

**Determining the main pre-fire condition factors of wildfire-induced socio-economic loss: A  
case study on the west coast of the United States during 2017 and 2022**

Ding Li (20866008)

GEOG 471

Nastaran Saberi

April 14, 2022

## **Abstract**

Due to climate change, more and more frequent extreme weather leads to more wildfires. In the United States (US), many wildfires are concentrated on the west coast area. The wildfires on the west coast of the US caused much socio-economic loss due to the astronomical costs of the firefighting and the destruction of human-built land use, etc. Therefore, the study aims to determine the most important pre-fire condition factors that may affect the amount of socio-economic loss caused by a wildfire. In the project, the pre-fire weather, vegetation, and topography are considered as the main theoretical pre-fire condition factors that might influence wildfire-induced socio-economic loss. From National Interagency Fire Center (NIFC) historical wildfire database, 20 typical wildfires with high socio-economic loss on the west coast of the US during 2017 and 2022 are selected. By utilizing multi-band time-series remote sensing data, remote sensing algorithms, and weather data from NASA POWER, the data on the pre-fire weather, vegetation, and topography of the 20 wildfires are gathered. Finally, by conducting the significance of the correlation coefficient test and training random forest models, it can be concluded that the areas that are more densely vegetated with low wind speed and steep topography are likely to have more expensive wildfires. The areas with lots of human-built land use, steep topography, and wet-bulk temperature have high risks of being destroyed.

## Content

1. Introduction.....	1
2. Methodology.....	3
2.1. Potential factors to the defined wildlife severity .....	3
2.2. Data sources.....	4
2.2.2. Remote sensing data.....	5
2.2.3. Weather datasets.....	7
2.2.4. Data summary table.....	8
2.3. Study area .....	9
2.4. Method for identifying land cover and land use.....	10
2.4.1. Literature review on classification methods.....	11
2.4.2. NDVI, NDBI, thresholding, and visualization .....	14
2.5. Elevation models .....	16
2.5.1. Literature review on elevation models .....	16
2.5.2 DEM.....	17
2.6. Feature importance ranking .....	18
2.6.1. Literature review on filter methods.....	19
2.6.2. Literature review on wrapper methods.....	21
2.6.3. Selection of methods used in the project.....	23
2.6.4. Pearson’s correlation coefficient and random forest.....	24
2.7. General workflow .....	25
3. Results.....	29
3.1. Pre-fire condition layers of Dixie .....	29

3.2. Socio-economic loss .....	32
3.3. Importance ranking for the pre-fire condition factors to the estimated cost of wildfires	34
3.4. Importance ranking for the pre-fire condition factors to the estimated cost of wildfires	40
4. Discussion .....	47
4.1. Interpretation of the results from example Dixie.....	47
4.2. Feature selection and importance ranking on estimated cost .....	51
4.3. Feature selection and importance ranking on estimated destroyed human-built area ....	55
4.4. Limitations .....	60
5. Conclusion .....	62
References .....	65

## **1. Introduction**

Wildfires, which refer to outdoor fires that are not under control, occur more and more frequently due to extreme weather brought by climate change (Bloom, 2018). As an essential consequence of climate change, adaptations should be proposed to minimize the impacts on the ecosystem and socioeconomic loss from wildfire occurrences (Hessburg, 2021). In particular, more concerns would be placed on the socioeconomic aspect in this study because wildfires have been proven to be a big issue for achieving the success of the third United Nations' Sustainable Development Goal (SDG), which is ensuring human health and well-being. First, biomass burning from wildfires leads to air pollution, which risks the public health of nearby residents (Yao et al., 2013; Stephens, 2014). Also, previous research (Ager et al., 2018) has found a robust link between wildfire smoke exposure, mortality, and respiratory illness morbidity. Additionally, exposed individuals to wildfire smoke had a significant positive association with exacerbations of asthma, chronic obstructive pulmonary disease, bronchitis, and pneumonia (Johnston et al., 2012). About 339000 annual death cases are caused by landscape fire smoke. Aside from mortality and direct damage to human health, wildfires could destroy a great number of homes, or even eliminate some population centers (National Geographic Society, 2019). Therefore, since wildfires cause non-communicable diseases, air pollution, and habitat destruction that relate to health and well-being closely, the main factors of the wildfire-induced socio-economic loss are worthy of being found and verified to help address problems with the third SDG (2016). In particular, knowing the most crucial pre-fire condition factors that contribute to the severity of wildfires can optimize and simplify the prevention and emergency strategies against wildfires, which can improve the efficiency and decrease the cost of fire fighting (Lovreglio et al., 2010). That is also a reduction of the socio-economic loss induced by wildfires. Therefore, to help

prevent wildfires from causing too much socio-economic loss, the research aims to determine the pre-fire condition factors that affect wildfire-induced socio-economic loss significantly.

To achieve the aim of the study, some objectives must be clarified. First, the severity of wildfires needs to be properly defined first to quantify how hazardous the wildfires are, which indeed does not have a unified standard and varies depending on the purposes of research (Han et al., 2021). For example, there are indices such as canopy index and composite burn index (CBI) introduced to quantify fire severity based on different concerns (Turner et al., 1994; Marcos et al., 2018). Nonetheless, based on the SDG goal that the study concerns (i.e. socio-economic loss) and the harms to the human community that wildfires lead to according to the previous discussion, the quantifications of wildfire severity defined in the study must include the cost of the wildfire, the area of the human-built land use destroyed during the wildfire, the number of death cases during the wildfire, the number of injury cases during the wildlife, and the air quality decreased due to the wildfire, etc. However, due to the limited time and data availability problems, the project only focuses on the cost of each wildfire and the area of the human-built land use destroyed during the wildfire. Therefore, in this study, the severity of each wildfire will be defined as the cost of each wildfire and the area of the human-built land use destroyed during the wildfire. Based on the definition, finding the main driving factors that contribute to the wildfire severity is the next step. In order to see what potential pre-fire condition factors to the severity of wildfires might be taken into account, a literature review on the mechanism of wildfires is in this study. Afterward, the multi-band time-series remote sensing data with appropriate temporal and spatial resolution will be collected, whose description will be included in the data source section in detail. Based on the remote sensing data and some properly selected algorithms, the pre-fire condition layers (e.g. pre-fire land surface temperature) and the post-fire

condition layers (e.g. post-fire human-built land use) will be generated. Finally, a series of statistical analyses, including Pearson's correlation coefficient and random forest, will be conducted to determine the most important pre-fire condition factors to the wildfire-induced socio-economic loss (i.e. the severity of the wildfire defined in this study) using the pre-fire condition layers, post-fire condition layers, and some extra attributes about the selected wildfires from non-remote-sensing sources.

## **2. Methodology**

As claimed before, in this section, the potential pre-fire condition factors that may influence wildfires theoretically will be summarized from a board literature review. The outlined pre-condition factors will all be considered in the project. Afterward, data sources and remote sensing techniques from which the pre-fire condition layers and the quantification of wildfire severity can be generated will be discussed. Finally, the statistical analyses based on the pre-fire condition layers and the quantification of wildfire severity will be conducted to determine the most important pre-fire condition factors to the severity of wildfires.

### **2.1. Potential factors to the defined wildlife severity**

First, the occurrence and the severity of wildfires are due to three features as conditions: fuel, oxygen, and heat (BC Wildfire Service, 2020 (a); Gale et al., 2021; Silverstein et al., 2010 (a). Particularly, in each position, the land cover and the humidity are supposed to correspond to the fuel condition, the geographic position and the latitude can directly determine the amount of oxygen in that place, and the heat can be measured by the land surface temperatures (LST) (Butsic et al., 2015; Jain, 2021; Peacock, 1998; Vlassova et al., 2014). Besides the initial environmental conditions, some induced behaviors from both human sources and natural sources

are also inevitable causes of wildfires (Ma et al., 2020) (a). For human behaviors, besides some factors that can hardly be accessed, camping fires and fallen power lines are considered as the most common causes (Muhs et al., 2021; Silverstein et al., 2010) (b). For natural conditions, lightning, rockfall, meteorite, and volcanoes are some theoretical reasons for wildfires (BC Wildfire Service, 2020) (b). Nonetheless, due to the irregularity and the inaccessibility of some data, only four types of data that are regularly available are considered in the study: weather, vegetation, and topography (Ma et al., 2020) (b). In summary, the pre-fire land cover, land surface temperature (LST), air temperature, altitude, air pressure, humidity, wind speed, wind direction, and precipitation in the perimeter of the wildfire are considered the theoretical factors that are important to the severity of wildfires.

## **2.2. Data sources**

The datasets used in the project are from National Interagency Fire Center (NIFC), the United States Geological Survey (USGS), and the Prediction of Worldwide Energy Resources department of the National Aeronautics and Space Administration (NASA POWER), from which the vector datasets, the remote sensing data, and the weather datasets that are related to the wildfires in the study are obtained. More detailed descriptions are as introduced below.

### ***2.2.1. Wildfire vector datasets***

NIFC is built next to the Boise Airport, where the equipment, support, and administrative offices for emergencies including wildfires are located. The organization is set to provide efficient and cost-effective wildfire management through policymaking, information collection, and educating personnel (NIFC, 1965). On the official website of NIFC, historical wildfire datasets are published and updated on time, among which are the Wildland Fire Locations Full History (WFL) dataset, and the Wildland Fire Perimeters Full History (WFP) dataset are used. In



the WFL dataset, the locations for all reported wildfires that occurred in the US from 2014 to 2022 are included in the dataset, which is in the form of a vector feature class consisting of points and some information about each wildfire. In the WFP dataset, the perimeters for all reported wildfires that occurred in the US from 2014 to 2022 are included in the dataset, which is in the form of a vector feature class consisting of polygons and the information about each wildfire that overlaps with the information recorded in the WFL dataset. The information on each wildfire included in the WFL and the WFP datasets is in the form of columns (i.e. attributes) in the datasets. According to the interest of this study, some specific columns are selected for use. The set of columns includes the discovery time of each wildfire, the end time of each wildfire, and name of each wildfire, the unique identification of each wildfire, the estimated cost of each wildfire, and the area of the perimeter of each wildfire (NIFC, 2021). Other columns are abandoned as redundant. Further steps for processing the datasets will be discussed in the methodology section.

### ***2.2.2. Remote sensing data***

About USGS, the organization is a science group belonging to the US Department of the Interior, which delves into hazards, climate change effects, land-use change, and the tools that learn about them (US Department of the Interior, 1879). From USGS, the datasets from Landsat 8 are available for the study. Specifically, Landsat includes a series of globe observation instruments, which return remotely sensed data about the Earth to USGS (Masek, 2021). So far, there have been 9 Landsat satellites in use, which are Landsat 1-9 respectively. From Landsat 1 to Landsat 9, a lot of improvements have been applied incrementally, which means the data is better for larger instrument codes. However, since Landsat 9 was launched in 2021, which is too new to record all the historical wildfires from 2017 to 2022, Landsat 8 which contains remotely

sensed data from 2013 to 2022 will be used. Another reason why Landsat 8 will be used is that a large majority of algorithms and current studies are based on Landsat 8, which will be shown in the below subsections. In this generation, Landsat 8 has less noise in the data, but more potential grey levels. The number of potential grey levels of Landsat 8 grows to 4096, whereas for previous generations, only 256 levels are available. There are also some additional improvements that increase the accuracy and the convenience of the remotely sensed data (USGS, 2022). From USGS, several datasets are available from Landsat 8, which can be grouped by level 1, level 2, level 3, collection 1, collection 2, and other data. In summary, level 1 data is raw data directly from the satellite without any correction or auxiliary, where level 2 data is corrected to eliminate atmospheric effects based on the level 1 data. Also, the level 2 data is corrected geometrically, radiometrically, with calibration, which enhance the accuracy or are the prerequisites of many algorithms (e.g. unsupervised classification) (Fraser et al., 1997; Kaasalainen et al., 2011; Topaloglu et al., 2016). However, for more mature data in level 3, the product will be too specific and indirect to the project. For example, the snow cover, burnt area, and water surface condition layers are useless to this project. About the difference between collection 1 and collection 2, collection 2 data is improved from collection 1 data. Some useful enhancement for the project involves the increase in precision and better consistency between level 1 and level 2 data. Therefore, the optimal selection of Landsat 8 data product is Landsat 8 collection 2 level 2 (L8 c2l2) dataset. In the dataset, only 5 bands are needed due to the purpose of this study, which are the Coastal aerosol band, the Red band, the Near Infrared (NIR) band, the SWIR 1 band, and the Thermal Infrared (TIRS) 1 band. Other bands will not be involved in the project.

### **2.2.3. Weather datasets**

The Prediction of Worldwide Energy Resources (POWER) Project is an applied science project of the Langley Research Center initiated in 2003 and funded by the NASA Earth Science Directorate Applied Science Program. This project provides detailed and accurate global weather data from Goddard's Global Modeling and Assimilation Office (GMAO) Modern Era Retrospective-Analysis for Research and Applications (MERRA-2) assimilation model products and GMAO Forward Processing – Instrument Teams (FP-IT) GEOS 5.12.4 Near Real Time (NRT) products. MERRA-2 data can cover weather data from 1981 to the present, but it is missing the last few months. To compensate for this missing data, the POWER group processes GEOS 5.12.4 data daily and adds the results to the period not covered by MERRA-2 to ensure real-time data availability, and successfully reduces the data delay to around two days (NASA & POWER, 1965) (a). GEOS 5.12.4 has the same grid resolution and similar parameters as MERRA-2, making it suitable as a complement to MERRA-2. POWER's ability to provide long-term, reliable data with high resolution and global coverage means that it is well suited for a wide range of weather studies, and it also has the unique advantage of being targeted at the study of wildfires. In the latest version of POWER 901, the latest data from NASA's GEWEX SRB version 4, CERES SYN 1-deg and FLASH Flux version 4A have been added, leading to further optimization of the resolution of the dataset and allowing POWER to provide detailed data on hourly time scales. The variability in wildfire development is complex, so coarse daily weather data are not competent in this study, while the hourly data provided by POWER can solve this problem by providing specific weather during the time of the wildfire. For accuracy, POWER data has been validated against the National Climatic Data Center (NCDC) and ensured to be accurate (NASA & POWER, 1965) (b). In summary, the conditions of POWER can meet the

needs of this study. To improve the efficiency of data use, some suitable attributes were selected, including temperature, relative humidity, wind direction, wind speed, barometric pressure, and precipitation, which are considered as possible influencing factors for wildfire development.

#### 2.2.4. Data summary table

To summarize the data sources used in the study, a data summary table is built as Table 2.1, where some extra useful information that is not mentioned in the previous subsections is listed. In particular, the L8 c2l2 data covers the whole globe, with a temporal resolution of 16 days. Additionally, the spatial resolution of the Red band and the NIR band is 30 meters, but the resolution of the TIR 1 band is 100 meters. The information is useful for the following methodology and discussion sections.

Data Publisher	Purpose	Content	Temporal range	Spatial range	Temporal resolution/unit	Spatial resolution/unit
NIFC	Select and overview study area	Locations, perimeters, and other information about wildfires	From 2014 to 2022	The US	Not applicable	The perimeter of each wildfire
USGS	Generate pre-fire and post-fire condition layers	Five bands: Coastal aerosol, Red, NIR, SWIR 1, TIRS 1	From 2013 to 2022	The earth	16 days	30 meters for Red, NIR, SWIR 1; 100 meters for TIRS 1
POWER	Analyze the impact of weather factors on wildfires	Hourly weather data when wildfires occur	From 2014 to 2022	The earth	1 hour	$\frac{1}{2}^{\circ}$ latitude by $\frac{5}{8}^{\circ}$ longitude

Table 2.1 Data summary table for the project. Information and data are provided by NIFC,

USGS, and NASA POWER.

### 2.3. Study area

This project focuses on the wildfires on the west coast of the US between 2017 and 2022 due to the extremely high frequency of wildfires with high socio-economic loss in the area during the period. Since some old wildfires might have already been studied by previous research, and more concerns are expected to be given to the recently happened wildfires, only the wildfires that occurred after 2017 are kept after filtering (Abatzoglou and Williams 2016; Calkin et al., 2015). Also, due to the flaw of the WFL and the WFP datasets, some records of wildfires lack crucial information, which can't be used. Also, the wildfires with perimeters smaller than 0.1 acres are not included in the WFP dataset, which is also considered too insignificant in this study. Hence, only the wildfires with full information and perimeters that are large enough are kept by the filter. Furthermore, since the study concerns the wildfires with high costs, only the top 20 wildfires with the highest estimated costs so far will be selected among the remaining wildfires. Thus, only the wildfires named Elkhorn, Slink, Doe, Silverado, Mineral, Castle, Dolan, Valley, Bighorn, Creek, Telegraph, Pinnacle, Sugar, Bootleg, Dixie, Tamarack, Caldor, Colony, Windy, and Alisal are left as the study area. The map of the wildfires is shown in Figure 2.2. According to Figure 2.2, it can be seen that most of the well-recorded wildfires with high costs and large perimeters are indeed concentrated on the west coast of the US.

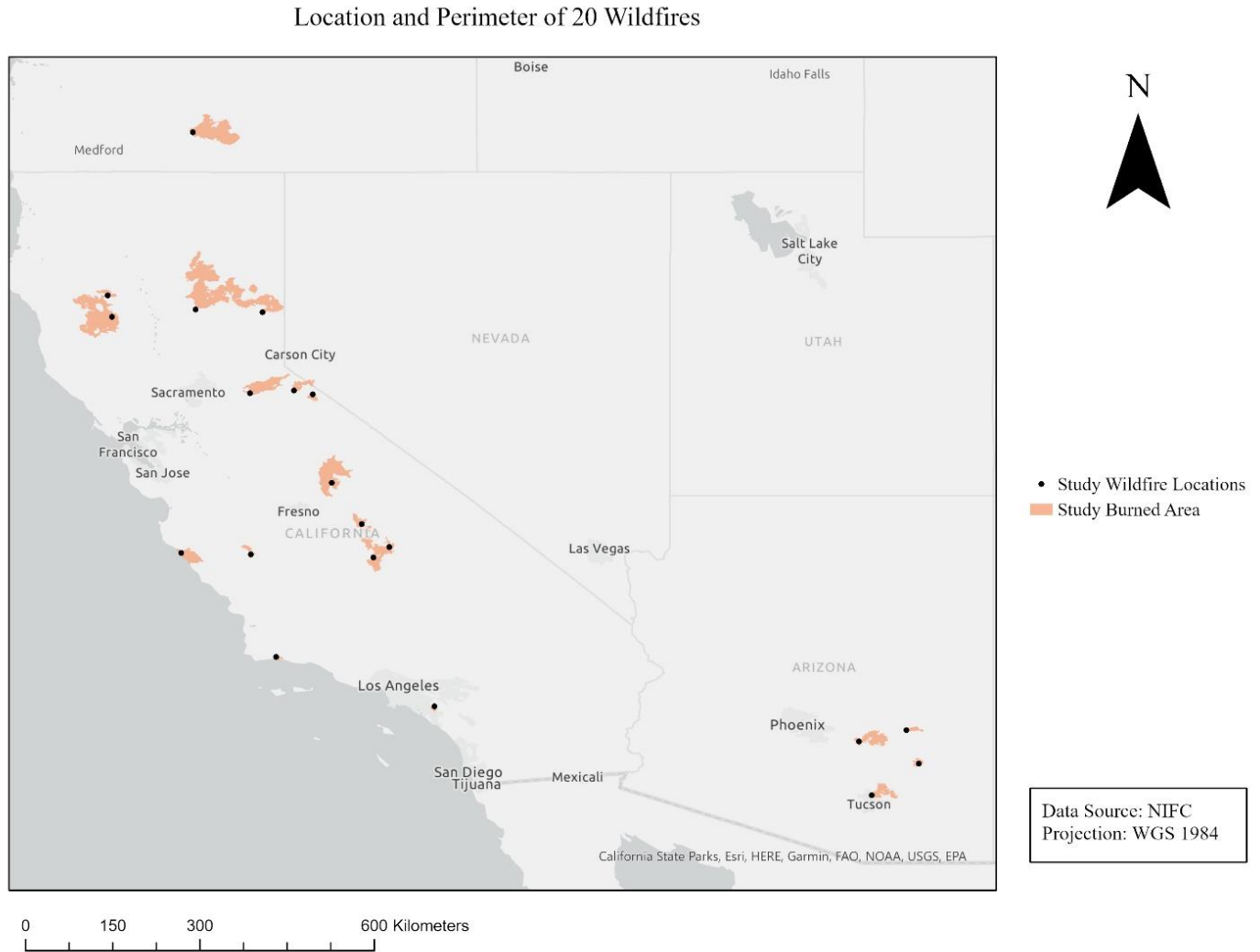


Figure 2.2 The map of selected 20 wildfires that occurred on the west coast of the US that will be studied in the project. The location and perimeter data of the wildfires are from NIFC as claimed in the data source section.

#### 2.4. Method for identifying land cover and land use

As claimed before, this study needs to get the pre-fire land cover, pre-fire human-built land use, and post-fire human-built land use. Hence, in this section, methods for the classification and identification of land cover and land use will be discussed.

### ***2.4.1. Literature review on classification methods***

For the classification of land use and land cover, many methods are available. The most commonly used approaches to classifying data into clusters to produce a classification can be split into three categories, which are supervised, semi-supervised, and unsupervised respectively (De Cáceres et al., 2015).

Unsupervised approaches use statistical and mathematical models (Kent, 2012) (a); Tich et al., 2014 (a)). Unsupervised classification (Puletti et al., 2014) (a) is a set of algorithms for classifying remote sensing data based on the nature of the values and relative placements of pixels in remote sensing pictures without the need of people, prior knowledge, or any other supplementary data (Lillesand et al., 2015) (a). According to multiple articles, Kmeans clustering is one of the best-unsupervised classification algorithms that can be used to detect human land use (Puletti et al., 2014); Lillesand et al., 2015) (b). Some research used unsupervised classification to divide each image into the soil, vegetation, and city with acceptable accuracy and excellent efficiency (Afify, 2011).

Supervised approaches allocate plot data to communities based primarily on the researcher's ecological knowledge and expertise gained through fieldwork observations (Tich et al., 2014) (b). For example, maximum likelihood classifier (MLC) is a supervised classification approach that requires training samples to create distinct classes. By using MLC as the decisive role, each pixel is assigned to its most similar category, and the two classified images are pixel-by-pixel compared. A bitmap is prepared for each category of each classified image for comparison in order to detect the transition from one category to another and to explain the change type of land cover category over the course of the survey period. It is frequently used as an auxiliary means of acquiring findings in earlier research (Sun et al., 2013). However,

preparing the training samples will require time and effort, which means that one of MLC's drawbacks is how time-consuming it is. Furthermore, the accuracy will be influenced by the spectral distinctness of the various classes. For example, the spectral distinctness of the urban area and the construction land may be comparable, resulting in misclassification (Evans, 1998).

To solve some of the discrepancies in many supervised methods, such as low repeatability, opaque grouping criteria, and inconsistent application, it is recommended to use semi-supervised strategies to recognize communities (Kent, 2012) (b). Since the results obtained by using only one classification method may not be accurate enough to meet the needs, Singh et al. (2014) used a hybrid classification strategy that included both unsupervised and supervised classification based on support vector machines (SVM) to accomplish land-use-landcover (LULC) estimation. Following the separation of the vegetation and non-vegetation groups, classification technology based on the training set is applied.

Other stream focuses on using NDVI with other land classification index or methods to get more accurate results. Analyses of satellite-sensed data for the NDVI are occasionally used to support abiotic response to climate change. Because it is closely associated with the fraction of photosynthetically active sunlight absorbed by plant canopies and hence leaf area, leaf biomass, and potential photosynthesis, NDVI can be used to proxy vegetation's reactions to climate change, according to (Zhou, et al., 2001) (a). On how to analyze land cover change, Zhang, et al (2009)'s study believes that although NDVI has been used to estimate vegetation productivity and rainfall in semi-arid areas, the normalized difference accumulation index (NDBI) has been developed to identify cities and built-up areas. Therefore, using NDVI and NDBI as alternative indicators of LULC can reveal the relationship between different indicators in the study. Similarly, NDVI is also used as the auxiliary index for the study of LULC. Borini, et al (2015)



(a)'s study aims to analyze the LULC dynamics in selected areas, and monitor and distinguish the trajectory of NDVI changes in the past. The results of Broini, et al (2015) (b) show that different trajectories related to NDVI and LULC help to better understand the dynamics of ecological processes and the changes of human impact in the region. In order to analyze the changes in LULC in the study area, NDVI and NDBI were used to determine. Similar to Zhang's 2009 study, NDVI and NDBI are indicators for vegetation extraction and detection of built-up areas. On the basis of predecessors, there are some innovations. In Guha, et al (2018) (a)'s research, these two indicators can be used to classify different types of LULC through appropriate thresholds. Similarly, Kulkarni, et al (2021) uses the formula to create an NDBI grid with band numbers 5 and 6. In order to obtain a more accurate classification, the research of Guha, et al (2018) (b) chose to use Boolean operators on the indexed spectral bands. For example,  $NDVI > 0.2$  and  $NDBI < 0$  are used together to extract vegetation, and  $0 < NDVI < 0.2$  and  $NDBI > 0.1$  are used together to extract built-up areas and bare land. In different studies, researchers also have different threshold selections.

Furthermore, when using NDVI and other indexes as a basis for classification, the numerical thresholds used will not only vary depending on the settings of different researchers, but Guha et al., (2018) (c) believe that these NDVI and NDBI thresholds used for classification may also vary depending on atmospheric conditions even within the same study area. The results of Wang and Tenhunen, (2004) also support using unsupervised k-means classification based on NDVI temporal profile metrics to provide an up-to-date land cover classification, because this method has high accuracy in the combination of various methods.

In summary, due to the large number of layers that need to be classified, there is a need for the efficiency of the classification method. Hence, supervised or semi-supervised algorithms

are too time-consuming for the project. Also, to further ease the workload, the land cover will only be classified into vegetation and others, which means there will not be a more detailed classification between different types of vegetation. In that case, the requirement for the accuracy of the classification method is not demanding. Therefore, in the project, only NDBI, thresholding, and visualization will be utilized in the study to recognize human-built land use.

#### ***2.4.2. NDVI, NDBI, thresholding, and visualization***

NDVI will be used in the study to identify vegetation and measure the vegetation coverage condition, and NDBI is used for recognizing human-built land use. The NDVI data captures the contrast between vegetation's red and near-infrared reflectance, which indicates the amount and energy absorption of leaf pigments like chlorophyll (Zhou, et al., 2001) (b). The index values of NDVI always range from negative one to positive one. If the index value is positive, it often means the cell has a plant cover. The bigger the plant's vegetative and density are, the higher the index value is expected to. Note that the formula for calculating NDVI is

$$NDVI = \frac{NIR-RED}{NIR+RED} \quad (2.3)$$

, which means the higher the NDVI value, the more NIR is present, indicating lush greenery that coincides with the feature of healthy vegetation. Hence, the density of vegetation in a pixel can be measured by NDVI. Particularly, Shrubs and grasslands, or senescing crops, and dense vegetation, or tropical rainforest, are usually obtained by the threshold criterion  $NDVI > 0.2$  (Purevdorj et al., 1998).

NDBI is invented based on the spectral response of built-up areas, which have a high reflectivity at intermediate infrared wavelengths and a low reflectivity at near-infrared wavelengths. As a result, when the NDBI algorithm is applied to a multi-band remote sensing image, urban or built-up regions seem dazzling white and have positive digital values, whilst

other areas appear dark and have negative or zero numerical values. By using the appropriate threshold values, these two indices can be used to classify land use and land cover (Chen et al., 2006) (a). As the rationale behind NDBI, more SWIR than NIR is reflected by built-up regions and bare soil. On the infrared spectrum, water does not reflect. The NIR spectrum reflects more than the SWIR spectrum on a greenie surface. Therefore, based on the rationale, the formula of NDBI was derived as

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR} \quad (2.4)$$

According to the calculation, NDBI values lay between negative one and positive one. The pixels with negative NDBI values represent water bodies whereas the pixels with higher values represent human build-up areas. Also, NDBI values for vegetation is lower than human-built land use, but usually positive (Zha et al., 2003). In this project, the thresholding criteria  $NDBI > 0.1$  will be used to extract built-up areas and bare land according to some previous research with a good accuracy (Yasin et al., 2020).

In summary, in this project, NDVI will be calculated by equation (2.1), NDBI will be calculated by equation (2.2). Pixels with the NDVI values larger than 0.2 will be classified as vegetation. Pixels with eh NDBI values larger than 0.1 will be classified as Human built-up land use. However, since the accuracy of the classification method is not guaranteed, and sometimes regarded as poor, some manual corrections of the classification will be implemented by visualization and manual modification, based on the fact that some human-built land use and vegetation can be very obvious under the visible bands, and is widely used in some research for recognizing vegetation land cover and built-up land use (Panigrahy et al., 2010; As-syakur et al., 2012).

## **2.5. Elevation models**

In this study, pre-fire elevations are considered as a theoretical factor that might influence wildfires according to section 2.5.1. Hence, a brief review on different types of digital elevation models (DEMs) will be given in 2.5.1. Afterward, the rationale behind the DEM model selected in the project will be discussed in section 2.5.2.

### ***2.5.1. Literature review on elevation models***

For representing elevation, lots of models are widely used, which include Digital Elevation Model (DEM), Digital Surface Model (DSM), Digital Terrain Model (DTM), and irregular triangulated network (TIN). A DEM is a digital simulation of a terrain surface by means of finite terrain elevation data, a digital representation of the elevation of a regional surface and a collection of elevation values representing the position of an object stored in digital form. DEM is a generic term for all forms of digital overlay data of terrain. Whereas Digital Surface Model (DSM) is a model that describes the top of reflections such as buildings and vegetation (Maune and Nayegandhi, 2018). DTM serves as a digital representation of the terrain giving a digital representation of the surface, slope, aspect, horizontal, vertical surface curvature and other terrain attributes. DTM is the bare ground DEM (Maune, 2001). Amongst others, DSM is particularly suitable for telecommunication management, aviation safety, forest management and 3D modelling and simulation (Pandjaitan et al., 2019). DTMs, on the other hand, can be directly extracted from the terrain, allowing for the observation of landscape features, so they are generally widely used in the analysis of terrain features (Thommeret et al., 2010).

The application of DEM is very extensive, as it can be used to extract contours or to fit terrain surfaces, as well as to create three-dimensional perspective views, which can help to better study the spatial pattern of the terrain. The contour lines obtained from the triangulated

DEM are obtained directly from the original observation data, avoiding the loss of interpolation accuracy, and thus the contour lines are more accurate (Li, 2009) (a). In addition, DEM can be scientifically and rationally used in mapping monolithic retouching to fundamentally improve the accuracy of retouching results and also improve detection efficiency (Li, 2018).

Hence, as most of wildfires occurred in wildland with some vegetation, DSM might be the most suitable model to represent elevation in that case (Van et al., 2010). Indeed, some research selected to use DSM (Zinck et al., 2010). However, the DEM can be conveniently obtained from L8 c2l2 dataset without any processing steps. And there are still some researchers use DEM to measure the pre-fire or post-fire elevations (Pelletier and Orem 2014; Sherman et al., 2007). Therefore, due to the large workload of the project, the convenience of obtaining DEM, and some evidence that shows the appropriation of the use of DEM for representing the elevation within fire perimeters, DEM will eventually be utilized to represent the pre-fire elevations.

### **2.5.2 DEM**

DEMs represent real surfaces with discrete sampling points and the DEM data is organised and calculated in the form of a simple matrix. The usual sources of DEM data are field survey data, topographic maps, and remotely sensed image data. Remote sensing imagery data is now generally used to construct DEMs due to its lower limitations, high accuracy, efficiency and the advantages of large data volumes. In fact, the construction of DEMs is largely influenced by interpolation methods, generally whole interpolation and point-by-point interpolation, as well as local block interpolation, which means that the elevation values are interpolated on the terrain plane based on coordinates and interpolation functions to obtain an accurate, ordered isomorphic and compact DEM with the same resolution as the source map (Hu et al., 2009). The regular grid

model, the contour model and the irregular triangulation model. It is the grid DEM and the irregular triangulation DEM that are commonly used and on which terrain analysis is based. The grid DEM is the most commonly used form, with data and organisation similar to image raster data, where the value of each image element is the elevation value, which is generated by interpolation of regular or irregular discrete data; the irregular triangulated network (TIN) model is another representation that reduces the disadvantages of data redundancy associated with regular grids, and is based on a finite set of points in a region that tends to be divided into a triangulated network of necklaces, where Any point in the region that falls on a vertex, edge or triangle of the triangular surface is continuous but not infinitesimal throughout the region (Li, 2009) (b). Therefore, the mean value of the DEM layer within the fire perimeter can well represent the average elevation of where fire occurred and spread. The standard deviation of the DEM layer within the fire perimeter shows the variability of the elevation, or how steep the terrain is of where the fire occurred and spread. Together, the two statistics well represent the distribution of the elevation of where the fire occurred and spread, which represent one of the pre-fire condition factors to wildfire-induced socio-economic loss in this project (Lee and Kim, 2008) (a).

## **2.6. Feature importance ranking**

In this section, the statistical and machine learning methods used to determine the most important pre-fire condition factors that influence the wildfire-induced socio-economic will be discussed. The methods will include filter-style methods and wrapper-style methods. Note that in machine learning, an independent variate or an explanatory variable can also be called a feature. Also, a label is equivalent to a responsive variate, or dependent variable in a machine learning

context (Kshetri et al., 2019). Therefore, in the following content, they may be used interchangeably depending on where the reviewed articles are published.

### ***2.6.1. Literature review on filter methods***

Filter methods are the first class of feature selection methods, which obtain a subset of features by removing insignificant features one by one. Recall that the pre-fire condition factors that are taken into account in the study involve NDVI, LST, air temperature, altitude, air pressure, humidity, wind speed, wind direction, and precipitation in the perimeter of the wildfire. Hereby, to model wildfires, too many variates should be considered by theories. Thus, finding a small subset of the most significant factors that contribute to the occurrences and severity of wildfires is necessary, so that only some important variates need to be considered (ZHAI et al., 2002). The reduction of the number of input variates makes models more plausible and more efficient (Urbanowicz et al., 2018) (a). Also, fewer input variates ease the problems about the unavailability of observed data (Kuhn et al., 2019); (Urbanowicz, 2018) (b). Hence, at least one of the filter methods must be implemented in this project to reduce the number of considered features.

First, filter methods usually conduct statistical tests about the correlation of every single feature and the responsive variate separately (Kohavi, 1997) (a). Based on the statistical result, each feature is assigned a score that represents the correlation of the feature with the responsive variate. Finally, the features are ranked based on the scores. Hence, the rank of each feature can reasonably represent the importance of the feature to some extent, since the more intensive the correlation is between the feature and the responsive variate, the more crucial the feature is to the responsive variate (Kohavi, 1997) (b); (Sánchez-Marono et al., 2007). Also, the optimal statistical tests vary depending on the types of the explanatory variate and the responsive variate

(Kaushik, 2016) (a). Unfortunately, for some study, the dataset that will be utilized to train the model will include different types of variates. For example, the weather attribute may be windy, cloudy, etc., which means the weather attribute is a categorical variate; Also, for the occurrence of wildfires, the variate is a count or a discrete variate. That is, for each pair of feature and responsive variate, different statistical methods may be different based on their variate types (Kaushik, 2016) (b). Nonetheless, in this project, all of the variates are continuous. Hence, only the statistical methods for two continuous variates need to be discussed.

For a continuous variate and a continuous variate, many univariate regression models can be suitable. First, a simple linear regression can be applied, which assumes a linear relationship between the explanatory variate and the responsive variate, and the coefficient of the linear model expression calculated based on the data points represents the strength of the linear relationship between the two variates. Also, by conducting linear regression hypothesis testing, whether there is indeed a linear relationship between the two variates can be revealed (Seber and Lee, 2012). Therefore, the score used by the filter methods can be computed in terms of the linear regression analysis results. Also, since filter methods only select one feature as the explanatory variate of the linear regression model each time, the sensitivity for the existence of multicollinearity among different explanatory variates can be eliminated (Alin, 2010). However, since many variates in the dataset are not in a linear relationship obviously, the method won't fit the data points well for some variates. Also, the model is not stable when the number of outliers is large (Yu et al., 2017). Additionally, linear regression tends to overfit according to some research (Hawkins, 2004). To solve the issues, non-linear regressions can be introduced (Amemiya, 1983). For example, usual non-linear regression methods include Gaussian functions, logarithmic functions, exponential functions, and so on (Kumar et al., 2015; Weber, 2002;



Tanaka et al., 1995). Nonetheless, since there will be lots of features in the dataset, and different features will have different types of relationships with the responsive variate, a scatterplot can be used. As a graphical summary, scatterplots can plot two-dimensional points on a graph, where the coordinates of the points are the feature and the responsive variate (Cleveland et al., 1985). By visualizing the scatterplot of each feature and the responsive variate, the type of relationship between them can be intuitive and assumed. Also, some other techniques, such as the Pearson correlation coefficients, can also perform well in quantifying the correlation between each feature and the responsive variate, where the variates are all continuous (Benesty et al., 2009). Therefore, for simplicity, simple linear regression and the Pearson correlation coefficients are available, and non-linear regression methods can be selected with the help of a scatterplot when the relationship between some specific feature and the responsive variate is obviously non-linear. Nonetheless, since non-linear regression methods have not been widely used in remote sensing study, only scatterplots and Pearson correlation coefficients will be used in the project.

To evaluate the class of methods, one of the advantages of filter methods is the computational efficiency. For filter methods, the number of computations of the correlation scores is equal to the number of features, which is much fewer than methods that need to traverse all subsets of features. Also, under the modern computation environment, each computation of the correlation of a pair of variates is cheap (Hastie et al., 2009). However, as a preprocessing step, the ranking might not be optimal, since the correlations between features and what model the study uses are ignored (Stork et al., 2001). Hence, wrapper methods should be introduced.

### ***2.6.2. Literature review on wrapper methods***

Unlike filter methods, wrapper methods obtain a compact subset of features by trying different subsets of features and selecting among the subsets. First, wrapper methods train a

model by using a subset of features. Afterward, the performance of the trained model is assessed. Based on the performance, features will be removed or added, and the whole process will be repeated. Finally, the algorithm will stop until adding or removing any feature won't affect the performance of the model. So far, the subset can be regarded as optimal (Kohavi et al., 1997) (c). Furthermore, wrapper methods can be divided into different classes based on the initial number of features and how the subsets are generated. In particular, there are coarsely two categories of wrapper methods: forward selection, backward selection, and recursive feature elimination (Guyon and Elisseeff, 2003); (Chen et al., 2006) (b). For forward selection, the initial set of features contains no features. Based on the performance of the trained model, the one feature that best improves the performance of the model will be finally added to the set. That is, for each addition of any feature, every feature that is not currently in the set will be tried (Blanchet et al., 2008). For backward selection, the initial feature set contains all features. For each feature in the set, the method will train a model without the feature. Based on the performance, the feature that has the least impact on the performance of the model will be eliminated from the set. That is, for each removal of any feature, every feature that is currently in the set will be tried (Sutter and Kalivas, 1993). For comparison, forward selection has better performance when the final optimal feature set contains very few features, whereas backward elimination is suitable when the final optimal feature set is big (Guyon et al., 2008).

As another irregular wrapper-like feature selection method, random forest is worthy to introduce separately. Overall speaking, random forest is a nonparametric model, so no previous assumptions are needed, which also include any linear relationship (Cafri et al., 2018). Also, a random forest can handle both regression work and classification work, which is exactly what

the study needs (Liaw and Wiener, 2002). Therefore, random forest is eventually used in our study. More details will be discussed in section 2.6.4.

### ***2.6.3. Selection of methods used in the project***

The methods described above are usual methods for feature importance ranking and feature selection, but the methods all lead to some challenges for this study. First, if filter methods are applied for feature importance ranking, the manual analysis work will be time-consuming. Because of the existence of different types of features and responsive variates, and different types of relationships between them, the optimal parametric statistical model must be carefully chosen for any pair of variates. Also, filter methods do not usually give optimal ranking, since the method is independent of the model people use. For the wrapper methods, the biggest problem is that the algorithms are too expensive. Therefore, for the feature ranking and selection work in our study, filter methods can hardly be optimal and will be inconvenient for people to conduct, whereas wrapper methods are inefficient for computers.

In summary, the methods all have some advantages and disadvantages, so the methods can be combined. For filter methods, the computation efficiency is much better than wrapper methods for computers. Also, wrapper methods train multivariate models, whereas filter methods only train univariate models, which avoids too high data dimensionality. However, there will be a heavier workload for people, and the ranking is hardly optimal for the trained model compared to wrapper methods. Additionally, there are more nonparametric models available for wrapper methods, where fewer assumptions are needed in advance compared to filter methods. As a wrapper-style method, a random forest has very good performance in handling nonlinear relationships between features and the responsive variates, while keeping some key advantages of wrapper methods, which is suitable for this study. Therefore, in our project, scatterplots and

Pearson correlation coefficients will be utilized as filter methods for dataset pre-processing, and a random forest will be conducted to rank features and finally select the important features.

#### ***2.6.4. Pearson's correlation coefficient and random forest***

As one of the most commonly used statistical method to determine the intensity off the relationship between any pair of continuous variates, the significance of the correlation coefficient test will be used in the project as a filter method. By inputting any pair of continuous variates, the Pearson correlation coefficient and the p-value of the correlation will be returned. For the Pearson correlation coefficient, the coefficient will always be within negative one and positive one. Positive coefficients mean the variates are positively correlated, which means if the value of one of the variates is high, the value of the other variate is also expected to be high. The closer the coefficient reaches one, the stronger the positive correlation is. On the other hand, negative coefficients mean the variates are negatively correlated, which means if the value of one of the variates is high, the value of the other variate is also expected to be low. The closer the coefficient reaches negative one, the stronger the negative correlation is. For a pair of variates with a coefficient close to 0, the correlation between them is considered very weak. Additionally, the p-values of the correlations indicates whether the correlation is significant. Specifically, if the p-value of the correlation is very small, the correlation is expected to be significant statistically. On the other hand, if the p-value of the correlation is very large, the correlation is not considered significant (Sedgwick, 2012) (a).

The rationale behind random forest and how it works in the project must be introduced. First, a decision tree should be introduced as the basic block of random forest. A decision tree is a tree-structured tool that classifies data points based on features. In each node, the data points in the node will be divided into different branches according to their categories of one specific

feature. Finally, the leave nodes represent the final classification result (Myles et al., 2004). Then, for a random forest, the dataset is resampled with replacement. Then, for each sample, a random subset of features is chosen. Based on the samples with the random subsets of features, many decision trees are trained parallelly (Biau and Scornet, 2016). Then, the likelihood for a subset of features to classify a data point wrongly is quantified by a metric called impurity, which usually refers to Gini impurity or some other impurity metrics (Zhi et al., 2018). Basically, the lower the purity is, the better the features can classify the data point. Hence, based on the ability of a feature to decrease the impurity, the importance of the feature can be quantified. Therefore, the remaining of the process reduces to a standard wrapper method. How random forest works in the study will be discussed in the following general workflow section.

## **2.7. General workflow**

In this section, the general workflow that combines methods discussed in the previous subsections is integrated. The workflow graph is as demonstrated in Figure 2.3. First, the records of the 20 wildfires are selected out using the criteria described in the study area subsection from the WFL and the WFP datasets. For each wildfire of the 20 wildfires, the L8 c212 remote sensing datasets within the perimeter of the wildfire before and after the occurrence of the fire are downloaded from USGS based on the perimeter, the occurrence time, and the end time of the wildfire recorded in the WFP dataset. The pre-fire vegetation condition layer is obtained using NDVI after computing the NIR band and the Red band of the pre-fire L8 c212 data using equation 2.3. The pre-fire LST layer is directly derived from Band-10 of the pre-fire L8 c212 dataset. The weather datasets are downloaded from NASA POWER based on the occurrence time and location of the wildfire from the WFL dataset, whose attributes include temperature, relative humidity, wind direction, wind speed, barometric pressure, and precipitation as

discussed above. Hereby, all the measures of the theoretical pre-fire factors that might influence wildfire severity are obtained. To estimate the area of the destroyed human-built land use during the wildfire, the difference between the areas of human-built land use before and after the wildfire will be used. To get the human-built land use, the pre-fire and post-fire NDBI layers are computed based on the SWIR 1 bands and the NIR bands of the pre-fire L8 c212 data and the post-fire L8 c212 data. Using an appropriate threshold (i.e. selecting out the pixels with  $NDBI > 0.1$ ) and visual interpretation, human-built ground objects before and after the wildfire are identified. Each of the areas of the human-built land use before and after the fire can be estimated by the number of pixels that are identified as human-built times the spatial resolution of L8 c212 (i.e.  $30m \times 30m$ ). Based on the calculated areas, the area of destroyed human-built land use during the wildfire will be estimated by the difference between the areas of human-built land use before and after the fire. For the estimate cost for each wildfire, the value can be accessed in the WFL dataset. So far, the estimated destroyed human-built land use and the estimated cost for each wildfire are obtained, which are all the considered socio-economic loss in this study. The obtained pre-condition factors, including the pre-fire NDVI mean value, NDVI standard deviation, LST mean value, LST standard deviation, DEM mean value, DEM standard deviation, air temperature, altitude, air pressure, humidity, wind speed, wind direction, and precipitation in the perimeters of the wildfires, will all be regarded as the features (i.e. explanatory variates) in the training set. The obtained socio-economic loss factors, including the estimated cost and the estimated area of destroyed human-built land use during the wildfires, will be input as the outcome variables (i.e. responsive variates) of the training set. To first conduct the filter method to drop the insignificant factors, scatterplot matrix of the features and the outcome variable are made, where the Pearson correlation coefficients between each pair of feature and outcome

variable will also be calculated. The correlation between each feature-outcome pair is thus revealed, where the features with insignificant correlations with the outcome variable will be dropped. After the filter process, a random forest model will be trained and tuned with an acceptable accuracy. Based on the random forest model and the built-in feature importance ranking function, the most important pre-fire condition factors for the wildfire-induced socio-economic loss will be ranked based on the estimated costs and the estimated areas of destroyed human-built land use respectively.

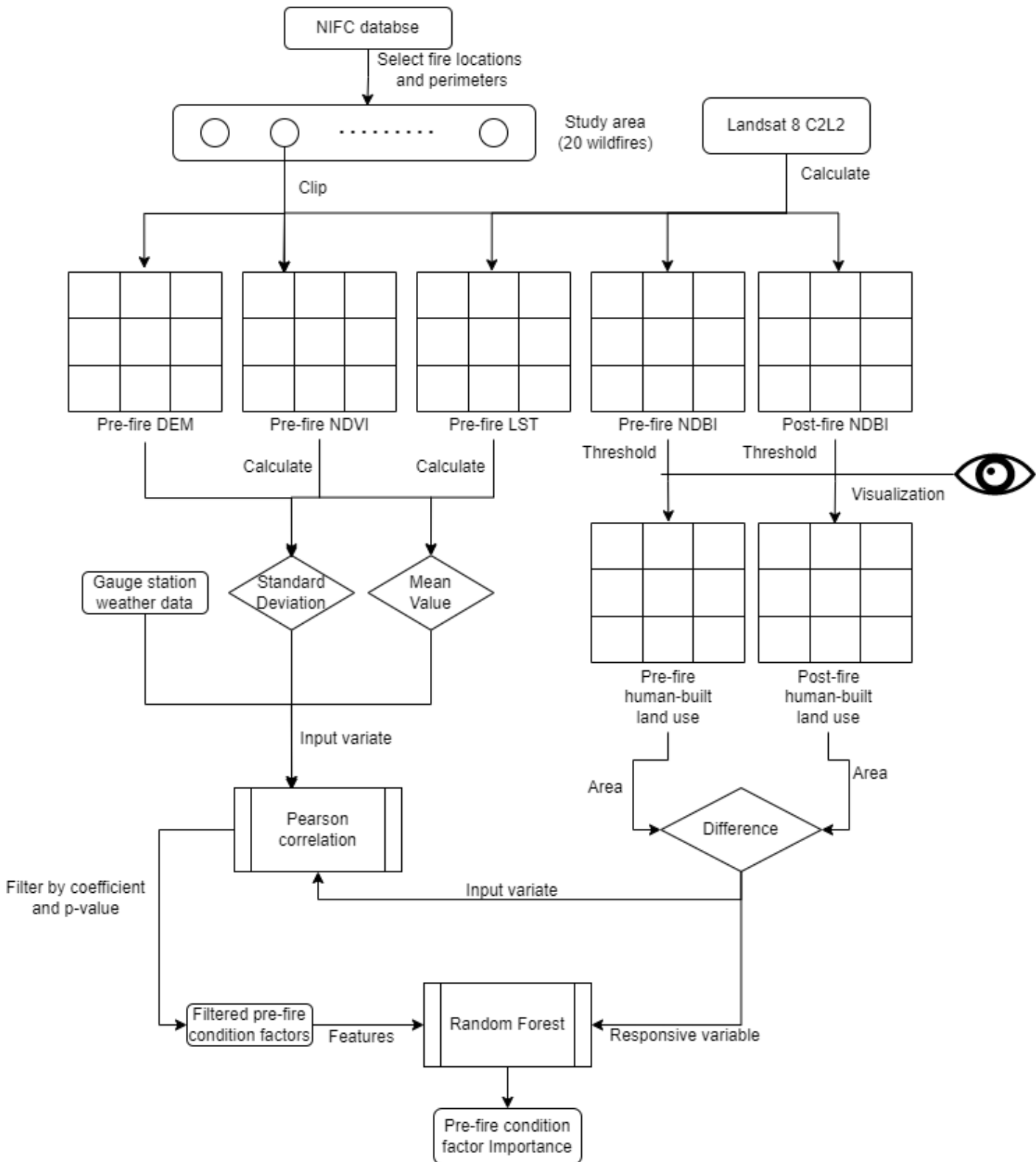


Figure 2.3 Workflow of the methods



### **3. Results**

To give an intuition of the characteristics of the pre-fire conditions of wildfires, the most expensive wildfire, Dixie, will be the example. Also, the socio-economic loss information will be demonstrated as well. Finally, the analysis results from Pearson correlation coefficients and random forest will be shown.

#### **3.1. Pre-fire condition layers of Dixie**

In Figure 3.1, the vegetation coverage within the perimeter of Dixie is classified using NDVI and visual interpretation is shown. The map shows that the vegetation coverage was much larger than other land covers some days before the occurrence time. Most of the land in the perimeter of Dixie was recognized as vegetation. The mean value of the NDVI layer is 0.23067, with a standard deviation of 0.0783. The mean value is higher than 0.2 that is the threshold for classifying a pixel into vegetation.

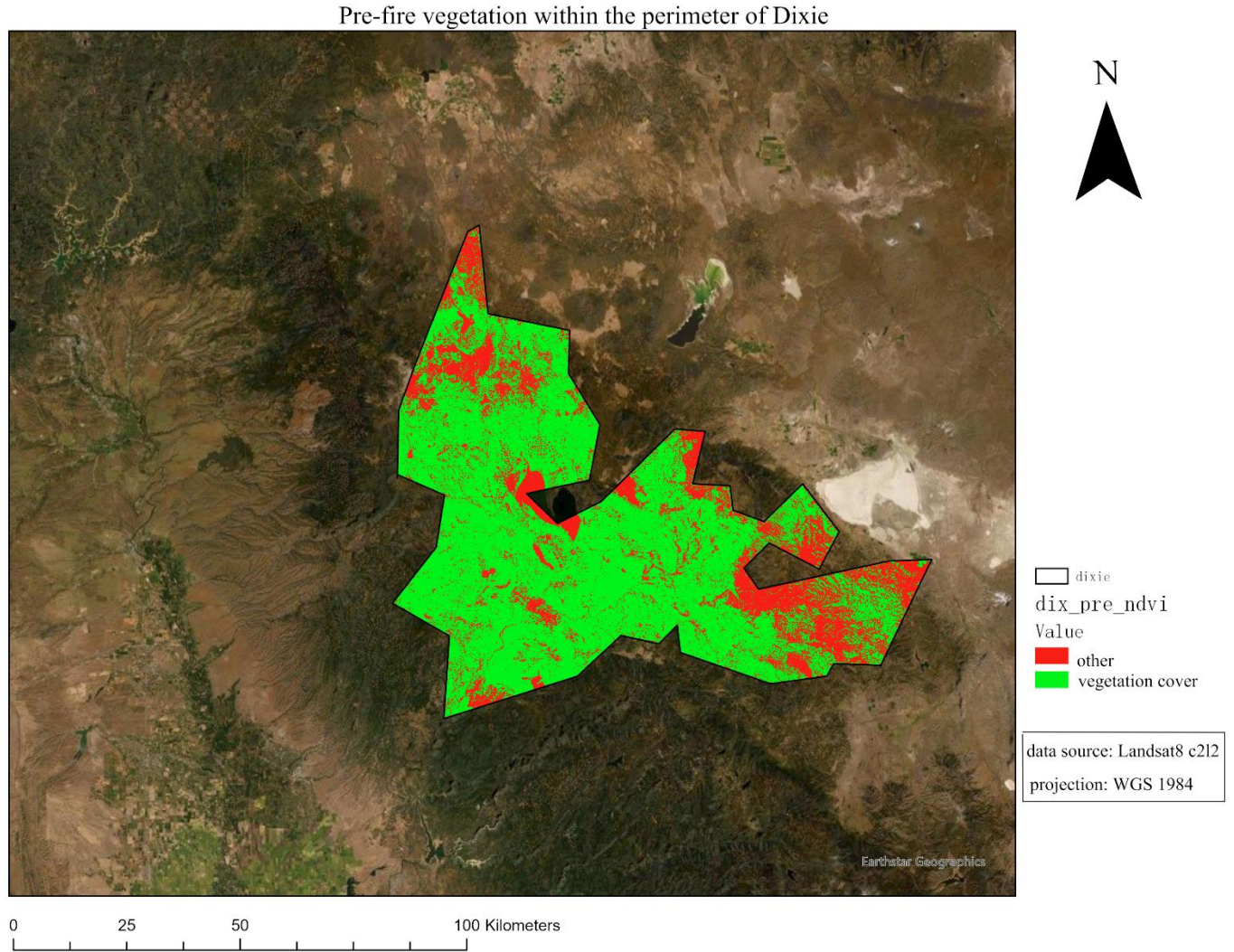


Figure 3.1 Pre-fire vegetation coverage of Dixie

In Figure 3.2, the DEM of the perimeter of Dixie is shown. According to the map, the elevation of the perimeter lies within the range of 430 meters and 3200 meters. The northwestern part and the southeastern part of the perimeter are in red and yellow, which means the elevations of the two parts are above 1600 meters, some of which might reach 3000 meters. The middle part and the southwestern part of the perimeter are almost in green, which means whose areas have lower elevation, which might be below 1000 meters but higher than 430 meters. The mean value

of the DEM layer is about 1677.29082 meters, with a standard deviation about 309.09423 meters.

## The DEM covering the Dixie fire perimeter

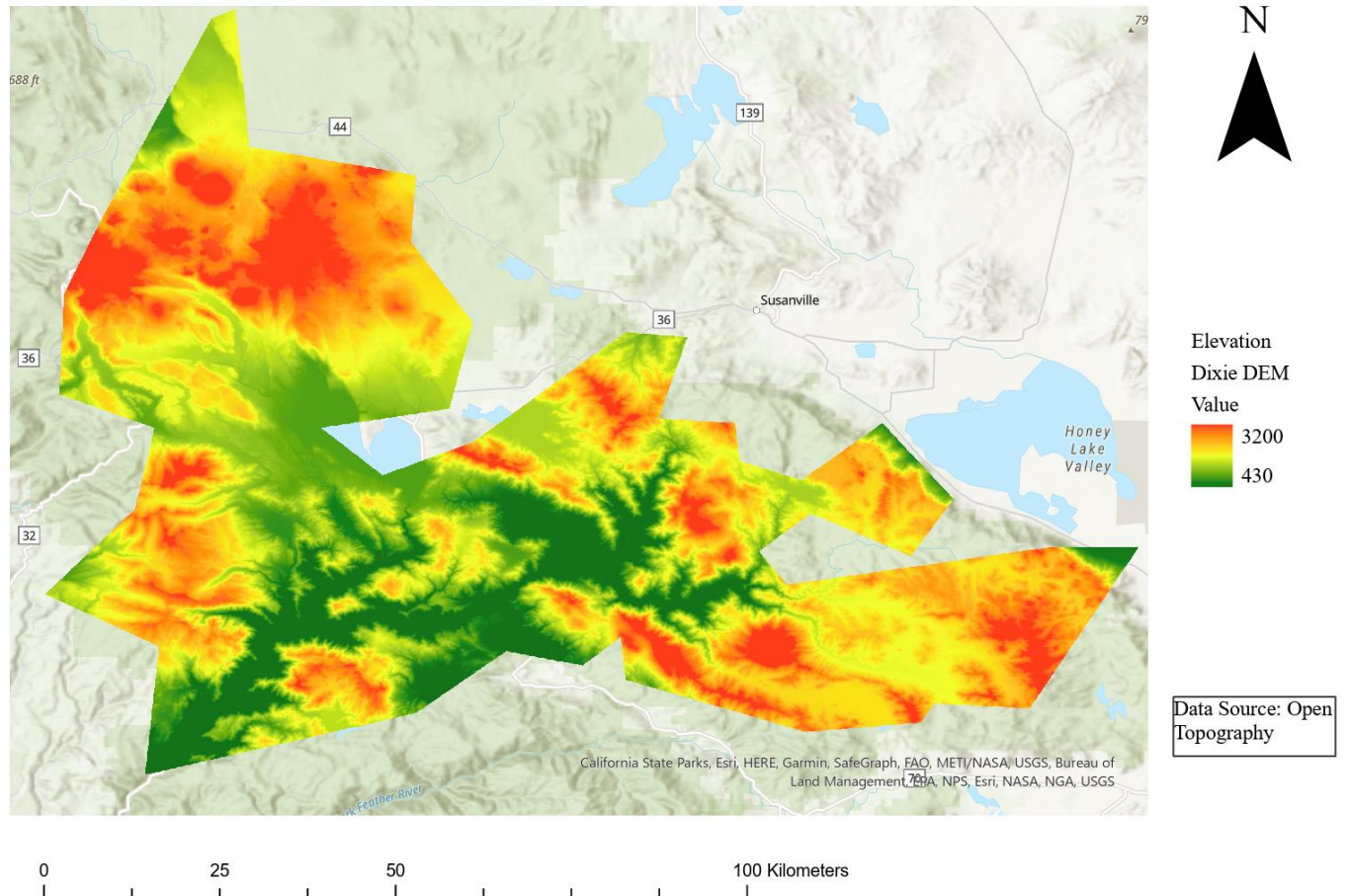


Figure 3.2 Pre-fire DEM within the perimeter of Dixie

In Figure 3.3, the pre-fire LST distribution map within the perimeter of Dixie is shown. The map shows that just before the occurrence time, the LST laid within the range between 295.428 Kelvin and 336.407 Kelvin. The northwestern and the eastern parts of the pre-fire LST are in red or yellow, which means they had relatively high temperatures above 310 Kelvin. The



middle part of the perimeter had lower temperatures, which is around 300 Kelvin but still higher than 295 Kelvin. The mean value of the LST is about 313.8870845 Kelvin with a standard deviation of about 5.469765104 Kelvin.

## Pre-fire Land Surface Temperature In Dixie Fire Perimeter

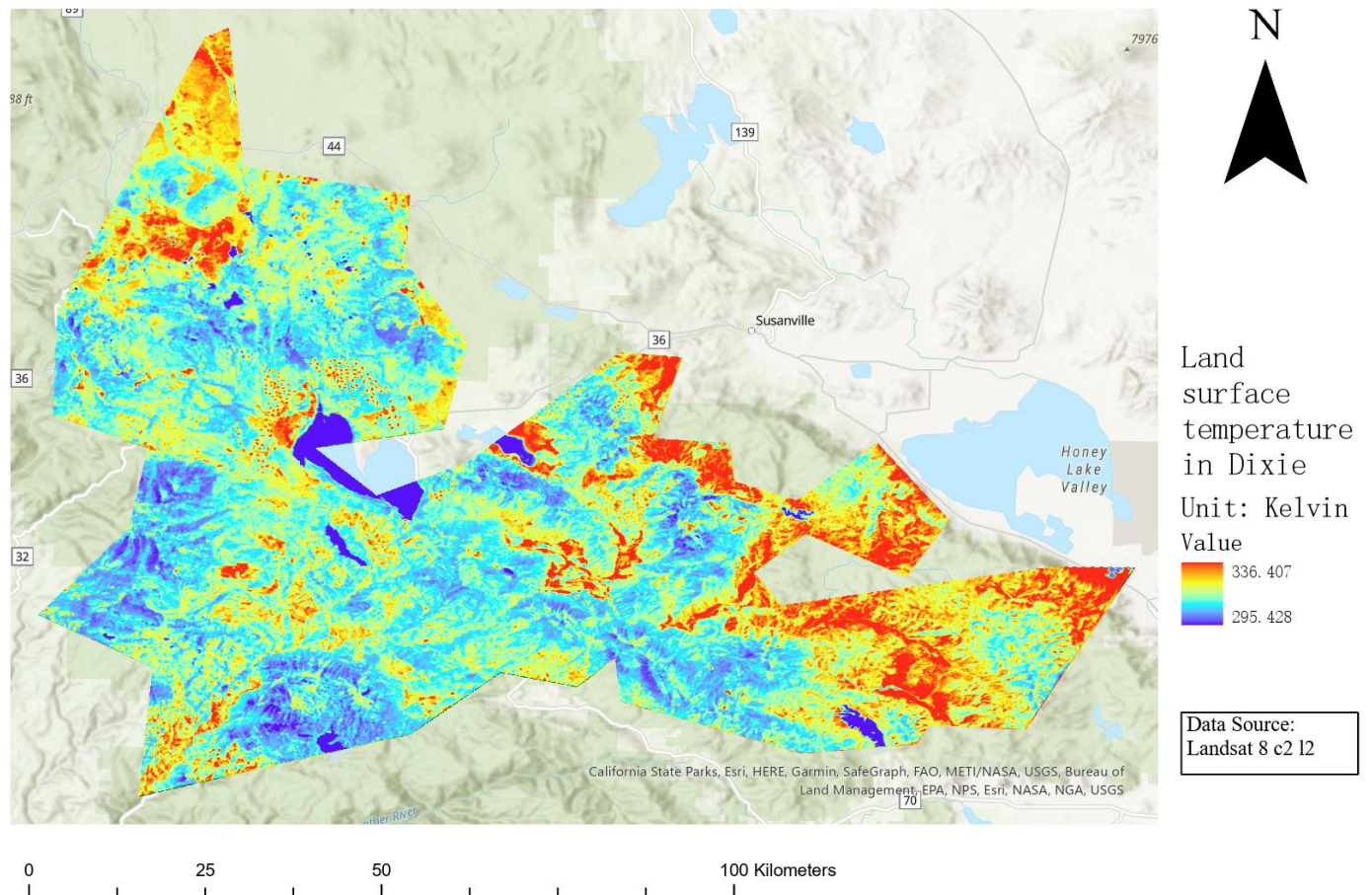


Figure 3.3 Pre-fire LST within the perimeter of Dixie

### 3.2. Socio-economic loss

The pre-fire human-built land use is shown in Figure 3.4 and the post-fire human-built land use is in Figure 3.5. According to Figure 3.4, there was an amount of obvious human-built

land use marked in red. However, after Dixie ended, the area of the human-built land use marked in red decreased significantly. Most of the existed human-built ground objects in the pre-fire map disappeared according to Figure 3.5. To be more specific, after Dixie, the human-built land use within the perimeter of Dixie decreased by about 656280900 square meters.

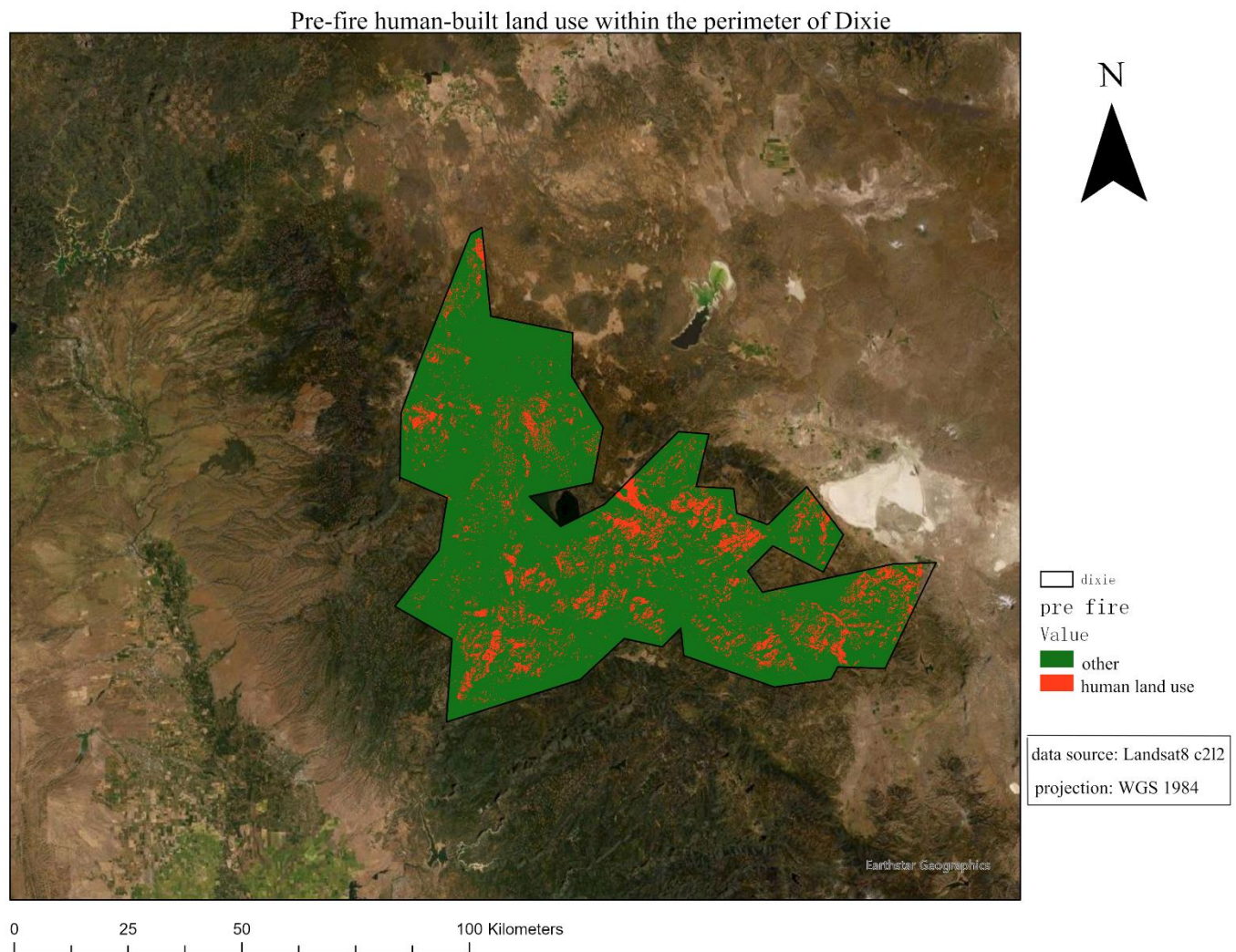


Figure 3.4 Pre-fire human-built land use within the perimeter of Dixie



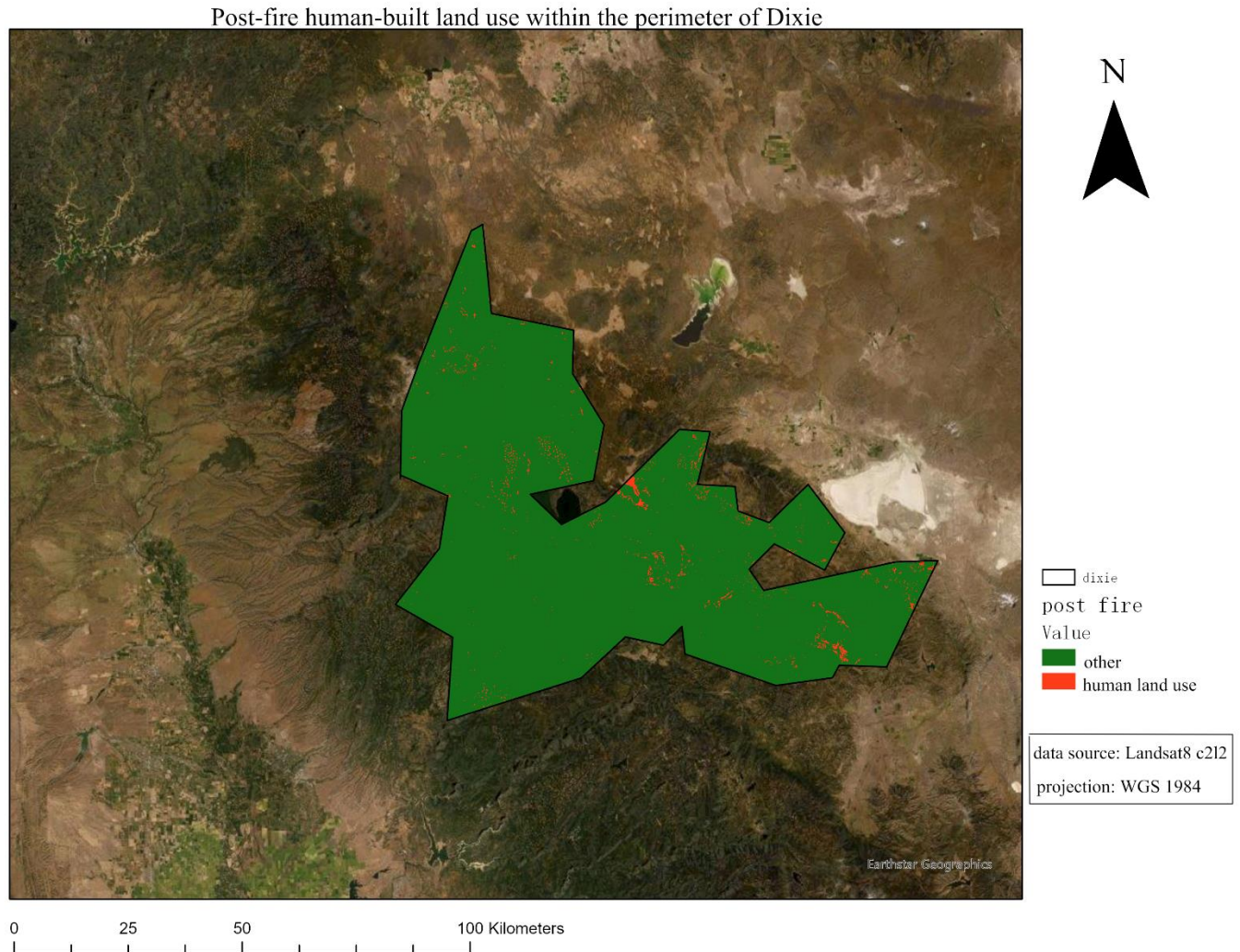


Figure 3.5 Post-fire human-built land use within the perimeter of Dixie

### 3.3. Importance ranking for the pre-fire condition factors to the estimated cost of wildfires

The frequency histogram of the estimated costs for the 20 selected wildfires is in Figure 3.6. According to the histogram, the distribution of the variate is extremely left-skewed, where there are 12 fires with costs lower than 100 million dollars. There are 5 fires with costs between 100 million and 200 million dollars. Besides those fires, 271147512 dollars were spent for fighting Caldor, 494401420 dollars were spent for fighting Sugar, and 637428216 dollars were

spent on Dixie, which were the top 3 wildfires with the highest costs that are recorded in full information in NIFC. The overall distribution of estimated cost is unimodal, but too left-skewed, which is not bell-shaped or symmetric.

The distribution of the log numbers of estimated cost is shown in Figure 3.7, which is unimodal, coarsely symmetric, and roughly bell-shaped. In the following analysis, the log number of estimate cost will replace estimated cost as the responsive variate. The interpretation will be claimed in the discussion section.

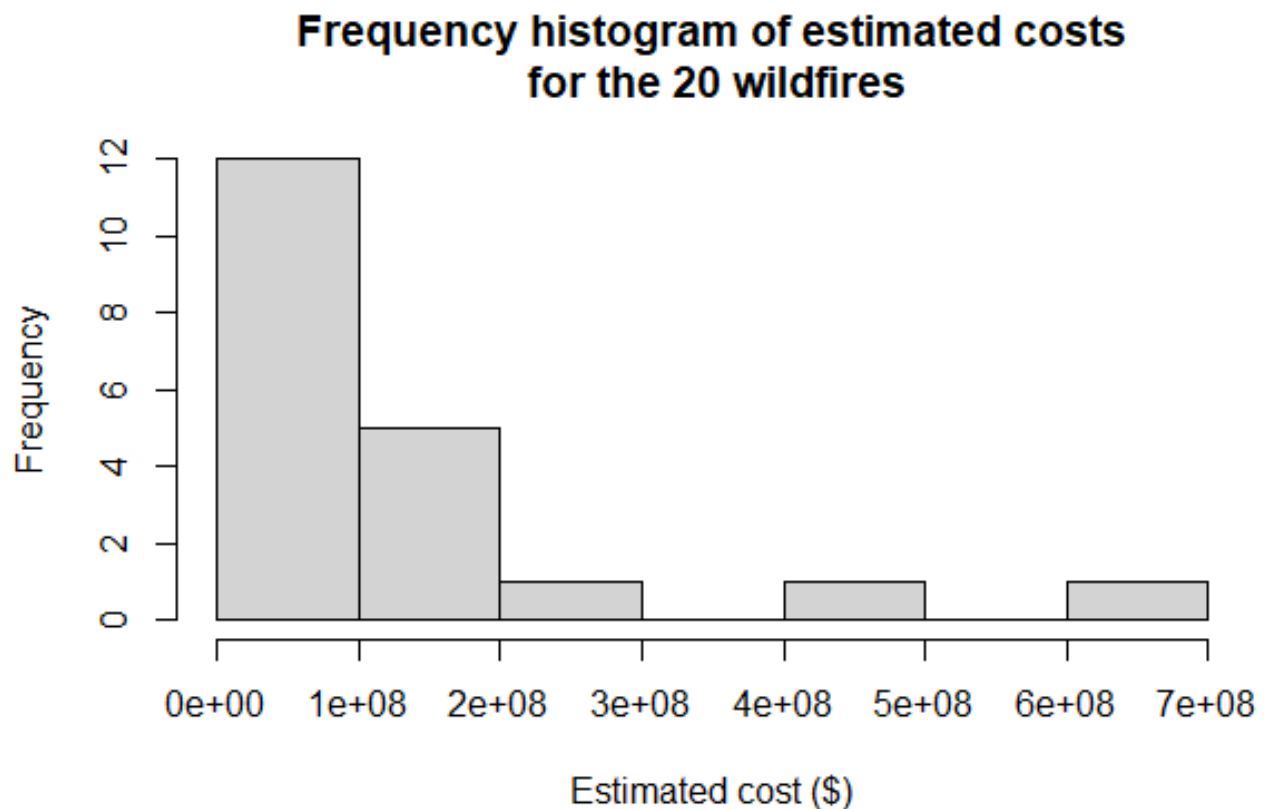


Figure 3.6 Frequency histogram of the estimated costs for the 20 selected wildfires. The data source is from the WFP dataset provided by NIFC.

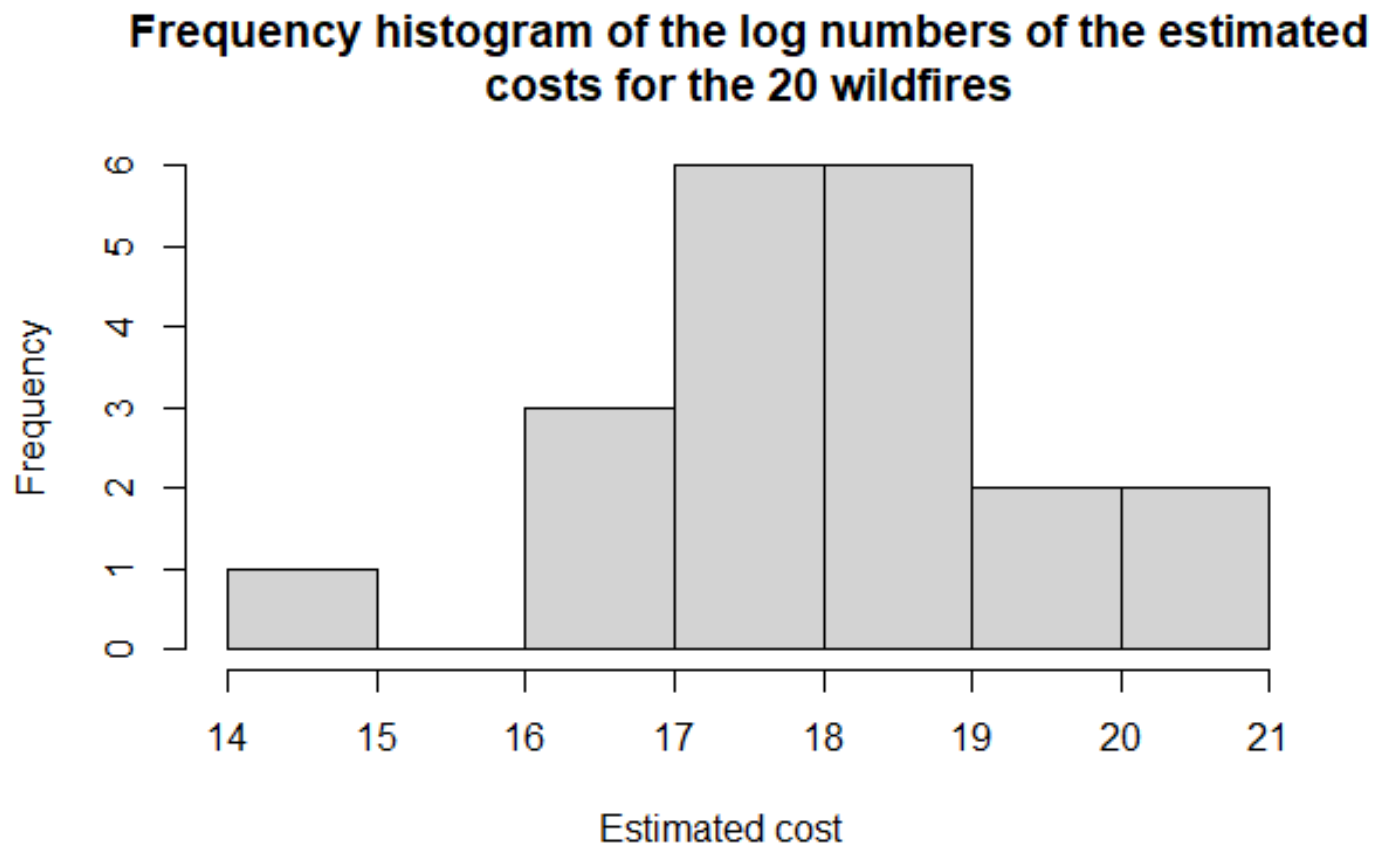


Figure 3.7 Frequency histogram of the log numbers of the estimated costs for the 20 selected wildfires

In this section, the selected results of the significance of the correlation coefficient test for estimated cost will be shown. The scatterplot matrix of the selected variates with estimated cost is demonstrated in Figure 3.8. The correlation coefficient between the standard deviation of pre-fire NDVI and estimated cost is about 0.6406396 with a p-value of 0.00234. The correlation coefficient between the mean value of pre-fire NDVI and estimated cost is about 0.5 with a p-value of 0.02517. The correlation coefficient between the standard deviation of pre-fire DEM and estimated cost is about 0.4275387 with a p-value of 0.06006. The correlation coefficient



between the pre-fire wind speed within 10 meters of the wildfire location and estimated cost is about -0.5433363 with a p-value of 0.01329. The correlation coefficient between the pre-fire wind speed within 50 meters of the wildfire location and estimated cost is about -0.5078705 with a p-value of 0.02225. In Figure 3.8, for the standard deviation of NDVI, mean value of NDVI, and DEM, the trends of the scatterplots with estimate cost are obviously from bottom-left to top-right without counting the outliers outside of the ellipse. For the scatterplot of wind speed and estimate cost, the trend is from top-left to bottom-right. For the standard deviation of NDVI, the frequency histogram is right-skewed compared with usual normal distributions. The frequency histogram of the mean value of NDVI is not bimodal, and the middle bar is even lower than the sidebars near it. Other factors, such as temperature, relative humidity, wind direction, barometric pressure, precipitation, and LST, have too small correlation coefficient with estimated cost and the p-value of the correlations are too high.

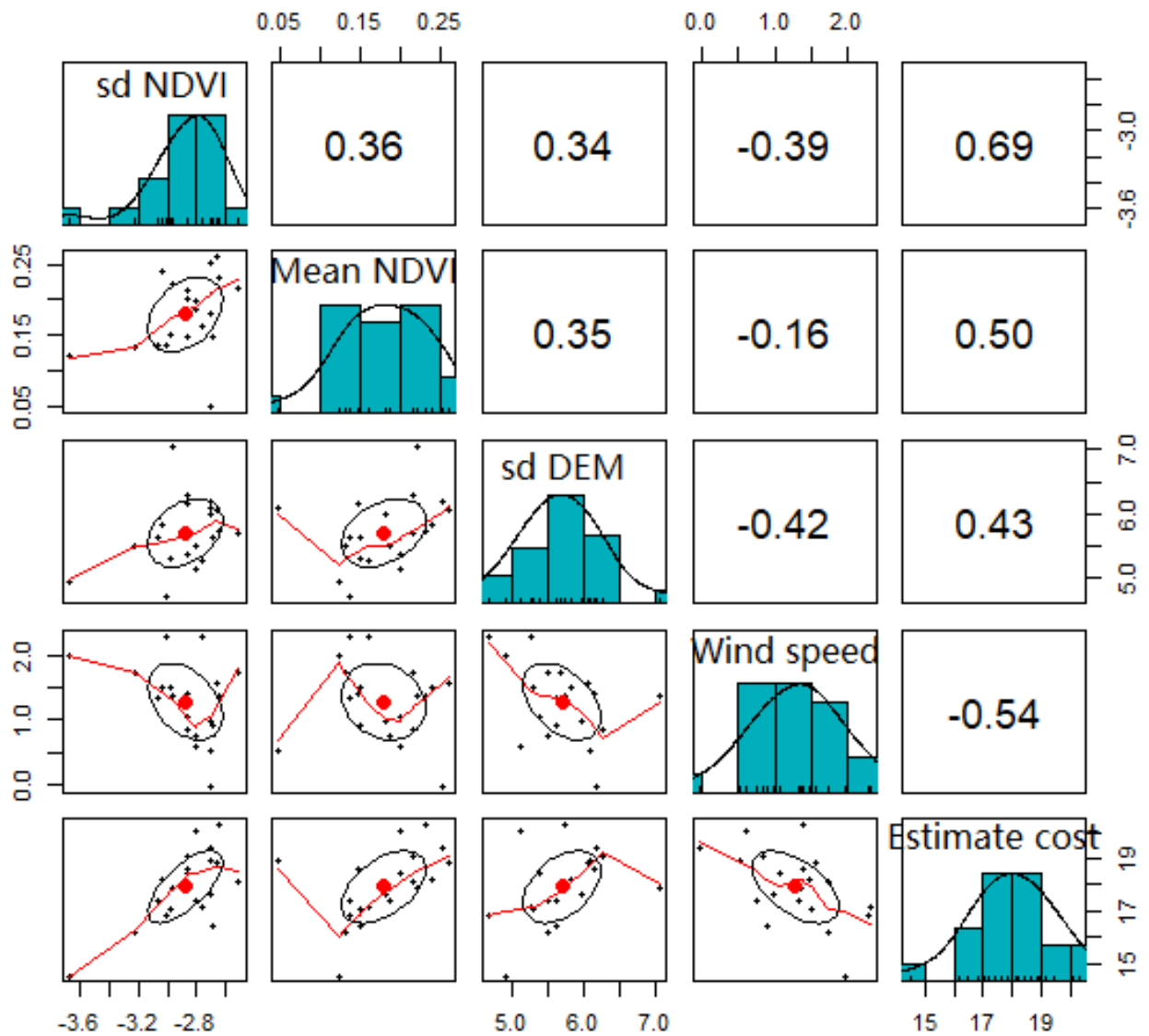


Figure 3.8 Scatterplot matrix of the variates that are strongly correlated to the estimated cost of the wildfires

After the filter method, only the five variates are kept for training a random forest model. According to the error-tree plot in Figure 3.9, the error stabilizes to the lowest level as the number of trees in the random forest increases. Therefore, the number of trees is set to 5000 as

optimal. After trying all possible parameter values for the number of variables tried at each split, 1 is set as optimal. Based on the parameters, a random forest with 37.72% variance explained and a mean of squared residuals of 1.129724 is trained. The model retunes the importance ranking among the mean value of the pre-fire NDVI, the pre-fire wind speed, the standard deviation of the pre-fire NDVI, and the standard deviation of the pre-fire DEM as shown in Figure 3.10. According to the figure, the mean value of the pre-fire NDVI is the most important variable under both criteria. The pre-fire wind speed, the standard deviation of the pre-fire NDVI, and the standard deviation of the pre-fire DEM are ranked the second, third, and fourth place respectively according to the criterion based on the MSE of the model increased by the variable. On the other hand, the standard deviation of the pre-fire DEM, the pre-fire wind speed, and the standard deviation of the pre-fire NDVI are ranked at the second, third, and fourth place respectively according to the Gini impurity criterion.

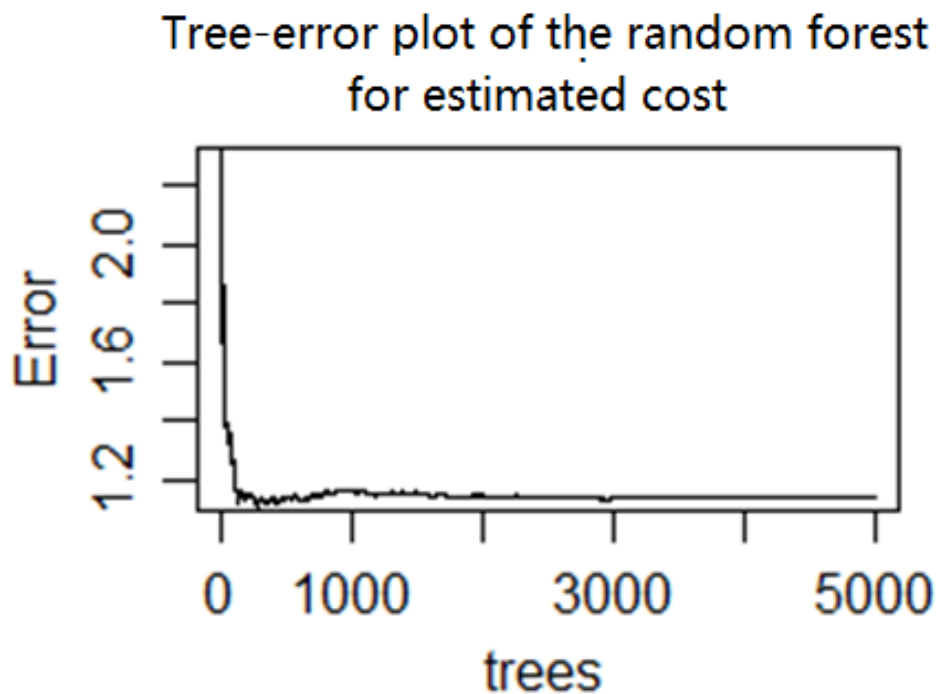


Figure 3.9 Tree-error plot of the random forest model for estimated cost

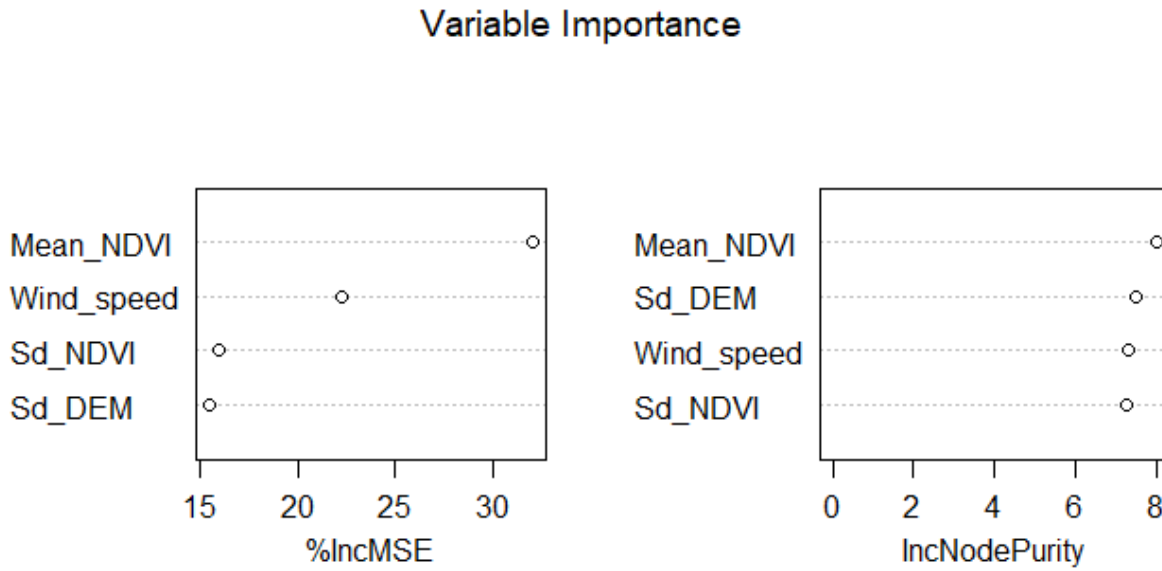


Figure 3.10 Plots of variable importance to estimated cost ranking generated by the random forest model based on both the increased MSE and increased Gini purity

### 3.4. Importance ranking for the pre-fire condition factors to the estimated cost of wildfires

The frequency histogram of the destroyed area of human-built land use for the 20 selected wildfires is in Figure 3.11. According to the histogram, there are 15 fires smaller than 100 square kilometers, where Alisal destroyed 970200 square meters, which destroyed the least area of human-built land use among the 20 wildfires. There are 3 fires that destroyed human-built land use larger than 100 square kilometers but smaller than 200 square kilometers. Besides those fires, 486.7083 square kilometers were destroyed by Doe and 656.2809 square kilometers were destroyed by Dixie, which were the wildfires that destroyed the largest area of human-built land use that are recorded in full information in NIFC.

The distribution of the log numbers of estimated destroyed human-built land use is shown in Figure 3.11, which is unimodal, coarsely symmetric, and roughly bell-shaped. In the following

analysis, the log number of estimate estimated destroyed human-built land use will replace estimated destroyed human-built land use as the responsive variate. The interpretation will be claimed in the discussion section.

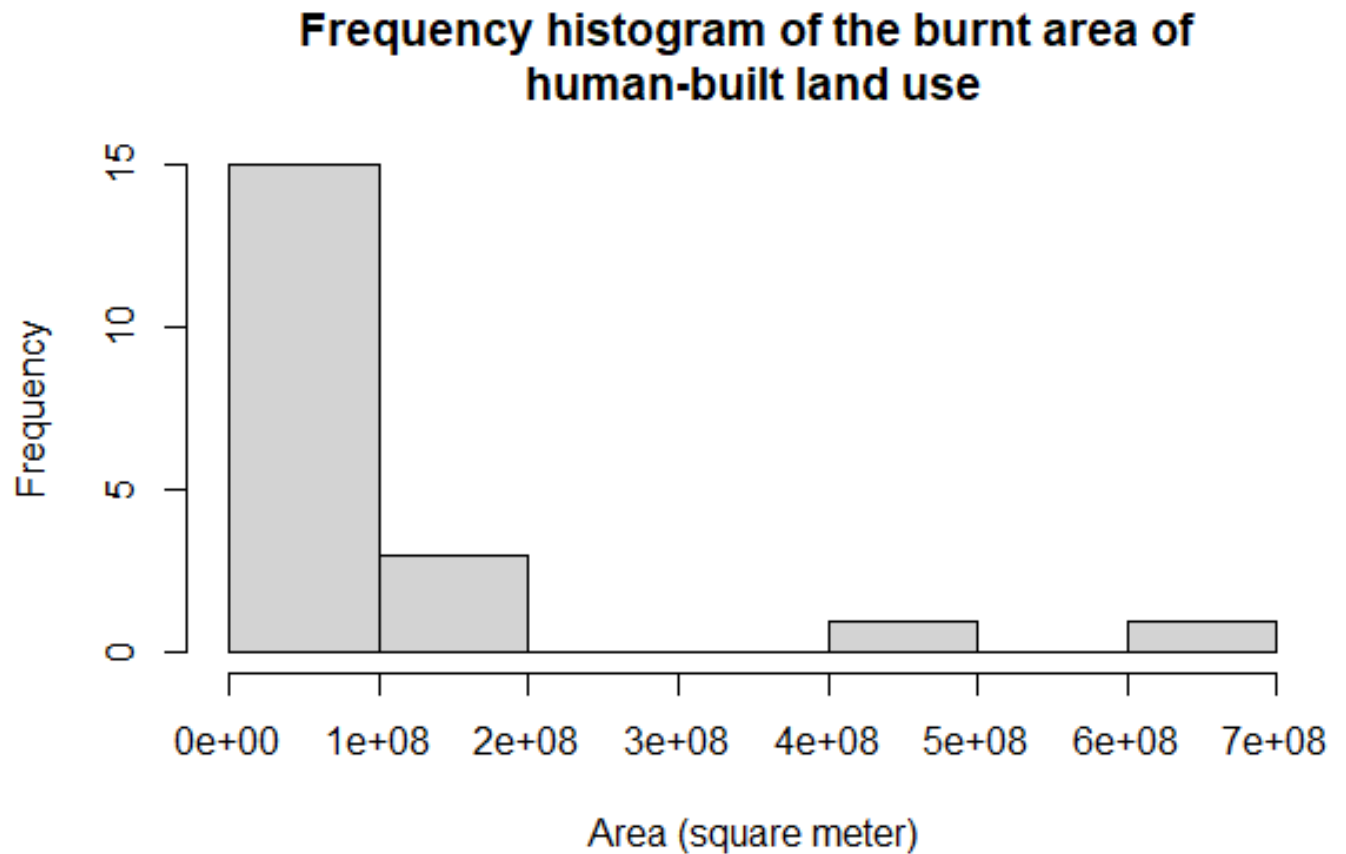


Figure 3.11 Frequency histogram of the burnt area of human-built land use of the 20 wildfires

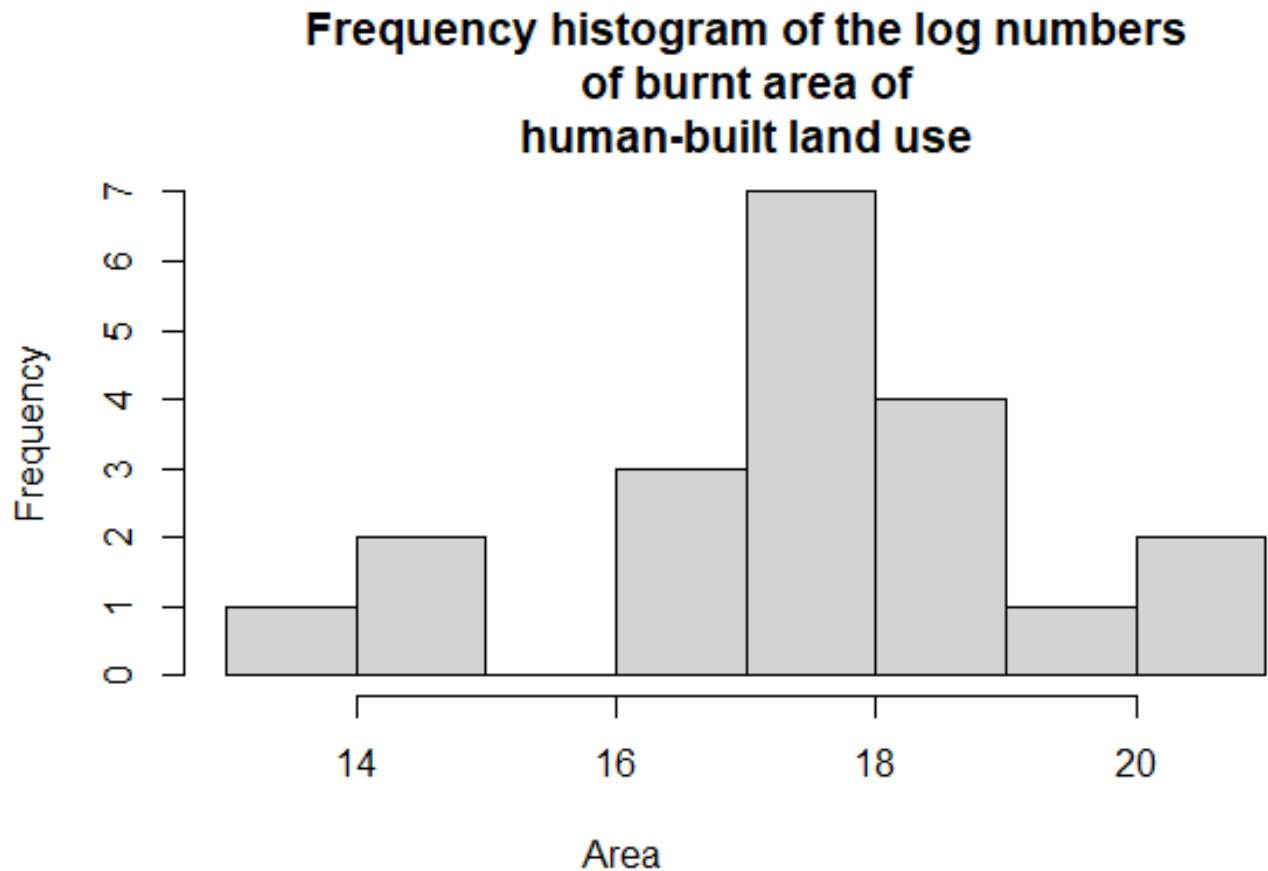


Figure 3.12 Frequency histogram of the log numbers of burnt area of human-built land use of the 20 wildfires

According to the results of the significance of the correlation coefficient test, the correlation coefficient between pre-fire air temperature within 2 meters of the wildfire location and the estimated destroyed area of human-built land use is about 0.45 with a p-value of 0.04648. The correlation coefficient between pre-fire wet-bulk temperature within 2 meters of the wildfire location and estimated destroyed area of human-built land use is about 0.39 with a p-value of 0.0907. The correlation coefficient between the mean value of DEM and estimated destroyed area of human-built land use is about 0.25 with a p-value of 0.2896. The correlation

coefficient between the standard deviation of DEM and estimated destroyed area of human-built land use is about 0.27 with a p-value of 0.248. In Figure 3.13, the scatterplot matrix of the variates is demonstrated. For the standard deviation of DEM and pre-fire air temperature within 2 meters of the wildfire location, the trends of the scatterplots with estimate cost are obviously from bottom-left to top-right. For the mean value of DEM, the trend is basically from bottom-left to top-right without counting the outlier outside of the ellipse, but the trend on the right side of the ellipse is from top-left to bottom-right. For pre-fire wet-bulk temperature within 2 meters of the wildfire location, the trend of the scatterplot is basically from bottom-left to top-right without counting the outlier outside of the ellipse, but the trends at both sides are from top-left to bottom-right. Other factors that are not mentioned in this section have too small correlation coefficient with estimated cost and the p-value of the correlations are too high.

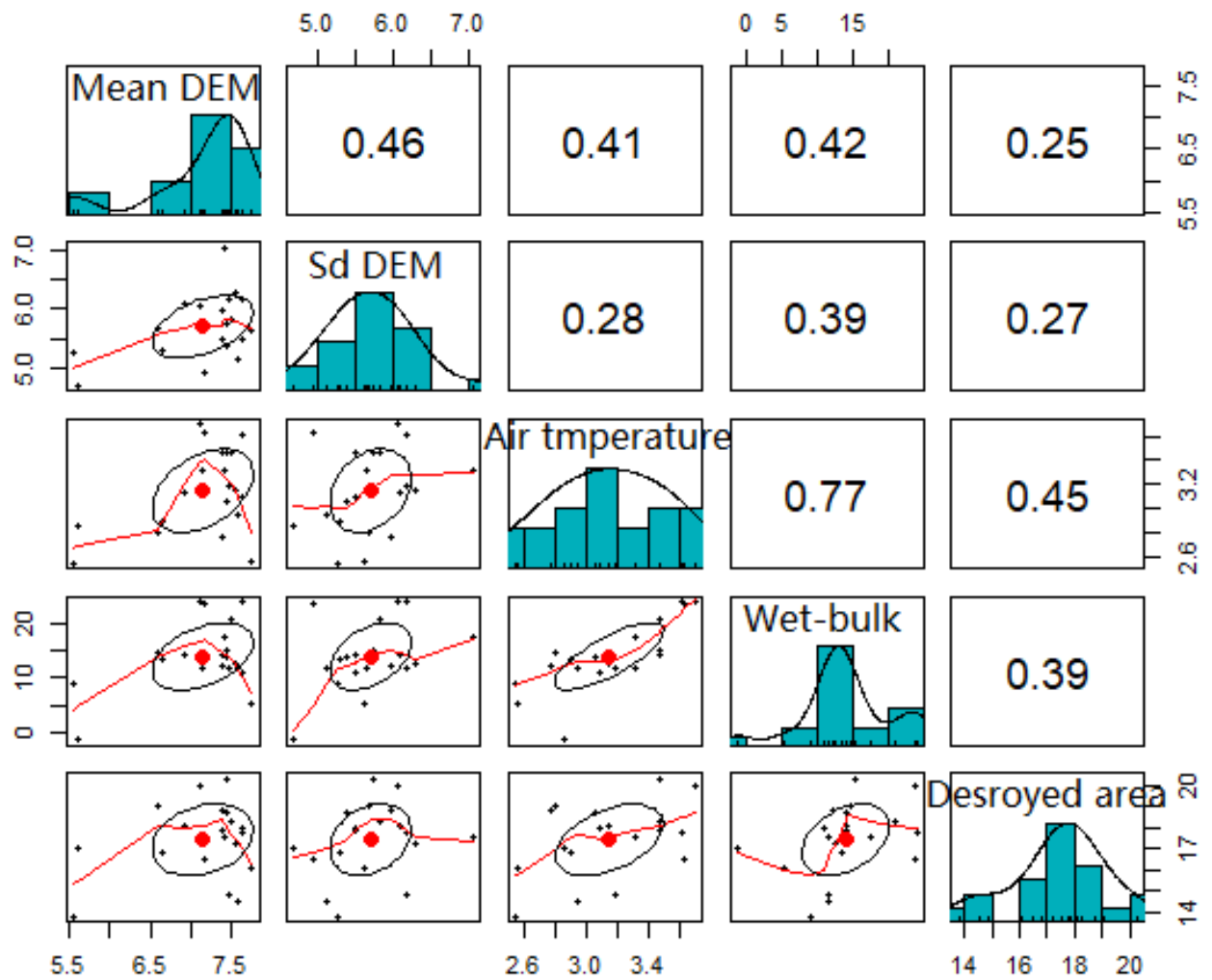


Figure 3.13 Scatterplot matrix of the variates that are strongly correlated to the estimated destroyed area of human-built land use

After the filter method, only the five variates are kept for training a random forest model. According to the error-tree plot in Figure 3.14, the error stabilizes to the lowest level as the number of trees in the random forest increases. Therefore, the number of trees is set to 5000 as optimal. After trying all possible parameter values for the number of variables tried at each split,



1 is set as optimal. Based on the parameters, a random forest with 12.46% variance explained and a mean of squared residuals of 2.385962 is trained. The model retunes the importance ranking among pre-fire air temperature within 2 meters of the wildfire location, pre-fire wet-bulk temperature within 2 meters of the wildfire location, the mean value of DEM, and the standard deviation of DEM as shown in Figure 3.15. According to the figure, the standard deviation of DEM and pre-fire wet-bulk temperature are evaluated as the most important variables based on the MSE of the model increased by the variables, whose scores are larger than 20. The importance score of pre-fire air temperature within 2 meters of the wildfire location and the mean value of DEM based on the MSE of the model increased by the variables are negative or close to 0. Hence, the two variables are not regarded as important as the standard deviation of DEM and pre-fire wet-bulk temperature to the estimated destroyed human-built area. On the other hand, pre-fire wet-bulk temperature within 2 meters of the wildfire location, the standard deviation of DEM, pre-fire air temperature within 2 meters of the wildfire location, and the mean value of DEM are ranked at the first, second, third, and fourth place respectively according to the Gini impurity criterion. And the importance scores are close.

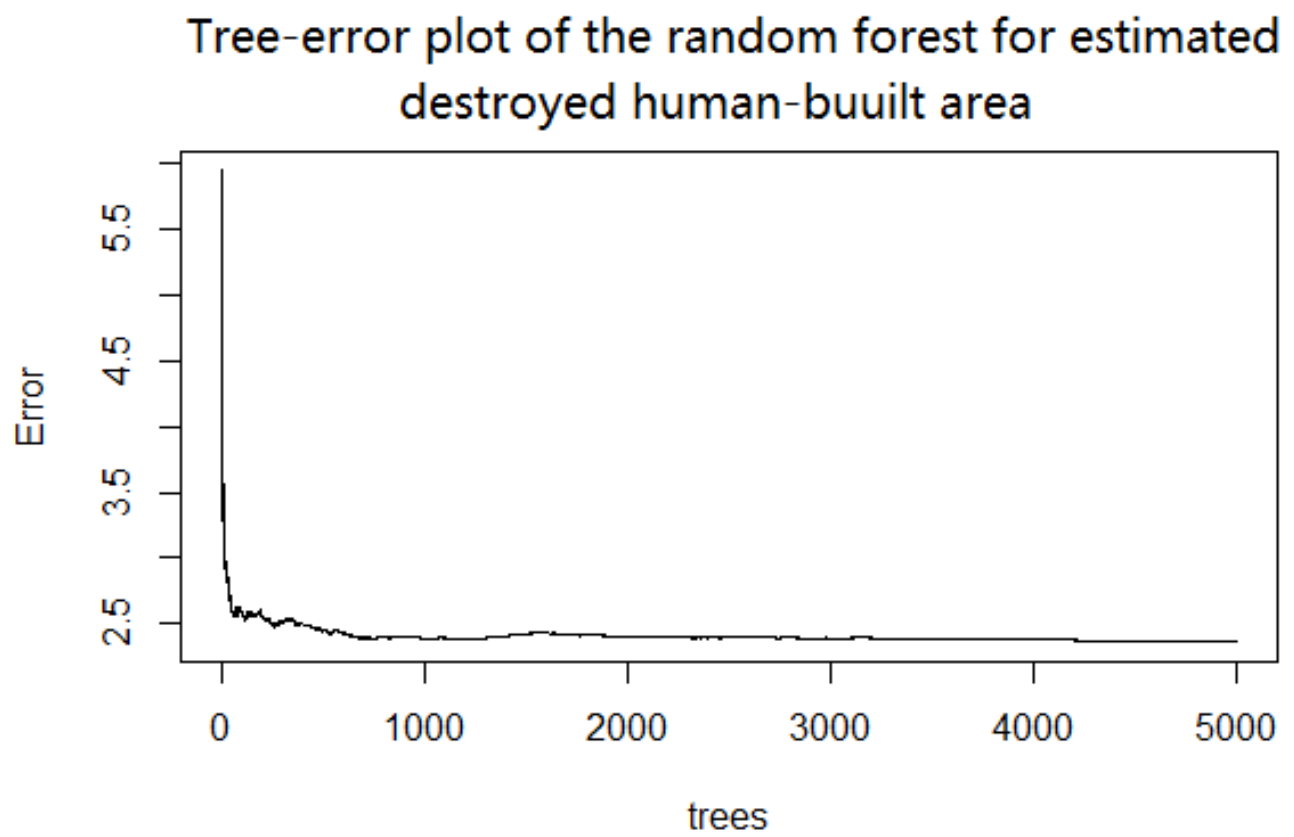


Figure 3.14 Tree-error plot of the random forest model for estimated destroyed human-built area

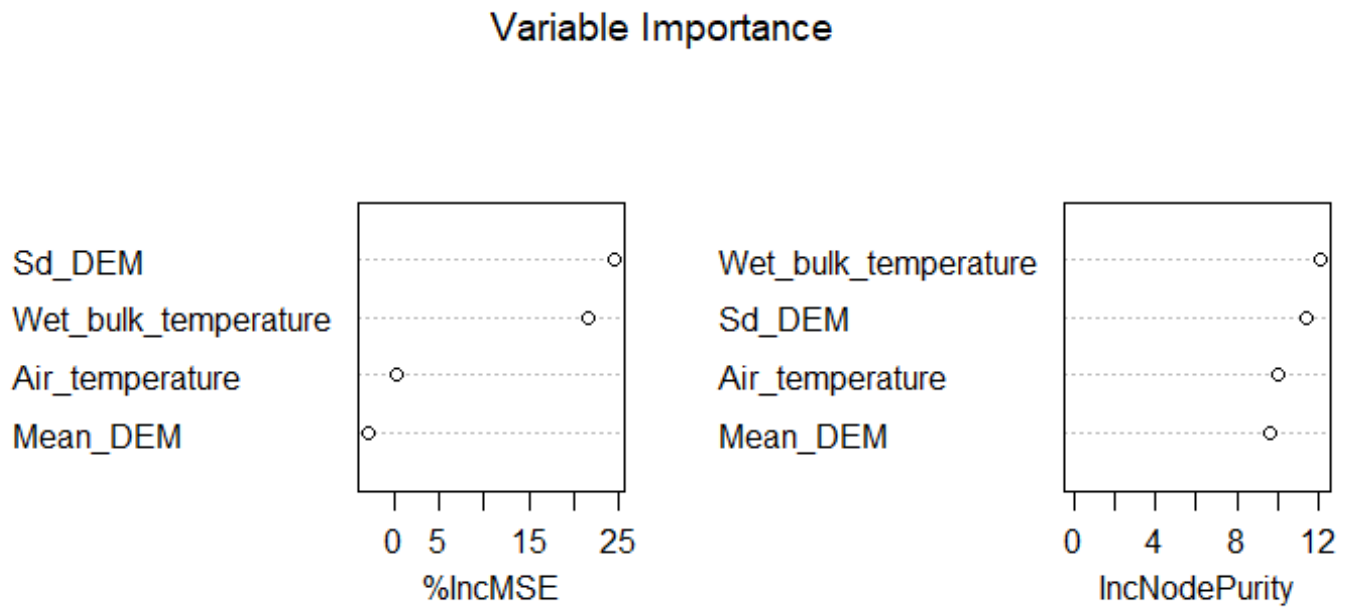


Figure 3.15 Plots of variable importance to estimated destroyed human-built area ranking generated by the random forest model based on both the increased MSE and Gini impurity

## 4. Discussion

In this section, the results shown in the previous section will all be analyzed. The information that can be derived from the pre-fire condition maps of Dixie as the example will be first analyzed. Moreover, the results of the feature selection and explanatory variable importance ranking on the estimated cost and estimated area of destroyed human-built land use will be interpreted respectively.

### 4.1. Interpretation of the results from example Dixie

In Figure 3.1, the pre-fire vegetation coverage map within the perimeter of Dixie shows that the vegetation coverage was large and took up a major part of the perimeter, which indicates the proportion of vegetation in the perimeter just before Dixie was very high. The mean value of the NDVI layer is 0.23067, which is higher than 0.2, so the land cover is vegetation dominated

and the vegetation is very dense on average within the perimeter. The dense and big amount of vegetation gave Dixie adequate fuel to burn, which might be the main reason why Dixie caused so much socio-economic loss. Moreover, in Figure 4.1, the distribution of the NDVI values of the pixels within the perimeter of Dixie are roughly unimodal, symmetric, and bell-shaped, which indicates that the NDVI values are normally distributed. Hence, according to the property of normal distributions, about 68% of pixels would have NDVI values within the range from 0.15237 and 0.30897 (Lee and Kim 2008) (b) based on the fact that the NDVI has a standard deviation of 0.0783 and a mean value of 0.23067. The high variation of the vegetation land cover might also contribute to the duration length and the difficulty of the firefighting, which increases the socio-economic loss. Therefore, the high density, volume, and enough variation of the vegetation within the perimeter of Dixie indeed gave the adequate prerequisite for the extremely terrible severity of Dixie theoretically.

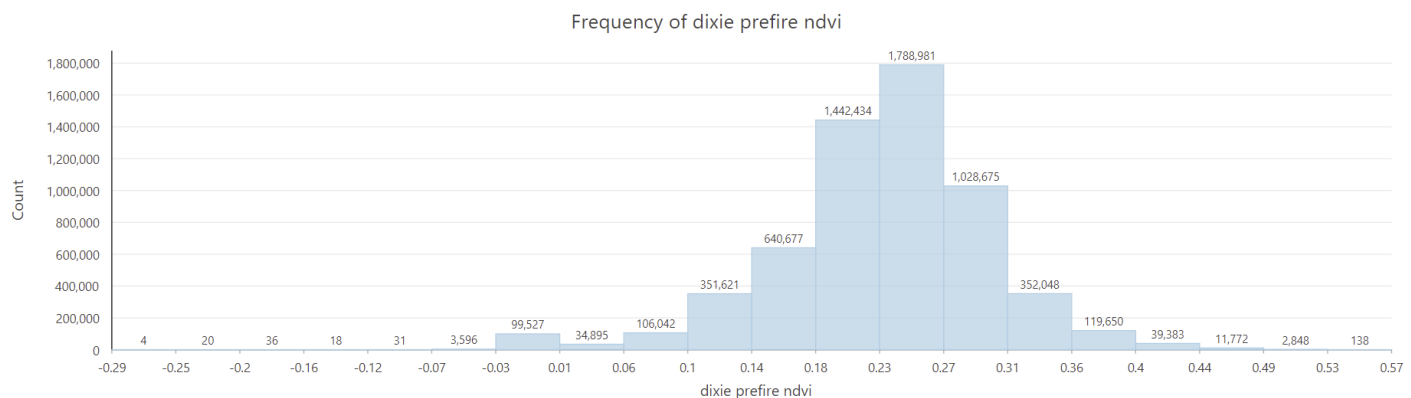


Figure 4.1 Frequency histogram of the pre-fire NDVI values within the perimeter of Dixie

According to a report published by Center for Wilderness Safety (CWS), the content of oxygen will decrease as the elevation increases due to the lower air pressure at higher elevation levels. Table 4.1 shows the oxygen content by altitude. In Figure 3.2, the DEM of the perimeter

of Dixie shows that the elevation of the perimeter lies within the range from 430 meters and 3200 meters, where the air at an elevation of 430 meters will have the most adequate oxygen that is very close to the oxygen content when the elevation is 0 meter. The air at an elevation of 3200 meters will still have about 14% oxygen content, which is much higher than the 6.9% that is the oxygen of the air at some plateau. The northwestern part and the southeastern part of the perimeter are at elevations between 1600 meters and 3000 meters, which are the highest parts in the perimeter of Dixie. The elevations of the two areas are categorized into moderate or high altitudes. In Table 4.1, the air of those areas will approximately contain oxygen between 14% and 17%. The middle part and the southwestern part of the perimeter are almost below 1000 meters but higher than 430 meters. Those elevations in those areas are considered low or moderate altitudes. In Table 4.1, the oxygen content in the air at those elevations is expected to be between about 18% and 20%. The mean value of the DEM layer is about 1677.29082 meters, so the average elevation of the perimeter can be categorized as moderate altitude, where the air contains oxygen content of around 17%. That being said, most of the area within the perimeter of Dixie has an elevation between 430 meters and 3000 meters, where the oxygen content of the air is between 14% and 20%, which can be considered adequate according to the report (Center for Wilderness Safety, 2022). Additionally, the standard deviation of the elevation of the perimeter of Dixie is about 309.09423 meters, which is quite high compared to the mean value of 1677.29082 meters. Therefore, the variation of the DEM within the perimeter is extremely significant, which corresponds to the topography of the place (Lee and Kim, 2008) (c). In particular, the topography within the perimeter is quite steep, which might add difficulty to the firefighting, since steep topography is hard for firefighters and firefighting equipment to proceed and access (Page et al., 2018). Therefore, the moderately high elevation within the perimeter of

Dixie contains adequate oxygen for Dixie to burn, and the steep typography might suppress the efficiency of firefighting and adds the cost of firefighting theoretically.

Altitude (ft)	Altitude (m)	Effective O2 %	Altitude Category	Example
0 ft	0 m	20.9 %	Low Altitude	Sea Level
1000 ft	305 m	20.1 %	Low Altitude	
2000 ft	610 m	19.4 %	Low Altitude	
3000 ft	914 m	18.6 %	Moderate Altitude	
4000 ft	1219 m	17.9 %	Moderate Altitude	
5000 ft	1524 m	17.3 %	Moderate Altitude	Boulder, CO (5328 ft)
6000 ft	1829 m	16.6 %	Moderate Altitude	Mt. Washington (6288 ft)
7000 ft	2134 m	16.0 %	Moderate Altitude	
8000 ft	2438 m	15.4 %	High Altitude	Aspen, CO (8000 ft)
9000 ft	2743 m	14.8 %	high Altitude	
10,000 ft	3048 m	14.3 %	High Altitude	
11,000 ft	3353 m	13.7 %	High Altitude	Mt. Phillips (11,711 ft)

Table 4.1 Oxygen Levels by Altitude. The table is created by CWS in 2022.

According to the pre-fire LST map in Figure 3.3, the LST lay within the range between 295.428 Kelvin and 336.407 Kelvin. Even for the lowest LST, 295.428 Kelvin is a moderately high temperature. The northwestern and the eastern parts of the pre-fire LST had relatively high temperatures above 310 Kelvin. The middle part of the perimeter had a lower temperature around 300 Kelvin but still higher than 295 Kelvin. The mean value of the LST is about 313.8870845 Kelvin, which is an extremely high temperature. Therefore, the very high LST might indeed promote the occurrence and the severity of Dixie theoretically. Also, the standard deviation of the pre-fire LST was about 5.469765104, which is relatively small compared to the

mean value of 313.8870845 Kelvin. Hence, the pre-fire LST within the perimeter of Dixie was stably high. Therefore, the stably high pre-fire LST within the perimeter of Dixie was expected to lead to the occurrence of the severe wildfire.

#### **4.2. Feature selection and importance ranking on estimated cost**

According to the frequency histogram of the estimated costs for the 20 selected wildfires is in Figure 3.6, 12 fires with costs lower than 100 million dollars. There are 5 fires with costs between 100 million and 200 million dollars. 271147512 dollars were spent for fighting Caldor, 494401420 dollars were spent for fighting Sugar, and 637428216 dollars were spent for Dixie, which were the top 3 wildfires with the highest costs that are recorded in full information in NIFC. The 20 wildfires all occurred and spread on the west coast of the US. Hence, wildfires on the west coast of the US indeed caused too much socio-economic loss by too expensive firefighting.

According to Figure 3.7, the distribution of estimated cost is too left-skewed, which violates the assumption of using the Pearson correlation coefficient that both variates need to be normally distributed. Hence, according to the methodology section, some increasing functions might help to normalize the variate to have a normal distribution. After taking the log numbers of the variate, the frequency histogram of the logarithmized estimated costs for the 20 selected wildfires is demonstrated in Figure 3.7. In this case, the distribution is unimodal, coarsely symmetric, and roughly bell-shaped. Hence, the log function of the estimated cost variate is well normally distributed. Therefore, the log number of the estimated costs is the replacement of the estimated costs for Pearson correlation coefficient.

According to the results of the significance of the correlation coefficient test, the correlation coefficient between the standard deviation of pre-fire NDVI and estimated cost is

about 0.6406396, which is much larger than 0 and close to 1. Hence, the correlation between the standard deviation of pre-fire NDVI and estimated cost is positive and intensive. The p-value of the significance of the correlation coefficient test between the standard deviation of pre-fire NDVI and estimated cost is 0.00234, which is smaller than 0.05. Thus, there is a significant positive correlation between the standard deviation of pre-fire NDVI and estimated cost, which means the higher the standard deviation of pre-fire NDVI is, the higher the estimated cost is expected to be. The correlation coefficient between the mean value of pre-fire NDVI and estimated cost is about 0.5, which is positive and much larger than 0, so there is an obvious positive correlation between the mean value of pre-fire NDVI and estimated cost. Also, the p-value of the significance of the correlation coefficient test between the mean value of pre-fire NDVI and estimated cost is 0.02517, which is smaller than 0.05. Thus, there exists a significant positive correlation between the mean value of pre-fire NDVI and estimated cost, which means the higher the mean value of pre-fire NDVI is, the higher the estimated cost is expected to be. Similarly, since the correlation coefficient between the standard deviation of pre-fire DEM and estimated cost is about 0.4275387, which is positive and much larger than 0, the correlation between the standard deviation of pre-fire DEM and estimated cost is obvious and positive. Also, since the p-value of the significance of the correlation coefficient test between the standard deviation of pre-fire DEM and estimated cost is 0.06006, which is close to 0.05 and much smaller than 0.1, there exists a significant positive correlation between the standard deviation of pre-fire DEM and estimated cost, which means the higher the standard deviation of pre-fire DEM is, the higher the estimated cost is expected to be. The correlation coefficient between the pre-fire wind speed within 10 meters of the wildfire location and estimated cost is about -0.5433363, which is negative and much smaller than 0, so the correlation between the pre-fire



wind speed within 10 meters of the wildfire location and estimated cost is obvious and negative. Also, since the p-value of the significance of the correlation coefficient test between the pre-fire wind speed within 10 meters of the wildfire location and estimated cost is 0.01329, which is smaller than 0.05, there exists a significant negative correlation between the pre-fire wind speed within 10 meters of the wildfire location and estimated cost, which means the higher the pre-fire wind speed within 10 meters of the wildfire location is, the lower the estimated cost is expected to be. For other variates, the correlation between each of the variates and estimated cost is too insignificant because the coefficient is too close to 0 and the p-value of the test is too large. Hence, the variates are not considered to be correlated to estimated cost. In that case, the variates are filtered out by the filter method, which will not be involved in the training set of the random forest model. Therefore, in light of the data on the 20 wildfires, more diverse and denser vegetation might lead to higher costs since higher standard deviation and mean value of the wildfire are likely to increase the cost of the wildfire as claimed before. Steeper topography around where a wildfire occurs is likely to increase the cost of the wildfire since a higher standard deviation leads to a higher estimated cost based on the observed data in this study. Finally, stronger wind around where the wildfire occurs might help decrease the cost of the wildfire since the higher the pre-fire wind speed within 10 meters of the wildfire location is, the lower the estimated cost is expected to be according to the data analysis in this project.

After the filter method, only four pre-fire condition variates are significantly correlated to estimated cost, which are the standard deviation of pre-fire NDVI, the mean value of pre-fire NDVI, the standard deviation of pre-fire DEM, and the pre-fire wind speed within 10 meters of the wildfire location. Thus, only these variates are input as the features in the random forest. After the first attempt to train the random forest, the error-tree plot in Figure 3.9 will be

generated, where the x-axis is the number of decision trees trained in the random forest model, and the y-axis is the MSE of the model, where lower MSE means higher accuracy of the model (Ghosal and Hooker, 2020) (a). As the number of decision trees increases, the means squared error (MSE) will decrease, and finally stabilize at an MSE around 1.13. Thus, adding more decision trees trained in the model will make the model more accurate according to the trend in Figure 3.9. Furthermore, since the training set of the model is very small, which only includes 20 observations and 5 variates, so setting a large number of decision trees won't be too expensive. Therefore, 5000 is set for the number of trained decision trees of the model as optimal. After fixing the number of trees in the model, all possible attempts for different parameter values for the number of variables tried at each split are conducted. Finally, 1 is set as optimal as the number of variables tried at each split since with that parameter value, the model has the highest accuracy. Therefore, the optimal random forest model is trained using the optimal parameters based on the training dataset. The model has a 37.72% variance explained, which is a high enough value for an accurate model in some research (Pang et al., 2006) (a). Hence, the results generated by the model should be considered robust enough.

By the results of the importance ranking among the mean value of the pre-fire NDVI, the pre-fire wind speed, the standard deviation of the pre-fire NDVI, and the standard deviation of the pre-fire DEM as shown in Figure 3.10, the mean value of the pre-fire NDVI is the most important variable under both criteria. Hence, taking the standard deviation of the pre-fire DEM into account will increase the most more accuracy to the model, so the variate is the most important one under the MSE-increase criterion. Also, taking the standard deviation of the pre-fire DEM in the model will increase the highest Gini purity to the model as well, which means the nodes of the trees will be split better, which means higher accuracy for classification.

However, since the project is training the random forest for regression, the index is just a reference and will be inappropriate compared to the MSE-increase criterion. For simplicity, the following discussion won't include Gini purity explicitly. The pre-fire wind speed, the standard deviation of the pre-fire NDVI, and the standard deviation of the pre-fire DEM are ranked the second, third, and fourth place respectively according to the criterion based on the MSE of the model increased by the variable, which means pre-fire wind speed, the standard deviation of the pre-fire NDVI, and the standard deviation of the pre-fire DEM are the second, third, and fourth most important variates in terms of adding the accuracy of the model. Since the importance of the mean value of the pre-fire NDVI, the pre-fire wind speed, the standard deviation of the pre-fire NDVI, and the standard deviation of the pre-fire DEM means how much variance of estimated cost they can explain, the importance of the pre-fire condition factors corresponding to the variates should also follow (Genuer et al., 2010). Thus, the average level of the pre-fire vegetation within the perimeter is much more important to the cost of the wildfire than other pre-fire condition factors. The pre-fire wind speed has moderate importance to the cost of the wildfire. The steepness of the topography and the diversity of the pre-fire vegetation have also a decent amount of importance to the cost of the wildfire according to Figure 3.10 but are ranked as the last two.

#### **4.3. Feature selection and importance ranking on estimated destroyed human-built area**

According to the frequency histogram of the estimated destroyed human-built land use for the 20 selected wildfires is in Figure 3.11, Alisal destroyed 970200 square meters, which is even the smallest area of human-built land use the 20 selected wildfires destroyed but still very expensive. For the top 2 severe wildfires in terms of the area of destroyed human-built land use, 486.7083 square kilometers were destroyed by Doe and 656.2809 square kilometers were

destroyed by Dixie, which are very large areas of human-built land use. The 20 wildfires all occurred and spread on the west coast of the US. Hence, wildfires on the west coast of the US indeed caused too much socio-economic loss by destroying too much area of human-built land use.

According to Figure 3.11, the distribution of estimated destroyed human-built land use is too left-skewed, which violates the assumption of using the Pearson correlation coefficient that both variates need to be normally distributed. Hence, according to the methodology section, some increasing functions might help to normalize the variate to have a normal distribution. After taking the log numbers of the variate, the frequency histogram of the logarithmized estimated destroyed human-built land use for the 20 selected wildfires is demonstrated in Figure 3.12. In this case, the distribution is unimodal, coarsely symmetric, and roughly bell-shaped. Hence, the log function of the estimated cost variate is well normally distributed. Therefore, the log number of the estimated costs is the replacement of the estimated costs for the Pearson correlation coefficient.

According to the results of the significance of the correlation coefficient test, the correlation coefficient between pre-fire air temperature within 2 meters of the wildfire location and the estimated destroyed area of human-built land use is about 0.45, which is positive and much larger than 0, so the correlation between pre-fire air temperature within 2 meters of the wildfire location and the estimated destroyed area of human-built land use is obvious and positive. The p-value of the correlation test between them is 0.04648, which is smaller than 0.05, so pre-fire air temperature within 2 meters of the wildfire location is significantly correlated to the estimated destroyed area of human-built land use significantly. The correlation coefficient between pre-fire wet-bulk temperature within 2 meters of the wildfire location and estimated

destroyed area of human-built land use is about 0.39, which is positive and much larger than 0, so the two variates are positively correlated. Even though the p-value of the correlation test between them is 0.0907, which is larger than 0.05. However, since pre-fire wet-bulk temperature within 2 meters of the wildfire location has already been the variate with the least p-value of the correlation with estimated destroyed area of human-built land use, and the p-value is smaller than 0.1, which is an acceptable p-value, the variate will be regarded as having some positive correlation with estimated destroyed area of human-built land use. For other variates, the correlation coefficients with estimated destroyed area of human-built land use are too small, and the p-values are too large. Hence, other variates are not considered to have a strong correlation with estimated destroyed area of human-built land use. However, 15 variates are filtered out by the Pearson correlation coefficient, which is too many. It would be much more reasonable to keep some extra variates with correlations coefficients that are not too low, even though the p-value might be very high. In that case, the mean value of DEM and the standard deviation of DEM are kept for the compromise. The correlation coefficient between the mean value of DEM and estimated destroyed area of human-built land use is about 0.25 with a p-value of 0.2896. The correlation coefficient is much larger than 0 but is still far from 0.5. Hence, there might be a weak positive correlation between the mean value of DEM and estimated destroyed area of human-built land use. Also, since the p-value is too large compared to 0.05, there is no evidence showing that the correlation is significant. However, by seeing Figure 3.13, the trend of the scatterplot between the mean value of DEM and estimated destroyed area of human-built land use is basically from bottom-left to top-right without counting the outlier outside of the ellipse, but the trend on the right side of the ellipse is from top-left to bottom-right. Thus, the outliers at the right side of the ellipse are the main reason why the correlation between the mean value of

DEM and estimated destroyed area of human-built land use is weak. However, since the Pearson correlation coefficient assumes a linear relationship between variates, which can hardly tolerate outliers, there might still be some relationships between the mean value of DEM and estimated destroyed area of human-built land use under some non-linear relationship (Sedgwick, 2012) (b). Therefore, the variate is kept for the following random forest model since random forest models handle non-linear relationships well according to section 2.6.3. The correlation coefficient between the standard deviation of DEM and estimated destroyed area of human-built land use is about 0.27, which is larger than 0 but still smaller than 0.5, so there might be a weak positive correlation between them. The p-value of the correlation test is 0.248, which is much larger than 0.05, the correlation between the standard deviation of DEM and estimated destroyed area of human-built land use is about 0.27 is weakly positive and insignificant. However, by seeing Figure 3.13, the trend of the scatterplot between the standard deviation of DEM and estimated destroyed area of human-built land use is obviously from bottom-left to top-right. The main reason why the correlation is weak and insignificant is that the outliers outside of the ellipse are distributed too dispersedly, which do not concentrate on a straight line. Thus, the assumption of Pearson correlation that the variates must have a linear relationship is violated. In that case, due to the trend and the relatively high coefficient, the standard deviation of DEM will be kept for the next random forest model. For other factors, their correlation coefficients with estimated destroyed area of human-built land use are too close to 0, so the correlations are too weak. Also, the p-values for the correlation tests are too large, so the correlations are insignificant. In other words, those variates are not considered to be correlated with estimated destroyed area of human-built land use. Therefore, in light of the data, higher pre-fire air temperature around 10 meters of where the wildfire occurs might lead to larger area of destroyed human-built land use.

Also, the higher the wet-bulk temperature around the wildfire location is, the larger the area of destroyed human-built land use during the wildfire is likely to be.

After the filter method, only four pre-fire condition variates are kept based on the intensity of their correlations with estimated destroyed area of human-built land use, which are pre-fire air temperature within 2 meters of the wildfire location, pre-fire wet-bulk temperature within 2 meters, the mean value of DEM, and the standard deviation of DEM. Thus, only these variates are input as the features in the random forest. After the first attempt to train the random forest, the error-tree plot in Figure 3.14 will be generated, where the x-axis is the number of decision trees trained in the random forest model, and the y-axis is the MSE of the model, where lower MSE means higher accuracy of the model (Ghosal and Hooker, 2020) (b). As the number of decision trees increases, the MSE will decrease, and finally stabilize at a MSE around 2.4. Thus, adding more decision trees trained in the model will make the model more accurate according to the trend of Figure 3.14. As previously discussed, a large number of decision trees won't be too expensive due to the size of the training set used in the project. Hence, 5000 is set for the number of trained decision trees of the model as optimal. After fixing the number of trees in the model, all possible attempts for different parameter values for the number of variables tried at each split are conducted. Finally, 1 is set as optimal as the number of variables tried at each split since with that parameter value, the model has the highest accuracy. Therefore, the optimal random forest model is trained using the optimal parameters based on the training dataset. The model has 12.46% variance explained, which is a high enough value for an accurate model in some research (Pang et al., 2006) (b). Hence, the results generated by the model should be considered as robust enough.

As claimed before, for regression work, the MSE of the model increased by the variable should be the most appropriate index for evaluating the importance of variables. Hence, the importance of the pre-fire condition factors to the estimated destroyed area of human-built land use will only be interpreted by the MSE of the model increased by the variable. According to Figure 3.15, the MSE of the model increased by the mean value of DEM is negative, which means the pre-fire condition is noise to the model. Hence, the mean value of DEM is not important to the estimated destroyed area of human-built land use at all (Dewi and Chen, 2019). Similarly, since the MSE of the model increased by pre-fire air temperature around the wildfire is too close to 0, the variate is not important to the estimated destroyed area of human-built land use. On the other hand, the MSE of the model increased by the standard deviation of DEM and pre-fire wet-bulk temperature are very high, which are about 16 and 24 respectively, so they are very important pre-fire condition factors to the estimated destroyed area of human-built land use. Therefore, only the steepness of the topography within the perimeter of wildfires and the pre-fire wet-bulk temperature are important to the area of destroyed human-built land use during the wildfire.

#### **4.4. Limitations**

There are some limitations and sources of error in the project, which will influence the credibility of the results and interpretations shown in the above sections. The limitations are basically from the sample size, flaws of the data source, the method for getting the pre-fire condition variates, and the violation of some assumptions for the statistical models.

Due to the big workload for processing a wildfire and the limited time, only 20 wildfires are included in the study area. The number of observations is too low compared to the number of explanatory variates (e.g. 20, 4). This will lead to the curse of dimensionality, which influences



the robustness of the trained model and the credibility of the results from the model (Köppen, 2000).

According to the data source section, there are some wildfires that occurred in the US but were not recorded in full information, which are then abandoned, which may be a bias in sample selection (Heckman, 1979). Also, the temporal resolution of L8 c2l2 is 16 days, so the pre-fire data for some wildfires are selected some days before the occurrence time. Also, the post-fire data for some wildfires are selected some days after the end time. The precedence or the delay are all controlled within one week. However, the compromise may generate some errors. Moreover, there exists some inaccurate information in the WFP and WFL datasets, especially for the locations of the wildfires. In particular, the locations might deviate from the real locations where the fire occurred (NIFC, 2022).

In the process of identifying the vegetation and human-built land use, NDVI, NDBI, thresholding, and visual interpretation are used. However, the accuracy is usually lower than supervised classification or object-oriented classification (Fugara et al., 2009). Therefore, the error will lead to inaccuracy in the estimated destroyed human-built land use variate and the map in Figure 3.1.

According to Figure 3.8, the frequency histogram of the standard deviation of NDVI is right-skewed compared with usual normal distributions. The frequency histogram of the mean value of NDVI is not bimodal, and the middle bar is even lower than the sidebars near it, which is different from usual normal distributions. Hence, the normality assumption of the Pearson correlation between the standard deviation of NDVI and estimated cost is violated, which makes the results unpalatable. Similarly, the normality assumption of the Pearson correlation between the standard deviation of NDVI and estimated destroyed area of human-built land use is also

violated, which makes the results unplausible. The frequency histogram of the mean value of DEM is too right-skewed compared with usual normal distributions. Hence, the normality assumption of the Pearson correlation between the mean value of DEM and estimated cost is violated, which makes the results unplausible. Similarly, the normality assumption of the Pearson correlation between the mean value of DEM and estimated destroyed area of human-built land use is also violated, which makes the results unplausible.

## **5. Conclusion**

In this section, the main findings of the project will be summarized. Based on the findings, the aim of the project will be satisfied by concluding the most important pre-fire condition factors to the wildfire-induced socio-economic loss. Finally, some future directions will be provided based on the limitations of the project.

According to the study area section, most of the well-recorded wildfires with high costs and large perimeters are indeed concentrated on the west coast of the US. Hence, it can be concluded that the likelihood for expensive wildfires to occur on the west coast of the US is very high. Based on the interpretation of the pre-fire condition maps of Dixie as the example, the high density, volume, and enough variation of the vegetation within the perimeter of Dixie indeed gave the adequate prerequisite for the extremely terrible severity of Dixie theoretically. Also, the moderately high elevation within the perimeter of Dixie contains adequate oxygen for Dixie to burn, and the steep typography might suppress the efficiency of firefighting and adds the cost of firefighting theoretically. Moreover, the stably high pre-fire LST within the perimeter of Dixie was expected to lead to the occurrence of the severe wildfire. From the interpretation of the socio-economic loss variates, it is concluded that the wildfires on the west coast of the US indeed

caused too much socio-economic loss by too expensive firefighting. Also, wildfires on the west coast of the US indeed caused too much socio-economic loss by destroying too much area of human-built land use. According to the statistical analysis, more diverse and denser vegetation might lead to higher cost. Steeper topography around where a wildfire occurs is likely to increase the cost of the wildfire. Stronger wind around where the wildfire occurs might help decrease the cost of the wildfire. Moreover, higher pre-fire air temperatures around 10 meters of where the wildfire occurs might lead to larger areas of destroyed human-built land use. Also, the higher the wet-bulk temperature around the wildfire location is, the larger the area of destroyed human-built land use during the wildfire is likely to be. Among the pre-fire condition factors that are important to the socio-economic loss, the average level of the pre-fire vegetation within the perimeter is much more important to the cost of the wildfire than other pre-fire condition factors. The pre-fire wind speed has moderate importance to the cost of the wildfire. The steepness of the topography and the diversity of the pre-fire vegetation have also a decent amount of importance to the cost of the wildfire. Only the steepness of the topography within the perimeter of wildfires and the pre-fire wet-bulk temperature are important to the area of destroyed human-built land use during the wildfire.

Therefore, the first pre-fire condition for severe wildfires is that the location is around the west coast of the US. For preventing high-cost wildfires, the areas on the west coast of the US in some specific periods with a high average level of vegetation density, low wind speed, and steep topography are worthy of attention. For preventing wildfires that may destroy a large amount of human-built land use, the areas in some specific periods with lots of human-built land use, steep topography, and wet-bulk temperature must be focused.

In future research, more wildfires in the US must be studied. Hereby, a larger sample set will be analyzed so that the final results and conclusions can be more plausible and robust. Also, classification methods with higher accuracy are expected to be implemented for identifying land use and land cover in more detail. Thus, some more detailed research can be conducted. For example, researchers can discuss what specific types of vegetation are more likely to cause more socio-economic loss. Furthermore, some other quantifications of wildfire-induced socio-economic loss can be studied, such as the mortality caused by wildfires, air quality damaged by wildfires, or some other disease caused by wildfires.

## References

- Abatzoglou, J. T., & Williams, A. P. (2016). Impact of anthropogenic climate change on wildfire across western US forests. *Proceedings of the National Academy of Sciences*, 113(42), 11770-11775.
- Afify, H. A. (2011). Evaluation of change detection techniques for monitoring land-cover changes: A case study in new Burg El-Arab area, *Alexandria Engineering Journal*, Volume 50, Issue 2, 2011, Pages 187-195, ISSN 1110-0168, <https://doi.org/10.1016/j.aej.2011.06.001>.
- Ager, A. A., Palaiologou, P., Evers, C. R., Day, M. A., & Barros, A. M. G. (2018).
- Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, 18(3), 91-93.
- Alin, A. (2010). Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3), 370-374.
- Amemiya, T. (1983). Non-linear regression models. *Handbook of econometrics*, 1, 333-389.
- Analysis of the Relationship between Land Surface Temperature and Wildfire Severity in a Series of Landsat Images. *Remote Sensing (Basel, Switzerland)*, 6(7), 6136–6162. <https://doi.org/10.3390/rs6076136>
- Assessing Transboundary Wildfire Exposure in the Southwestern United States. *Risk Analysis*, 38(10), 2105–2127. <https://doi.org/10.1111/risa.12999>
- As-syakur, A., Adnyana, I., Arthana, I. W., & Nuarsa, I. W. (2012). Enhanced built-up and bareness index (EBBI) for mapping built-up and bare land in an urban area. *Remote sensing*, 4(10), 2957-2970.
- Atmosphere, 12(1), 109. <https://doi.org/10.3390/atmos12010109>
- Badia, Pallares-Barbera, M., Valldeperas, N., & Gisbert, M. (2019). Wildfires in the wildland-urban interface in Catalonia: Vulnerability analysis based on land use and land cover change. *The Science of the Total Environment*, 673, 184–196. <https://doi.org/10.1016/j.scitotenv.2019.04.012>
- BBC. (2018, November 23). California wildfires: Camp Fire nearly fully contained. BBC News. Retrieved February 4, 2022, from <https://www.bbc.com/news/world-us-canada-46315029>
- BC Wildfire Service. (2020, June 12). Wildfire causes. Province of British Columbia.
- Belgiu, & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing* (pp. 1-4). Springer, Berlin, Heidelberg.

- Bento-Gonçalves, & Vieira, A. (2020). Wildfires in the wildland-urban interface: Key concepts and evaluation methodologies. *The Science of the Total Environment*, 707, 135592–135592. <https://doi.org/10.1016/j.scitotenv.2019.135592>
- Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, 25(2), 197–227.
- Blanchet, F. G., Legendre, P., & Borcard, D. (2008). Forward selection of explanatory variables. *Ecology*, 89(9), 2623–2632.
- Bloom, Flower, A., Medler, M., & DeChaine, E. G. (2018). The compounding consequences of wildfire and climate change for a high-elevation wildflower (*Saxifraga austromontana*). *Journal of Biogeography*, 45(12), 2755–2765. <https://doi.org/10.1111/jbi.13441>
- Borini Alves, Pérez-Cabello, F., & Rodrigues Mimbrero, M. (2015). Land-use and land-cover dynamics monitored by NDVI multitemporal analysis in a selected southern amazonian area (Brazil) for the last three decades. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, 40(7), 329–335. <https://doi.org/10.5194/isprsarchives-XL-7-W3-329-2015>
- Bucsela, Celarier, E. ., Wenig, M. ., Gleason, J. ., Veeffkind, J. ., Boersma, K. ., & Brinksma, E. . (2006). Algorithm for NO<sub>2</sub> vertical column retrieval from the ozone monitoring instrument. *IEEE Transactions on Geoscience and Remote Sensing*, 44(5), 1245–1258. <https://doi.org/10.1109/TGRS.2005.863715>
- Butsic, V., Kelly, M., & Moritz, M. A. (2015). Land Use and Wildfire: A Review of Local Interactions and Teleconnections. *Land* 2015, 4(1), 140–156. <https://doi.org/https://doi.org/10.3390/land4010140>
- Cafri, G., Li, L., Paxton, E. W., & Fan, J. (2018). Predicting risk for adverse health events using random forest. *Journal of Applied Statistics*, 45(12), 2279–2294.
- Calkin, D. E., Thompson, M. P., & Finney, M. A. (2015). Negative consequences of positive feedbacks in US wildfire management. *Forest Ecosystems*, 2(1), 1–10.
- CATEGORIZATION. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W7, 167–173. <https://doi.org/10.5194/isprs-annals-IV-2-W7-167-2019>
- Chen, X. W., & Jeong, J. C. (2007, December). Enhanced recursive feature elimination. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)* (pp. 429–435). IEEE.
- Cleveland, W. S., & McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716), 828–833.
- COHEN, & GOWARD, S. N. (2004). Landsat's Role in Ecological Applications of Remote Sensing. *Bioscience*, 54(6), 535–545. [https://doi.org/10.1641/0006-3568\(2004\)054\[0535:LRIEAO\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0535:LRIEAO]2.0.CO;2)

- Connelly, L. (2020). Logistic regression. *Medsurg Nursing*, 29(5), 353-354.
- de Bie, C. A. J. M., Khan, M. R., Smakhtin, V. U., Venus, V., Weir, M. J. C., & Smaling, De Cáceres, M., Chytrý, M., Agrillo, E., Attorre, F., Botta-Dukát, Z., Capelo, J., Wiser, S. K. (2015). A comparative framework for broad-scale plot-based vegetation classification. *Applied Vegetation Science*, 18(4), 543– 560. <https://doi.org/10.1111/avsc.12179>
- Detle, H., & Biedermann, S. (2003). Robust and efficient designs for the Michaelis–Menten model. *Journal of the American Statistical Association*, 98(463), 679-686.
- Dewi, C., & Chen, R. C. (2019). Random forest and support vector machine on features selection for regression analysis. *Int. J. Innov. Comput. Inf. Control*, 15(6), 2027-2037.
- E. M. A. (2011). Analysis of multi-temporal SPOT NDVI images for small-scale land-use mapping. *International Journal of Remote Sensing*, 32(21), 6673–6693. <https://doi.org/10.1080/01431161.2010.512939>
- Ekström, J. (2011). The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule Debate. *Environmental Health Perspectives*, 120(5), 695–701. <https://doi.org/10.1289/ehp.1104422>
- Evans, F. H. (1998). An investigation into the use of maximum likelihood classifiers, decision trees, neural networks and conditional probabilistic networks for mapping and predicting salinity.
- Everitt, B. S. (1992). *The analysis of contingency tables*. CRC Press.
- Fraser, R. S., Bahethi, O. P., & Al-Abbas, A. H. (1977). The effect of the atmosphere on the classification of satellite observations to identify surface features. *Remote Sensing of Environment*, 6(3), 229-249.
- Fugara, A., Mohammed, A., Pradhan, B., & Ahmed Mohamed, T. (2009). Improvement of land-use classification using object-oriented and fuzzy logic approach. *Applied Geomatics*, 1(4), 111-120.
- Gale, Cary, G. J., Van Dijk, A. I. J. ., & Yebra, M. (2021). Forest fire fuel through the lens of remote sensing: Review of approaches, challenges and future directions in the remote sensing of biotic determinants of fire behaviour. *Remote Sensing of Environment*, 255, 112282–. <https://doi.org/10.1016/j.rse.2020.112282>
- Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14), 2225-2236.
- Ghosal, I., & Hooker, G. (2020). Boosting random forests to reduce bias; one-step boosted forest and its variance estimate. *Journal of Computational and Graphical Statistics*, 30(2), 493-502.

- Guha, S., Govil, H., Dey, A., & Gill, N. (2018). Analytical study of land surface temperature with NDVI and NDBI using Landsat 8 OLI and TIRS data in Florence and Naples city, Italy. *European Journal of Remote Sensing*, 51(1), 667–678. <https://doi.org/10.1080/22797254.2018.1474494>
- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), 1157-1182.
- Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (Eds.). (2008). *Feature extraction: foundations and applications* (Vol. 207). Springer.
- Han, D., Di, X., Yang, G., Sun, L., & Weng, Y. (2021). Quantifying fire severity: a brief review and recommendations for improvement. *Ecosystem Health and Sustainability*, 7(1), 1973346.
- Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.
- Hawkins, D. M. (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1), 1-12.
- Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153-161.
- Herman, Bhartia, P. K., Torres, O., Hsu, C., Seftor, C., & Celarier, E. (1997). Global distribution of UV-absorbing aerosols from Nimbus 7/TOMS data. *Journal of Geophysical Research: Atmospheres*, 102(D14), 16911–16922. <https://doi.org/10.1029/96JD03680>
- Hessburg, Prichard, S. J., Hagmann, R. K., Povak, N. A., & Lake, F. K. (2021). Wildfire and climate change adaptation of western North American forests: a case for intentional management. *Ecological Applications*, 31(8). <https://doi.org/10.1002/eap.2432>
- Home. GloVis. (n.d.). Retrieved February 5, 2022, from <https://glovis.usgs.gov/> Jain, Castellanos-Acuna, D., Coogan, S. C. P., Abatzoglou, J. T., & Flannigan, M. D.
- Ismail, Muhamad Ludin, A. N., & Hosni, N. (2020). Comparative Assessment of the Unsupervised Land Use Classification by Using Proprietary GIS and Open Source Software. *IOP Conference Series. Earth and Environmental Science*, 540(1), 12020–. <https://doi.org/10.1088/1755-1315/540/1/012020>
- Jain, P., Castellanos-Acuna, D., Coogan, S. C., Abatzoglou, J. T., & Flannigan, M. D. (2022). Observed increases in extreme fire weather driven by atmospheric humidity and temperature. *Nature Climate Change*, 12(1), 63-70.
- Janitza, & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PloS One*, 13(8), e0201904–e0201904.
- Johnston, F., Henderson, S., Chen, Y., Randerson, J., Marlier, M., DeFries, R., et al. (2012). Estimated global mortality attributable to smoke from landscape fires.



- Kaasalainen, S., Pyysalo, U., Krooks, A., Vain, A., Kukko, A., Hyypä, J., & Kaasalainen, M. (2011). Absolute radiometric calibration of ALS intensity data: Effects on accuracy and target classification. *Sensors*, 11(11), 10586-10602.
- Kaushik, S. (2016, December 1). *Feature selection methods: Machine learning*. Analytics Vidhya. Retrieved February 23, 2022, from <https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/>
- Kent, M. (2012). *Vegetation description and data analysis: A practical approach*, 2nd ed. Oxford, UK: Wiley-Blackwell.
- Kleinbaum, D. G., Dietz, K., Gail, M., Klein, M., & Klein, M. (2002). *Logistic regression* (p. 536). New York: Springer-Verlag.
- Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence*, 97(1-2), 273-324.
- Köppen, M. (2000, September). The curse of dimensionality. In 5th online world conference on soft computing in industrial applications (WSC5) (Vol. 1, pp. 4-8).
- Kornbrot, D. (2014). Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*.
- Kshetri, K., Kunjan KshetriKunjan Kshetri 32722 gold badges33 silver badges99 bronze badges, KodiologistKodiologist 19.1k22 gold badges3636 silver badges6868 bronze badges, Marco StamazzaMarco Stamazza 32933 silver badges77 bronze badges, aditya raiaditya rai 1111 bronze badge, & JohnMountfortJohnMountfort 111 bronze badge. (2019, September). *What is the difference between independent variable and a feature?* Cross Validated. Retrieved February 22, 2022, from <https://stats.stackexchange.com/questions/138740/what-is-the-difference-between-independent-variable-and-a-feature>
- Kuhn, M., & Johnson, K. (2019). *Feature Selection Methods*. In *Applied predictive modeling*. essay, Springer.
- Kulkarni, K., Vijaya, P. (2021) NDBI Based Prediction of Land Use Land Cover Change. *J Indian Soc Remote Sens* 49, 2523–2537 . <https://doi.org/10.1007/s12524-021-01411-9>
- Kumar, G. R., Mangathayaru, N., & Narasimha, G. (2015, September). An improved k-Means Clustering algorithm for Intrusion Detection using Gaussian function. In *Proceedings of the The International Conference on Engineering & MIS 2015* (pp. 1-7).
- Lee, Y. G., & Kim, S. Y. (2008). *Introduction to statistics*. Yulgokbooks, Korea, 342-351.
- Li, D., Meng, X., Song, J., Li, M., & Lyu, K. (2022). Main Factor Analysis to the Wildfire Caused Socioeconomic Loss Based on Remote Sensing and Random Forest (proposal).
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.

- Lillesand, Kiefer, R. W., & Chipman, J. W. (2015). Remote sensing and image interpretation (Seventh edition.). John Wiley & Sons, Inc.
- Lovreglio, R., Leone, V., Giaquinto, P., & Notarnicola, A. (2010). Wildfire cause analysis: four case-studies in southern Italy. *iForest-Biogeosciences and Forestry*, 3(1), 8.
- Ma, W., Feng, Z., Cheng, Z., Chen, S., & Wang, F. (2020). Identifying forest fire driving factors and related impacts in China using random forest algorithm. *Forests*, 11(5), 507. <https://doi.org/10.3390/f11050507>
- Malik, A., Rao, M. R., Puppala, N., Koouri, P., Thota, V. A., Liu, Q., Chiao, S., & Gao, J. (2021). Data-driven wildfire risk prediction in Northern California.
- Marcos, E., Fernández -García, V., Fernández -Manso, A., Quintano, C., Valbuena, L., Tárrega, R., Luis -Calabuig, E., & Calvo, L. (2018). Evaluation of composite burn index and land surface temperature for assessing soil burn severity in Mediterranean fire-prone pine ecosystems. *Forests*, 9(8), 494. <https://doi.org/10.3390/f9080494>
- Masek, J. G. (Ed.). (2021, December 17). *Landsat Then and Now*. NASA. Retrieved April 11, 2022, from <https://landsat.gsfc.nasa.gov/about/>
- Maxwell, S. K., & Sylvester, K. M. (2012). Identification of “ever-cropped” land (1984– 2010) using Landsat annual maximum NDVI image composites: Southwestern Kansas case study. *Remote Sensing of Environment*, 121(Complete), 186–195. <https://doi.org/10.1016/j.rse.2012.01.022>
- McKenzie, D., Gedalof, Z. E., Peterson, D. L., & Mote, P. (2004). Climatic change, wildfire, and conservation. *Conservation biology*, 18(4), 890-902.
- Miller, & Thode, A. E. (2007). Quantifying burn severity in a heterogeneous landscape with a relative version of the delta Normalized Burn Ratio (dNBR). *Remote Sensing of Environment*, 109(1), 66–80. <https://doi.org/10.1016/j.rse.2006.12.006>
- Muhs, Parvania, M., Nguyen, H. T., & Palmer, J. A. (2021). Characterizing Probability of Wildfire Ignition Caused by Power Distribution Lines. *IEEE Transactions on Power Delivery*, 36(6), 3681–3688. <https://doi.org/10.1109/TPWRD.2020.3047101>
- Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6), 275-285.
- Myoung, B., Kim, S., Nghiem, S., Jia, S., Whitney, K., & Kafatos, M. (2018). Estimating live fuel moisture from Modis satellite data for Wildfire Danger Assessment in Southern California USA. *Remote Sensing*, 10(2), 87. <https://doi.org/10.3390/rs10010087>
- NASA. (2022, February 5). Earth Science Data Systems (ESDS) program. NASA. Retrieved February 5, 2022, from <https://earthdata.nasa.gov/esds>

National Aeronautics and Space Administration & Prediction Of Worldwide Energy Resources. (2020, October 14). *POWER Data Methodology*. Retrieved April 11, 2022, from <https://power.larc.nasa.gov/docs/methodology/>

National Aeronautics and Space Administration & Prediction Of Worldwide Energy Resources. (2020, March 9). *About the Prediction Of Worldwide Energy Resources (POWER) Project*. Retrieved April 11, 2022, from <https://www.arcgis.com/home/item.html?id=52116d331ff64e468fe9351fc1c76423>

National Geographic Society. (2019, July 15). Wildfires. National Geographic Society.

National Interagency Fire Center. (1965). *What is NIFC?* Welcome to the Nation's Logistical Support Center. Retrieved April 10, 2022, from <https://www.nifc.gov/about-us/what-is-nifc>

National Interagency Fire Center. (2021, July 6). *WFIGS - wildland fire perimeters full history*. National Interagency Fire Center. Retrieved April 10, 2022, from <https://data-nifc.opendata.arcgis.com/datasets/nifc::wfigs-wildland-fire-perimeters-full-history/about>

National Interagency Fire Center. (2021, November 10). National Interagency Fire Center. Wildland Fire Open Data. Retrieved February 5, 2022, from <https://data-nifc.opendata.arcgis.com/>

O'Brien, G., O'keefe, P., Rose, J., & Wisner, B. (2006). Climate change and disaster management. *Disasters*, 30(1), 64-80.

Page, W. G., & Butler, B. W. (2018). Fuel and topographic influences on wildland firefighter burnover fatalities in Southern California. *International journal of wildland fire*, 27(3), 141-154.

Pang, H., Lin, A., Holford, M., Enerson, B. E., Lu, B., Lawton, M. P., ... & Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, 22(16), 2028-2036.

Panigrahy, R. K., Kale, M. P., Dutta, U., Mishra, A., Banerjee, B., & Singh, S. (2010). Forest cover change detection of Western Ghats of Maharashtra using satellite remote sensing based visual interpretation technique. *Current Science*, 657-664.

Peacock, A. J. (1998). ABC of oxygen: Oxygen at high altitude. *BMJ*, 317(7165), 1063– 1066. <https://doi.org/10.1136/bmj.317.7165.1063>

Pelletier, J. D., & Orem, C. A. (2014). How do sediment yields from post-wildfire debris-laden flows depend on terrain slope, soil burn severity class, and drainage basin area? Insights from airborne-LiDAR change detection. *Earth Surface Processes and Landforms*, 39(13), 1822-1832.

Puletti, N., Perria, R. & Storch, P. (2014). Unsupervised classification of very high remotely sensed images for grapevine rows detection. *European Journal of Remote Sensing*, 47(1), 45–54. <https://doi.org/10.5721/EuJRS20144704>

- Purevdorj, T.S., Tateishi, R., Ishiyama, T., & Honda, Y. (1998). Relationships between percent vegetation cover and vegetation indices. *International Journal of Remote Sensing*, 19(18), 3519–3535.
- Roy, D. P., Boschetti, L., & Trigg, S. N. (2006). Remote Sensing of Fire severity: Assessing the performance of the normalized burn ratio. *IEEE Geoscience and Remote Sensing Letters*, 3(1), 112–116. <https://doi.org/10.1109/lgrs.2005.858485>
- Sánchez-Marono, N., Alonso-Betanzos, A., & Tombilla-Sanromán, M. (2007, December). Filter methods for feature selection—a comparative study. *In International Conference on Intelligent Data Engineering and Automated Learning* (pp. 178-187). Springer, Berlin, Heidelberg.
- Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.
- Sedgwick, P. (2012). Pearson’s correlation coefficient. *Bmj*, 345.
- Sherman, W. R., Penick, M. A., Su, S., Brown, T. J., & Harris, F. C. (2007, March). Vrfire: an immersive visualization experience for wildfire spread analysis. In 2007 IEEE Virtual Reality Conference (pp. 243-246). IEEE.
- Silverstein, V. B., & Nunn, L. S. (2010). *Wildfires: the science behind raging infernos*. Enslow Publishers.
- Sinaga, & Yang, M.-S. (2020). Unsupervised K-Means Clustering Algorithm. *IEEE Access*, 8, 80716–80727. <https://doi.org/10.1109/ACCESS.2020.2988796>
- Singh, S. K., Srivastava, P. K., Gupta, M., Thakur, J. K., & Mukherjee, S. (2014). Appraisal of land use/land cover of mangrove forest ecosystem using support vector machine. *Environmental Earth Sciences*, 71(5), 2245–2255. <https://doi.org/10.1007/s12665-013-2628-0>
- Sneep, de Haan, J. F., Stammes, P., Wang, P., Vanbaucse, C., Joiner, J., Vasilkov, A. P., & Levelt, P. F. (2008). Three-way comparison between OMI and PARASOL cloud pressure products. *Journal of Geophysical Research: Atmospheres*, 113(D15), D15S23–n/a. <https://doi.org/10.1029/2007JD008694>
- Soto, Bermudez, J., Happ, P. ., & Feitosa, R. . (2019). A COMPARATIVE ANALYSIS of UNSUPERVISED and SEMI-SUPERVISED REPRESENTATION LEARNING for REMOTE SENSING IMAGE CATEGORIZATION. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4(2), 167–173. <https://doi.org/10.5194/isprs-annals-IV-2-W7-167-2019>
- Stephens, Burrows, N., Buyantuyev, A., Gray, R. W., Keane, R. E., Kubian, R., Liu, S., Seijo, F., Shu, L., Tolhurst, K. G., & van Wagendonk, J. W. (2014). Temperate and boreal forest mega-fires: characteristics and challenges. *Frontiers in Ecology and the Environment*, 12(2), 115–122. <https://doi.org/10.1890/120332>
- Stork, D. G., Duda, R. O., Hart, P. E., & Stork, D. (2001). *Pattern classification. A Wiley-Interscience Publication*.

- Sun, J., Yang, J., Zhang, C., Yun, W., & Qu, J. (2013). Automatic remotely sensed image classification in a grid environment based on the maximum likelihood method. *Mathematical and Computer Modelling*, 58(3), 573-581. <https://doi.org/10.1016/j.mcm.2011.10.063>
- Sutter, J. M., & Kalivas, J. H. (1993). Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection. *Microchemical journal*, 47(1-2), 60-66.
- Tanaka, H., Ishibuchi, H., & Yoshikawa, S. (1995). Exponential possibility regression analysis. *Fuzzy Sets and Systems*, 69(3), 305-318.
- Tichý, L., Chytrý, M., & Botta-Dukát, Z. (2014). Semi-supervised classification of vegetation: Preserving the good old units and searching for new ones. *Journal of Vegetation Science*, 25(6), 1504– 1512. <https://doi.org/10.1111/jvs.12193>
- Topaloğlu, R. H., Sertel, E., & Musaoğlu, N. (2016). ASSESSMENT OF CLASSIFICATION ACCURACIES OF SENTINEL-2 AND LANDSAT-8 DATA FOR LAND COVER/USE MAPPING. *International archives of the photogrammetry, remote sensing & spatial Information Sciences*, 41.
- Torres, Tanskanen, A., Veihelmann, B., Ahn, C., Braak, R., Bhartia, P. K., Veeffkind, P., & Levelt, P. (2007). Aerosols and surface UV products from Ozone Monitoring Instrument observations: An overview. *Journal of Geophysical Research: Atmospheres*, 112(D24), D24S47–n/a. <https://doi.org/10.1029/2007JD008809>
- Tsai, C. J., Lee, C. I., & Yang, W. P. (2008). A discretization algorithm based on class-attribute contingency coefficient. *Information Sciences*, 178(3), 714-731.
- Turner, M. G., Hargrove, W. W., Gardner, R. H., & Romme, W. H. (1994). Effects of fire on landscape heterogeneity in Yellowstone National Park, Wyoming. *Journal of Vegetation Science*, 5(5), 731–742. <https://doi.org/10.2307/3235886>
- United Nations. (2016, January 1). Goal 3 | Department of Economic and Social Affairs.
- United Nations. Retrieved February 4, 2022, from <https://sdgs.un.org/goals/goal3> Vandaele, Hermans, C., Fally, S., Carleer, M., Colin, R., Mérienne, M. -F., Jenouvrier, A.,
- Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, 85, 189-203.
- US Department of the Interior. (1879, March 3). *About Us*. U.S. Geological Survey. Retrieved April 10, 2022, from <https://www.usgs.gov/about/about-us>
- USGS. (2022, April 11). *Landsat 8*. Landsat 8 | U.S. Geological Survey. Retrieved April 11, 2022, from <https://www.usgs.gov/landsat-missions/landsat-8>
- Van Der Maaten, L., Postma, E., & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71), 13.

- Van Leeuwen, W. J., Casady, G. M., Neary, D. G., Bautista, S., Alloza, J. A., Carmel, Y., ... & Orr, B. J. (2010). Monitoring post-wildfire vegetation response with remotely sensed time-series data in Spain, USA and Israel. *International Journal of Wildland Fire*, 19(1), 75-93.
- Vandaele, A. C., Hermans, C., Fally, S., Carleer, M., Colin, R., Merienne, M. F., ... & Coquart, B. (2002). High-resolution Fourier transform measurement of the NO<sub>2</sub> visible and near-infrared absorption cross sections: Temperature and pressure effects. *Journal of Geophysical Research: Atmospheres*, 107(D18), ACH-3.
- Vlassova, Pérez -Cabello, F., Mimbbrero, M., Llovería, R., & García -Martín, A. (2014).
- Wang, Li, D., & Wang, Y. (2019). Realization of remote sensing image segmentation based on K-means clustering. *IOP Conference Series. Materials Science and Engineering*, 490(7), 72008–. <https://doi.org/10.1088/1757-899X/490/7/072008>
- Wang, Q., & Tenhunen, J. D. (2004). Vegetation mapping with multitemporal NDVI in North Eastern China Transect (NECT). *International Journal of Applied Earth Observation and Geoinformation*, 6(1), 17–31. <https://doi.org/10.1016/j.jag.2004.07.002>
- Wang, Rich, P. M., Price, K. P., & Kettle, W. D. (2005). Relations between NDVI, Grassland Production, and Crop Yield in the Central Great Plains. *Geocarto International*, 20(3), 5–11. <https://doi.org/10.1080/10106040508542350>
- Weber, K. (2002). Students' Understanding of Exponential and Logarithmic Functions.
- Wei, Xie, Y., Wang, X., Jiao, J., He, S., Bie, Q., Jia, X., Xue, X., & Duan, H. (2020). Land cover mapping based on time-series MODIS-NDVI using a dynamic time warping approach: A case study of the agricultural pastoral ecotone of northern China. *Land Degradation & Development*, 31(8), 1050–1068. <https://doi.org/10.1002/ldr.3502>
- WOLFRAM. (2022). Weather Data-Wolfram language documentation. Wolfram Language & System Documentation Center. Retrieved February 5, 2022, from <https://reference.wolfram.com/language/guide/WeatherData.html.en?source=footer>
- Yao, J., Brauer, M., & Henderson, S. B. (2013). Evaluation of a wildfire smoke forecasting system as a tool for public health protection. *Environmental Health Perspectives*, 121(10), 1142. <https://link.gale.com/apps/doc/A351948514/AONE?u=uniwater&sid=bookmark-AONE&xid=3aee60d3>
- Yasin, Abdullah, J., Noor, N. M., & Yusoff, M. M. (2020). Land Cover and NDBI analysis to map built up area in Iskandar Malaysia. *IOP Conference Series. Earth and Environmental Science*, 540(1), 12073–. <https://doi.org/10.1088/1755-1315/540/1/012073>
- Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics-Simulation and Computation*, 46(8), 6261-6282.
- Zha, Y., Gao, J., & Ni, S. (2003). Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. *International Journal of Remote Sensing*, 24(3), 583–594.

ZHAI, G., SATO, T., SEO, K., FUKUZONO, T., & IKEDA, S. (2002). RISK FACTOR ANALYSIS AND EVALUATION OF NATURAL DISASTERS: APPLICATION OF THE RIFAE FRAMEWORK TO THE 2000 TOKAIFLOOD DISASTER IN JAPAN. In *Computational Intelligent Systems For Applied Research* (pp. 208-217).

Zhi, T., Luo, H., & Liu, Y. (2018). A Gini impurity-based interest flooding attack defence mechanism in NDN. *IEEE Communications Letters*, 22(3), 538-541.

Zinck, R. D., Johst, K., & Grimm, V. (2010). Wildfire, landscape diversity and the Drossel–Schwabl model. *Ecological Modelling*, 221(1), 98-105.