Modeling Wildfire Ignition Distribution and Making Prediction of Human-caused Wildfire

Weichen Ouyang\(^1\), Chia Yung Han\(^1\), and Susanna T.Y. Tong\(^2\)

\(^1\)Department of Electrical Engineering and Computing Systems, University of Cincinnati, Cincinnati, Ohio, U.S.A
\(^2\)Department of Geography, University of Cincinnati, Cincinnati, Ohio, U.S.A

ouyangwn@mail.uc.edu, han@ucmail.uc.edu, tongty@ucmail.uc.edu

Abstract – This paper proposes a further exploration of machine learning algorithms within the context of modelling the spatial distribution patterns of the human-caused wildfires over a Southern California landscape. In this research, the wildfire distribution problem is defined as a Binary Classification task conducted on a cellular lattice overlay the study area. A fifteen-year historical wildfire occurrence data as target variable was used, along with eight independent variables derived from anthropogenic factors such as distance measure to road-network and Wildland-Urban Interface. Meteorological factors such as temperature and humidity have also been used in the model training process. Both of the two machine learning algorithms, the Conditional Inference Tree and the Random Forest methods, combining with the Synthetic Minority Over-Sampling Technique, demonstrate a significant improvement over traditional method. And the predicted result shows that the location with high proximity with WUI and road tend to be more vulnerable towards wildfire incidence.

Keywords: Machine learning, Wildfire, Random Forest, Ignition modeling, SMOTE

1 Introduction

Wildfire occurrence and fire intensity have been increasing in many states in western USA, especially in California, where is suffering from the record-breaking long drought climate. According to the National Interagency Fire Center (NIFC), in 2015, over 10 million acres of land have been burned in the U.S., making the year the worst in the wildfire history [1].

The deteriorating situation has brought much concern about wildfires and called for a better understanding of the cause and distribution of wildfire activity. Predictive models play a vital role in designing preventive strategies in wildfire management, whereas pure after-incidence suppression is costly and resource-limited. Therefore, comprehending driving forces of fire ignition and predicting where fires are likely to happen are becoming the research hotspot among scientific community.

Human-wildland interaction has been recognized as the dominant source of wildfire ignition over natural causes (lightning). During last decade, about 90% of wildfires in U.S. are caused by human, and for instance, in 2015, there were 58916 human-caused wildfires comparing the 9235 lightning-caused fires [1]. As the wildland–urban interface (WUI), the community built at the fringe of human developments and undeveloped vegetation expands, the proportion and the number of human-induced fires continue to rise [2]. Because the interactions between human and wildland tend to be spatially concentrated around WUI, it is important to understand how WUI and other human constructions around WUI, such as road-network, affect the wildfire frequency in a given region. It is natural then to consider using the distance measure between fire occurrence and both WUI and road-network to quantify the human influence on the frequency and spatial pattern of fire, as well as where fire risk is highest on a landscape. Several research work on the corresponding ignition-distribution models quantifying the relationship between anthropogenic factors and fire occurrence have been reported [3] [4].

Conceptually speaking, the basic approach of modelling ignition-distribution is to analyze wildfire ignition locations in relation to environmental variables that are assumed to influence the spatial distribution of ignitions. The model will estimate the response of wildfire ignitions to predictors (features) derived from these environmental variables. In the design of a model, the occurrence of an ignition event at a point in space will be encoded as a binary response (0 and 1), accompanied with absence data (points in space with minimal likelihood for fire occurrence). Then statistical or machine learning algorithms can be used to generate a classifier to classify every space unit in the study area to be either class 0 or 1, corresponding to low or high probability for a wildfire to happen at that corresponding point. However, because it is generally impossible to identify places where no ignition can occur, the definition of true absent points is not feasible. Thus, in most cases, when the fire presence locations (positive cases) are compared against negative cases for a given set of background environmental conditions [5], a severe “imbalanced dataset” problem will most certainly occur, wherein the number of data points belonging to one class, the majority class, is much more than the other, the minority class.

Over the past decade, several methods for building wildfire occurrence-distribution prediction systems or wildfire
risk assessment systems have been traditionally developed using the Binary Logistic Regression method, based on different variables, and scales. Some of which are widely accepted by the wildfire management authorities [3] [4] [6] [7]. Meanwhile, because machine learning algorithms have shown a high predictive accuracy and adaptability in Data Mining, AI and various other fields, some researcher also started to use those algorithms to model spatial distribution of wildfire occurrence or ignition. Amatulli [8] used Decision Trees to assess and generate a risk map of wildfire of Gargano Peninsula on Italy’s East Coast; Artificial Neural Networks (ANN) was used to classify the space units of central Portugal into 5 class of risk ranks [9]; and more recently, an ensemble machine learning algorithm, Random Forest, has been used to model the fire distribution in a national forest in Michigan [10].

In comparison with the traditional method – Logistic Regression, the machine learning algorithms have their advantages. ANN has its merits in capturing non-linear relationship between the feature set and the response variable; Decision Trees, specifically the Conditional Inference Trees, are more adaptive to imbalance dataset and are more resistant to the overfitting problem than Logistic Regression; Random Forest, as an ensemble learning algorithm, usually outperforms other classification algorithms in terms of its predictive accuracy. The advantages of utilizing machine learning algorithms in modeling spatial distribution of wildfire over Logistic Regression (LR) will be further explored in this research. Two algorithms, namely, Random Forest (RF) and Conditional Inference Tree (CIT) are chosen in the comparison study. All three will be applied to building the predictive model of wildfire occurrence in the Santa Monica National Recreational Area, which is located at the West Shore of Los Angeles.

To be specific, the main objectives of this research are (1) identify and collect a combination of environmental and anthropogenic explanatory variables and then apply machine learning algorithms to produce predictive models for predicting patterns of the spatial distribution of human-caused wildfire over the landscape of a national forest – Santa Monica Mountains National Recreation Area; (2) build the model using the two widely-accepted classification algorithms, the Conditional Inference Tree and Random Forest, and investigate the results and compare them with the result from the model built with the Logistic Regression; and 3) explore different oversampling and under-sampling methods to find a usable algorithm to solve the “imbalance dataset” problem that exists in the related previous works.

2 Method

2.1 Study area

This research is conducted based on data collected for the region in the Santa Monica National Recreation Area plus an adjacent town, Thousand Oaks (Figure 1). This mountainous area resides on the west shore of Los Angeles, on the shore line of Pacific, across the county of Ventura and L.A. The study area includes a substantial amount of WUIs, and has experienced a series of wildfire occurrences over last decade. The whole area covers about 1 billion square meters, to facilitate the model building process, a fishnet (cellular lattice or grid) is created, consisting of approximately 1 million spatial cells of size of 30m x 30m (meter) each.

Figure 1. The Study Area – Santa Monica Mountain, CA

2.2 Problem definition

To model the spatial distribution of wildfire over this study area, a Binary Classification Problem is defined over the 1 million cells that constitute the study area. Each cell will be measured by eight features, five anthropogenic and three weather measurements (detailed in later sections), and each cell will eventually be classified by the classifier, which is generated by one of the three classifier construction algorithms (Logistic Regression, CIT and RF), to be one of the two classes: 1, representing high probability of fire occurrence and 0, the opposite. This classification result is a form of long-term risk assessment, that is, the evaluation of the proneness to wildfire incident over a period of time, based on the characteristics of climate and impact of anthropogenic factors in the near neighborhood of that specific cell.

2.3 Response variable

2.3.1 Response variable – fire ignitions

The Santa Monica ignition dataset, acquired from the USGS historical wildfire record from 2000 to 2014, contains 210 geo-coordinated points [11]. These 210 points are resampled using the given grid. By doing a careful check, there are no two points residing within a common cell. To prepare a dataset for constructing classifier, 790 random points are generated using ArcGIS at a minimal 300 meters away from each other. This is to guarantee that the 790 points are scattered across the whole region, to capture the variance in the 8 features in the study area. These random points will serve as the negative (0) cases, and together with the 210 positive cases, a 1000-data-points dataset is formed. Figure 2 shows the ignition points on this map. Also shown are the road network
and the WUI polygons whose borders are the human nature interface.

2.4 Features

2.4.1 Independent variables – anthropogenic

Human-caused ignitions were concentrated close to roads, in high road density areas, and near the WUI interface [12]. As shown in Figure 2, the Ignition data in Santa Monica Mountains also demonstrate a trend that they are to be found near human developments and communities. In this research, three types of human construction artefacts are recognized: the road network, human development (man-made construction units), communities (in terms of WUI polygon). The road network data is acquired by downloading TIGER/Line files updated for “US census 2010” from the US Census Bureau. Human development data is derived from the land coverage data of the study area, provided by the National Park Service (NPS). The land coverage data classifies every 30-meter by 30-meter cell to be a kind of the coverage types, four of which are human development areas. The cells are extracted from the file and formed the map layer of human development. WUI data is provided by the Silvis Project from University of Wisconsin-Madison [2].

Accordingly, five Independent variables are derived from the above three: distance to road, distance to WUI, distance to human development, level of development (derived by constructing a kernel density estimate of human development over 1000 meters), population density (derived by constructing a kernel density estimate of population density measure of each WUI over 1000 meters). The 1000-meter threshold is chosen to represent the max range people usually get into the remote wildland. Detail description are in Table 1.

2.4.2 Independent variables – climate

Weather plays a vital role in fire ignition process, low temperature, heavy precipitation and high humidity will significantly diminish the chance of fire occurrence. Thus, to capture impact to ignition distribution from the biophysical characteristics of the study region of chosen, 30 weather stations scattered within and surrounding the area had been chosen and 3 variables had been collected and aggregated over the 15 years span to reflect the local climate (2000 to 2014): mean January temperature, mean July temperature and mean annual humidity. The 30-point data are then used to generate, using a 5-point interpolation, a continuous surface with all 3 variables being populated across the whole landscape. Details of the variables used in this study are listed in Table 1.

2.4.3 Data manipulation

Since all the cells in our grid system is of the size 30 meter by 30 meter, all of the above 8 independent variables are resampled to this 30-m resolution. Consequently, all the one million cells would now possess a feature vector of dimension eight, and then the 790 geo-coordinated negative points can be used to extract the corresponding feature vector from the cells.

Figure 2. Ignition points (year 2000-2014), road network and WUI
Table 1. List of variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Resolution</th>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variable</td>
<td>point</td>
<td>USGS</td>
<td>N=210, from 2000 to 2014</td>
</tr>
<tr>
<td>Ignition points point</td>
<td>30m</td>
<td>US Census TIGER/Line files</td>
<td>Nearest Perpendicular distance</td>
</tr>
<tr>
<td>Independent variables</td>
<td>30m</td>
<td>Silvis Project mapping data</td>
<td>Nearest Perpendicular distance</td>
</tr>
<tr>
<td>Distance to Development</td>
<td>30m</td>
<td>NPS data store</td>
<td>Nearest Perpendicular distance</td>
</tr>
<tr>
<td>Development level</td>
<td>1000-m</td>
<td>NPS data store</td>
<td>Kernel Density Estimate</td>
</tr>
<tr>
<td>WUI population Density</td>
<td>1000-m</td>
<td>Silvis Project WUI mapping data</td>
<td>Kernel Density Estimate</td>
</tr>
<tr>
<td>Mean January Temperature</td>
<td>30m</td>
<td>Mesowest weather system</td>
<td>5-neighbor interpolation</td>
</tr>
<tr>
<td>Mean July Temperature</td>
<td>30m</td>
<td>Mesowest weather system</td>
<td>5-neighbor interpolation</td>
</tr>
<tr>
<td>Mean Relative Humidity</td>
<td>30m</td>
<td>Mesowest weather system</td>
<td>5-neighbor interpolation</td>
</tr>
</tbody>
</table>

2.5 Create a balance dataset

The dataset thus generated is a typical “imbalance dataset” since the positive: negative rate is 1:4, which means that it would greatly decrease the overall sensitivity of the classifier. Because we are aiming to build a classifier that identify the vulnerable spot that is most likely encounter wildfire in this study, sensitivity is an essential performance measurement in the model evaluation process. Under the 1 to 4 ratio of positive to negative cases, the model will be most likely tuned towards recognizing every case to be negative. This is common in the modeling spatial distribution of wildfire problem. In [7], Syphard also used a dataset with 1:5.5 positive to negative rate, yielding an Area Under Curve (AUC) of 0.71 in the Receiver Operating Characteristic (ROC) test. However, in this study, the original imbalance dataset yields a 13% sensitivity classifier using CIT algorithm, basically meaning the classifier can only recognize one case in every ten positive cases, making the generated classifier useless.

To overcome this, certain over-sampling and under-sampling technique is needed to create a new “balanced” dataset as the input to our training procedure. The Synthetic Minority Over-Sampling Technique (SMOTE) is used to over-sampling the minority class [13]. The synthetic samples are created in a less application-specific manner, by operating in “feature space” rather than “data space”. The minority class is over-sampled by taking each minority class sample and introducing synthetic samples along the line segments joining any or all of the k minority class nearest neighbors. That is, the samples are generated as follows. First, take the difference between the feature vector (sample) under consideration and its nearest neighbor. Multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration. This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. For our dataset, SMOTE is executed on the original 1000-data-points dataset, a new balanced dataset which contains 2730 data points, 1470 positive and 1260 negative. And the positive to negative ratio is altered to be 1.167: 1.

2.6 Model building

Before entering the model building phase, an exploratory analysis is conducted to eliminate feature variables that are highly collinearly related to other features. Features such as “development level” and “mean July temperature” were discarded. Thus, the feature dimension had been reduced to be 6. And the dataset containing 2730 data points, will be divided into training set and test set based on a ratio of 7:3.

2.6.1 Logistic regression

Logistic Regression models are statistical models which provide insights into the relationship between a qualitative dependent variable, dichotomous in the present case, and one or more independent explanatory variables, whether qualitative or quantitative. The mathematical expression of LR models is:

\[ y_i = \frac{e^{(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}}{1 + e^{(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})}} \]  \hspace{1cm} (1)

In this work, the Logistic Regression model is developed using a forward stepwise procedure in which the independent variables are introduced into the model one by one. According to the result of evaluating improvement in the model by adding the variable, as measured by the Akaike Information Criterion (AIC), a decision is made to keep or discard that variable.

2.6.2 Conditional Inference Tree

Both traditional classification and regression tree (CART) algorithm and the CIT method recursively perform univariate splits of the dependent variable based on values of a set of covariates. CART and related algorithms usually employ information measures for selecting the current covariate. It uses GINI coefficient to evaluate the likelihood of a random sample chosen at this node being misclassified in the child-nodes.

According to [14], CIT avoided the following variable selection bias of the CART methods: They tend to select variables that have many possible splits or many missing
values. Unlike the others, CIT uses a significance test procedure in order to select variables, instead of selecting the variable that maximizes an information measure (e.g., Gini coefficient).

In this study, CIT is parameterized to use 5% as the maximum allowed p-value for permutation test conduct at each splitting node, a minimum node size of 20 for inner tree nodes, a minimum node size of 7 for leave nodes, and no restrictions on the tree depth.

2.6.3 Random Forest

Random Forest is an ensemble Machine Learning technique which works in a parallel fashion similar to other meta-algorithms like bootstrapping [15]. The classification result in RF is generated by averaging the predictions of many individual decision trees, each is constructed based on a subset of the training data. This algorithm is proven to be more stable and accurate than any single decision tree algorithm.

RF can be parameterized according to the number of trees averaged in the forest (ntrees), the number of predictor variables randomly selected at each iteration (mtry), and the minimum number of observations at end nodes (nodesize), which can decrease the length of the tree branches and thus simplifying the structure of the trees. All combinations of five ntrees levels (500, 1000, 1500 and 2000) and three mtry levels (from 1 to 3) were tested. The nodesize parameter was left at its default value. The values of the parameters in the final model were mtry = 2 and ntrees = 500. Models with higher values of these parameters did not improve accuracy significantly.

3 Result

3.1 Model evaluation and comparison

To compare the prediction accuracy among the three models, the area under the curve (AUC) of the receiver operating characteristic (ROC) chart is used along with three other common measures: accuracy, sensitivity, and specificity. Sensitivity refers to the proportion of ignitions that is correctly predicted, and specificity the proportion of non-ignitions correctly predicted. AUC ranges from 0.5 to 1, in general, an AUC value above 0.8 indicates a good model performance.

In summary, unlike previous works, there are two datasets used in this study, a synthetic one generated by SMOTE with 2730-data-points, and the original sample of 1000 points, the latter is a subset of the former. The performance of models on both datasets will be evaluated.

3.1.1 Performance over synthetic dataset

Both the CIT and the traditional method LR show a good classification accuracy over the synthetic dataset, however, the RF yields a result of approximately 95% of correct overall prediction in the test set, as shown in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>CIT</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.718</td>
<td>0.755</td>
<td>0.95</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.819</td>
<td>0.754</td>
<td>0.962</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.577</td>
<td>0.755</td>
<td>0.933</td>
</tr>
<tr>
<td>AUC</td>
<td>0.756</td>
<td>0.804</td>
<td>0.983</td>
</tr>
</tbody>
</table>

In terms of sensitivity, LR outperforms CIT by about 6%, which may be due to the synthetic data points’ strength in the specificity of the decision plane causing the increased accuracy of linear relation. This phenomenon will increase the LR classification success rate and will decrease CIT’s rate. At the other hand, RF, as an ensemble algorithm will keep the stability by average out the variance caused by high specificity issue caused by synthetic points. Thus, Random Forest has the best performance in every way for synthetic dataset.

3.1.2 Performance over original dataset

The correctness of classification has much more significance in the original dataset, since the result will be directly used to generate the predicting map for preventive strategy design and research.

<table>
<thead>
<tr>
<th></th>
<th>LR</th>
<th>CIT</th>
<th>RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.576</td>
<td>0.753</td>
<td>0.936</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.800</td>
<td>0.719</td>
<td>0.971</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.516</td>
<td>0.762</td>
<td>0.927</td>
</tr>
<tr>
<td>AUC</td>
<td>0.717</td>
<td>0.802</td>
<td>0.988</td>
</tr>
</tbody>
</table>

As shown in Table 3, all three algorithms suffer some loss in the overall accuracy, while LR suffers the most, which may be due to the imbalance nature of the original dataset reflecting in terms of low linear separability. But RF shows its robustness against imbalance issue. Hence it is the best classification algorithm for use.

Furthermore, both machine learning algorithms show the greater predicting power in terms of AUC, as shown in Figures 3, 4, and 5. The AUC value of 0.756 is obtained for the LR method, while CIT’s AUC values is about 0.8. And the AUC value for the RF reaches above 0.98. This measure for the three methods clearly shows that their better identification ability of the actual ignition data, which demonstrate their value in replacing the traditional LR algorithm in construct predicting model of wildland fire distribution.
3.2 Variable Importance

Figure 6 shows the variable importance graph generated by using the RF method.

Figure 6. Variable Importance Chart in Random Forest model

The two features that show the highest importance value, indicating their predictive power in the model generated by Random Forest are Distance to Road and WUI population density kernel density estimate. The result clearly reflects the fact that human-caused wildfire tend to occur at locations within a convenient distance from roads and of higher residential human traffic.

4 Conclusion

Machine learning models improve the prediction accuracy of traditional regression methods. Either RF or CIT models yield an improvement in accuracy over LR methods for wildfire occurrence assessment, according to AUC values.

More specifically, the RF algorithm seems to be the best choice not only due to its higher accuracy, but also because its calibration is easier as it involves fewer parameters. Another advantage of RF is its cartographic outputs, which seem to be more realistic than those from other models due to RF’s higher spatial variability and therefore greater spatial discriminatory power. This enables RF to provide a better reflection of variability in wildfire occurrence linked to heterogeneity of landscapes and human activities.

Prediction of wildfire occurrences may benefit from using multiple approaches, yielding a range of predictions in a concurrent manner rather than relying on a single map. Regardless of the method considered, both the Distance to Road and the surrounding WUI population density have proved to be the variables most closely related to fire occurrence, and hence presenting themselves to be the most important predictors in the models. In any case, we can draw the conclusion that fire occurrence in Santa Monica mountains is mainly related to the expansion of human settlement pressuring on wildlands and to accidents or negligence in the course of human-wildland interaction.
5 References


