MONASH University

Exploring bushfire risk across Victoria with open data sources and open-source software

Yunfang Wu

Under the supervision of Professor Di Cook and Dr Kate Saunders

Monash University in 2022

Monash Business School

Department of Econometrics and Business Statistics

Contents

Chapter 1

Introduction

1.1 Introduction

Bushfires are a natural phenomenon that creates damaging disruptions to natural ecosystems and the economy, which can cause loss of life, property and infrastructure. The State of Victoria is one of the most fire-prone regions in the world due to its weather and conditions. The recent Black Summer fire, which lasted from November 2019 to February 2020, destroyed more than 300 homes and 6,632 heads of stock and burned more than 1.5 million hectares of public and private land. Approximately 3,050 insurance was generated, and the estimated insured loss was around 18.6 million. The weather was recorded to be much hotter and drier than in other years (Black Summer bushfires, SA, 2019-20 — Australian Disaster Resilience Knowledge Hub, 2022). There are many more damaging bushfires in history, on 16 February 1983, over 100 fires swept across Victoria and South Australia, killing 75 people and causing widespread damage, where high

temperatures, intense winds, and low summer rainfall caused a high fire danger in Victoria's eucalyptus forests (Ash Wednesday 1983, 2022).

1.2 Motivation

Expanding our understanding of bushfire risks can be extremely beneficial for us to reduce both environmental and economic harm during the fire seasons in the future in Victoria. The knowledge can be used to discover the risk of bushfires in different areas of Victoria, as well as understand how different environmental factors contribute to the risk of bushfires. This project aims to use open-source data and create a model that allows better understanding of the bushfire in the state of Victoria, Australia. By systematically analysing the risk of bushfire, the risk from bushfires can be then understood better and managed more effectively. Finding out area with high bushfire risk allow resources to be allocated to the area and time where it is needed.

1.3 Research Objectives

1.3.1 What part of Victoria has higher bush fire density?

This research will use machine learning methods to understand the risk of bushfires in different areas of Victoria by incorporating Himawari-8 satellite fire hotspot data obtained from the Japan Aerospace Exploration Agency

(JAXA). A usual data analysis workflow involves collecting, cleaning, examining, and modelling the data to be able to incorporate different factors, in particular the Himawari-8 satellite fire hotspot data, into our bushfire risk prediction model; there are several questions to consider in each step of the data analysis process.

Incorporating satellite hotspot data allow us to obtain more accurate information regarding fire location and time, however, processing the fire hotspot data accurately can be challenging. The fire hotspot data is an example of high-frequency and high-resolution data, which contains all fire hotspots' location information that was captured at a ten minutes interval. Therefore, to be able to analyze Himawari-8 satellite fire hotspot data, the first step is to process the fire hotspot data so that it reflects interpretable information about bushfires in Victoria. The first question is, how do we process the Himawari-8 satellite fire hotspot data so that we can obtain information on the difference in the distribution of historical bushfires in different areas of Victoria, Australia. By investigating ways to clean and examine fire hotspot data and comparing methods to estimate the distribution of historical fire hotspots, we can utilise the information from JAXA to improve our overall understanding of bushfire risk in Victoria.

1.3.2 What factors are more important for predicting the bushfire density?

The next research question is to find out important variables that can be used to explore the risk of bushfires. There are many other existing literature pointing out the important factors impact the risks of bushfire. However, it is a challenge collect all relevant data that are relevant in our case. As the project is limited to source data from open sources.

A good way to find out different contributing factors to the density of bushfire risk will be specifying a model that can have the dependent variable be the density of bushfire and the variables be the relevant factors that contributes to bushfire. To gain insights into the contributing factors of fire risk, a model that can represent the relationship between bushfire and different factors that influence bushfire needs to be specified. However, finding out a good model to do this can be challenging. There are also many factors to consider across different stages of the data modelling process. Since the predicted value is the density of bushfires, the model for consideration has to be able to capture a non-linear relationship. Aside from non-linear parametric models, some other machine learning models are worth exploring.

1.3.3 How do changes in factors affect the bushfire risk of a particular location?

After we have specified a good model to capture the relationship between the fire density and all the relevant factors that impact bushfires, The next stage would be to figure out how different Variables affect the fire density estimate.

In the case of linear regression, examining the contribution of each explanatory variable to the bushfire density can challenging. In the case of regression model, the interaction between each explanatory variable needs to be considered, as the correlation between variables can affect our ability to measure how much each variable affects the prediction accurately. the relevance of variables needs to be carefully considered, as including irrelevant variables can be detrimental to linear model. Furthermore, it is

good to include as many relevant variables as possible since omitting relevant variables can affect model performance when the omitted variables correlate with included explanatory variables.

Chapter 2

Literature Review

Four sets of the literature are reviewed in this section in order to gain a better understanding of the previous studies and theories developed by other authors, as well as conducting spacial analysis using the R Studios packages.

The first part focus on the clustering algorithms to identify fire ignition points from satellite hotspot data, which will enable us to analyse patterns in historical bushfires.

Second part of literature review include a brief review on the concepts and implementation of special data analysis, focusing on the area of point analysis in statistical packages in R. There will be an introduction on different data types and concepts in the area spacial data analysis. As understanding concepts in spacial analysis is important for correctly utilising open-platform software to assist this study.

Third set of literature reviews the methods to estimate hotspot density and the process of conducting point analysis. This will also provide an direction on how to conduct point analysis using the open source packages on statistical software R.

A third set of literature reviews the application of machine learning algorithms and the comparison of model performance between different methods, in particular the area of bushfire prediction.

2.1 A Brief Literature Review on Hotspot Data Processing

To be able to identify the number of fires started, the ignition points of fire need to be identified form the Himawari-8 satellite fire hotspot data. Weihao Li has developed a spatial-temporal clustering algorithm to detect fire ignition points from fire hotspot data(Li, 2020). There are several considerations that he points out to be important. Firstly, the fire hotspot data are too sensitive which may capture irrelevant fire hotspots that may not necessarily symbolise a bushfire. Therefore, Li suggested that hotspots need to be filtered first base on the firepower. Anything below 100 is less likely to be a fire, therefore, should be eliminated before modelling to avoid misidentifying. Secondly, hotspots that are isolated in location and time are less likely to be a fire; therefore, they are considered as noise and removed from the final bushfire clusters.

A method of spatial-temporal clustering is developed by Weihao Li (2020). Two types of clustering algorithms inspire his clustering algorithm, that is, the Density-Based Spatial Clustering of Applications with Noise and

Fire Spread Reconstruction. Since the Density-Based Spatial Clustering of Applications cluster hotspot on both time dimensions, it is not applicable in the case of finding ignition points from fire hotspot data, as fire usually happens at a one-time point and is continuous by moving along one direction of the timeline. Fire Spread Reconstruction take into account only the time dimension for clustering, which means it will merge two hotspots that are far from each geographically as long as they appeared around a similar time. Therefore, it is not suitable for fire ignition point detection either. Another limitation of FSR is that it lacks detailed consideration of parameter tuning. Taking account of both the Density-Based Spatial Clustering of Applications with Noise and Fire Spread Reconstruction methods, Weihao Li has developed a clustering algorithm that accounts for both the temporal and spatial characteristics of bushfires.

Weihao Li designed a clustering process for hotspot data, it takes the information of hotspot latitude and longitude with time, where firstly, all hotspots on the same time dimension are clustered based on location for each time-slot, based on the parameter called ActiveTime. Then, hotspots in the same time dimension are merged based on the parameter called AjdDist, which represents the distance that a fire may spread a 10 minute time-space. Then a membership id is assigned to each cluster. Then clusters are merged on the time dimension base on membership id to identify the ignition time of each fire, the earliest observed hotspot is decided to be the fire ignition point, and the information of such earliest hotspots is considered to be the ignition time and location of the corresponding clustered bushfire. As a result of the clustering, each bushfire will have a unique bushfire ID, time and location.

In Li and Liu's research, a classification model was built to predict the cause of bushfire ignitions in Victoria during the 2019-2020 bushfire season

(Li, 2020). Variables including climate data, road maps, vegetation information, fire stations and recreation sites information were used to predicte the cause of fire recorded by the Victorian Department of Environment, Land, Water and Planning. There are two main types of models considered, statistical models including Generalized linear mode, generalized additive model and computational models including Random forest model, support vector machine, and artificial neural networks. They compared the classification results with Fire cause data provided by the Victorian Department of Environment, Land, Water and Planning, and found that the random forest model has the best performance overall. The final model showed good predictive ability with 90.5 per cent accuracy in lightning-caused bushfires and 74.95 per cent overall accuracy. Weihao Li predicted the causes of 2019-2020 Australian bushfires using the model and found lightning was the main cause (82 per cent), while arson only took up a fraction of the total cases (3.62 per cent).

2.2 A Brief Review on Hotspot Density Estimation

After hotspot data has been clustered and historical fire ignition points has been identified, a non-parametric method for estimating hotspot density should be used to identify the density of historical fire in Victoria. By using non-parametric method and treating fire ignition point as a random variable, less assumption is made about the distribution of fire ignition point, but instead we determined the probability distribution directly from the sample of observed historical fire ignition location data.

Two dimensional kernel density estimation can be used to understand the which part of Victoria has higher fire risk compared to other parts in Victoria. Two dimensional kernel density estimation is a non-parametric method to estimate the probability density function of a random variable based on kernels as w eights. The benefit of using a non-parametric is that fewer assumption is made about the distribution of data, but instead, the density distribution is determined directly from the observed data. There are existing packages on spacial point analysis in R, especially the area of kernal density estimation. It is usually calculated on the basis of dividing the overall observation area to grid cells, that is referred to as the raster object in R.

The kernel density is calculated by the density of points around each output raster cell.Conceptually, a smoothly curved surface is fitted over each point. I can be a bi-variate Gaussian distribution or other types of two-dimensional probability distribution. The surface value is highest at the location of the point and diminishes with increasing distance from the point, reaching zero at the search radius distance from the point. The search radius is decided by the bandwidth matrix, where the first column of this matrix represents the search radius on the x-axis and the second column represents the search radius on the y axis. It can be think as a matrix that decides the area where we want to add up the two-dimensional Gaussian distribution base on each point as the center. The density at each output raster cell is calculated by adding the values of all the kernel surfaces where they overlay the raster cell centre. The kernel function is based on the function described in Silverman(1986).

2.3 A Brief Literature Review on Spatial Data Processing in R

Spacial data can be data associated with locations, locations on the earth can be more specific and called data geospatial data. There are two main types of spatial data in R which is relevant to this study, that is, raster data and vector data.

Raster data is designed to store geospatial information that is continuous across space. So it is the most common kind of data for things like climate variables or satellite imagery. Values are stored in different cells in a girded area. Any imagery can be thought as a grid of pixels, and with each pixel, there is a value of each pixel, which means it is essentially a matrix of numbers. This can be useful to represent fire density values in different areas of Victoria. In R, useful packages that deal with raster data include Star and Raster.

Other types of data are vector data. These can be either Point, Line or Polygon. A point is represented by a set of XY coordinates, in our study, the hotspot data are a type of vector data. A line consists of a series of points with XY coordinates. And polygons are a series of lines where the first point is connected to the last point and makes a close shape. The Victoria boundaries map can be an example of a polygon. The package called sf can be used to interpret vector data in R.

2.4 A Brief Review on Modeling Methods

After hotspot data has been clustered and historical fire ignition points has been identified, a non-parametric method for hotspot density should be used to identify the density of historical fire in Victoria. By using non-parametric method and treating fire ignition point as a random variable, less assumption is made about the distribution of fire ignition point, but instead we determined the probability distribution directly from the sample of observed historical fire ignition location data.

The US fire-fighting technical society separated the measurement of bushfire risk by the probability of fire occurrence and the result of the fire occurrence. As we are using fire ignition information, the question here focuses more on the factors which contribute to a fire ignition. The most common choice for bushfire risk modelling is the generalised additive model (GAM), which is used by predict the number of lightning ignitions in Western Australia. Other parametric models have also been used to predict the risk of bushfires, including multiple linear regression, and generalized logistic regression.

In Li and Liu's research, he has built a classification model to predict the cause of bushfire ignitions in Victoria during the 2019-2020 bushfire season (Li, 2020). There are two main types of model considered, statistical models including Generalized linear mode, generalized additive model and computational models including Random forest model, support vector machine, and artificial neural networks. They compared the classification results with Fire cause data provided by the Victorian Department of Environment, Land, Water and Planning, and found that the random forest model has the best performance overall.

Chapter 3

Data

This section focuses on explaining the variables used in this study. This project utilises data from various sources that are open to the public. A total of 1605 fire ignition points data are included in this study. The 1605 fire ignition is generated during the 2019 to 2020 fire season, which spans from October 2019 to March 2020.

The variable used has been divided into four main categories, that is Himawari-8 satellite data, environmental factors, vegetation information, and proximity to human activities variables.

3.1 Himawari-8 satellite data

Himawari-8 from the Japan Aerospace Exploration Agency: JAXA, offer much improved capabilities of monitoring large-scale weather events in the

atmosphere and phenomena on the Earth's surface. Himawari-8 is equipped with sensors with 16 spectral bands: 3 visible bands at 0.5 – 1 km resolution, 3 near-infrared bands and 10 infrared bands at 1 – 2 km. The time interval of Himawari-8 fire hotspot observations is 10 minutes. With such a high frequency or temporal resolution, Himawari-8 imagery provides valuable information for the exploration and monitoring of bushfires.

Fire hotspot data is obtained from the Japan Aerospace Exploration Agency FTP site. Each fire hotspot has a unique location and time, other information regarding the fire hotspot like the fire power is also relevant for our analysis. Firstly, hotspot data are filtered base on location, where only hotspots in Victoria, Australia are include. The satellite is very sensitive usually capture irrelevant hotspots that may not necessarily symbolise fire. As suggested by Li and Liu (2020), fire hotspot data from Himiwari-8 is firestly filter to only include hotspot that has a fire power of above 100, as anything below 100 are not likely to be a fire. Using the fire hotpot data clustering algorithm developed by Weihao Li (2020), satellite fire hotspots are clustered into fire ignition points. This is achieved by a spacial-temporal clustering method considering the characteristic of bushfire. Where hotspots were firstly sliced base on the variable ActiveTime, then it is merged base on geographic location by distance, hotspots that are isolated in location and time are less likely to be a fire; therefore, they are considered as noise and removed from the final bushfire clusters. Finally, hotspots are grouped into clusters of fire, and each fire has a unique Id, as well as the location information that is the latitude and longitude.

The fire ignition points are then filtered base on their predicted cause of fire generated by Li and Liu (2020). Only fire that was caused by lightning are included; where fire ignitions that are deemed to be caused by burning off, accident, arson are disregard. This simplifies the modeling process as including different cause of fire may add more noise to the model and reduce

the predictability of bushfire. Even though burning off, accident, arson fires are disregarded; as bushfires are mostly caused by lightning, we still have the majority of fire ignition data, which is 1348 ignition hotspot caused by lightning, where the original sample hotspot data set contains a total of 1605 fire ignition points.

3.2 Environmental Factors

Data representing environmental factors including climate data, water properties, are obtained for each fire ignition point via various open source platforms. Climate data from Bureau of Meteorology and Commonwealth Scientific and Industrial Research Organisation. Maximum temperature, minimum temperature, rainfall and solar exposure are retrieved via an open source R package "bomrang", which is a data client of Bureau of Meteorology. The station-based wind speed are obtained from Commonwealth Scientific and Industrial Research Organisation. A data dictionary is provided in Table 3.2. For each average available, for example, average rainfall, 8 different daily average based the past 7, 14, 28, 60, 90, 180, 360 and 720 days are included. For average wind speed, average wind speed of the pase 1, 3, 6, 12 month are included.

3.3 Vegetation Information

The vegetation information can be obtained from nationwide forest data. There are 2018 nationwide forest data set compiled by r package forest.

Table 3.1: Environmental Factors

Variable Name	Description	Unit
rf	Rainfall on that day	mm
arf7	Average rainfall in the past 7/14... days	mm
arf...	Average rainfall in the past ... days	mm
maxt	Maximum temperature on that day	Celsius degree
amaxt7	Average maximum temperature in the past 7 days	Celsius degree
amaxt14	Average maximum temperature in the past 14 days	Celsius degree
amaxt...	Average maximum temperature in the past ... days	Celsius degree
amint	Minimum temperature on that day	Celsius degree
amint7	Average minimum temperature in the past 7days	Celsius degree
amint...	Average minimum temperature in the past ...days	Celsius degree
se	Global solar exposure on that day	MJ/m^2
ase..	Average global solar exposure in past ... days	MJ/m^2
ws	Average wind speed on that day	m/s
aws_m0	Average wind speed on that month	m/s
aws_m1	Average wind speed in last month	m/s

Table 3.2: Environmental Factors

Variable Name	Description
FOR TYPE	Forest type. Eg. Acacia, Callitris, Casuarina, etc.
COVER	Forest crown cover
HEIGHT	Forest height class

This is the fifth and the latest national State of the Forest Report. Previous national State of the Forest Report were published in 1998-2013 and superseded.

3.4 Proximity to Human Activities Variable

Information regarding1,797,217 roads in Australia downloaded from OpenStreetMap (https://www.openstreetmap.org) by Weihao Li (2020). Since

TABLE 3.3: Proximity to Human Activities Variable

Variable Name	Description	Unit
dist cfa	Distance to the nearest CFA station	m
dist camp	Distance to the nearest camp site	m
dist road	Distance to the nearest road	m

bushfires are influenced by anthropogenic factors, the comprehensive open-source OpenStreetMap map represents the reachability of bushfire ignition locations (Li, 2020). Roads belong to 27 different road classes in Australia, and the road map is one of the layers in the full archive. Data of Country Fire Authority (CFA) fire stations are retrieved from the CFA website . Since camping activities may be associated with bushfires. Victorian recreation sites data are included. The dataset contains 417 camping locations in Victoria. The distance to the nearest recreation area site and fire station and road are computed for every historical fire origin by Weihao Li (Li,2020). A data dictionary is provided in Table 3.3.

Chapter 4

Methodology

4.1 Hotspot Clustering

The first stage of this project is to compute fire ignition points from Himawari-8 fire hotspot data using the spotoroo package. To be able to use the function set out in R, hotspot data is firstly

To identify fire hotspots within Victoria from Himawari-8 fire hotspot data

The order to identify fire hotspots in Victoria from satellite image data, We have two firstly convert the latitude and longitude from the original satellite data, as well as the Victoria map into a recognisable Special object in R. Spatial data that is not represented by raster, are vector data. This study uses points data and polygon data. A point is represented by a set of XY coordinates; in our study, the hotspot data are a type of vector data. A line consists of a series of points with XY coordinates. And polygons are

a series of lines where the first point is connected to the last point and makes a close shape. The Victoria boundaries map can be an example of a polygon. The package called sf can be used to interpret vector data in R. An important concept regarding spatial data analysis is the Coordinate System. Spatial data is always created in a coordinate system. Coordinates can be expressed in many different ways, such as decimal degrees, feet, meters, or kilometres. A geographical coordinate, the usual latitude-longitude pairs is a format of the coordinate system measured in degrees. When plotting a map, the coordinates in degrees need to be projected on a two-dimensional surface, which means a projected coordinate system (PCS) is then used. A PCS Linear measurements are used for the coordinates rather than angular degrees. When interpreting different spatial data from different sources, it is important to check the PCA of each data to ensure accuracy when binding explaining variables to the hotspot data. Using a function in the R package sf, all fire hotspots in the format of points can be identified in the Vicotria polygon.

Then, using the package Spotoroo, which takes three parameters, that is, longitude, latitude and time, fire ignitions were successfully identified. Using the random forest classification model by Weihao Li (2020), ignition causes of ignitions were identified as arson, accident, burn off and lightning. Ignitions caused by lightning were selected.

4.2 Two Dimensional Density Estimation

In our case, we derived the density estimation of lightning-caused fire ignitions from historical fire ignition points from 2019 to 2020. There were several important considerations to be made before coming up with a less

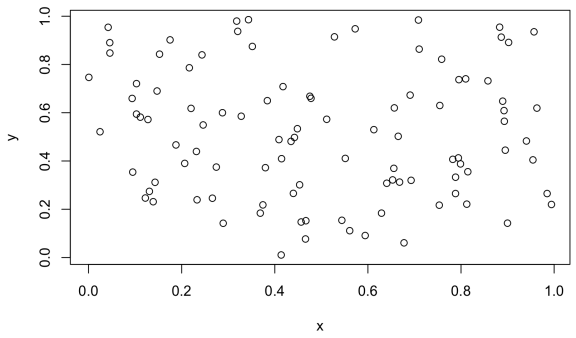

FIGURE 4.1: Random Points

biased estimation of density. In Silverman (1986)'s formula, the choice of the kernel does not affect the overall estimation much. However, the choice of bandwidth parameter H, can significantly affect the overall result of density estimation; since Epanechnikov proved that statistical results were not (significantly) affected by the choice of kernel function, most of the authors have emphasized the fact that bandwidth's choice is the crucial issue in this problem. To illustrate the influence of different bandwidths choice on the resulting density estimation, 100 random points were computed in Figure 4.1 below .

Using the kde2d function in R, with the choice of kernel set to a default two-dimensional Gaussian density distribution, and the bandwidth is set to 0.5 and 1 resprectively. The reuslt of these two density estimates represent density estimates in the format of a heat map in Figure 4.2 and Figure 4.3 below.

More red means the cell has a higher density estimate. The estimate with with bandwidth one on the x and y axis has a much smoother than the left one, smoother meaning the estimated density varies less for a given size of area. Where the one on the left, when the bandwidth is set to a lower value 0.5, the density estimate varies more in the same area, and we can see two

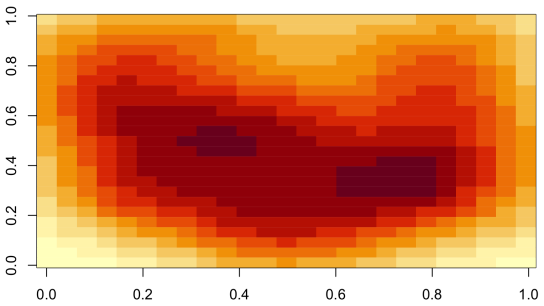

FIGURE 4.2: Left Heat Map

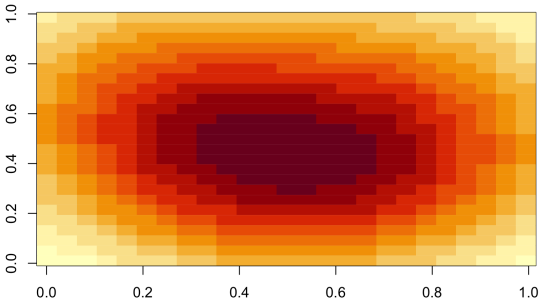

FIGURE 4.3: Right Heat Map

clusters of different densities for the figure on the left. Overall, a smaller bandwidth will result in a more wiggly estimate.

The Bias of density estimation of edges of Victoria map regions also needs to take into account. The bias of density estimation of edges happens because of the fact that we do not have what the density looks like outside the map, but still need to average the density base areas outside the map. Since the data region outside the boundary of Victoria is unknown, and sometimes, the boundary is connected to the sea, there are problems resulting from this boundary bis. When estimating fire density close to the frontier, since the bandwidth area that we consider may include an arbitrary area in the sea,

resulting in an underestimating of the fire density along the edge of the map. To address the problem of the boundary bias, K-functions, introduced in Ripley (1976), can be used to compute quantities with an edge correction, taking into account the boundary configurations of Victoria. The univariate kernel density estimator for a uniform kernel – also called 'moving histogram' – is defined as:

$$\widehat{f}_h(x) = \frac{1}{nh}\sum^{n} \mathbf{1}(d(x, X_i) \leq h), \text{ where } d(x, X_i) = |x - X_i|.$$

Which the choice of kernel, where only observations within a distance h of an arbitrary point (x,y) were considered, a "proper edge correction method" is:

$$\widehat{f}_h(z) = \frac{1}{nh}\sum_{i=1}^{n} \frac{1}{\omega_i(z)}\mathbf{1}(d(z, Z_i) \leq h), \text{ where } d^2(z, Z_i) = (x - X_i)^2 + (y - Y_i)^2,$$

where the weight,(z) is defined as the proportion of a circumference of a circle centered at point z that lies within the study area$S$. This method is called Ripley's circumference.

To compute the appreciate bandwidth for the estimation, in the context of product of (symmetric) kernels, one can prove using Taylor's expansion, that:

$$\mathbb{E}[\widehat{f}_h(z)] \sim f(z) + \alpha_1\left(\frac{h_X^2}{2}\frac{\partial^2 f}{\partial x^2}f(z) + \frac{h_Y^2}{2}\frac{\partial^2 f}{\partial y^2}f(z)\right) \text{ and } \mathrm{Var}[\widehat{f}_h(z)] \sim \frac{\alpha_2}{nh_X h_Y}f(z)^2,$$

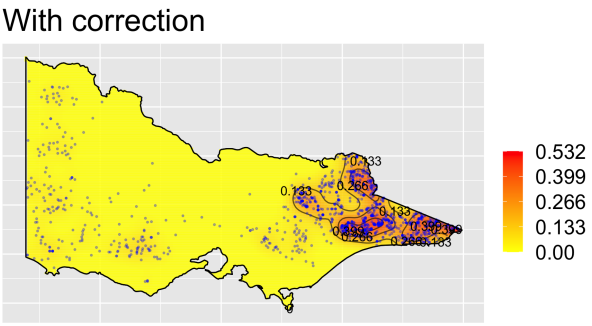

FIGURE 4.4: Density estimation result

where α_2 and α_2 are parameters related to the shape of the kernel function. The mean integrated squared error is then:

$$\text{MISE}(h) = \mathbb{E}\left[\int [\widehat{f}_h(z) - f(z)]^2 dz\right] \sim \alpha_3 h^4 + \frac{\alpha_4}{nh}$$

so $h = argmin\{MISE(h)\}$is $\alpha n - \frac{1}{5}$ for some constant . In the case where the true density f is a Gaussian distribution, with a diagonal variance matrix, Silverman's rule of thumb can be used:

$$h_i^\star \sim \left(\frac{2}{3}\right)^{\frac{1}{6}} \cdot \sigma_i \cdot n^{-1/6}.$$

Using R to apply the above methods, the resulting estimation is plotted in graph fi.demo1, the computed optimal bandwidth is 0.02606 terms of x and y-axis. Since our x and y-axis represent longitude and latitude in degrees, it is equivalent to around 2.9 kilometers in length on a two dimensional surface.

To compare different effect of different bandwidth, a bandwidth of 0.02606/2 is used to compute density estimate shown in Figure 4.5. This is equivalent

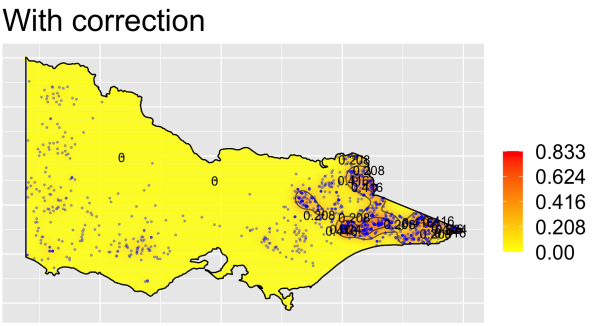

FIGURE 4.5: Density estimation result using small bandwidth

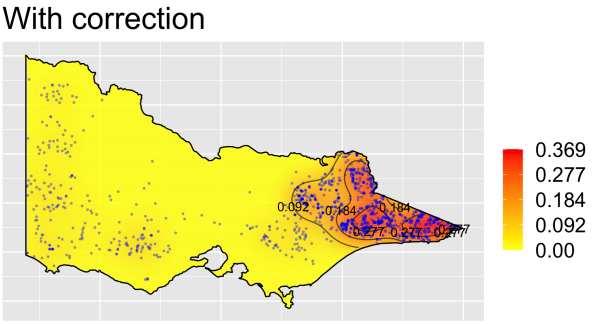

FIGURE 4.6: Density estimation result using large bandwidth

as a bandwidth of 1.45 kilometers in length on a two dimensional surface. Notice that the range of the density has been reduced, it is now 0 0.369, where previously when bandwidth was double (2.9 km in radius), the density estimate had a range of 0 0.532, when the bandwidth was about half. As bandwidth parameter increase, the overall density estimation is more smooth.

Another comparison is made by increasing the bandwidth to $0.02606 * 2$ as shown in Figure 4.6, this is twice the optimal bandwidth set out in the Silverman's rule of thumb. It is now equivalent to a distance of 5.8 kilometers. The estimation appears more wiggly. After comparing all the

different bandwidth choice, sticking with the optimal bandwidth set out in the Silverman's rule of thumb seems to be more sensible. it gives more detail than using a 8km but better than 1.45km where the result are really wiggly.

4.3 Modeling

After computing the estimated the risk value for each lighting ignitions in Victoria, the next step is to use different models to discover the relationship between different variables to the the density estimate. There are three main type of model that is included on this study. The first model is a decision tree model, and the second model is a random forest model. The performance of models are compared base on their testing set predictability. This also prevent the models from over fitting, that is, the situation where the noise or random fluctuations in the training data is picked up and learned as concepts by the model. Which will prevent us from finding out relationships between different variables role in contributing to the fire density, therefore preventing up to drawn a fair conclusion later.

4.3.1 Decision Tree Model

A decision tree makes decisions by splitting nodes into sub-nodes. This process is performed multiple times during the training process until only homogeneous nodes are left. Node splitting, or simply splitting, is the process of dividing a node into multiple sub-nodes to create relatively pure nodes. This process is performed multiple times during the training process until only homogeneous nodes are left. For predicting continuous output,

Reduction in Variance is used to decide node purity. There are several steps involved in the tree model splitting process. Firstly, for each split, the variance of each child node is calculated. Secondly, the variance of each split is calculated from the weighted average variance of child nodes. Then, the split with the lowest variance is selected. These three steps are repeated until completely homogeneous nodes are achieved.

Even though it is good to be able to interpret the model in a straightforward manner, tree models usually suffer from low variability; trees can have higher variability across different training data sets. To explore how a change in training can affect the tree model being produced, another decision tree model is fitted on a different training set sampled from the fire ignition data.

Looking at the visualisation of the new tree model being produced, we can see the model has significantly changed. With the root node being the average maximum temperature and the end nodes being completely different from the other decision tree model.

4.3.2 Random Forest Model

The last model included in this study is the Random Forest model, it address the disadvantage of the tree model by having two feature of randomness. Firstly, it prevents over fitting by using multiple trees sampled from the original fire ignition data set and produces the density estimation base on an average of different decision trees. Another instance of randomness is called feature bagging. It randomly selects subsets of explanatory variables used in each data sample, which adds more diversity to the dataset and reduces the correlation among decision trees. To fit a random forest model,

we first select the number of trees to fit. In this case, n is firstly selected to be the default value 500. R will bootstrap samples from the original data set 500 times, creating 500 training and testing sets. The algorithm will then fit a decision tree to each training set and predict the result by averaging all predicted values based on all trees. The number of trees is then increased to 1000 to compare the effect of the overall performance of these two models. The MSE of the testing set result shows the average mean squared residuals, which can be used to compare the performance of the smaller random forest model (500 trees) and the bigger random forest model (1000 trees). As a result, the smaller tree has an MSE of 0.001337, and the MSE for the larger random forest is 0.001367, which means the model is only slightly improved after fitting double the amount of trees.

Our research question is to find out the important variables that can impact the fire density estimate. For the random forest model, we can discover this by using a percentage increase in MSE as a measurement. %IncMSE is the most robust and informative measure for accessing the importance of different variables in the case of a regression treer random forest. It is the increase in mse of predictions as a result of a variable values randomly shuffled. This is done in R the following steps: R will firstly grow the grow regression forest. Compute testing set mse, name this will be MSE_0. Then, for variable j, permute their values, R will predict and compute the MSE_j again on testing sets. This step is repeated for all variables. The formula of calculating percentage increase in MSE is: $(MSE_j - MSE_0)/MSE_0 * 100\%$. The plot Figure 4.7 shows the variable importance ranked based on percentage increase in MSE.

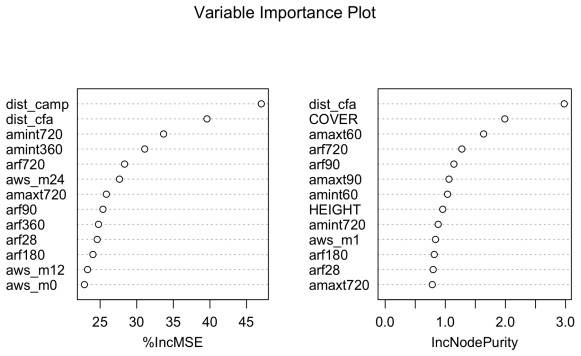

FIGURE 4.7: Variance Importance Plot

Chapter 5

Result

5.1 Result From Density Estimation

The result of density estimation is visualised in This shows the density estimate of all historical lightning ignitions. Each red dot represent a fire caused by lightning, they are plotted on the Victoria map. A lower density estimate is represented by a color gradient to represent a lower density of historical ignitions caused by lightning. From the estimation result, we can see higher fire density on the east side of Victoria. The highest density is the boundary region where Victoria is connected to New South Wales. The maximum density is 0.532. On the west side, fewer fire ignitions are caused by lightning.

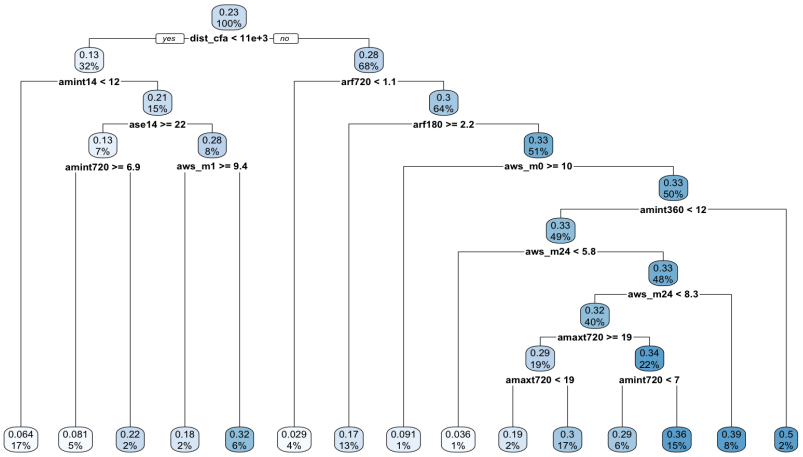

FIGURE 5.1: First Tree Model

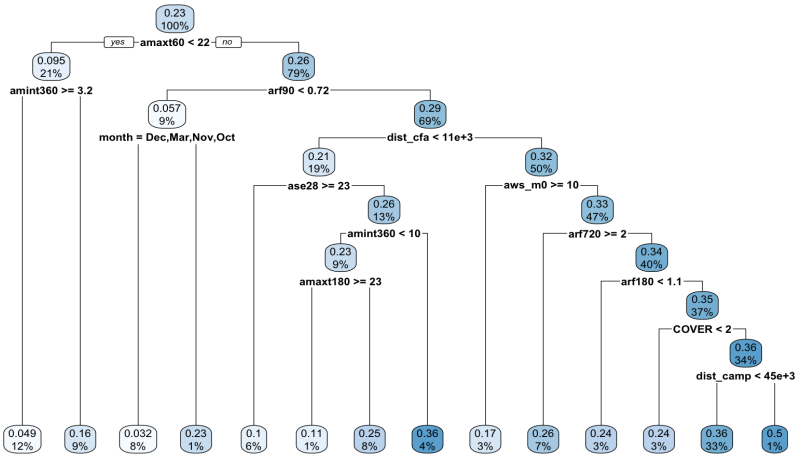

FIGURE 5.2: Second Tree Model

5.2 Result From Tree Model

Figure **??**

Looking at the two tree models in Figure 5.1 and Figure 5.2, each of them

has included different types of variables. These two models included mostly environmental variables, with one proximity to human activity variable, that is, the distance to the CFA station. The environmental variables included in this model are Average minimum temperature in the past 14 days, average rainfall in the past 720 days, average rainfall in the past 180 days, average solar explosion in the past 14 days, average wind speed in the past months, average wind speed in the current months, average minimum temperature in the past 360 days, average wind speed in the past 24 months, average maximum temperature in the past 720 days average minimum temperature in the past 720 days average maximum temperature in the past 720 days. We can see that average minimum temperature in the past 720 days, average wind speed in the past 24 months and average maximum temperature in the past 720 days were in the first tree model twice.

Looking at the second tree model, It included one proximity to human activity factor, that is distance to campsites. It also includes one vegetation information which is the crown cover. The rest of the variable that is important are all environmental variables. Variables included in the second model are the average maximum temperature in the past 60 days, average rainfall in the past 90 days, the month, distance to CFA Station, average solar exposure in the 28 days, average minimum temperature in the past 360 days, the average maximum temperature in the past 180 days, average wind speed in the current months, average rainfall in the past 720 days, average rainfall in the past 130 days. In the two different trees, we can see that the variables are different; however, they both included environmental factors and distance to CFA station. Since these two different decision tree model has a similar Mean squared error, We may say that including different type of environmental factors in general produce similar output in terms of explaining the variation in The density estimate.

Table 5.1: Comparison of RFM

no. of trees	MSE	% Variation explained
150	0.00140681	92.93
500	0.001354913	93.19
1000	0.001398205	92.98

On the other hand, because the environmental factors included in each of the tree models is different base on different training set data, It will be biased to rely on the result of this tree model. Even though these factors are included in these particular 23 models that we have produced. This means that if we go on and fit more tree models. We may not obtain similar variables as what these two models have included. Therefore, referring to the random forest model, finding the variable importance of many different trees can reduce this bias.

5.3 Result From the Random Forest Model

Using the default parameter in r for fitting a random Forest model, There are a total of 500 trees, and at each split, 19 variables were tried. The overall mean squared error for the default Random Tree Forest is 0.001354. The mean squared error is quite small, This is because the maximum density is around 0.5. The variation explained by the 500 trees random Forest model is 93.9 10%. This result shows that the random Forest model performs significantly better than the decision tree model. The decision tree model has a mean-square error of 0.0038.

To test out another possible number of trees, the RFM Figure **??** can be used as a guide. It shows the decrease in mean squared error as we increase the number of trees in our Random Forest model. And from Figure **??** we

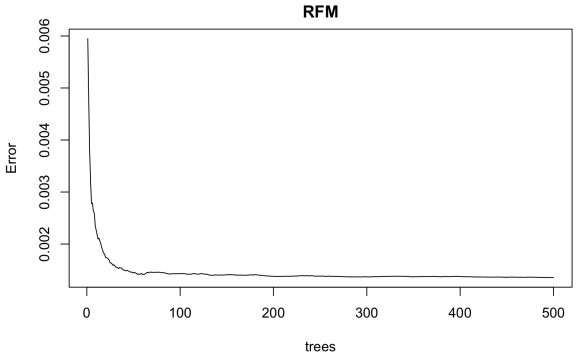

FIGURE 5.3: FRM plot

can see that after 200 the line seems pretty straight. This indicates that 200 number trees are sufficient enough to explain the variations in density, As adding more trees will not result in a better estimation of the density value.

The result shows that a Random Forest model with 150 trees has a mean squared error of 0.0014, and the variation explained is 92.93%. This is similar to when we fit 500 numbers of trees into the Random Forest model.

To test out other possible Random Forest models for performance comparison, another model with 1000 trees is built using R by changing the default parameter "ntrees" to 1000. Since we are fitting more trees, the number of variables tried at each split is set to decrease to 10 (previously, this is 19). This is to reduce the correlation between the different trees in the Random Forest model. The resulting MSE has then increased. This shows that adding more trees to the Random Forest model does not necessarily increase the out-of-sample performance of the model, in other trees, the predictability/ explaining power of the model. After comparing the performance of the three Random Forest Models, the one with 200 trees is the best model; since it provides fairly good estimation, and is less likely to be overfitted to our sample data. The resulting variance importance

plot is shown in Figure 4.7. The 20 most important variable in terms of percentage increase in mean squared error are: Distance to campsite, distance to CFA Station, average minimum temperature in the past 360 days, average minimum temperature in the past 720 days, average rainfall in the past 720 days, average wind speed in the past 24 months, average rain fall in the past 28 days, average rainfall in the past 360 days, average rainfall in the past 180 days, average wind speed in the past 12 months, average maximum temperature in the past 720 days, average rainfall in the past 90 days, average wind speed in the past month, average solar exposure in the past 360 days, average minimum temperature in the past 180 days, average solar exposure in the past 720 days, average solar exposure in the past 60 days distance to Road average wind speed in the current months.

To conclude the top 20 most important variables, the top two variables are both proximity to human activity variables. That is the distance to camp site variable and the distance to CFA station variable. They both have a value of percentage increase in the mean square error of more than ten per cent. This meaning when permuting these two variables, the resulting mean squared error of the new Random Forest model has increased by more than ten per cent. The next two most important variables are both related to the average minimum temperature, that is two different versions of the average, one is the average for 360 days and the other one is the average for 720 days. They also both have a percentage increase in mean squared error of more than ten percent. The other important variable are related to average minimum temperature as well as rainfall wind speed and average maximum temperature.

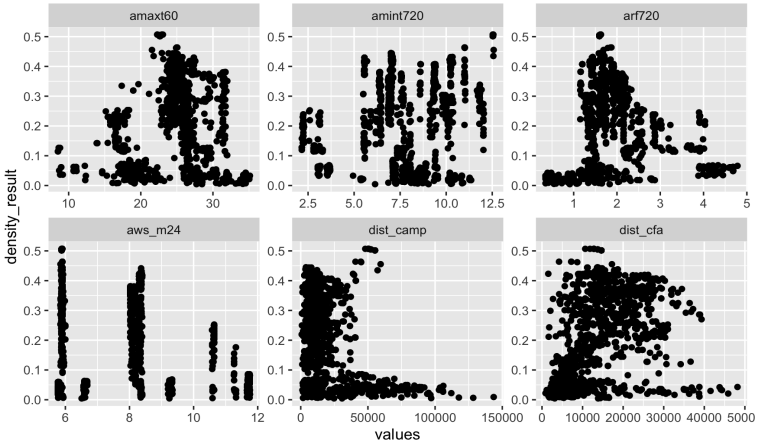

FIGURE 5.4: Scatter Plot

5.4 Explore the Effect of Change in Important Factors on the Density Estimation

5.4.1 Interpreting from Scatter Plot

A simple way to explore the relationship between the density estimate and the important variables pointed out in the previous section is by looking at a scatter plot. Setting the value on the x-axis to represent the value of exploratory variables and the value on the y-axis to the estimated density, a relationship is shown in Figure 5.4. As for the environmental factors, there is no obvious relationship to be shown in the plot. This may be because the relationship between these variables to the density estimate is not linear. There are no obvious patterns from the scatter plot for the environmental variables.

However, we do see some pattern between the proximity of human activity variables to the estimated density. For the distance to campsite plot, we

can see two clusters of scatter plots. This cluster on the left somehow shows as the distance to the campsite increase, the density variable tends to be higher. However, because we have another class here that does not comply with this relationship, where the density does not change much as the distance to the campsite increases. It is hard to make a sound interpretation. This might be because, in the middle of the Victoria map, we have a lot of hot spots with really low-density estimation, and their distance to campsites varies, creating the cluster on the bottom of the plot.

Looking at the scatter plot of distance to CFA. The plot is heteroskedastic, therefore the relationship between the distance to CFA station, and the density estimate, is again, non-linear. In some cases, as the distance increase, the estimated density increase; in some other observations, the increase in the distance to CFA does not increase the estimated density. Therefore, more investigation is needed to look at how these variables interact with the overall density estimate.

5.4.2 Interpreting from Random Forest Model

First, select a sample where the estimated density value is 0.0258139. This represents the fire ignition point that is located on a slot of the map where the estimated fire density value is 0.0258139. The average rainfall in the past 720 days for this particular ignition point used to be $1.301111 MJ/m^2$; changing it 5, the resulting predicted density estimation becomes 0.07566. This is rather not expected because when rainfall increase, the likelihood of fire ignition should decrease. However, this may be related to the fact that all the fire ignition point in the data set has been restrained to lightning. As this value represent the average rainfall in the past two years, the rain earlier may have produced more vegetation. When it is dry closer to the

fire ignition time, the large amount of vegetation may generate more combustible material, resulting higher density of fire estimate. To confirm this assumption, since the previous average rainfall in the past 6 days for this particular ignition point is more than 5, if we reduce it to 1, it is expected that the density estimation is likely to reduce. After changing the average rainfall in past 6 days variable to 1, the density estimate has then increased to 0.07701 from the previous value 0.0258139. This somehow proved our hypothesis earlier. However, to fully investigate the interaction between vegetation and the average rainfall in the past 720 days, more research should be done to make a better conclusion.

Chapter 6

Discussion and Limitation of the Study

There are several limitations due to the time and nature of this project. Firstly, even though density estimates were computed based on Ripley's correction method, this is still not the best outcome. The density estimate could have been more accurate if fire hotspots outside Victoria were included in the density estimating proceed. That way, we would make fewer assumptions about the hotspot density outside the boundary of Victoria. Due to the time limitation of the research, the relationship between the independent variable and the dependent variable fire density estimate could not be further investigated. As mentioneded in the last part of the result section, the interaction between different environment variable to the overall fire density could be better explained.

Bibliography