Reducing Poaching Risk through Land Use and Patrol Routes Planning using Data Driven Optimization

Yiwen Yuan

CMU-CS-21-117

May 2021

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Dr. Fei Fang(Advisor)
Dr. George Chen

Submitted in partial fulfillment of the requirements for the degree of Master in Computer Science.

Keywords: Anti-poaching, Data-driven Optimization, Forest Zoning

Abstract

The forest, along with the many products it provides, is an important source of income to the local population of the Congo Basin. Specifically, legal logging generates revenue for the local government and creates jobs for residents. However, the increase of human intrusion on the forest threatens the livelihood of ecosystems. Studies have found that the roads built by logging companies to transport logs have facilitated poaching activity in the area. Adequate land planning can be a solution to this issue. We mostly focus our work on the two tasks. First, we focus on evaluating the capacity and limits of state-of-art machine learning models on predicting poaching risk using geological features. Second, given historical data, we aim at designing future land use assignments and patrol routes that would induce the least amount of poaching risk in the area. In our work, we make the following contributions: 1) we train and test several models on the task of predicting poaching risk in the Congo with multiple data sets. Our results show that with different synthetically labeled data sets, the models' performance can achieve around 0.74 in AUC. 2) we propose a suitability based forest zoning optimization problem that can assign an area with multiple land uses. 3) we propose a data-driven optimization problem to determine a set of logging sites and patrol routes that maximizes revenue while reducing poaching risk in the area. For both 2) and 3), our experiments show that our calculated solution can induce much less poaching risk in the area compared to current practice.

Acknowledgments

First and foremost I am extremely grateful to my advisor, Prof. Fei Fang, for her invaluable advice, continuous support, and patience during my Master's study. Their immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. In addition, I would like to thank Claire Wang, Dr. David Morgan and William Winston. Without their assistance and dedicated involvement in every step throughout the process, this project would have never been accomplished. Last but not the least, I would like to say thank you to Prof. George Chen, who is kind enough to be my committee member and help me on my thesis.

Contents

1 Introduction **1**

2 Related Work **5**

3 Poaching Risk Prediction **9**
- 3.1 Data Set 9
  - 3.1.1 Area of Interest 9
  - 3.1.2 Features 10
- 3.2 Methodology 13
  - 3.2.1 Machine Learning Models 13
- 3.3 Evaluation 14
  - 3.3.1 Label Sampling 14
  - 3.3.2 Results 17

4 Suitability Based Forest Zoning **21**
- 4.1 Problem Formulation 21
  - 4.1.1 Objective Function 22
  - 4.1.2 Constraints 23
- 4.2 Land Use Suitability Calculation 25
- 4.3 Methodology 27
- 4.4 Experiments and Evaluations 28

5 Minimizing Poaching Risk through Land-Use Assignment and Patrol Route Design **31**
- 5.1 Problem Formulation 31
  - 5.1.1 Logging Sites Selection 32
  - 5.1.2 Patrol Routes Planning 34
- 5.2 Methodology 35
- 5.3 Evaluation and Results 36

6 Conclusion and Discussion **39**
- 6.1 Conclusion 39
- 6.2 Future Work 39

Bibliography **41**

List of Figures

3.1 The visualization of forest cover map of year 2016 in a section of area of interest. The area covered by forest green color is area covered by trees. The roads and rivers are the area covered by white/blue. 10

3.2 A cropped portion of original road data provided by the logging companies vs. added roads. Original roads are outlined in red. Added roads are outlined in green. The black in the background represents land with forest cover. 12

3.3 Green refers to intact forest landscape; blue refers to logging sites in year 2017; yellow refers to patrolled grids and red refer to poaching found in patrolled grids. The method used for simulating poaching and patrolling layers are: 1) top left: rule-based/sigmoid 2) top right: uniform/sigmoid 3) bottom left: rule-based/rule-based 4) bottom right: uniform/rule-based 16

4.1 The surrounding grids of area of interest. Green refers to logging sites, purple refers to conservation and blue is reserved for future logging. 24

4.2 Green refers to logging sites, purple refers to conservation and blue is reserved for future logging. The sampling combination(poaching, patrolling) for each : 1) top left: sigmoid, rule-based. 2) top right: sigmoid, uniform. 3) rule-based, rule-based. 4) rule-based, uniform 29

5.1 A comparison between simulated land-use and patrol decisions and optimized land-use and patrol decision results. Left: simulated data from Gaussian process. Right: Optimized decision. Green Area denotes the intact forest landscape; Blue Area denotes the logging sites; Red area denotes poaching activity; Yellow area denotes patrol routes. 35

5.2 The decision result for each sampling combination of patrolling and poaching: 1) top left: rule-based/sigmoid 2) top right: uniform/sigmoid 3) rule-based/rule-based 4) rule-based/uniform 36

List of Tables

3.1 Example data point, with feature abbreviations. 11
3.2 The poaching risk prediction performance of each model and each sampling method for poaching. The data used for training is from year 2017 and the data used for testing is from year 2018. The sampling method for patrolling is uniform. 18
3.3 The poaching risk prediction performance of each model and each sampling method for poaching. The data used for training is from year 2017 and the data used for testing is from year 2018. The sampling method for patrolling is rule-based. 19

4.1 The suitability score prediction performance of each individual land use model. The data set uses rule based sampling for patrolling and poaching layer. 27
4.2 The total suitability score of the result assignment of our model and actual assignment. 30

5.1 The average poaching risk in IFL area. The first element of the tuple is the poaching risk of synthetic 2018 assignments. The second element of the tuple is the poaching risk of our proposed logging sites. 37
5.2 The average poaching risk in each patrolling grid. The first element of the tuple is the poaching risk of synthetic 2018 patrol routes. The second element of the tuple is the poaching risk of our proposed routes 38

Chapter 1

Introduction

Studies from Maisels et al. and Blake et al. have shown a sharp decline in the population of forest elephants in central Africa since the early 2000s [13] [2]. The increase in human activities in the forest (including illegal poaching, hunting, and logging) has become a major threat to the forest elephants. Such human-elephant interactions are inevitable as usage of the forest is an important source of income for locals, as logging not only contributes to government revenue, but also creates jobs for local residents.

In the Kabo region in particular, logging activities dating back to the 1980s have changed the region's landscape. Each year, the local government and the logging company decide upon an area (the annual "Allowable Cut") in the remaining IFL (Intact Forest Landscape) used for logging. Roads and trails are then opened in that area. According to domain experts, the trails and roads left out by loggers have facilitated poaching activity, which have accelerated the decline of forest elephant numbers since 1980s. To address this issue, local stakeholders have sent out patrol teams over the years to collect information on poaching sites, campsites, species distribution, etc. Ultimately however, patrol resources are limited, so it is hard to cover all areas with high frequency and rigor. In addition, multiple reports of gunshots reveal that poachers are aware of the rangers' patrol patterns and tend to avoid the rangers by either going into small trails created by logging or getting out at night. Therefore, even though substantial amount of effort has been put into patrols, the number of poaching cases found still increases each year. Nevertheless, Breuer et al. suggest making structured and prospective scenario planning through land use management [5]. Our work follows their suggestion and focuses on the cause of the challenge. First, we test the capacity of multiple machine learning models on their capabilities in predicting poaching risk. Second, we investigate the task of making future land use assignments given historical data and practices. We develop a data-driven forest zoning model for the forest that achieves a sustainable fulfillment of both social needs and ecological continuity. Finally, we formulate an optimization problem to select the best logging sites that would induce least poaching threat to the environment and to determine the best patrol routes for the rangers.

Despite the abundance in previous work on poaching threat detection, there are several main challenges to our prediction task. First, the landscape (allowable cuts and intact forest landscape) changes yearly. According to domain experts, elephants tend to stay 10 kilometers away from allowable cut regions due to noise generated by logging activity. Therefore, no poaching occurs in the allowable cut each year. Prior studies have utilized Markov Random Fields [11] or clustering techniques [15] to model the temporal change in poaching pattern. However, these are not viable in our case as the landscape feature vector changes yearly, from the openings of small trails in the forest to the shift in logging sites to the shrink in intact forest landscape.

Another challenge we faced is that existing conservation software and technologies are not a good fit for our problem. Popular methods like Marxan and Zonation takes in the assumption that a planning unit is associated with a fixed amount of species population. The total amount of population to be reserved is then calculated by summing up the population amount in each selected units. And this value is lower bounded by a target amount in the constraint. This is not true in our case as elephant distribution is sparse in the vast forest. In addition, GPS telemetry data from Blake et al. shows that an elephant's yearly residence area can be as large as $1000km^2$ [3]. Reaching the constraint on the total amount of population might cause the program to take too many planning units for reservation. To tackle the problem, we take inspiration from existing forest planning techniques and perform data-driven optimization to find the optimal assignment of land use. We split the area of interest into small grid cells. Each grid cell is associated with a suitability score to each land use. Each pair of the grid cells is associated with a compatibility score for each pair of use it can take. The optimization objective is thus the summation of suitability and compatibility. To penalize fragmented assignments, a penalty term is added to the objective. The suitability score is learned using the data set constructed by the machine learning model we trained for predicting poaching risk on a yearly basis. The simulated annealing search algorithm is used to efficiently find solution to the problem as the number of planning units scales into thousands. The resulting performance varies with the land use, with the park use model performing the strongest with respect to predicting the ground truth land use on a constructed data set.

However, the suitability based zoning has some intrinsic drawbacks. The main issue is the bias in calculating the suitability score. Since there is no ground truth label for suitability and it is determined by manual labeling based on domain knowledge, its validity can hardly be verified or measured. We thus propose a new optimization problem formulation, where our overall goal is to maximize revenue gained from logging while minimizing total risk induced by the logging activities. We formulate it as an integer linear programming problem and solve it using Gurobi software. In addition, we also plan out patrol routes for the rangers so they can focus their patrols on high poaching risk areas. The logging sites selected by our work would induce lower poaching risk in the surrounding area and the patrol routes planned out by our work would cover areas with twice as much average poaching risk as compared to selecting patrol areas uniformly at random or based on simple rules.

We are working with the Wildlife Conservation Society (WCS) and local stakeholders to take the results of our model into consideration in the planning of new logging sites in the Congo Basin.

Chapter 2

Related Work

Prior studies have been done on gauging poaching threat given geological data. Ferreguetti et al. discovered that poacher occupancy increases near water resources and forest edges [22]. Poacher detectability also increases near these areas, human settlements, areas dense with game species, and during periods of bright moonlight. Results from Shaffer et al. also reinforce this, using GIS software and statistical analysis to find that poaching incidents are more common near roads and water features; closer proximity to roads facilitates easier escape, and proximity to water sources increase the likelihood of finding wildlife [16]. Work by Fang et al. in Queen Elizabeth National Park in Uganda makes use of nearest distance to the park boundary, roads, rivers and lakes, towns and villages, and patrol posts; topographical slope; wetness; and animal density estimates of the local species [10].

The objective of zoning is to assign areas for human activities while maintaining conservation objectives. Many technologies and softwares have been developed for solving the land use planning problem. In general, there are two approach to the problem. 1) Minimize total cost in constructing a conservation network. 2) Maximize the suitability and compatibility of the land use assignment. For the first type of problem, there exists many mature softwares and techniques, such as Marxan[1] and Zonation[2]. Marxan solves a minimum representation problem of finding the lowest cost set of planning units to conserves a specific amount of the target feature [1]. Zonation uses priority rankings to obtain a set of high value planning units(i.e. cells with high animal diversity and density). Both Marxan and Zonation can only solve the binary problem, i.e. whether a planning unit should be conserved or not. Marxan with Zones, implemented by Watts, allows for multi-use land planning through adding a term in the optimization objective that penalizes the objective if a feature cannot reach its target conservation amount under the current assignment of planning units [19]. As mentioned in the introduction section, the amount constraint on the conservation target in the first approach does not quite align with our problem. In studies on the second approach, multicriteria decision analysis (MCDA) is a common tool. Eastman proposed procedures in GIS to perform multi-criteria evaluation for suitability and multi-objective decision making on the assignment [9]. In Verdiella et al.'s work, the maximization objective is the sum of the suitability score of all planning units and compatibility score of all pairs of planning units [17]. The compatibility score is multiplied by the inverse of distance between the pair. Our zoning model is structured after their work, but the constraints and objective functions are modified based on the logging scenario. In addition, we adopted a data-driven method in calculating the suitability score.

Prior work has also been done to learn the suitability of various geographic locations for different types of land usage using machine learning. A variety of computational methods have been used to tackle the land use suitability task, i.e., classifying the suitability for various areas of land the most applicable use for it (from a discrete set of use cases), i.e., habitat for certain wildlife species, urban development, agricultural usage, etc. In the last century, the calculation of suitability score, which is usually referred as multi-criteria evaluation, is achieved within GIS (Geological Information System) using methods like weighted linear combination (WLC), or-

[1]http://www.uq.edu.au/marxan/

[2]https://www.helsinki.fi/en/researchgroups/digital-geography-lab/software-developed-in-cbig

dered weighted average (OWA), as mentioned by Eastman [9]. Such calculations are restricted by the functionalities of the GIS and machine learning is used. Mokarram et al. use an Adaboost and rotating forest-inspired algorithm called RotBoost to boost performance of decision trees in classifying land on the Shavur plain into 3 different levels of agricultural suitability [14]. They use a method proposed by the UN's Food and Agriculture Organization to generate ground truth labels for training. Zhang et al. use a feed-forward neural network to evaluate suitability for construction in Hangzhou, using the Delphi expert labelling method to obtain ground truth labels [21]. Djuric et al. assess urban development suitability in Belgrade, Serbia using support vector machines [8]. Brown et al.'s work does not discuss AI methods for suitability analysis, but lists the following as relevant factors in determination of logging suitability [6]:

- timber density (volume of timber produced per acre)
- ability of reforestation in logging area

Though for our task, evidently, we would want to consider more sustainability-focused features as well.

Chapter 3

Poaching Risk Prediction

The task of predicting poaching threat has always been important as it can not only provide both government officials and forest rangers useful information on the area's current situation, but also facilitate future decision making and planning. In this section, we investigate into the design of a poaching threat prediction pipeline, starting from data collection, to machine learning prediction. We first construct a data set using open source datasets and shapefiles. Then we test several models on their capacity in predicting poaching threat sampled from various methods. The constructed data set and trained prediction model would enable the downstream tasks of land use and patrolling decision making which we will explain in the next chapters.

3.1 Data Set

3.1.1 Area of Interest

The area of interest in our study is the Kabo forest in Republic of Congo. There are three types of land. Annual Allowable Cut (AAC) stands for the logging sites in the year. Intact Forest Landscape (IFL) stands for the area that borders forest and shows no major signs of human activity. National Park is the conserved area.

Our region of interest is a rectangular region with top right lat-long coordinate $(17.282496, 2.63123)$ and bottom left lat-long coordinate $(16.01047, 1.217691)$. The region is chosen so that it covers all past allowable cuts(areas for logging) and extends into the Nouabalé-Ndoki National Park and the town of Bomassa that are adjacent to the allowable cuts.

3.1.2 Features

We split the area of interest into square grid cells of 1km by 1km. In total there are 4350 grid cells. Each grid cell is associated with a set of geographical covariates. There are two kinds of features for each grid cell: 1) distance to a variety of landscapes, such as roads, villages, rivers, parks 2) geological information including area of tree cover, elevation, area of water cover. The full set of features is shown in Tab 3.1.

We obtain the first set of features from the shapefiles provided by the logging company and

Figure 3.1: The visualization of forest cover map of year 2016 in a section of area of interest. The area covered by forest green color is area covered by trees. The roads and rivers are the area covered by white/blue.

domain experts. The shapefiles contain multi polygons of various landscapes and land use areas such as intact forest landscapes, allowable cuts, rivers, villages, roads etc. One challenge of

Feature	Example 1
distance to roads (dist2road)	5.89m
distance to rivers (dist2river)	6.68km
distance to village (dist2village)	5.32km
distance to annual allowable cuts (dist2AAC)	13.00km
distance to intact forest landscape (dist2IFL)	41.81km
is the grid cell in a national park? (in_park)	0
highest elevation (high_elev)	437m
lowest elevation (low_elev)	426m
median elevation (median_elev)	432m
mean elevation (mean_elev)	431.5m
area covered by forest (forest_cover)	0.2197km^2
area covered by water (water_cover)	0km^2

Table 3.1: Example data point, with feature abbreviations.

obtaining the distance to roads feature from the shapefiles is that the shapefiles are incomplete. Specifically, to be able to transport the logs, the logging companies open new trails in the forest in the area they log each year. However, many trails are omitted in their report to the conservation organization. Calculating distance directly from the given shapefiles would cause bias. Our approach to correct this bias is to identify the trails using the global forest cover data set created by Hansen et al.(2018) [12]. The data set records global forest cover in 2000 and yearly forest loss/gain till 2018. We can access this data set as a `FeatureCollection` in Google Earth Engine Code. Using the built in methods, we accumulate the forest loss each year in the area of interest to the original forest map in 2000 to retrieve a forest cover map of each year, as shown in Fig 3.1. Using the forest data map constructed, we are able to identify and label the missing roads in the original shapefile each year as shown in Fig 3.2. Since we are looking at a fairly small area with only around 130 roads in total, we manually draw the missing roads. This would not be a viable option for larger areas or years after 2018. However, we can use computer vision tools on satellite imageries to directly retrieve roads data. Nachmany et al. [20]'s work indicates that regionally trained models perform well on identifying roads using state-of-art models. This can be the approach in those cases.

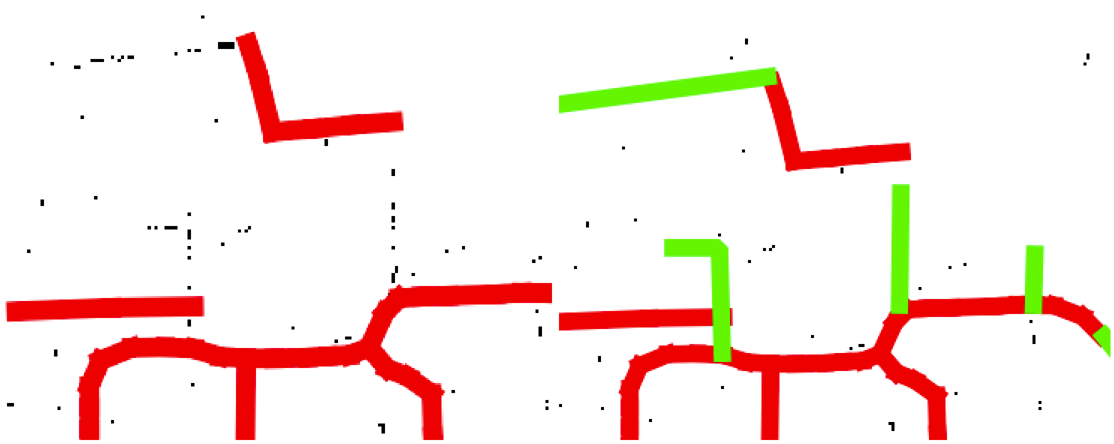

Figure 3.2: A cropped portion of original road data provided by the logging companies vs. added roads. Original roads are outlined in red. Added roads are outlined in green. The black in the background represents land with forest cover.

The values of the first set of distance features are calculated using the Pyshp package in Python. Note that we use the center point of a grid to calculate its distance to landscapes outside that grid. For landscapes that locate within the grid, then the distance to them is 0.

The second set of geological features are collected from a variety of open source data set. The elevation data is collected from The Shuttle Radar Topography Mission (SRTM) digital elevation data set. The tree cover data is collected from the yearly forest cover data set we have constructed out of the Hansen Global Forest Change data set. All the features are calculated using the `reduce` function on Google Earth Engine.

A data point is assigned to a positive label if poaching has occurred in the grid and rangers have patrolled in the grid. A data point is assigned to a negative label if poaching has not occurred in the grid and rangers have patrolled in the grid.

3.2 Methodology

3.2.1 Machine Learning Models

We train the following models on the training set and assess their accuracy, precision, recall, F1 score, and area under the curve (AUC) on the testing set. Let $\approx$ represent the feature vector associated with each data point.

- *Multilayer Perceptron* We use a Multilayer Perceptron with 200 hidden nodes and ReLU activation. W_1 is the weight matrix of size $\texttt{feature\_size} \times \texttt{hidden\_size}$. W_2 is the weight matrix of size $\texttt{hidden\_size} \times \texttt{1}$. b_1 is the bias vector of size $\texttt{hidden\_size}$. b_2

is the bias term of size `1`. MLP is specified by

$$\mathbf{h}(\mathbf{t}) = W_1\mathbf{t} + b_1$$

$$\mathbf{a_1}(\mathbf{h}) = \text{ReLU}(\mathbf{h})$$

$$o(\mathbf{a_1}) = W_2\mathbf{a_1} + b_2$$

$$a_2 = \text{ReLU}(o)$$

- *Logistic Regression* Logistic regression is specified by the sigmoid function. k is the weight vector of size `feature_size`. b is a bias term.

$$\sigma(\mathbf{t}) = \frac{1}{1 + e^{-k\mathbf{t}+b}}$$

- *Gaussian Process* A Gaussian process is specified by its mean function $m(\mathbf{t})$ and its covariance function $K(\mathbf{t}, \mathbf{t}')$. Here we are using a constant mean and *squared exponential* (SE) covariance. σ_z is the signal variance and β is the length scale.

$$K(\mathbf{t}, \mathbf{t}') = \sigma_z^2 \exp(-\frac{1}{2\beta^2}\|\mathbf{t} - \mathbf{t}'\|_2^2)$$

- *XGBoost* We use a XGBoost Classifier with 100 estimators and a max depth of 5. Instead of having a closed form solution, XGBoost is an ensemble of decision trees.

XGBoost, Gaussian process and MLP Classifiers are models that have achieved good performance in related studies. Logistic regression is picked because it has a closed form solution and can be incoporated in optimization model.

3.3 Evaluation

3.3.1 Label Sampling

To test the capacity and robustness of our proposed model, we simulate patrolling and poaching activities in area of interest for each year. We assume that 1) the poaching and patrolling activities follow some distribution/rule with respect to the geological features 2) the distribution/rule does not change over the years, but the geological features do. Therefore, the same grid cell might have different poaching risk over the years, as the geological landscape over it or near it has changed. Our sampling process involves sampling two layers, the poaching layer and the patrolling layer. Both the poaching layer and patrolling layer are binary maps on the grid cells, The poaching layer indicates the occurrence of poaching activities, whereas the patrolling layer indicates if the ranger has visited the grid cell. A poaching activity on a grid cell can only be observed if the grid cell has been patrolled. We simulate the two layers based on domain knowledge and adopt different ways to test the capacity of our machine learning models. According to domain experts, the patrol team typically patrol less than 10% of the grid cells each year and they have found less than 200 elephant carcasses. Given the domain knowledge, the number of positive grids in the patrolling layer we sample is 400 and the number of positive grid cells in the poaching layer we sample is 20% of total grid cells. The layered data is shown in Fig 3.3

We use two ways to sample the patrolling layer. The first method is just sampling the grid cells uniformly at random, without looking at any conditions. The second method is a rule based sampling method. Based on domain knowledge, rangers would patrol in trucks along the roads or in boats along the rivers. To sample according to this knowledge, we first filter out grids that are beyond 1km of rivers or roads. Then we sample uniformly at random in the remaining grids.

We use three ways to sample the poaching layer. The first method is a ruled based sampling method, similar to the rule based sampling of the patrolling layer. First, we filter out the grid cells that do not satisfy one or more of the following conditions. 1) Within 1 km of rivers or roads or inside IFL 2) Tree density $> 30\%$ 3) At least 10 km away from logging sites. Those constraints capture the ideal setting for poaching to occur as suggested by domain experts. We then sample uniformly at random in the remaining grid cells.

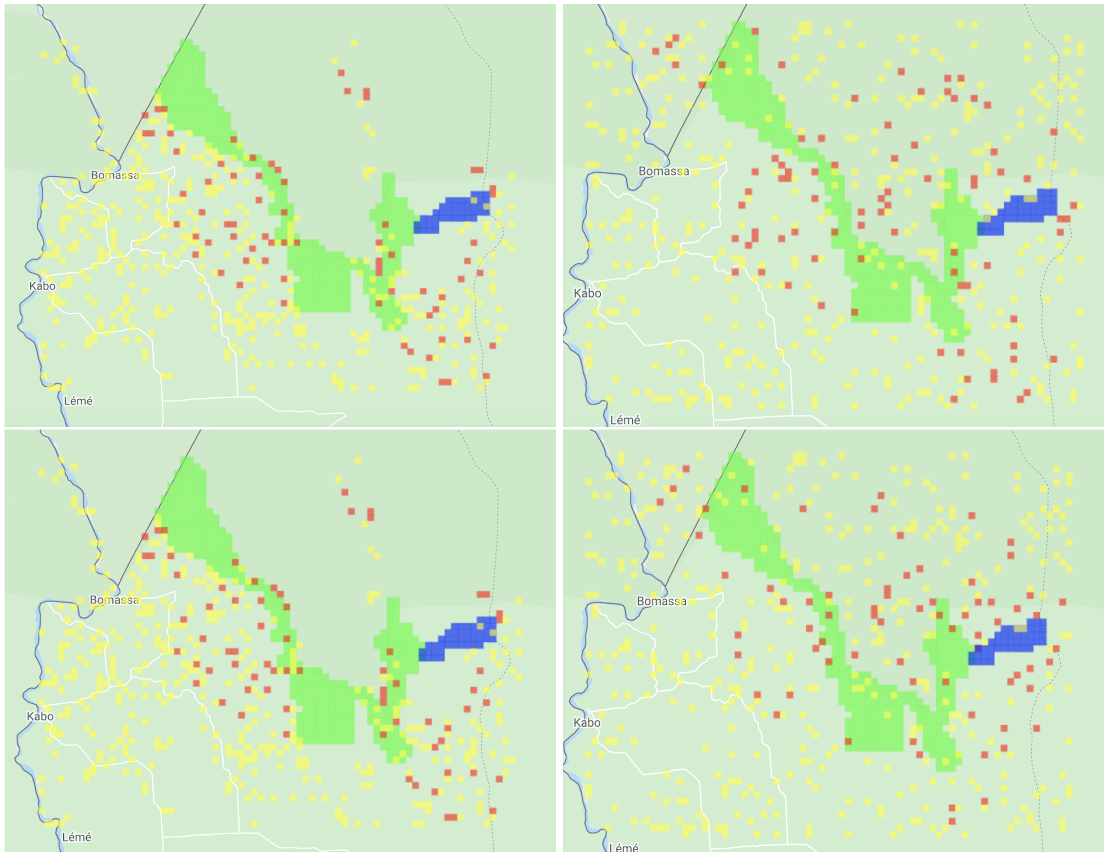

Figure 3.3: Green refers to intact forest landscape; blue refers to logging sites in year 2017; yellow refers to patrolled grids and red refer to poaching found in patrolled grids. The method used for simulating poaching and patrolling layers are: 1) top left: rule-based/sigmoid 2) top right: uniform/sigmoid 3) bottom left: rule-based/rule-based 4) bottom right: uniform/rule-based

The second method is by the sigmoid function with constructed weights.

$$\sigma(\mathbf{t}) = \frac{1}{1 + e^{-k\mathbf{t}+b}}$$

We split the features into two sets based on how the value of those features would impact poaching likelihood in a grid cell. The first set consists of *dist2roads, dist2rivers, dist2IFL, water_cover*. As the values of features in the first set increase, poaching is less likely to occur. The second set consists of *dist2village, dist2AAC, forest_cover*. As the values of features in the second set increase, poaching is more likely to occur. We assign negative weights to features in the first set and positive weights to features in the second set. The weights we apply are based on the scale. Features with larger values are assigned with a small random weights in absolute value, while features with smaller values are assigned with a large random weights in absolute value. We filter out the points whose function values are below 0.5, and then sample uniformly in the remaining grid cells.

In the last method, we direct sample from a multivariate Gaussian process. Gaussian process is a popular tool for nonparametric function estimation. Berger and Wang(2016) [18] propose a method to sample from shaped constrained Gaussian process by imposing rejection on its derivative process. This enables us to sample from a multivariate distribution which ensuring the monotonicity of the sampled data. Therefore allowing the sampled poaching risk to always decrease with respect to certain features such as the distance to roads. This would one of the future directions. For this part, we use a Gaussian process specified by the following mean and covariance function.

$$Z(\mathbf{t}) \backsim m(\mathbf{t}) = \mathbf{1}$$
$$K(\mathbf{t}, \mathbf{t}') = \sigma_z^2 \exp(-\frac{1}{2\beta^2}(\|\mathbf{t} - \mathbf{t}'\|_2^2)$$

where $\sigma_z = 1$ and $\beta = 1$. We filter out the points whose sampled values are below 0.5, and then sample uniformly in the remaining grids.

3.3.2 Results

Since there are two ways to sample the patrolling layer and three ways to sample the poaching layer, in total we have six ways to sample our data set. For each set of sampling methods, we sample two sets of data, a training set based on 2017 geographical information and a testing set based on 2018 geographical information. Our training set is constructed based on 2017 data and

Model	Sampling Method	Accuracy	Precision	Recall	F1	ROC-AUC
MLP	Rule-based	0.757-0.765	0.395-0.411	0.631-0.638	0.488-0.498	0.711-0.714
LR	**Rule-based**	**0.673-0.723**	**0.273-0.372**	**0.547-0.763**	**0.364-0.5**	**0.623-0.739**
GP	Rule-based	0.752-0.789	0.376-0.441	0.613-0.627	0.47-0.513	0.703-0.720
XGB	Rule-based	0.825-0.836	0.714-1.0	0.063-0.067	0.115-0.125	0.528-0.533
MLP	Sigmoid	0.750-0.848	0.285-0.300	0.289-0.029	0.288-0.056	0.514-0.582
LR	**Sigmoid**	**0.636-0.697**	**0.279-0.411**	**0.615-0.768**	**0.358-0.525**	**0.627-0.706**
GP	**Sigmoid**	**0.689-0.765**	**0.317-0.361**	**0.4-0.681**	**0.354-0.435**	**0.617-0.706**
XGB	Sigmoid	0.820-0.840	0.333-0.500	0.031-0.053	0.058-0.093	0.512-0.517
MLP	**GP**	**0.710-0.753**	**0.224-0.237**	**0.101-0.314**	**0.142-0.262**	**0.509-0.551**
LR	GP	0.546-0.587	-0.193-0.211	0.382-0.848	0.272-0.287	0.511-0.551
GP	GP	0.563-0.719	0.190-0.246	0.191-0.514	0.215-0.278	0.522-0.543
XGB	GP	0.800-0.831	0.417-0.555	0.056-0.071	0.102-0.122	0.522-0.526

Table 3.2: The poaching risk prediction performance of each model and each sampling method for poaching. The data used for training is from year 2017 and the data used for testing is from year 2018. The sampling method for patrolling is uniform.

our testing set is constructed based on 2018 data. The prediction task aims at predicting the future poaching risk from the historical data as our downstream optimization problem focuses on making decisions of future given historical data. Only the patrolled grid cells are used as data points in both training and testing set. A data point is assigned to a positive label if poaching has occurred in the grid and rangers have patrolled in the grid. A data point is assigned to a negative label if poaching has not occurred in the grid and rangers have patrolled in the grid. We run each sampling method for three times and record the best and worst performance on each metric. The results are shown in Tab 3.2 and Tab 3.3. In Tab 3.2, uniform sampling is used for patrolling layer. In Tab 3.3, rule based sampling is used for patrolling layer. Our results show that over

Model	Sampling Method	Accuracy	Precision	Recall	F1	ROC-AUC
MLP	Rule-based	0.613-0.674	0.455-0.5	0.664-0.919	0.571-0.608	0.672-0.692
LR	Rule-based	0.553-0.726	0.409-0.552	0.824-0.828	0.546- 0.663	0.622-0.752
GP	**Rule-based**	**0.742-0.743**	**0.576-0.589**	**0.706-0.791**	**0.642-0.667**	**0.733-0.755**
XGB	Rule-based	0.675-0.701	0.519-0.789	0.103-0.112	0.172-0.196	0.528-0.549
MLP	Sigmoid	0.652-0.716	0.262-0.342	0.418-0.759	0.322-0.471	0.596-0.691
LR	**Sigmoid**	**0.642-0.702**	**0.272-0.321**	**0.731-0.773**	**0.397-0.453**	**0.678-0.731**
GP	Sigmoid	0.642-0.700	0.270-0.277	0.537-0.727	0.365-0.393	0.634-0.676
XGB	Sigmoid	0.815-0.818	0.143-0.235	0.030-0.061	0.049-0.096	0.498-0.512
MLP	GP	0.709-0.757	0.2-0.204	0.105-0.288	0.138-0.239	0.505-0.538
LR	**GP**	**0.313-0.546**	**0.169-0.193**	**0.557-0.848**	**0.281-0.287**	**0.530-0.551**
GP	GP	0.502-0.812	0.151-0.2	0.061-0.368	0.093-0.215	0.451-0.507
XGB	GP	0.815-0.844	0.5-0.538	0.053-0.106	0.095-0.177	0.520-0.544

Table 3.3: The poaching risk prediction performance of each model and each sampling method for poaching. The data used for training is from year 2017 and the data used for testing is from year 2018. The sampling method for patrolling is rule-based.

all data sets, logistic regression and Gaussian process have more consistent performance across the data sets. Specifically, Gaussian process has achieved an average of 0.771 in accuracy and an average of 0.712 in ROC-AUC using rule based poaching sampling and uniform patrolling sampling. It has also achieved an average of 0.743 in accuracy and an average of 0.744 in ROC-AUC using rule based poaching sampling and rule based patrolling sampling. As for the two different patrolling sampling method, choosing patrolling sites uniformly at random can improve the performance of models, potentially due to the diversity of data collected. However, this approach might not be possible in real life scenario because of the difficulties and inefficiencies to collect data in certain areas in the forest. When labels are generated by Gaussian process, none of the models can produce a prediction significantly better than random guess. Since there is not a model that performs significantly better than others across different data sets, in real world scenario, we suggest testing all the models and selecting the one with the best prediction results.

In the next two sections, we test the proposed optimization problem on four synthetically labeled data sets, whose patrolling and poaching layers sampling method are : (uniform, rule-based), (uniform, sigmoid), (rule-based, rule-based), (rule-based, sigmoid).

Chapter 4

Suitability Based Forest Zoning

In this section, we use the traditional forest zoning method to find the optimal land use assignment of Kabo forests. As mentioned in the introduction section, the land use of a planning unit should be determined based on two factors–the suitability of this planning unit for a specific use and the compatibility of the uses of neighboring units.

4.1 Problem Formulation

The zoning problem is formulated as finding the assignment of land uses to a set of planning units that produces the maximum suitability and compatibility. The planning units are square grids of 1km by 1km. There are multiple land uses in Congo Basin, including logging, hunting, human residence, forest conservation, and intact forest landscape. Our studies focuses on the intact forest landscape area that borders the national park(forest) and logging sites. The objective of the study is to figure out which part of the land should be considered into the national park, which part of the land should be considered for logging and which part of the land should be reserved for logging in the future. Correspondingly, there are three types of land use: 1) park, i.e. forest conservation 2) logging 3) intact forest landscape (IFL), i.e. reserved for logging in the future.

To set up the problem, we let X be a matrix of size number of grids by the number of distinct types of land use. Each row of the matrix is a one hot vector, indicating the land use of that grid.

$$X_{gu} = \begin{cases} 1 & \text{if the grid } g \text{ is assigned to use } u \\ 0 & \text{if the grid } g \text{ is not assigned to use } u \end{cases}$$

4.1.1 Objective Function

The objective function consists of three parts. The first part expresses the total suitability of the grids for each use.

In specific, each grid has a suitability score associated with each land use. Details on calculating the suitability score is explained in the next section. Let G be the set of all planning units, U be the set of all land uses, S be the suitability matrix. The suitability term in the objective function is defined as $\sum_{g \in G} \sum_{u \in U} S_{gu} X_{gu}$.

An obvious drawback for only using the first term as our optimization object is that the best solution is just assigning each grid to the use that has the highest suitability score for that grid. This ignores the fact that adjacent grids should be assigned to more compatible uses and conflicting uses should be located far away from each other. Assigning forest conservation areas right next to logging sites is bad practice as the noise created by logging draws away elephants that reside in the forest. On the other hand, ifls are neutral to forest conservation and logging sites. We define a compatibility C for different use of land and our goal is to maximize $\sum_{g \in G, m \in G, g \neq m} \sum_{u,n \in U} \frac{C_{un}}{dist_{gm}} X_{gu} X_{mn}$, where X_{gu} and X_{mn} represents grid g and m being assigned to use u and n. $dist_{gm}$ represents the distance between grid g and grid m. The compatibility term is determined based on domain knowledge and it is commutative, i.e. $C_{un} = C_{nu}$.

In addition to the suitability and compatibility, another desirable feature of the land use assignment problem is the connectivity. Highly fragmented sites for logging is obviously not feasible to the logging company as it increases the cost of transportation. Therefore, fragmented assignments should be penalized in the objective function. On way to measure the connectivity of the assignment is to consider the number of adjacent grid pairs that are assigned to different use. Let $N(g)$ denote the set of neighbor grids of g. The penalty term can be written as $\sum_{g \in G, m \in N(g)} \sum_{u \in U} \mathbb{1}_{X_{gu} \neq X_{mu}}$, where $\mathbb{1}_{X_{gu} \neq X_{mu}}$ is an indicator random variable that equals to 1 if $X_{gu} \neq X_{mu}$ and 0 otherwise.

The final maximization objective is the combined sum of the three parts, with weighting constant α and β.

4.1.2 Constraints

In Verdiella et al.'s work, a preference matrix P of size number of land uses is multiplied to the suitability product as $S_{gu}PuX_{gu}$ [17]. However, in our case, instead of having a preference over the different land uses, the government provides the logging company a quota on amount to log each year. To preserve the forest and allow for sustainability, only one to three trees are being cut down per hectare of land in forest. This knowledge enables us to have a strict constraint on the economical value produced by setting a logging site. Let R_{min} and R_{max} be the bound for annual revenue gained from logging, t be the revenue gained for setting a planning unit to be a logging cite. The economical constraint is as follows:

$$R_{\text{min}} \leq t \sum_{g \in G} \mathbb{1}_{X_{g,\text{logging}}=1} \leq R_{\text{max}}$$

The area of interest is next to parks, intact forest landscapes and allowable cuts in the previous

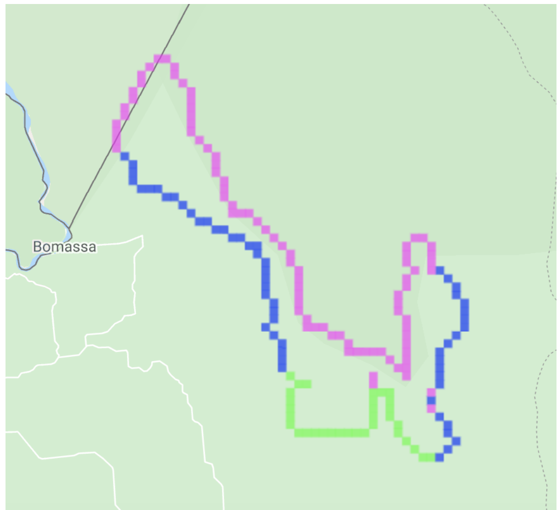

Figure 4.1: The surrounding grids of area of interest. Green refers to logging sites, purple refers to conservation and blue is reserved for future logging.

years. Fig 4.1 shows the use of bordering grids of our area of interest. The land use of those areas, though cannot be changed, does have an influence on compatibility of our land use assignment. A natural way to deal with those units is to include them in the set of planning units but assign a 0 suitability score, i.e. only counting their influence on compatibility and connectivity. Let E be the set of units directly bordering our area of interest.

$$X_{gu} = 1 \quad \text{for } g \in E \text{ and the corresponding } u$$

$$S_{gu} = 0 \quad \text{for } g \in E$$

Summing up, our final optimization problem is

$$
\begin{aligned}
\text{maximize} & \sum_{g \in G} \sum_{u \in U} S_{gu} X_{gu} \\
& + \alpha \sum_{g \in G, m \in G, g \neq m} \sum_{u, n \in U} \frac{C_{un}}{dist_{gm}} X_{gu} X_{mn} \\
& - \beta \sum_{g \in G, m \in N(g)} \sum_{u \in U} \mathbb{1}_{X_{gu} \neq X_{mu}} \\
\text{subject to } & X_{gu} = \{0, 1\} \forall g, u \\
& \sum_{u} X_{gu} = 1 \\
& t \sum_{g \in G} \mathbb{1}_{X_{g,\text{logging}}=1} \geq R_{\text{min}} \\
& t \sum_{g \in G} \mathbb{1}_{X_{g,\text{logging}}=1} \leq R_{\text{max}} \\
& X_{gu} = 1 \quad \text{for } g \in E \text{ and the corresponding } u \\
& S_{gu} = 0 \quad \text{for } g \in E
\end{aligned}
$$

4.2 Land Use Suitability Calculation

The suitability of each grid is associated with its historical land use and poaching risk. To compute the yearly suitability of individual grids in our region of interest for each of the three possible land uses (designated the land as AAC, IFL, or protected park area), we train three separate XGBoost models using a joint data set consisting of historical data and sampled data constructed via domain rules. This process is detailed below.

We first generate "ground truth" binary suitability labels for each 1km by 1km-sized grid in our region of interest using historical data - that is, to compute the suitability for some year X, we label grid g as a positive training point for use u if g was utilized for use u in a year preceding

X.

Since this mostly only generates positive data points, we next construct a data set based on rules inspired by domain knowledge for unlabelled data. Specifically, for each use u, if some grid g had never historically been used for u, we add it to a secondary constructed dataset and label it according to the rules below.

For the AAC use in a given year, a grid is:

- **suitable** if it has a high poaching risk and was part of IFL land 2 years ago.
- **unsuitable** if it has a high poaching risk and is part of AAC land that same year.

For the IFL use in a given year, a grid is:

- **suitable** if it has a low poaching risk.
- **unsuitable** if it has a high poaching risk and is part of IFL land that same year.

And for the park use in a given year, a grid is:

- **suitable** if it has a high poaching risk, as it would be better protected as a national park.
- **unsuitable** if it has a low poaching risk, as it may serve better as land dedicated to generating revenue.

Poaching risk is computed by the model with the best performance as shown in Fig 3.2 and 3.3 for each data set.

From this rule-constructed data set, we sample a number of points equal to the existing number of data points in the historical data set, creating an aggregated training data set for each land use. We finally upsample the training data set for each use such that there are equal numbers of positive and negative training points.

We use these to train the models to predict a suitability score between 0 and 1. The input fea-

Land Use	Accuracy	Precision	Recall	F1	ROC-AUC
AAC	0.955	0.957	0.953	0.955	0.955
IFL	0.831	0.845	0.81	0.827	0.831
Park	0.959	0.959	1.0	0.666	0.707

Table 4.1: The suitability score prediction performance of each individual land use model. The data set uses rule based sampling for patrolling and poaching layer.

tures to the model are geographic features we have deemed relevant: the swamp and forest cover density (measured in area) within each grid, the animal presence probability from our predictive model, and the distances to the nearest roads, rivers, and villages. We were advised by domain experts that the swamp and forest densities are particularly relevant because they both determine how appropriate a piece of land is for logging, and building roads to facilitate logging activity.

We first split the joint constructed data sets into a training set (70% of the entire data set) and test set (30%), training each land use suitability model on the former and assessing its accuracy, precision, recall, F1 score, and area under the curve (AUC) against the ground truth labels. The performance of each individual suitability model on the test set constructed for 2017 using rule based sampling for patrolling layer and poaching layer, is shown in Tab 4.1

4.3 Methodology

The optimization problem is an integer programming problem. Simulated annealing algorithm is used here as it is time efficient and can provide stable assignments close to optimal solution.[4] The algorithm itself is shown in Alg 1.

Algorithm 1: Simulated Annealing Algorithm

Randomly sample land uses to grids while constraints hold true. Let the assignment be x.;

Calculate the optimization objective $f(x)$;

Fix temperature t;

for *T iterations* **do**

  Randomly sample a grid i in x, denote the use of grid i as u;

  Randomly sample a new use u' from $U \setminus \{u\}$ while the constraints hold true. Set the new use of grid i to be u'. Let the new assignment be x' ;

  Calculate the difference in objective function f. $\delta = f(x') - f(x)$;

  if $\delta > 0$ **then**

    $x = x'$;

  else

    With probability $p = \exp(\delta/t)$, $x = x'$;

  end

end

4.4 Experiments and Evaluations

We solve the optimization problem and produce the optimal land use assignments of year 2018 on four different data sets. We pick $T = 50000$ and $t = 1$ after careful tuning. The suitability scores incorporate poaching risk and therefore is used to measure the performance of our model. From Tab 4.2, we can see that the total sum of suitability of the assignment produced by our model exceeds the actual assignment. The assignment result of year 2017 is shown in Fig 4.2. Even though our proposed assignment has higher overall suitability. One major problem is that since we are using the objective function to control connectivity rather than a constraint. The optimal solution does not yield a feasible plan in real life scenarios. The logging sites selected for each data set are very scattered and could induce a lot of transportation cost. Inspired by this, we propose our next section of work, in which we consider connectivity as a hard constraint.

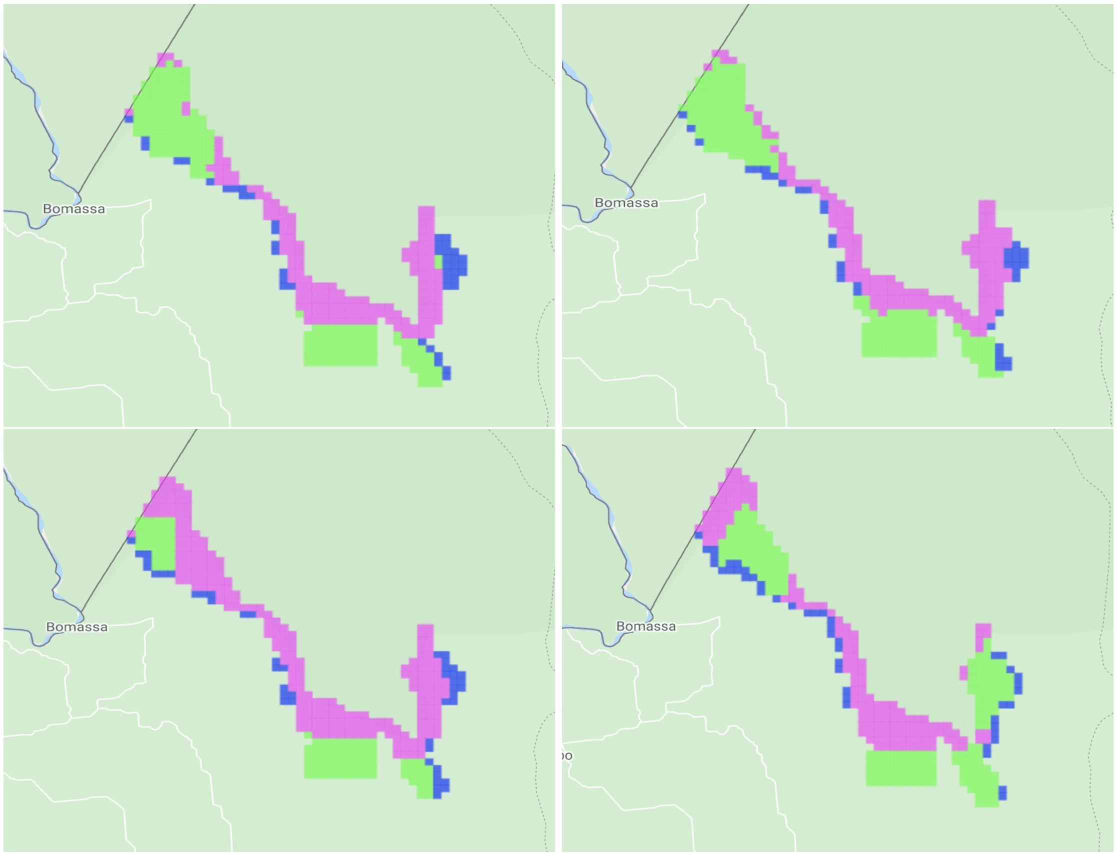

Figure 4.2: Green refers to logging sites, purple refers to conservation and blue is reserved for future logging. The sampling combination(poaching, patrolling) for each : 1) top left: sigmoid, rule-based. 2) top right: sigmoid, uniform. 3) rule-based, rule-based. 4) rule-based, uniform

	original assignment	calculated assignment
Rule Based Poaching Sampling Uniform Patrolling Sampling	154.86	196.17
Sigmoid Poaching Sampling Uniform Patrolling Sampling	180.59	178.48
Rule Based Poaching Sampling Rule Based Patrolling Sampling	160.78	168.57
Sigmoid Poaching Sampling Rule Based Patrolling Sampling	180.97	178.13

Table 4.2: The total suitability score of the result assignment of our model and actual assignment.

Chapter 5

Minimizing Poaching Risk through Land-Use Assignment and Patrol Route Design

In previous section, we have proposed a fine grained method for determining the land-use assignment. However, the method has two major drawbacks. First, the labels used for the suitability scores are determined by domain knowledge. Therefore, their validity is hard to verify. Second, we do not consider the patrolling efforts of the rangers, which, by our previous analysis, could help reduce the poaching risk in a region. Therefore, in this section, we consider solving the task of land use planning and patrol routes designing together.

5.1 Problem Formulation

We formulate the problem of patrol route selection as an optimization problem. We first introduce some assumptions and definitions. Let S be the set of grids we're interested in and i be the index of each grid. Let z_i be a vector of environmental covariates in grid i. We have two sets of decision variables–d's for land-use assignment decision and x's for patrolling routes decision.

The government's objective is to gain as much revenue as possible while minimizing the potential poaching risk. On the other hand, the ranger's objective is to pick a subset of the grids to regularly patrol upon given that their human resource is limited, i.e. the number of grids they can regularly patrol on is bounded. Therefore, x_i is the indicator random variable of whether the patrol team frequently patrols over grid i, and d_i is the indicator random variable of whether the government has leased this grid to the logging company.

p_{poach} represents the poaching risk. The value of p_{poach} is dependent on two factors. 1) Environmental covariates 2) The logging sites. Note here our $p_{\text{poach}}(z_i, d_i)$ is different from the one in previous studies. We can consider p_{poach} is a black box function. In our optimization setting, we use the sigmoid function with parameters trained from fitting logistic regression in the training data to simulate p_{poach}. The reasons are as follows: 1) sigmoid function has a simple closed form solution, making it easier to be used in optimization problem formulation. 2) As shown in Tab 3.2 and Tab 3.3, logistic regression has a stable performance as compared to other methods for different sampling method.

5.1.1 Logging Sites Selection

Our optimization model is a two step process. First, we determine the land-use assignment d_i's by solving the following optimization problem.

$$\max_{d} \sum_{i \in S} R_i d_i - \alpha \sum_{i \in S} p_{\text{poach}}(z_i, d_{N(i)})$$
$$\text{s.t. } C(d_i) \leq 0$$

Here R_i represents the revenue gained by logging in grid i. α is a constant parameter, adjusting the ratio between logging revenue and poaching risk incurred. The optimization objective states that the government's goal is the increase the revenue while reducing total poaching risk across the grids. We choose the parameters based on domain knowledge and historical data. The areas that can be considered as candidate logging sites are grids that have not been logged in the past thirty years. In addition, they have to be in the intact forest landscape area according to convention. Furthermore, only 1 tree would be cut down in 1 hectare of area in the forest in order to preserve the wood density. Therefore, we can assume the logging revenue R is the same across all the candidate grids. We set α to be 2.

$C(d_i)$ denotes the constraints on the grids used for logging. To reduce the transportation cost for the logging company, the grids selected should be connected and clustered. We use the notion of minimal node cut set to enforce the connectivity constraint, proposed by [7]. Let the grid have node set V and edge set E,

$$\Gamma(u, v) = \{S \subseteq V \setminus \{u, v\} : S \text{ is a minimal } uv\text{-node cut}\}$$

To impose that the selected grids for logging are connected, the following inequalities suffice:

$$\sum_{S} d \geq d_u + d_v - 1 \quad \forall S \in \Gamma(u, v), \forall u, v \in V\{u, v\} \notin E$$

It is possible to further extend the shape of logging sites, in terms of size of each connected components, etc. The inequalities needed are explained in [7].

Secondly, the number of grids used for logging should be bounded. Let M_{min} and M_{max} be the bound for annual revenue gained from logging, T be the revenue gained for setting a planning unit to be a logging site. The economical constraint is as follows:

$$M_{\text{min}} \leq \sum_{g \in G} R_i d_i \leq M_{\text{max}}$$

According to convention, we set M_{min} to be 10% of IFL grids and M_{max} to be 12% of total number of IFL grids.

Putting things together, our final optimization problem for deciding land use assignment is

$$
\begin{aligned}
\text{maximize} & \sum_{i \in S} R_i d_i - \alpha \sum_{i \in S} p_{\text{poach}}(z_i, d_{N(i)}) \\
\text{subject to } & \sum_{S} d \geq d_u + d_v - 1 \quad \forall S \in \Gamma(u,v), \forall u, v \in V\{u,v\} \notin E \\
& \sum_{g \in G} R_i d_i \geq M_{\text{min}} \\
& \sum_{g \in G} R_i d_i \leq M_{\text{max}}
\end{aligned}
$$

5.1.2 Patrol Routes Planning

Once we have determined the logging sites assignments, we use the following optimization objective to find the patrol decisions.

$$
\begin{aligned}
& \max_{x} \sum_{i \in S} p_{\text{poach}}(z_i, d_i) x_i \\
& \text{s.t. } c(x_i) \leq 0
\end{aligned}
$$

$c(x_i)$ denotes the constraints on the patrol routes. For our setting, it only has a constraint on the total number of grids being patrolled. Let H be a parameter on the human patrolling resource of the rangers.

$$\sum_{i} x_i \leq H$$

H is set to 10% of the total grids. The parameters of the sigmoid function are taken directly from our trained model of 2017 data.

Putting things together, our final optimization problem for determining patrol routes is

$$\begin{aligned} \text{maximize} & \sum_{i \in S} p_{\text{poach}}(z_i, d_{N(i)}) x_i \\ \text{subject to} & \sum_i x_i \leq H \end{aligned}$$

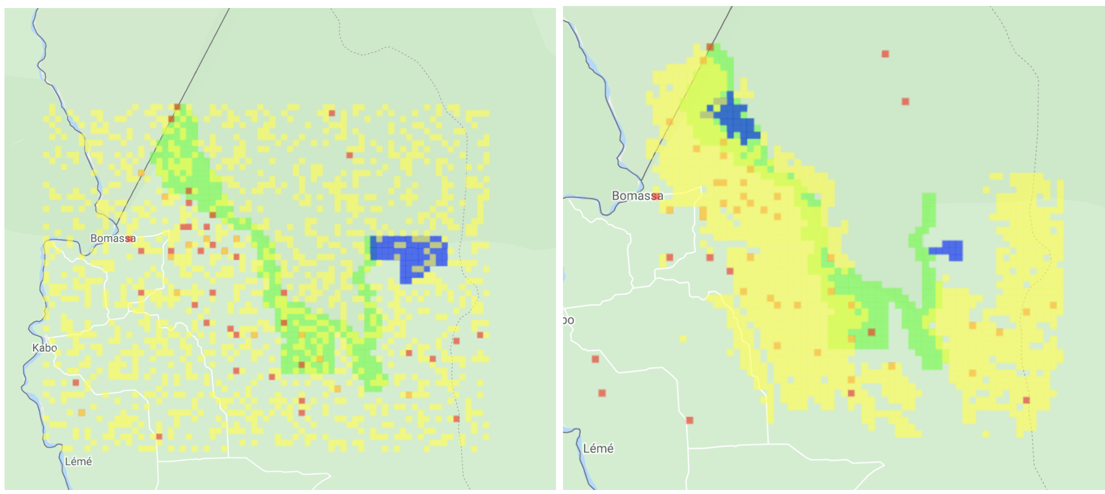

Figure 5.1: A comparison between simulated land-use and patrol decisions and optimized land-use and patrol decision results. Left: simulated data from Gaussian process. Right: Optimized decision. Green Area denotes the intact forest landscape; Blue Area denotes the logging sites; Red area denotes poaching activity; Yellow area denotes patrol routes.

5.2 Methodology

The logging sites selection problem can be considered as an integer linear programming problem. The pipeline works as follows. First, we train the logistic regression model on 2017 data, and then use the trained parameters of the sigmoid function in our optimization objective. We then solve the optimization problem with branch and bound algorithm implemented in Gurobi library to get the optimal logging sites for year 2018. Once the logging sites have been selected, the

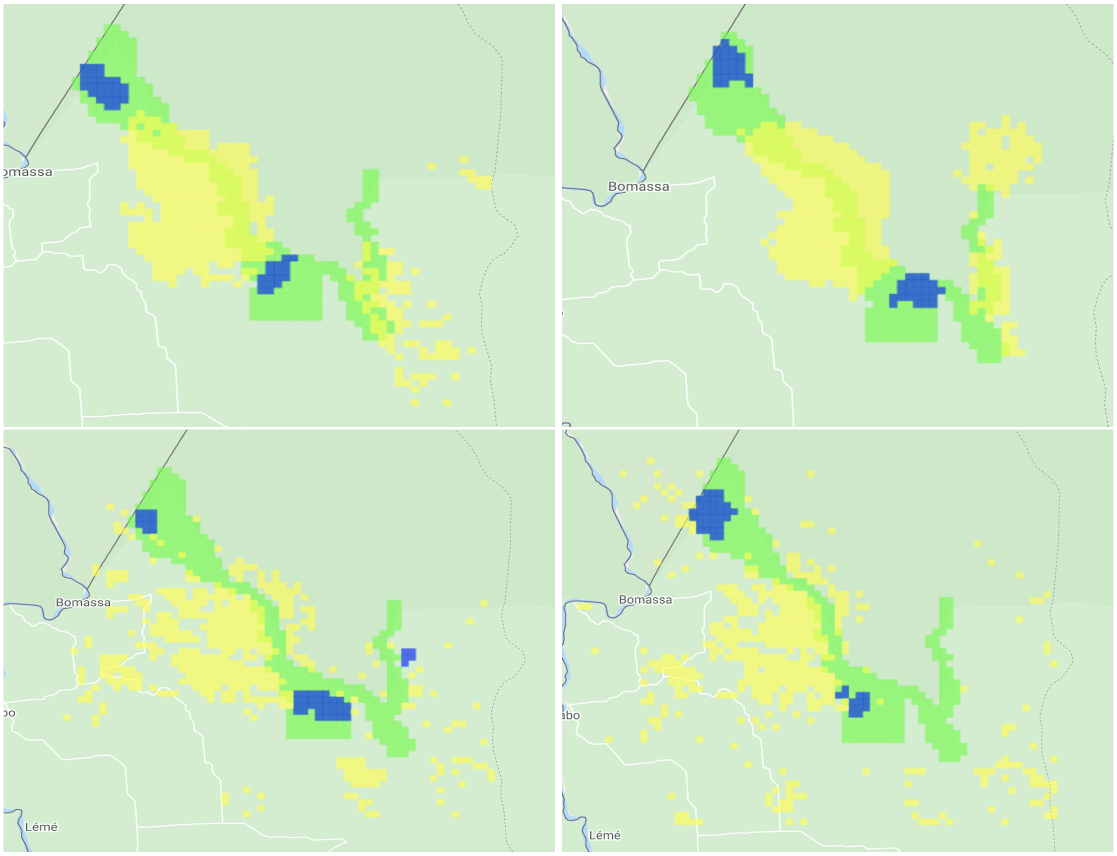

Figure 5.2: The decision result for each sampling combination of patrolling and poaching: 1) top left: rule-based/sigmoid 2) top right: uniform/sigmoid 3) rule-based/rule-based 4) rule-based/uniform

patrol routes planning part is simply selection the grids with top 10% p_{poach} value given the assignment.

5.3 Evaluation and Results

We run the optimization on each dataset for three times. Fig 5.2 shows the resulting assignments of the four data sets. Tab 5.1 and 5.3 shows the comparison of real world assignment to our assignment. The optimal logging sites calculated by our model induces much less poaching threat as compared to the actual logging sites. Specifically, for the data set with sigmoid poaching

Sampling Combination	Run #1	Run #2	Run #3
Rule Based Poaching Sampling Uniform Patrolling Sampling	0.442, 0.284	0.539, 0.445	0.497 ,0.289
Sigmoid Poaching Sampling Uniform Patrolling Sampling	0.575 ,0.471	0.681, 0.953	0.364, 0.556
Rule Based Poaching Sampling Rule Based Patrolling Sampling	0.373, 0.306	0.357, 0.168	0.322, 0.233
Sigmoid Poaching Sampling Rule Based Patrolling Sampling	0.492, 0.491	0.483, 0.480	0.536, 0.497

Table 5.1: The average poaching risk in IFL area. The first element of the tuple is the poaching risk of synthetic 2018 assignments. The second element of the tuple is the poaching risk of our proposed logging sites.

sampling and uniform patrolling sampling, our calculated logging sites can reduce the average poaching risk in the intact forest landscape area by 18%. (We are only looking at the IFL area because other areas are far away from the logging sites and the potential poaching risk would not be likely impacted by logging). Our calculated patrol routes also outperforms the sampled patrol routes for 2018. The routes our model proposes can cover grids with twice as much poaching risk on average as compared to the original sampled routes.

Sampling Combination	Run # 1	Run # 2	Run # 3
Rule Based Poaching Sampling Uniform Patrolling Sampling	0.417, 0.793	0.547, 0.939	0.148, 0.348
Sigmoid Poaching Sampling Uniform Patrolling Sampling	0.365, 0.812	0.351, 0.547	0.666, 0.927
Rule Based Poaching Sampling Rule Based Patrolling Sampling	0.478, 0.793	0.353, 0.722	0.527, 0.352
Sigmoid Poaching Sampling Rule Based Patrolling Sampling	0.448, 0.839	0.339, 0.625	0.302, 0.535

Table 5.2: The average poaching risk in each patrolling grid. The first element of the tuple is the poaching risk of synthetic 2018 patrol routes. The second element of the tuple is the poaching risk of our proposed routes

Chapter 6

Conclusion and Discussion

6.1 Conclusion

In the work, we propose a pipeline process for protecting forest elephants in the Kabo forest of Republic of Congo. In the first part of the work, we construct a data set of geological landscape features from publicly available data sets and landscape shape files. We test the capacity of several machine learning models on predicting poaching. Then, we propose two ways to make land use assignments. In the first suitability based land use assignment study, we use forest zoning to determine the best use for each part of the land, so that different areas can be assigned with different levels of conservation protection. In the second study, we use integer programming to select the best logging sites that induces the least poaching risk. Then, we select the best patrol routes based on the poaching risk of the logging sites. For both approaches, the comparison results have shown that the solution proposed by our model can provide better land use assignments.

6.2 Future Work

The major next step would be testing the machine learning models on predicting real world patrolling and poaching data in the Kabo area. We would also calculate and evaluate our proposed land use assignments solved with real world data.

Further more, we are looking forward to evaluating how patrolling can affect the poaching risk and using bi-level optimization to solve for the optimal land use and patrolling plans.

Bibliography

[1] I.R. Ball and H.P. Possingham. Marxan (v. 1.8.6): Marine reserve design using spatially explicit annealing. user manual. 2000. URL `http://www.uq.edu.au/marxan`. 2

[2] Stephen Blake, Samantha Strindberg, Patrick Boudjan, Calixte Makombo, Inogwabini Bila-Isia, Omari Ilambu, Falk Grossmann, Lambert Bene-Bene, Bruno de Semboli, Valentin Mbenzo, Dino S'hwa, Rosine Bayogo, Liz Williamson, Mike Fay, John Hart, and Fiona Maisels. Forest elephant crisis in the congo basin. *PLoS Biology*, 2007. 1

[3] Stephen Blake, Sharon L. Deem, Samantha Strindberg, Fiona Maisels, Ludovic Momont, Inogwabini-Bila Isia, Iain Douglas-Hamilton, William B. Karesh, and Michael D. Kock. Roadless wilderness area determines forest elephant movements in the congo basin. *PLoS One.*, 2008. 1

[4] Jan Bos. Zoning in forest management: a quadratic assignment problem solved by simulated annealing. *Journal of Environmental Management*, 1993. 4.3

[5] Thomas Breuer, Fiona Maisels, and Vicki Fishlock. The consequences of poaching and anthropogenic change for forest elephants. *Conserv Biol*, 30(5), 2016. doi: 10.1111/cobi.12679. 1

[6] Christopher J. Brown and Jodi Wetzstein. Timber suitability, long-term sustained yield capacity, and allowable sale quantity report. 08 2011. 2

[7] Rodolfo Carvajal, Miguel Constantino, Marcos Goycoolea, Juan Pablo Vielma, and Andres Weintraub. Imposing connectivity constraints in forest planning models. *Operations*

Research, 61(4), 2013. 5.1.1

[8] Uros Djuric, Miloš Marjanović, Vladimir Susic, Rastko Petrović, Biljana Abolmasov, Snezana Zecevic, and Irena Basaric. Land-use suitability analysis of belgrade city suburbs using machine learning algorithm. 01 2013. 2

[9] Jiang H. Eastman, J.R. and J.. Toledano. Multi-criteria and multi-objective decision making for land allocation using gis. 1998. 2

[10] Fei Fang, Thanh Nguyen, Rob Pickles, Wai Lam, Gopalasamy Clements, Bo An, Amandeep Singh, Milind Tambe, and Andrew Lemieux. Deploying paws: Field optimization of the protection assistant for wildlife security, 2016. URL `https://www.aaai.org/ocs/index.php/IAAI/IAAI16/paper/view/11814`. 2

[11] Shahrzad Gholami, Benjamin Ford, Fei Fang, Andrew Plumptre, Milind Tambe, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Mustapha Nsubaga, and Joshua Mabonga. Taking it for a test drive: A hybrid spatio-temporal model for wildlife poaching prediction evaluated through a controlled field test. *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2017. 1

[12] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend. High-resolution global maps of 21st-century forest cover change. *Science*, 342(6160):850–853, 2013. ISSN 0036-8075. doi: 10.1126/science.1244693. URL `https://science.sciencemag.org/content/342/6160/850`. 3.1.2

[13] Fiona Maisels and et al. Devastating decline of forest elephants in central africa. *PloS one*, 8, 2013. 1

[14] M. Mokarram, Saeid Hamzeh, F. Aminzadeh, and Abdol Zarei. Using machine learning for land suitability classification. *West African Journal of Applied Ecology*, 23:63–73, 09

2015. 2

[15] Parinaz Rashidi, Tiejun Wang, Andrew Skidmore, Anton Vrieling, Roshanak Darvishzadeh, Bert Toxopeus, Shadrack Ngene, and Patrick Omondi. Spatial and spatiotemporal clustering methods for detecting elephant poaching hotspots. *Ecological Modelling*, 297:180 – 186, 2015. ISSN 0304-3800. doi: https://doi.org/10.1016/j.ecolmodel.2014.11.017. 1

[16] Michael J. Shaffer and Joseph A. Bishop. Predicting and preventing elephant poaching incidents through statistical analysis, gis-based risk analysis, and aerial surveillance flight path modeling. *Tropical Conservation Science*, 9:525 – 548, 2016. 2

[17] Sabatini M. Maciela M. C. Verdiella, A. and R.M. Rodriguez Iglesiasa. A mathematical model for zoning of protected natural areas. *International Transactions In Operational Research*, 2005. 2, 4.1.2

[18] Xiaojing Wang and James O. Berger. Estimating shape constrained functions using gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 9:1 – 25. 3.3.1

[19] M. E. Watts. Marxan with zones: Software for optimal conservation based land- and sea-use zoning. *Environmental Modelling & Software*, 2009. 2

[20] Hamed Alemohammad Yoni Nachmany. Accurate natural trail detection using a combination of a deep neural network and dynamic programming. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 83 – 89, 2019. 3.1.2

[21] L. Zhang, J. Li, C. Kong, L. Qu, J. Zhu, Z. Chen, and Y. Luo. Using neural network to evaluate construction land use suitability. In *2010 Second International Workshop on Education Technology and Computer Science*, volume 2, pages 331–334, 2010. 2

[22] Átilla Colombo Ferreguetti, Juliane Pereira-Ribeiro, Jayme Augusto Prevedello, Walfrido Moraes Tomás, Carlos Frederico Duarte Rocha, and Helena Godoy Bergallo. One

step ahead to predict potential poaching hotspots: Modeling occupancy and detectability of poachers in a neotropical rainforest. *Biological Conservation*, 227:133 – 140, 2018. ISSN 0006-3207. doi: https://doi.org/10.1016/j.biocon.2018.09.009. 2