Taking it for a Test Drive: A Hybrid Spatio-temporal Model for Wildlife Poaching Prediction Evaluated through a Controlled Field Test

Shahrzad Gholami*[1], Benjamin Ford[1], Fei Fang[2], Andrew Plumptre[3], Milind Tambe[1], Margaret Driciru[4], Fred Wanyama[4], Aggrey Rwetsiba[4], Mustapha Nsubaga[5], and Joshua Mabonga[5]

[1] University of Southern California{benjamif,sgholami,tambe}@usc.edu
[2] Harvard University, Boston, MA, 02138, fangf07@seas.harvard.edu
[3] Wildlife Conservation Society, New York City, NY, 10460, aplumptre@wcs.org
[4] Uganda Wildlife Authority, Kampala, Uganda,
{margaret.driciru,fred.wanyama,aggrey.rwetsiba}@ugandawildlife.org
[5] Wildlife Conservation Society, Kampala, Uganda, {mnsubuga,jmabonga}@wcs.org

Abstract. Worldwide, conservation agencies employ rangers to protect conservation areas from poachers. However, agencies lack the manpower to have rangers effectively patrol these vast areas frequently. While past work has modeled poachers' behavior so as to aid rangers in planning future patrols, those models' predictions were not validated by extensive field tests. In this paper, we present a hybrid spatio-temporal model that predicts poaching threat levels and results from a five-month field test of our model in Uganda's Queen Elizabeth Protected Area (QEPA). To our knowledge, this is the first time that a predictive model has been evaluated through such an extensive field test in this domain. We present two major contributions. First, our hybrid model consists of two components: (i) an ensemble model which can work with the limited data common to this domain and (ii) a spatio-temporal model to boost the ensemble's predictions when sufficient data are available. When evaluated on real-world historical data from QEPA, our hybrid model achieves significantly better performance than previous approaches with either temporally-aware dynamic Bayesian networks or an ensemble of spatially-aware models. Second, in collaboration with the Wildlife Conservation Society and Uganda Wildlife Authority, we present results from a five-month controlled experiment *where rangers patrolled over 450 sq km across QEPA*. We demonstrate that our model successfully predicted (1) where snaring activity would occur and (2) where it would not occur; in areas where we predicted a high rate of snaring activity, rangers found more snares and snared animals than in areas of lower predicted activity. These findings demonstrate that (1) our model's predictions are selective, (2) our model's superior laboratory performance extends to the real world, and (3) these predictive models can aid rangers in focusing their efforts to prevent wildlife poaching and save animals.

Keywords: Predictive models, Ensemble techniques, Graphical models, Field test evaluation, Wildlife protection, Wildlife poaching

* Shahrzad Gholami & Benjamin Ford are both first authors of this paper.

1 Introduction

Wildlife poaching continues to be a global problem as key species are hunted toward extinction. For example, the latest African census showed a 30% decline in elephant populations between 2007 and 2014 [2, 13]. Wildlife conservation areas have been established to protect these species from poachers, and these areas are protected by park rangers. These areas are vast, and rangers do not have sufficient resources to patrol everywhere with high intensity and frequency.

At many sites now, rangers patrol and collect data related to snares they confiscate, poachers they arrest, and other observations. Given rangers' resource constraints, patrol managers could benefit from tools that analyze these data and provide future poaching predictions. However, this domain presents unique challenges. First, this domain's real-world data are few, extremely noisy, and incomplete. To illustrate, one of rangers' primary patrol goals is to find wire snares, which are deployed by poachers to catch animals. However, these snares are usually well-hidden (e.g., in dense grass), and thus rangers may not find these snares and (incorrectly) label an area as not having any snares. Second, poaching activity changes over time, and predictive models must account for this temporal component. Third, because poaching happens in the real world, there are mutual spatial and neighborhood effects that influence poaching activity. Finally, while field tests are crucial in determining a model's efficacy in the world, the difficulties involved in organizing and executing field tests often precludes them.

Previous works in this domain have modeled poaching behavior with real-world data. Based on data from a Queen Elizabeth Protected Area (QEPA) dataset, [7] introduced a two-layered temporal graphical model, CAPTURE, while [5] constructed an ensemble of decision trees, INTERCEPT, that accounted for spatial relationships. However, these works did not (1) account for both spatial and temporal components nor (2) validate their models via extensive field testing.

In this paper, we provide the following contributions. (1) We introduce a new hybrid model that enhances an ensemble's broad predictive power with a spatio-temporal model's adaptive capabilities. Because spatio-temporal models require a lot of data, this model works in two stages. First, predictions are made with an ensemble of decision trees. Second, in areas where there are sufficient data, the ensemble's prediction is boosted via a spatio-temporal model. (2) In collaboration with the Wildlife Conservation Society and the Uganda Wildlife Authority, we designed and deployed a large, controlled experiment to QEPA. Across 27 areas we designated across QEPA, rangers patrolled approximately 452 kilometers over the course of five months; to our knowledge, this is the largest controlled experiment and field test of Machine Learning-based predictive models in this domain. In this experiment, we tested our model's selectiveness: is our model able to differentiate between areas of high and low poaching activity?

In experimental results, (1) we demonstrate our model's superior performance over the state-of-the-art [5] and thus the importance of spatio-temporal modeling. (2) During our field test, rangers found over three times more snaring activity in areas where we predicted higher poaching activity. When accounting for differences in ranger coverage, rangers found twelve times the number of findings per kilometer walked in those areas. These results demonstrate that (i) our model is selective in its predictions and (ii) our model's superior predictive performance in the laboratory extends to the real world.

2 Background and Related Work

Spatio-temporal models have been used for prediction tasks in image and video processing. Markov Random Fields (MRF) were used by [12, 14] to capture spatio-temporal dependencies in remotely sensed data and moving object detection, respectively.

Critchlow et al. [3] analyzed spatio-temporal patterns in illegal activity in Uganda's Queen Elizabeth Protected Area (QEPA) using Bayesian hierarchical models. With real-world data, they demonstrated the importance of considering the spatial and temporal changes that occur in illegal activities. However, in this work and other similar works with spatio-temporal models [9, 10], no standard metrics were provided to evaluate the models' predictive performance (e.g., precision, recall). As such, it is impossible to compare our predictive models' performance to theirs. While [4] was a field test of [3]'s work, [9, 10] do not conduct field tests to validate their predictions in the real-world.

In the Machine Learning literature, [7] introduced a two-layered temporal Bayesian Network predictive model (CAPTURE) that was also evaluated on real-world data from QEPA. CAPTURE, however, assumes one global set of parameters for all of QEPA which ignores local differences in poachers' behavior. Additionally, the first layer, which predicts poaching attacks, relies on the current year's patrolling effort which makes it impossible to predict future attacks (since patrols haven't happened yet). While CAPTURE includes temporal elements in its model, it does not include spatial components and thus cannot capture neighborhood specific phenomena. In contrast to CAPTURE, [5] presented a behavior model, INTERCEPT, based on an ensemble of decision trees and was demonstrated to outperform CAPTURE. While their model accounted for spatial correlations, it did not include a temporal component. In contrast to these predictive models, our model addresses both spatial and temporal components.

It is vital to validate predictive models in the real world, and both [4] and [5] have conducted field tests in QEPA. [5] conducted a one month field test in QEPA and demonstrated promising results for predictive analytics in this domain. Unlike the field test we conducted, however, that was a preliminary field test and was not a controlled experiment. On the other hand, [4] conducted a controlled experiment where their goal, by selecting three areas for rangers to patrol, was to maximize the number of observations sighted per kilometer walked by the rangers. Their test successfully demonstrated a significant increase in illegal activity detection at two of the areas, but they did not provide comparable evaluation metrics for their predictive model. Also, our field test was much larger in scale, involving 27 patrol posts compared to their 9 posts.

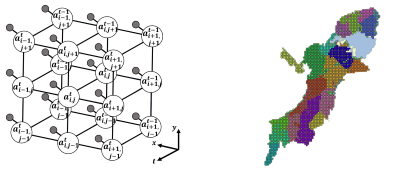
3 Wildlife Crime Dataset: Features and Challenges

This study's wildlife crime dataset is from Uganda's Queen Elizabeth Protected Area (QEPA), an area containing a wildlife conservation park and two wildlife reserves, which spans about 2,520 square kilometers. There are 37 patrol posts situated across QEPA from which Uganda Wildlife Authority (UWA) rangers conduct patrols to apprehend poachers, remove any snares or traps, monitor wildlife, and record signs of illegal activity. Along with the amount of patrolling effort in each area, the dataset contains 14 years (2003-2016) of the type, location, and date of wildlife crime activities.

Rangers lack the manpower to patrol everywhere all the time, and thus illegal activity may be undetected in unpatrolled areas. Patrolling is an imperfect process, and there is considerable uncertainty in the dataset's negative data points (i.e., areas being labeled as having no illegal activity); rangers may patrol an area and label it as having no snares when, in fact, a snare was well-hidden and undetected. These factors contribute to the dataset's already large class imbalance; there are many more negative data points than there are positive points (crime detected). It is thus necessary to consider models that estimate hidden variables (e.g., whether an area has been attacked) and also to evaluate predictive models with metrics that account for this uncertainty, such as those in the Positive and Unlabeled Learning (PU Learning) literature [6]. We divide QEPA into 1 square kilometer grid cells (a total of 2,522 cells), and we refer to these cells as targets. Each target is associated with several static geospatial features such as terrain (e.g., slope), distance values (e.g., distance to border), and animal density. Each target is also associated with dynamic features such as how often an area has been patrolled (i.e., coverage) and observed illegal activities (e.g., snares).

(a) Snare (b) QEPA grid

Fig. 1: Photo credit: UWA ranger



(a) Spatio-temporal model (b) Geo-Clusters

Fig. 2: Geo-clusters and graphical model

4 Models and algorithms

4.1 Prediction by Graphical models

Markov Random Field (MRF) To predict poaching activity, each target, at time step $t \in \{t_1, ..., t_m\}$, is represented by coordinates i and j within the boundary of QEPA. In Figure 2(a), we demonstrate a three-dimensional network for spatio-temporal modeling of poaching events over all targets. Connections between nodes represent the mutual spatial influence of neighboring targets and also the temporal dependence between recurring poaching incidents at a target. $a^t_{i,j}$ represents poaching incidents at time step t and target i, j. Mutual spatial influences are modeled through first-order neighbors (i.e., $a^t_{i,j}$ connects to $a^t_{i\pm1,j}$, $a^t_{i,j\pm1}$ and $a^{t-1}_{i,j}$) and second-order neighbors (i.e., $a^t_{i,j}$ connects to $a^t_{i\pm1,j\pm1}$); for simplicity, the latter is not shown on the model's lattice. Each random variable takes a value in its state space, in this paper, $\mathcal{L} = \{0, 1\}$.

To avoid index overload, henceforth, nodes are indexed by serial numbers, $\mathcal{S} = \{1, 2, ..., N\}$ when we refer to the three-dimensional network. We introduce two random fields, indexed by $\mathcal{S}$, with their configurations: $\mathcal{A} = \{\boldsymbol{a} = (a_1, ..., a_N) | a_i \in$

$\mathcal{L}, i \in \mathcal{S}\}$, which indicates an *actual* poaching attack occurred at targets over the period of study, and $\mathcal{O} = \{\boldsymbol{o} = (o_1, ..., o_N) | o_i \in \mathcal{L}, i \in \mathcal{S}\}$ indicates a *detected* poaching attack at targets over the period of study. Due to the imperfect detection of poaching activities, the former represents the hidden variables, and the latter is the known observed data collected by rangers, shown by the gray-filled nodes in Figure 2(a). Targets are related to one another via a neighborhood system, $\mathcal{N}_n$, which is the set of nodes neighboring n and $n \notin \mathcal{N}_n$. This neighborhood system considers all spatial and temporal neighbors. We define neighborhood attackability as the fraction of neighbors that the model predicts to be attacked: $u_{\mathcal{N}_n} = \sum_{n \in \mathcal{N}_n} a_n / |\mathcal{N}_n|$.

The probability, $P(a_i | u_{\mathcal{N}_n}, \boldsymbol{\alpha})$, of a poaching incident at each target n at time step t is represented in Equation 1, where $\boldsymbol{\alpha}$ is a vector of parameters weighting the most important variables that influence poaching; $\boldsymbol{Z}$ represents the vector of time-invariant ecological covariates associated with each target (e.g., animal density, slope, forest cover, net primary productivity, distance from patrol post, town and rivers [3, 8]). The model's temporal dimension is reflected through not only the backward dependence of each a_n, which influences the computation of $u_{\mathcal{N}_n}$, but also in the past patrol coverage at target n, denoted by c_n^{t-1}, which models the delayed deterrence effect of patrolling efforts.

$$p(a_n = 1 | u_{\mathcal{N}_n}, \boldsymbol{\alpha}) = \frac{e^{-\boldsymbol{\alpha}[\boldsymbol{Z}, u_{\mathcal{N}_n}, c_n^{t-1}, 1]^\intercal}}{1 + e^{-\boldsymbol{\alpha}[\boldsymbol{Z}, u_{\mathcal{N}_n}, c_n^{t-1}, 1]^\intercal}} \tag{1}$$

Given a_n, o_n follows a conditional probability distribution proposed in Equation 2, which represents the probability of rangers detecting a poaching attack at target n. The first column of the matrix denotes the probability of not detecting or detecting attacks if an attack has not happened, which is constrained to 1 or 0 respectively. In other words, it is impossible to detect an attack when an attack has not happened. The second column of the matrix represents the probability of not detecting or detecting attacks in the form of a logistic function if an attack has happened. Since it is less rational for poachers to place snares close to patrol posts and more convenient for rangers to detect poaching signs near the patrol posts, we assumed dp_n (distance from patrol post) and c_n^t (patrol coverage devoted to target n at time t) are the major variables influencing rangers' detection capabilities. Detectability at each target is represented in Equation 2, where $\boldsymbol{\beta}$ is a vector of parameters that weight these variables.

$$p(o_n | a_n) = \begin{bmatrix} p(o_n = 0 | a_n = 0) & p(o_n = 0 | a_n = 1, \boldsymbol{\beta}) \\ p(o_n = 1 | a_n = 0) & p(o_n = 1 | a_n = 1, \boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} 1, & \frac{1}{1 + e^{-\boldsymbol{\beta}[dp_n, c_n^t, 1]^\intercal}} \\ 0, & \frac{e^{-\boldsymbol{\beta}[dp_n, c_n^t, 1]^\intercal}}{1 + e^{-\boldsymbol{\beta}[dp_n, c_n^t, 1]^\intercal}} \end{bmatrix} \tag{2}$$

We assume that $(\boldsymbol{o}, \boldsymbol{a})$ is pairwise independent, meaning $p(\boldsymbol{o}, \boldsymbol{a}) = \prod_{n \in \mathcal{S}} p(o_n, a_n)$.

EM algorithm to infer on MRF We use the Expectation-Maximization (EM) algorithm [1] to estimate the MRF model's parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$. For completeness, we provide details about how we apply the EM algorithm to our model. Given a joint distribution $p(\boldsymbol{o}, \boldsymbol{a} | \boldsymbol{\theta})$ over observed variables $\boldsymbol{o}$ and hidden variables $\boldsymbol{a}$, governed by

parameters $\boldsymbol{\theta}$, EM aims to maximize the likelihood function $p(\boldsymbol{o}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. To start the algorithm, an initial setting for the parameters $\boldsymbol{\theta}^{old}$ is chosen. At E-step, $p(\boldsymbol{a}|\boldsymbol{o}, \boldsymbol{\theta}^{old})$ is evaluated, particularly, for each node in MRF model:

$$p(a_n|o_n, \boldsymbol{\theta}^{old}) = \frac{p(o_n|a_n, \boldsymbol{\beta}^{old}).p(a_n|u^{old}_{\mathcal{N}_n}, \boldsymbol{\alpha}^{old})}{p(o_n)} \tag{3}$$

M-step calculates $\boldsymbol{\theta}^{new}$, according to the expectation of the complete log likelihood, $\log p(\boldsymbol{o}, \boldsymbol{a}|\boldsymbol{\theta})$, given in Equation 4.

$$\boldsymbol{\theta}^{new} = \arg\max_{\boldsymbol{\theta}} \sum_{a_n \in \mathcal{L}} p(\boldsymbol{a}|\boldsymbol{o}, \boldsymbol{\theta}^{old}). \log p(\boldsymbol{o}, \boldsymbol{a}|\boldsymbol{\theta}) \tag{4}$$

To facilitate calculation of the log of the joint probability distribution, $\log p(\boldsymbol{o}, \boldsymbol{a}|\boldsymbol{\theta})$, we introduce an approximation that makes use of $u^{old}_{\mathcal{N}_n}$, represented in Equation 5.

$$\log p(\boldsymbol{o}, \boldsymbol{a}|\boldsymbol{\theta}) = \sum_{n \in \mathcal{S}} \sum_{a_n \in \mathcal{L}} \log p(o_n|a_n, \boldsymbol{\beta}) + \log p(a_n|u^{old}_{\mathcal{N}_n}, \boldsymbol{\alpha}) \tag{5}$$

Then, if convergence of the log likelihood is not satisfied, $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$, and repeat.

Dataset preparation for MRF To split the data into training and test sets, we divided the real-world dataset into year-long time steps. We trained the model's parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ on historical data sampled through time steps $(t_1, ..., t_m)$ for all targets within the boundary. These parameters were used to predict poaching activity at time step t_{m+1}, which represents the test set for evaluation purposes. The trade-off between adding years' data (performance) vs. computational costs led us to use three years ($m = 3$). The model was thus trained over targets that were patrolled throughout the training time period (t_1, t_2, t_3). We examined three training sets: 2011-2013, 2012-2014, and 2013-2015 for which the test sets are from 2014, 2015, and 2016, respectively.

Capturing temporal trends requires a sufficient amount of data to be collected regularly across time steps for each target. Due to the large amount of missing inspections and uncertainty in the collected data, this model focuses on learning poaching activity only over regions that have been continually monitored in the past, according to Definition 1. We denote this subset of targets as $\mathcal{S}_c$.

Definition 1. ***Continually vs. occasionally monitoring:*** *A target i, j is continually monitored if all elements of the coverage sequence are positive; $c^{t_k}_{i,j} > 0, \forall k = 1, ..., m$ where m is the number of time steps. Otherwise, it is occasionally monitored.*

Experiments with MRF were conducted in various ways on each data set. We refer to a) a *global* model with spatial effects as **GLB-SP**, which consists of a single set of parameters $\boldsymbol{\theta}$ for the whole QEPA, and b) a *global* model without spatial effects (i.e., the parameter that corresponds to $u_{\mathcal{N}_n}$ is set to 0) as **GLB**. The spatio-temporal model is designed to account for temporal and spatial trends in poaching activities. However, since learning those trends and capturing spatial effects are impacted by the variance in local poachers' behaviors, we also examined c) a *geo-clustered* model which consists

of multiple sets of local parameters throughout QEPA with spatial effects, referred to as **GCL-SP**, and also d) a *geo-clustered* model without spatial effects (i.e., the parameter that corresponds to $u_{\mathcal{N}_n}$ is set to 0) referred to as **GCL**.

Figure 2(b) shows the geo-clusters generated by Gaussian Mixture Models (GMM), which classifies the targets based on the geo-spatial features, $\boldsymbol{Z}$, along with the targets' coordinates, $(x_{i,j}, y_{i,j})$, into 22 clusters. The number of geo-clusters, 22, are intended to be close to the number of patrol posts in QEPA such that each cluster contains one or two nearby patrol posts. With that being considered, not only are local poachers' behaviors described by a distinct set of parameters, but also the data collection conditions, over the targets within each cluster, are maintained to be nearly uniform.

4.2 Prediction by Ensemble models

A **Bagging ensemble model** or **B**ootstrap **agg**regation technique, called Bagging, is a type of ensemble learning which bags some weak learners, such as decision trees, on a dataset by generating many bootstrap duplicates of the dataset and learning decision trees on them. Each of the bootstrap duplicates are obtained by randomly choosing M observations out of M with replacement, where M denotes the training dataset size. Finally, the predicted response of the ensemble is computed by taking an average over predictions from its individual decision trees. To learn a Bagging ensemble, we used the *fitensemble* function of MATLAB 2017a. **Dataset preparation** for the Bagging ensemble model is designed to find the targets that are liable to be attacked [5]. A target is assumed to be attackable if it has ever been attacked; if any observations occurred in the entire training period for a given target, that target is labeled as attackable. For this model, the best training period contained 5 years of data.

4.3 Hybrid of MRF and Bagging ensemble

Since the amount and regularity of data collected by rangers varies across regions of QEPA, predictive models perform differently in different regions. As such, we propose using different models to predict over them; first, we used a Bagging ensemble model, and then improved the predictions in some regions using the spatio-temporal model. For global models, we used MRF for all continually monitored targets. However, for geo-clustered models, for targets in the continually monitored subset, $\mathcal{S}_c^q$, (where temporally-aware models can be used practically), the MRF model's performance varied widely across geo-clusters according to our experiments. q indicates clusters and $1 \leq q \leq 22$. Thus, for each q, if the average Catch Per Unit Effort (CPUE), outlined by Definition 2, is relatively large, we use the MRF model for $\mathcal{S}_c^q$. In Conservation Biology, CPUE is an indirect measure of poaching activity abundance. A larger average CPUE for each cluster corresponds to more frequent poaching activity and thus more data for that cluster. Consequently, using more complex spatio-temporal models in those clusters becomes more reasonable.

Definition 2. ***Average CPUE*** *is* $\sum_{n \in \mathcal{S}_c^q} o_n / \sum_{n \in \mathcal{S}_c^q} c_n^t$ *in cluster q.*

To compute CPUE, effort corresponds to the amount of coverage (i.e., 1 unit = 1 km walked) in a given target, and catch corresponds to the number of observations. Hence,

for $1 \leq q \leq 22$, we will boost selectively according to the average CPUE value; some clusters may not be boosted by MRF, and we would only use Bagging ensemble model for making predictions on them. Experiments on historical data show that selecting 15% of the geo-clusters with the highest average CPUE results in the best performance for the entire hybrid model (discussed in the Evaluation Section).

5 Evaluations and Discussions

5.1 Evaluation metrics

The imperfect detection of poaching activities in wildlife conservation areas leads to uncertainty in the negative class labels of data samples [5]. It is thus vital to evaluate prediction results based on metrics which account for this inherent uncertainty. In addition to standard metrics in Machine Learning (e.g., precision, recall, F1) which are used to evaluate models on datasets with no uncertainty in the underlying ground truth, we also use the L&L metric introduced in [6], which is a metric specifically designed for models learned on Positive and Unlabeled datasets. L&L is defined as $L\&L = \frac{r^2}{Pr[f(Te)=1]}$, where r denotes the recall and $Pr[f(Te) = 1]$ denotes the probability of a classifier f making a positive class label prediction and is estimated by the percentage of positive predictions made by the model on a given test set.

5.2 Experiments with real-world data

Evaluation of models' attack predictions are demonstrated in Tables 1 and 2. To compare models' performances, we used several baseline methods, i) Positive Baseline, **PB**; a model that predicts poaching attacks to occur in all targets, ii) Random Baseline, **RB**; a model which flips a coin to decide its prediction, iii) Training Label Baseline, **TL**; a model which predicts a target as attacked if it has been ever attacked in the training data. We also present the results for Support Vector Machines, **SVM**, and AdaBoost methods, **AD**, which are well-known Machine Learning techniques, along with results for the best performing predictive model on the QEPA dataset, INTERCEPT, **INT**, [5]. Results for the Bagging ensemble technique, **BG**, and RUSBoost, **RUS**, a hybrid sampling/boosting algorithm for learning from datasets with class imbalance [11], are also presented. In all tables, **BG-G*** stands for the best performing model among all variations of the hybrid model, which will be discussed in detail later. Table 1 demonstrates that **BG-G*** outperformed all other existing models in terms of L&L and also F1.

Table 2 provides a detailed comparison of all variations of our hybrid models, **BG-G** (i.e., when different MRF models are used). When **GCL-SP** is used, we get the best performing model in terms of L&L score, which is denoted as **BG-G***. The poor results of learning a global set of parameters emphasize the fact that poachers' behavior and patterns are not identical throughout QEPA and should be modeled accordingly.

Our experiments demonstrated that the performance of the MRF model within $\mathcal{S}_c^q$ varies across different geo-clusters and is related to the CPUE value for each cluster, q. Figure 3(a) displays an improvement in L&L score for the **BG-G*** model compared to **BG** vs. varying the percentile of geo-clusters used for boosting. Experiments with the

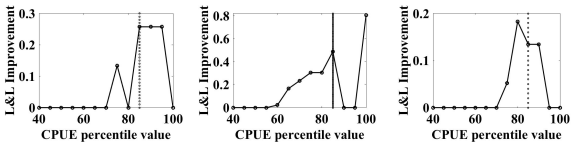
Test set	2014					2015					2016				
Models	**PB**	**RB**	**TL**	**SVM**	**BG-G***	**PB**	**RB**	**TL**	**SVM**	**BG-G***	**PB**	**RB**	**TL**	**SVM**	**BG-G***
Precision	0.06	0.05	0.26	0.24	0.65	0.10	0.08	0.39	0.4	0.69	0.10	0.09	0.45	0.45	0.74
Recall	1.00	0.46	0.86	0.3	0.54	1.00	0.43	0.78	0.15	0.62	1.00	0.44	0.75	0.23	0.66
F1	0.10	0.09	0.4	0.27	0.59	0.18	0.14	0.52	0.22	0.65	0.18	0.14	0.56	0.30	0.69
L&L	1.00	0.43	4.09	1.33	6.44	1.00	0.37	3.05	0.62	4.32	1.00	0.38	3.4	1.03	4.88
Models	**RUS**	**AD**	**BG**	**INT**	**BG-G***	**RSU**	**AD**	**BG**	**INT**	**BG-G***	**RUS**	**AD**	**BG**	**INT**	**BG-G***
Precision	0.12	0.33	0.62	0.37	0.65	0.2	0.52	0.71	0.63	0.69	0.19	0.53	0.76	0.40	0.74
Recall	0.51	0.47	0.54	0.45	0.54	0.51	0.5	0.53	0.41	0.62	0.65	0.54	0.62	0.66	0.66
F1	0.19	0.39	0.58	0.41	0.59	0.29	0.51	0.61	0.49	0.65	0.29	0.53	0.68	0.51	0.69
L&L	1.12	2.86	6.18	5.83	6.44	1.03	2.61	3.83	3.46	4.32	1.25	2.84	4.75	2.23	4.88

Table 1: Comparing all models' performances with the best performing BG-G model

Test set	2014				2015				2016			
MRF models	**GLB**	**GLB-SP**	**GCL**	**GCL-SP**	**GLB**	**GLB-SP**	**GCL**	**GCL-SP**	**GLB**	**GLB-SP**	**GCL**	**GCL-SP**
Precision	0.12	0.12	0.63	0.65	0.19	0.19	0.69	0.69	0.18	0.19	0.72	0.74
Recall	0.58	0.65	0.54	0.54	0.52	0.58	0.65	0.62	0.50	0.46	0.66	0.66
F1	0.20	0.20	0.58	0.59	0.28	0.29	0.65	0.65	0.27	0.27	0.69	0.69
L&L	1.28	1.44	6.31	6.44	0.99	1.14	4.32	4.32	0.91	0.91	4.79	4.88

Table 2: Performances of hybrid models with variations of MRF (BG-G models)

2014 test set show that choosing the 85^{th} percentile of geo-clusters for boosting with MRF, according to CPUE, (i.e., selecting 15% of the geo-clusters, with highest CPUE), results in the best prediction performance. The 85^{th} percentile is shown by vertical lines in Figures where the **BG-G*** model outperformed the **BG** model. We used a similar percentile value for conducting experiments with the MRF model on test sets of 2015 and 2016. Figure 3(b) and 3(c) confirm the efficiency of choosing an 85^{th} percentile value for those test sets, as well. Also, Table 1 demonstrates that for **BG-G*** recall increased up to almost 10% for the 2015 test set which would result in marking roughly 10% more vulnerable targets as attacked and thus protecting more endangered animals.

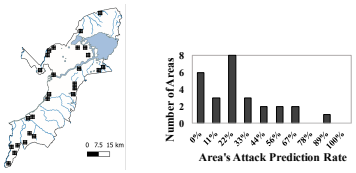


(a) Test set 2014 (b) Test set 2015 (c) Test set 2016

Fig. 3: L&L improvement vs. CPUE percentile value; BG-G* compared to BG

6 QEPA Field Test

While our model demonstrated superior predictive performance on historical data, it is important to test these models in the field.



(a) Patrolled areas (b) Prediction rates

Fig. 4: Patrol Area Statistics

The initial field test we conducted in [5], in collaboration with the Wildlife Conservation Society (WCS) and the Uganda Wildlife Authority (UWA), was the first of its kind in the Machine Learning (ML) community and showed promising improvements over previous patrolling regimes. Due to the difficulty of organizing such a field test, its implications were limited: only two 9-sq km areas (18 sq km) of QEPA were patrolled by rangers over a month. Because of its success, however, WCS and UWA graciously agreed to a larger scale, controlled experiment: also in 9 sq km areas, but rangers patrolled 27 of these areas (243 sq km, spread across QEPA) over five months; this is the largest to-date field test of ML-based predictive models in this domain. We show the areas in Figure 4(a). Note that rangers patrolled these areas in addition to other areas of QEPA as part of their normal duties.

This experiment's goal was to determine the selectiveness of our model's snare attack predictions: does our model correctly predict both where there are and are not snare attacks? We define attack prediction rate as the proportion of targets (a 1 km by 1 km cell) in a patrol area (3 by 3 cells) that are predicted to be attacked. We considered two experiment groups that corresponded to our model's attack prediction rates from November 2016 - March 2017: High (group 1) and Low (group 2). Areas that had an attack prediction rate of 50% or greater were considered to be in a high area (group 1); areas with less than a 50% rate were in group 2. For example, if the model predicted five out of nine targets to be attacked in an area, that area was in group 1. Due to the importance of QEPA for elephant conservation, we do not show which areas belong to which experiment group in Figure 4(a) so that we do not provide data to ivory poachers.

To start, we exhaustively generated all patrol areas such that (1) each patrol area was 3x3 sq km, (2) no point in the patrol area was more than 5 km away from the nearest ranger patrol post, and (3) no patrol area was patrolled too frequently or infrequently in past years (to ensure that the training data associated with all areas was of similar quality); in all, 544 areas were generated across QEPA. Then, using the model's attack predictions, each area was assigned to an experiment group. Because we were not able to test all 544 areas, we selected a subset such that no two areas overlapped with each other and no more than two areas were selected for each patrol post (due to manpower constraints). In total, 5 areas in group 1 and 22 areas in group 2 were chosen. Note that this composition arose due to the preponderance of group 2 areas (see Table 3). We provide a breakdown of the areas' exact attack prediction rates in Figure 4(b); areas with rates below 56% (5/9) were in group 2, and for example, there were 8 areas in group 2 with a rate of 22% (2/9). Finally, when we provided patrols to the rangers, *experiment group memberships were hidden to prevent effects where knowledge of predicted poaching activity would influence their patrolling patterns and detection rates*.

Experiment Group	Exhaustive Patrol Area Group Memberships	Final Patrol Area Group Memberships
High (1)	50 (9%)	5 (19%)
Low (2)	494 (91%)	22 (81%)

Table 3: Patrol Area Group Memberships

6.1 Field Test Results and Discussion

The field test data we received was in the same format as the historical data. However, because rangers needed to physically walk to these patrol areas, we received additional data that we have omitted from this analysis; observations made outside of a designated patrol area were not counted. Because we only predicted where snaring activity would occur, we have also omitted other observation types made during the experiment (e.g., illegal cattle grazing). We present results from this five-month field test in Table 4. To provide additional context for these results, we also computed QEPA's park-wide historical CPUE (from November 2015 to March 2016): 0.04.

Experiment Group	Observation Count(%)	Mean Count(std)	Effort(%)	CPUE
High (1)	15 (79%)	3 (5.20)	129.54 (29%)	0.12
Low (2)	4 (21%)	0.18 (0.50)	322.33 (71%)	0.01

Table 4: Field Test Results: Observations

Areas with a high attack prediction rate (group 1) had significantly more snare sightings than areas with low attack prediction rates (15 vs 4). This is despite there being far fewer group 1 areas than group 2 areas (5 vs 22); on average, group 1 areas had 3 snare observations whereas group 2 areas had 0.18 observations. It is worth noting the large standard deviation for the mean observation counts; the standard deviation of 5.2, for the mean of 3, signifies that not all areas had snare observations. Indeed, two out of five areas in group 1 had snare observations. However, this also applies to group 2's areas: only 3 out of 22 areas had snare observations.

We present Catch per Unit Effort (CPUE) results in Table 4. When accounting for differences in areas' effort, group 1 areas had a CPUE that was over ten times that of group 2 areas. Moreover, when compared to QEPA's park-wide historical CPUE of 0.04, it is clear that our model successfully differentiated between areas of high and low snaring activity. The results of this large-scale field test, the first of its kind for ML models in this domain, demonstrated that our model's superior predictive performance in the laboratory extends to the real world.

7 Conclusion

In this paper, we presented a hybrid spatio-temporal model to predict wildlife poaching threat levels. Additionally, we validated our model via an extensive five-month field test in Queen Elizabeth Protected Area (QEPA) where rangers patrolled over 450 sq km across QEPA — the largest field-test to-date of Machine Learning-based models in this

domain. On real-world historical data from QEPA, our hybrid model achieves significantly better performance than prior work. On the data collected from our field test, we demonstrated that our model successfully differentiated between areas of high and low snaring activity. These findings demonstrated that our model's predictions are selective and also that its superior laboratory performance extends to the real world. Based on these promising results, future work will focus on deploying these models as part of a software package to UWA to aid in planning future anti-poaching patrols.

Acknowledgments: This research was supported by MURI grant W911NF-11-1-0332, NSF grant with Cornell University 72954-10598 and partially supported by Harvard Center for Research on Computation and Society fellowship. We are grateful to the Wildlife Conservation Society and the Uganda Wildlife Authority for supporting data collection in QEPA. We thank the rangers and wardens for their contributions in collecting patrolling data, and we also thank Donnabell Dmello for her help in data processing.

References

1. Bishop, C.M.: Pattern recognition. Machine Learning 128, 1–58 (2006)
2. Census, G.E.: The great elephant census — a paul g. allen project. Press Release (Aug 2016)
3. Critchlow, R., Plumptre, A., Driciru, M., Rwetsiba, A., Stokes, E., Tumwesigye, C., Wanyama, F., Beale, C.: Spatiotemporal trends of illegal activities from ranger-collected data in a ugandan national park. Conservation Biology 29(5), 1458–1470 (2015)
4. Critchlow, R., Plumptre, A.J., Alidria, B., Nsubuga, M., Driciru, M., Rwetsiba, A., Wanyama, F., Beale, C.M.: Improving law-enforcement effectiveness and efficiency in protected areas using ranger-collected monitoring data. Conservation Letters (2016)
5. Kar, D., Ford, B., Gholami, S., Fang, F., Plumptre, A., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A.: Cloudy with a chance of poaching: Adversary behavior modeling and forecasting with real-world poaching data (2017)
6. Lee, W.S., Liu, B.: Learning with positive and unlabeled examples using weighted logistic regression. In: ICML. vol. 3 (2003)
7. Nguyen, T.H., Sinha, A., Gholami, S., Plumptre, A., Joppa, L., Tambe, M., Driciru, M., Wanyama, F., Rwetsiba, A., Critchlow, R., et al.: Capture: A new predictive anti-poaching tool for wildlife protection. pp. 767–775. AAMAS (2016)
8. O'Kelly, H.J.: Monitoring conservation threats, interventions, and impacts on wildlife in a cambodian tropical forest. Imperial College, London p. 149 (2013)
9. Rashidi, P., Wang, T., Skidmore, A., Mehdipoor, H., Darvishzadeh, R., Ngene, S., Vrieling, A., Toxopeus, A.G.: Elephant poaching risk assessed using spatial and non-spatial bayesian models. Ecological Modelling 338, 60–68 (2016)
10. Rashidi, P., Wang, T., Skidmore, A., Vrieling, A., Darvishzadeh, R., Toxopeus, B., Ngene, S., Omondi, P.: Spatial and spatiotemporal clustering methods for detecting elephant poaching hotspots. Ecological Modelling 297, 180–186 (2015)
11. Seiffert, C., Khoshgoftaar, T.M., Van Hulse, J., Napolitano, A.: Rusboost: A hybrid approach to alleviating class imbalance. IEEE SMC-A: Systems and Humans 40(1), 185–197 (2010)
12. Solberg, A.H.S., Taxt, T., Jain, A.K.: A markov random field model for classification of multisource satellite imagery. IEEE TGRS 34(1), 100–113 (1996)
13. on International Trade in Endangered Species of Wild Fauna, C., Flora: African elephants still in decline due to high levels of poaching. Press Release (Mar 2016)
14. Yin, Z., Collins, R.: Belief propagation in a 3d spatio-temporal mrf for moving object detection. In: IEEE CVPR. pp. 1–8. IEEE (2007)