work and provide pointers to the technical details which have been published elsewhere.

II. Automated Data Cleaning

Figure 1 shows signals from four air temperature thermometers deployed on a single instrument tower at the H. J. Andrews Experimental Forest in Oregon. Data are reported every 15 minutes, so the 24-hour daily cycle is clearly visible. However, the 1.5m sensor has a broken sun shield, so it heats up too quickly and reports incorrectly-elevated readings during the middle of the day. Figure 2 shows another situation in which two thermometers have become buried in snow, so that the reported values are no longer air temperature values. We seek an automated method that can detect these kinds of anomalies and also impute accurate values for the damaged readings. Furthermore, we want a general purpose method that can detect novel sensor failures, rather than a method that is only able to diagnose a fixed set of known failure modes.

To solve this problem, we have pursued a probabilistic modeling approach. At each time step t, let X_i^t denote the true temperature being measured by sensor i and O_i^t be the observed value reported by the sensor. We introduce a discrete "sensor status" variable S_i^t whose value is ok if the sensor is functioning correctly and $broken$ otherwise. We then define the probabilistic model

$$O_i^t \sim \begin{cases} \text{Normal}(X_i^t, \sigma_{ok}^2) & \text{if } S_i^t = ok \\ \text{Normal}(0, \sigma_{broken}^2) & \text{if } S_i^t = broken. \end{cases}$$

Here, σ_{ok}^2 is a small value (e.g., 0.1), while σ_{broken}^2 is a large value (e.g., 1000.0). According to this model, if the sensor is ok, then the observed value O_i^t is equal to the true value X_i^t with additive Gaussian noise specified by σ_{ok}^2. However, if the sensor is broken, then the value being reported is no longer related to the true temperature. The large variance σ_{broken}^2 allows the model to explain a broad range of anomalous behavior.

To extend this model to handle the time series of observations, we introduce a first-order Markov process whose conditional distribution, $P(X_i^t | X_i^{t-1})$, is assumed to be a linear Gaussian: $\text{Normal}(\beta_0 + \beta_1 X_i^{t-1}, \sigma_i^2)$. This turns the model into a Kalman filter with a switched observation distribution (similar to a switching linear dynamical system). In previous work [2], we scaled up this approach to handle relationships among multiple sensors. This was done by fitting a multivariate conditional linear Gaussian model to represent the joint conditional distribution $P(X^t | X^{t-1})$.

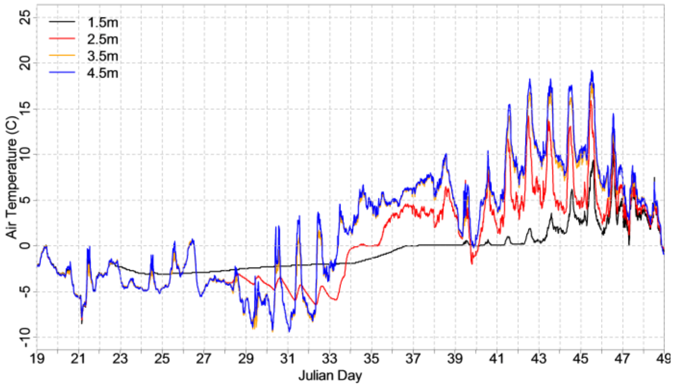

Figure 2. Temperature readings showing 1.5m thermometer buried in snow from day 23 to day 39.

Although this approach is very elegant, it raises difficult computational problems. First, the single-sensor switched Kalman filter is computationally intractable because at each time step, we must consider the two values of S_i^t, ok and $broken$. In a standard Kalman filter, the only hidden variable at each time step is X_i^{t-1}, and it can be marginalized out of the joint distribution $P(X_i^t, X_i^{t-1})$ to produce a single Gaussian distribution. However, when there is a discrete hidden variable such as S_i^{t-1} and it is marginalized out, the result is a mixture of two Gaussians (corresponding to the two values $S_i^{t-1} = ok$ and $S_i^{t-1} = broken$). Hence, to exactly represent the probability distribution $P(X_i^t | O_i^{1:t-1})$ requires a mixture of Gaussians with 2^{t-1} components.

We resolve this problem via the following "argmax" approximation. At each time step, we compute the most likely state of the sensor, $\hat{s}_i^t = \arg\max_s P(S_i^t | S_i^{1:t-1}, O_i^{1:t})$. We then assert that $S_i^t = \hat{s}_i^t$, which reduces the resulting distribution over X_i^t to a single Gaussian, which can then be processed using the standard Kalman filter update.

This solves the problem for a single sensor, but a similar problem arises with multiple sensors within a single time step, because the sensors are coupled through the joint conditional distribution $P(X^t | X^{t-1})$. With N sensors, there are 2^N possible configurations of the S_t variables. To apply the argmax approximation, we need to compute the most likely such configuration, which requires time exponential in N. We have evaluated several state of the art approximate methods for this computation including Rao-Blackwellized particle filters and expectation propagation, but the most effective method is an algorithm we call SearchMAP. This method starts by assuming that all sensors are working at time t and scores the likelihood of this: $P(O_{1:N}^t | S_{1:N}^t = \vec{s}, O^{1:t-1}, S_{1:N}^{t-1})$. It then considers all one-step "moves" that flip the status of one sensor, scores the likelihood of each of these, and keeps the configuration of maxi-

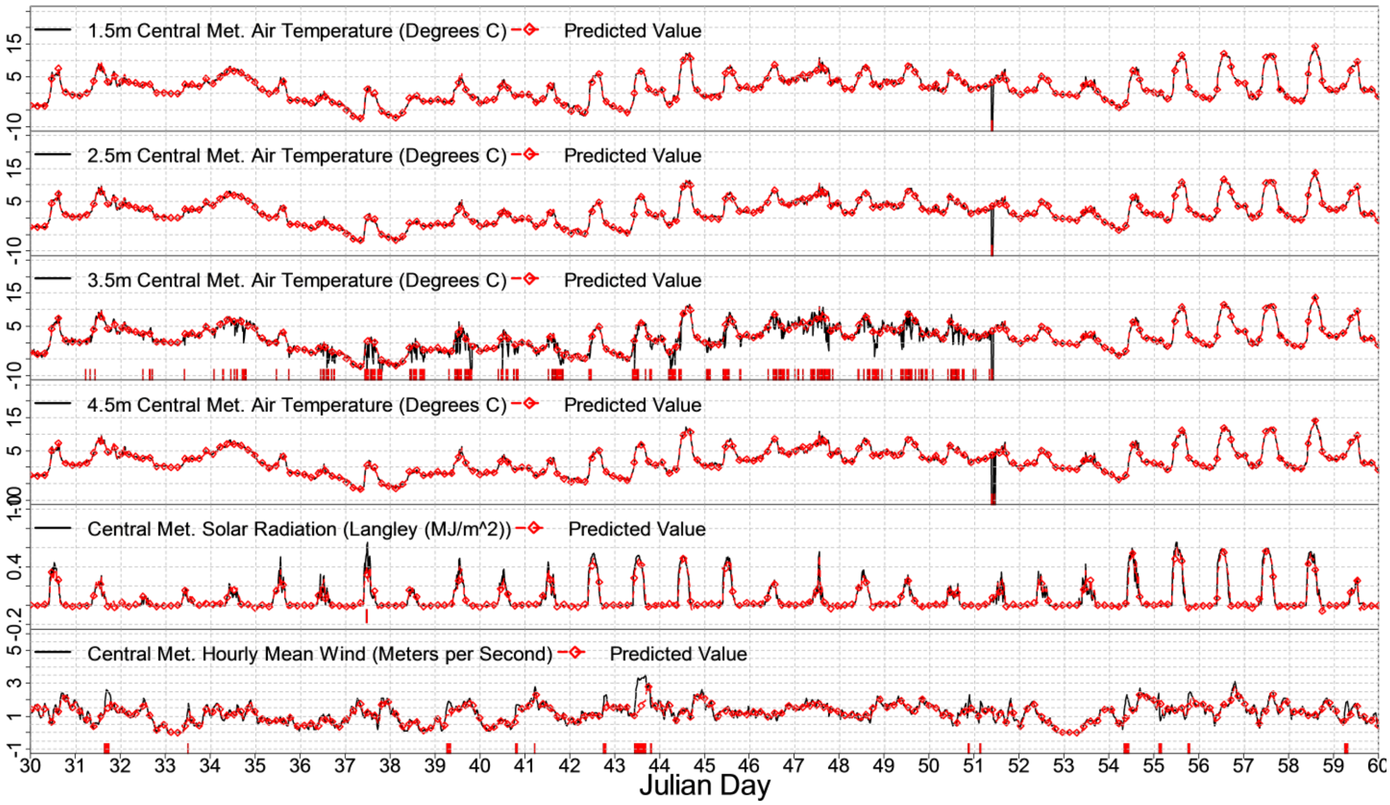

Figure 3. Signals from six sensors on a single tower. Solid black is the raw signal; red ticks at the base of each signal trace indicate points declared by the model

mum likelihood. This hill-climbing search is repeated until a local optimum is found.

Figure 3 shows some of the results of applying this method to 30 days of raw data from a network of 18 sensors (including air temperature, wind speed, and solar radiation). The model was trained on clean-only data provided to us by a domain expert, and then used to QC a new dataset containing raw, unchecked observations. In model training, we learn a set of conditional dependencies among the sensors that explains how their observations are spatially and temporally correlated. Once the model is trained, we apply the SearchMAP algorithm to perform inference on new observations. Our QC approach successfully identified a data-logger malfunction on day 51 that affected all air temperature sensors. It also flagged erratic measurements from the 3.5m air temperature sensor between days 36 and 50, which were caused by faulty voltage readings at the sensor. Though the model only produced a few false positives in the wind speed and solar radiation data, it failed to detect when the anemometer (wind speed sensor) froze on days 33 and 53.

Many challenges remain for automated data cleaning. The method described above only operates at a single temporal scale, so it misses anomalies (e.g., long-term trends) that are only detectable at broader scales. Multi-scale methods are needed that can find such anomalies and that can deal with data collected at widely-varying temporal and spatial scales. Moreover, our model cannot capture different weather regimes, such as storm systems, cold air drainage, and temperature inversions, which may alter the correlations among the sensors. Finally, the model could be improved by employing non-Gaussian distributions for quantities such as wind speed and precipitation, which are poorly modeled by Gaussians.

III. Fitting Models to Citizen Science Data

The second computational problem that we will consider is the challenge of fitting ecological models to citizen science data. Many important ecological phenomena occur at the scale of continents or the whole planet. Hence, we need to collect data at such scales in order to study these phenomena. Aside from satellite-based remote sensing, most data is collected by small teams of scientists working in small study areas. Citizen scientists can address this problem by fanning out across the planet to collect data.

We are collaborating with the Cornell Lab of Ornithology's eBird project (www.ebird.org), in which bird watchers report checklists of bird species that they observe at a particular time and place. Each month, eBird receives many thousands of uploaded checklists, and these more than one million species observations.

Given this data, there are many different questions that we can ask. In this section, we describe some of the work we have done fitting species distribution models (SDMs) to eBird data. A species distribution model for a species is a function that, given a description of a site (e.g., vegetation type, land use, distance to water, human population density, distance to roads, elevation, annual temperature and precipitation) predicts

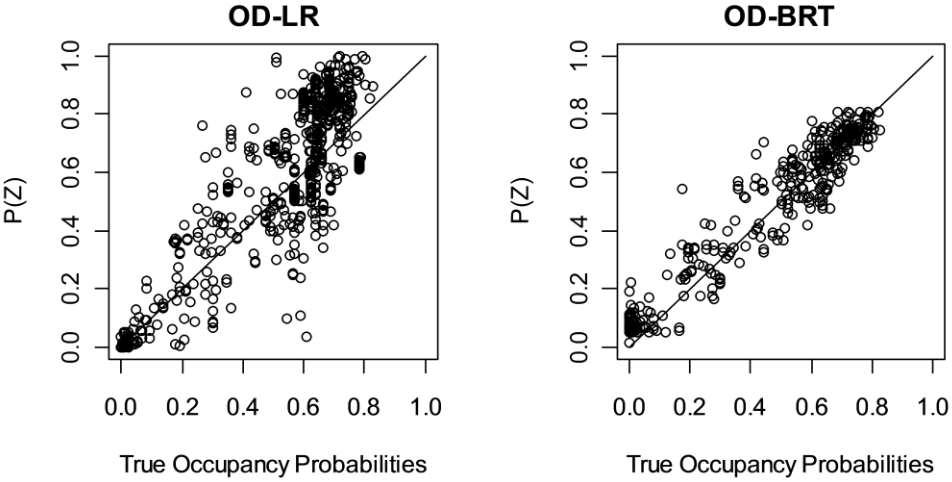

Figure 4. Left: OD model fit with log-linear component models; Right: OD model fit with boosted regression trees for the component models. Each circle is a

whether the species will be found at that site. Such models are useful as a first step toward understanding the habitat requirements and population dynamics of the species, and they can guide the design of conservation reserves and other policy decisions for protecting threatened and endangered species.

Citizen science data presents three major challenges. First, there is the potential for sampling bias, since bird watchers tend to go birding close to home and they tend to select locations where they believe the birds will be found. Second, many bird species are hard to detect, so the fact that the birder did not report the species does not mean that the species was not present. Third, volunteers may have different levels of expertise. Some birders have many years of experience and can identify many species from their songs and calls without any visual information. Other birders can only reliably identify species by sight.

We are pursuing research to address all three of these problems. To deal with sample selection bias, we are extending recent work in density ratio estimation [10]. If our goal is to build a map of bird species distribution across the continent, then we seek a model that is accurate for sites chosen uniformly at random in space. Call this distribution P_{target}, since it is the target distribution. However, our data is collected by birdwatchers visiting sites according to a different tion P_{source}. One approach to sampling bias correction is to reweight each source observation by the ratio $r(x) = P_{target}(x)/P_{source}(x)$. Estimating $P_{target}$and P_{source} separately is difficult, because density estimation in high-dimensional spaces is a challenging statistical problem. But estimating the ratio of two densities is much easier, and many methods are now available [10].

The problem of partial detection has been studied previously by wildlife biologists. The problem can be addressed by making multiple independent visits to a site and combining the information from those visits. Specifically, let Z_i be 1 if site i is occupied by the species and 0 otherwise, and let X_i be a vector of site attributes that describe the site. Then the species distribution model can be written as $F(X_i) = P(Z_i|X_i;\theta)$, where θ are the parameters of the model. Unfortunately, we do not directly observe Z_i, instead in each visit t to site i, we obtain an observation Y_{it} which is 1 if the species was detected and 0 otherwise. Let $P(Y_{it}|Z_i, W_{it};\psi)$ represent the detection model, where W_{it} is a vector of attributes describing the observation conditions at time t (e.g., time of day, weather, duration of the observation, density of the vegetation, etc.). If we assume no false detections, then we can write $P(Y_{it} = 1|Z_i = z, W_{it} = w;\psi) = zG(W_{it})$, where $G(w)$ is the probability of detecting the species given that it is present $(z = 1)$. If $z = 0$, then $Y_{it} = 0$ with probability 1. We call this the Occupancy-Detection (OD) model.

If we make multiple independent visits to the site, then if the site is occupied, visit t has probability $G(W_{it})$ of detecting the species. Given the assumption of no false detections, if we detect the species on any one of the visits, then we know $Z_i = 1$. If we never detect the species, then our belief about the value of Z_i depends on the probability of detection as determined by $G(W_{it})$. If the species is hard to detect, then even if we fail to detect it in several visits, this does not provide definitive evidence that the site is unoccupied.

This model was first introduced by MacKenzie et al. [7] in a formulation where both F and G are assumed to be either constant or to have the form of a log-linear model (like logistic regression). The model can be fit to data of the form $\{(X_i, W_{it}, Y_{it})\}$ via maximum likelihood estimation.

The log-linear model assumption creates many practical difficulties for applying the OD model at continental scales. Log linear models assume that each attribute of the site (or the visit) makes an independent contribution to the log probability of occupancy (or detection, respectively) and that these contributions scale linearly. If there are non-linearities, the data must be transformed to remove them. If there are interactions and dependencies among variables, these must be manually included in the model. While this is possible for well-understood situations, we need a more automated approach that works on large data sets at continental scales.

One of the most user-friendly machine learning models is based on classification and regression trees [1]. These models make no linearity assumptions, they automatically discover interactions among attributes, and they handle other imperfections (e.g., missing values) robustly. Although a single decision tree often has mediocre predictive accuracy, ensembles of trees achieve state-of-the-art performance in many domains. A drawback of tree methods is that they are non-probabilistic models. However, in 2000, Friedman showed how to incorporate trees into logistic regression classifiers via a form of boosting [5].

Building on Friedman's work and our own previous work on incorporating trees into conditional random fields [3], we have developed an algorithm (called OD-BRT) for integrating trees into the OD model. This is the first time that trees have been combined into a model with latent (hidden) variables. Figure 4 shows the results of a simulation experiment in which we applied OD-BRT to synthetic data where the true values of $F(X_i)$ and $G(W_{it})$ are known. The scatterplot shows that the OD-BRT model provides much more accurate predictions for the occupancy probabilities than the standard log-linear OD model (denoted OD-LR) [6].

We have also extended the OD model to include the expertise of each citizen scientist, a model that we call ODE (Occupancy, Detection, and Expertise). In the ODE model, there is an additional model component $P(E_u|B_u)$ that predicts the expertise level of birder u from background information B_u about the birder. Then when birder u visits site i at time t, the probability that $Y_{uit} = 1$ depends on Z_i, W_{it}, and B_u. Using a log-linear approach, we found that the ODE model was better at modeling eBird data than the OD model [11].

IV. Fitting Models of Population Dynamics to Count Data

The third problem we will discuss is the challenge of fitting a continent-scale model of bird migration to eBird observational data. Bird migration is poorly understood because it takes place at very large scale and because most migration occurs at night.

Our approach to modeling bird migration is to define a grid over North America and learn a Hidden Markov model (HMM) over the cells in this grid. Suppose we have a population of N birds. Let $Z_i^t = c$ be the grid cell c in which bird i is located at time t. Then the state of the system can be represented by the vector Z^t that specifies the location of each bird, and a Markov model of the migration dynamics can be defined by $P(Z^t|Z^{t-1}, X^{t-1})$, where X^{t-1} is a matrix of attributes describing such things as wind speed and direction and how the wind aligns with the headings between pairs of cells (c, c'), the distance between pairs of cells, the habitat in each cell, and so on.

Learning this model would be straightforward if we could put a GPS tracker on each bird. But instead, all we have are field observations $Y^t(c)$ of the number of birds observed in cell c at time t. Hence, we obtain an HMM with observation distribution $P(Y^t|Z^t, W^t)$, where $W^t(c)$ provides attributes of the observation conditions and effort in cell c at time t. Learning and inference for HMMs is well understood. However, in this case, the state of the HMM consists of the location of each bird in the population at each time step, and there are more than a billion birds in North America. So applying standard HMM methods is completely intractable.

Because none of the birds is tagged, we do not need to keep track of the location of each bird. Instead, it suffices to define a derived HMM that we call the Collective Hidden Markov Model (CHMM) [9]. Let $n^t(c)$ denote the number of birds in cell c at time t, and let $\boldsymbol{n}^t$ be the vector of these counts over all cells. Then we can define the transition distribution $P(\boldsymbol{n}^t|\boldsymbol{n}^{t-1}, X^{t-1})$ and the observation distribution $P(Y^t|\boldsymbol{n}^t, W^t)$. If we are willing to assume that each bird's migration decisions are independent and identically distributed, then this collective model is equivalent to the original HMM, but its state space is much smaller. Furthermore, let $\boldsymbol{n}^{t-1,t}$ denote the matrix of transition counts, such that $n^{t-1,t}(c, c')$ is the number of birds that moved from cell c to cell c' between time $t-1$ and time t. If we know the values of the counts $\boldsymbol{n}^t$ and $\boldsymbol{n}^{t-1,t}$ (or if we can estimate them via probabilistic inference in the CHMM), then these provide the sufficient statistics needed to estimate the transition probabilities in the original HMM. Hence, by reasoning in the CHMM, we can learn the parameters of the HMM.

Unfortunately, exact inference in this CHMM is still intractable, because we must consider all ways of partitioning N birds among the C cells, which is still an immense state space. To perform approximate inference, we have developed a Gibbs sampling algorithm that can draw samples in time that is independent of the population size [8]. We are currently applying this algorithm to fit the CHMM to eBird observations.

Once we have the fitted CHMM, we plan to apply it in several ways. First, we plan to provide a nightly bird migration forecast (a "Bird Cast"). This will be useful for managing low-altitude air traffic and wind farms. Second, we will analyze the fitted model to develop and test scientific hypotheses about the factors that control migration. This has the potential to help us understand how land-use changes and global climate change may affect future bird populations and migration patterns.

V. Concluding Remarks

Robust ecosystem management requires good models of the dynamics of ecosystems. This paper has described initial steps toward three aspects of such systems. First, we considered the problem of observing the current values of environmental variables such as temperature, wind, and solar radiation. Our approach relies on discovering and exploiting correlations among multiple, spatially-distributed sensors, so that we can isolate and recover from anomalies caused by sensing failures. Even this apparently simple problem poses computational challenges that had to be addressed by introducing approximations. Second, we considered the problem of fitting species distribution models and dynamical migration models to citizen science data. Although such data provide us with unprecedented spatial and temporal coverage, they also raise many challenges including spatial sampling bias, imperfect detection, and highly-variable observer expertise. Finally, we described our work in progress on fitting collective hidden Markov models to understand bird migration. These models promise to provide many scientific insights into migration phenomena.

Of course it is not enough to have good system models. We also need algorithms for computing optimal control policies using these models. Furthermore, there are many urgent ecosystem management problems where we must act immediately, before we have good models of the systems. Even our very best models are highly imperfect and fail to capture all of the complexity of these systems. Hence, in order to manage the earth's ecosystems well, our control policies must address two critical factors. First, they must balance achieving ecological goals with the need to collect additional observations that allow us to improve our models. Second, they must be robust to both the known unknowns (i.e., the explicit uncertainties represented by our probabilistic models) and the unknown unknowns (i.e., the unmodeled, or not-yet-modeled, aspects of the systems). Can we invent optimization methods that can meet these challenges?

REFERENCES

[1] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees. Boca Raton, FL: Chapman and Hall/CRC, 1984.

[2] E. Dereszynski, T. G. Dietterich, "Spatiotemporal models for anomaly detection in dynamic environmental monitoring campaigns," ACM Transactions on Sensor Networks, 8(1), 3:1-3:26, 2011.

[3] T. G. Dietterich, G. Hao, A. Ashenfelter, "Gradient tree boosting for training conditional random fields," Journal of Machine Learning Research, (9), 2113-2139. 2008.

[4] D. F. Doak, J. A. Estes, B. S. Halpern, U. Jacob, D. R. Lindberg, J. Lovvorn, D. H. Monson, "Understanding and predicting ecological dynamics: are major surprises inevitable?" Ecology, 89(4), 952–961, 2008.

[5] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," The Annals of Statistics, 29(5), 1189-1232, 2011.

[6] R. A. Hutchinson, L. Liu, T. G. Dietterich, "Incorporating boosted regression trees into ecological latent variable models," Twenty-Fifth AAAI Conference on Artificial Intelligence (pp. 1343-1348). 2011.

[7] D. MacKenzie, J. Nichols, G. Lachman, S. Droege, J. A. Royle, C. A. Langtimm, "Estimating site occupancy rates when detection probabilities are less than one," Ecology, 83(8), 2248-2255, 2002.

[8] D. Sheldon, T. G. Dietterich, "Collective graphical models," NIPS 2011. 2011.

[9] D. Sheldon, M. A. S. Elmohamed, D. Kozen, "Collective inference on Markov models for modeling bird migration," Advances in Neural Information Processing Systems, 20, 1321–1328, 2007.

[10] M. Sugiyama, T. Suzuki, T. Kanamori, Density ratio estimation in machine learning, 2012, New York, NY: Cambridge University Press.

[11] J. Yu, W. K. Wong, R. Hutchinson, "Modeling experts and novices in citizen science data for species distribution modeling," IEEE International Conference on Data Mining. Sydney, Australia: IEEE, 2010.