

# **Using Remote Sensing Data to Understand Fire Ignitions in Victoria during the 2019-2020 Australian Bushfire Season**

A thesis submitted for the degree of

Bachelor of Commerce (Honours)

by

Weihao Li

28723740



Department of Econometrics and Business Statistics

Monash University  
Australia

November 2020

# Contents

<b>Acknowledgement</b>	<b>1</b>
<b>Declaration</b>	<b>3</b>
<b>Abstract</b>	<b>5</b>
<b>1 Introduction</b>	<b>7</b>
1.1 2019-2020 Australia bushfires . . . . .	7
1.2 Remote sensing data . . . . .	8
1.3 Research objectives and contribution . . . . .	9
1.4 Scope of the thesis . . . . .	10
<b>2 Review of literature</b>	<b>13</b>
2.1 Spatio-temporal clustering . . . . .	13
2.2 Bushfire modelling . . . . .	15
<b>3 Data</b>	<b>19</b>
3.1 Data Sources . . . . .	19
3.2 Data processing for historical bushfire ignitions . . . . .	23
3.3 Compiled data . . . . .	24
3.4 Exploratory data analysis of historical bushfire ignitions . . . . .	24
<b>4 Detecting bushfire ignitions from hotspot data</b>	<b>29</b>
4.1 Overview of the hotspot data . . . . .	29
4.2 Outline of the algorithm . . . . .	30
4.3 Clustering results . . . . .	33
4.4 Data integration for ignition points in 2019-2020 . . . . .	36
<b>5 Classification of ignition causes</b>	<b>37</b>
5.1 Model description . . . . .	37
5.2 Feature selection . . . . .	38
5.3 Hyperparameter tuning and candidate model selection . . . . .	41
5.4 Results . . . . .	41
5.5 Predicting ignition cause for 2019-2020 season . . . . .	43
5.6 Fire risk map . . . . .	47

<b>6 Discussion and conclusion</b>	<b>51</b>
6.1 Policy Implications . . . . .	51
6.2 Limitation and possible extensions . . . . .	52
<b>A Covariate information</b>	<b>55</b>
<b>B Modified breadth-first search algorithm used in step 2 of the clustering algorithm</b>	<b>59</b>
<b>C Effects of parameter choices in the clustering algorithm</b>	<b>61</b>
<b>D Hyperparameter tuning for ignition classifiers</b>	<b>65</b>
<b>E Model performance</b>	<b>67</b>
<b>F Error rate map</b>	<b>69</b>
<b>G Supplementary material</b>	<b>71</b>
<b>Bibliography</b>	<b>73</b>



# Acknowledgement

I would like to thank my supervisors, Professor Dianne Cook and Emily Dodwell, for their invaluable support and guidance throughout this year.

R 3.6.3 (R Core Team, 2020), RStudio 1.3.959 (RStudio Team, 2020) and Python 3.7.1 (Van Rossum and Drake, 2009) are used for data analysis in this thesis.

R packages used in this research include tidyverse 1.3.0 (Wickham et al., 2019), sf 0.9.5 (Pebesma, 2018), lubridate 1.7.9 (Golemund and Wickham, 2011), rnaturalearth 0.1.0 (South, 2017), here 0.1 (Müller, 2017), raster 3.3.13 (Hijmans, 2020), bominr 0.7.0 (Sparks et al., 2020), geodist 0.0.4 (Padgham and Sumner, 2020), progress 1.2.2 (Csárdi and FitzJohn, 2019), GGally 2.0.0 (Schloerke et al., 2020), naniar 0.5.2 (Tierney et al., 2020), ggridges 0.5.2 (Wilke, 2020), ggthemes 4.2.0 (Arnold, 2019), caret 6.0.86 (Kuhn, 2020), lime 0.5.1 (Pedersen and Benesty, 2019), nnet 7.3.12 (Venables and Ripley, 2002), randomForest 4.6.14 (Liaw and Wiener, 2002), mgcv 1.8.31 (Wood, 2011), fastDummies 1.6.2 (Kaplan, 2020), xgboost 1.1.1.1 (Chen et al., 2020), plyr 1.8.6 (Wickham, 2011) and pROC 1.16.2 (Robin et al., 2011).

Python packages used in this research include tqdm 4.48.2 (Costa-Luis, 2019) and numpy 1.19.1 (Harris et al., 2020).

This thesis is produced using the R package bookdown 0.20 (Xie, 2020) with the R Markdown template for the honours thesis in the Department of Econometrics & Business Statistics, Monash University (Hyndman, 2020).



# **Declaration**

I declare that this thesis contains no material which has been submitted in any form for the award of any other degree or diploma in any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

.....  
Weihao Li

4<sup>th</sup> November 2020



# Abstract

Debate about whether lightning or arson was the cause of the devastating 2019-2020 Australian bushfire season, was fierce on social media. To study the cause of the recent fires, we collate data on historical bushfire origins, proximity to roads, campsites, CFA stations, leading weather conditions, vegetation types, along with new remotely sensed hotspot data. We develop a spatio-temporal clustering algorithm to detect bushfire ignitions from hotspot data. A random forest model, classifying cause, is trained on the historical data, and used to predict the cause of the 2019-2020 ignitions. The model, primarily using distance from roads, CFA stations and wind to achieve overall accuracy of 75% (91% for lightning). Lightning is the predicted cause in 82% of the recent Victorian fires, with arson at just 4%. The analysis using open data and open source software, is available to help fire authorities monitor and investigate future bushfire risk and causes.

*Key words:* machine learning, statistics, spatio-temporal data, cluster analysis, exploratory data analysis



# **Chapter 1**

## **Introduction**

### **1.1 2019-2020 Australia bushfires**

In Australia, bushfires have historically caused massive loss of property and life. Since the 1967 Hobart bushfires, insurance claims for building losses have exceeded 10 million Australia dollars (McAneney, Chen, and Pitman, [2009](#)). The 2019-2020 Australia bushfires had a more devastating impact on the environment and property compared with other major bushfires in history. According to a report by Lisa Richards, Nigel Brew and Smith ([2020](#)) published by the Parliament of Australia, 3094 houses were destroyed and over 17M hectares of land burned. These two figures represent the greatest in history. Fortunately, fewer lives were lost: 33 (including firefighters) compared to 173 in the 2009 Black Saturday fires and 47 in 1983 Ash Wednesday fires (Lisa Richards, Nigel Brew and Smith, [2020](#)).

Debate about the cause of this crisis was a focal point on social media at the beginning of the bushfire season. Some argued that climate change had a major impact on the catastrophic and unprecedented bushfires. They used hastag *#ClimateEmergency* on Twitter to convey their beliefs and promote action on climate change (Graham and Keller, [2020](#)). Climate Council, one of the biggest climate organizations in Australia made up of climate scientists and experts, claimed that climate change not only worsened the bushfire season but also increased the cost of fighting fires (Council, Climate, [2019](#)). Meanwhile, in the first few weeks of 2020, another group presented a completely different argument and claimed

the cause of this bushfire season was arson instead of climate change. They attempted to overtake the #ClimateEmergency hashtag with the #ArsonEmergency hashtag (Graham and Keller, 2020). Although police contradicted this controversial claim immediately (Knaus, 2020), the spread of this theory did not slow down at all. Research conducted by Dr. Timothy Graham and Dr. Tobias R. Keller (2020) from the Queensland University of Technology assessed around 300 Twitter accounts driving the #ArsonEmergency hashtag and found that a third of them were suspicious due to automated and inauthentic behaviour. They believed that the accounts were very likely run by disinformation campaigns. However, there was no direct evidence to support this hypothesis, particularly without knowledge of the owners of the accounts.

There is limited information on the cause of this catastrophic hazard in 2019-2020 (Lisa Richards, Nigel Brew and Smith, 2020). Bushfire investigation usually takes a certain amount of time and the result is not guaranteed. According to Beale and Jones's work (2011), the cause is known for only 58.9% of fires in history. Among the known cases, about 50% are due to suspicious and deliberate ignitions, 35% by accidents, and 6% by natural causes such as lightning. There is considerable variation in rates due to location, time, and difficulties of counting fire origins in a bushfire season.

Understanding the cause of 2019-2020 Australia bushfires may provide some help in developing effective strategies for mitigating bushfire impact. If the majority of bushfires are ignited by lightning, fuel management such as planned burns could be delivered more often and at a larger scale. Alternatively, if arson plays an important role in the ignitions, improvements in public legal education and research on environmental criminology may be required.

## 1.2 Remote sensing data

Remote sensing data are collected by remote sensors carried by a satellite or an aircraft. These sensors can detect energy reflected from the earth and record it as signals (Schowengerdt, 2006), which are then sent back to the Mission Control Center on the ground for further processing and the development of high-quality products, such as high-resolution colour images of Earth and sea surface temperature maps.

---

Contributing to the problem is that many bushfires start in very remote areas – locations deep into the temperate forests ignited by lightning – that are virtually impossible to access or to monitor. Remote sensing data, particularly satellite hotspot data, provide a possible solution to this challenge and may be useful in detecting ignitions and movement of bushfires. Newly available hotspot data collected from the Himawari-8 satellite (P-Tree System, 2020) has a 10-minute time resolution that could be used to detect fires almost in real-time.

Knowing the precise ignition location and ignition time allows us to reconstruct detailed information about conditions at the time of a fire's start, such as weather and distribution of vegetation. This may be helpful in understanding the cause of bushfires.

### **1.3 Research objectives and contribution**

The overall objectives of this research are to (i) develop an algorithm to detect time and location of ignitions for the 2019-2020 Australia bushfires in Victoria from satellite hotspot data, and (ii) build a predictive model to classify the potential causes of bushfire ignitions.

Clustering algorithms represent an unsupervised learning method to collect similar data points into the same group. To organize satellite hotspot data, this research proposes a spatio-temporal clustering algorithm inspired by two existing clustering algorithms, Density Based Spatial Clustering of Applications with Noise (**DBSCAN**) (Ester et al., 1996) and Fire Spread Reconstruction (**FSR**) (Loboda and Csiszar, 2007).

The core functionality of our clustering algorithm determines whether a hotspot is a new ignition point or a branch of an existing bushfire. This requires that the algorithm runs recursively in a temporal manner with reference to the bushfire dynamics. In contrast to traditional spatio-temporal clustering algorithms (Kisilevich et al., 2009), our algorithm slices the data by its temporal dimension and therefore divides the overall spatio-temporal clustering task into thousands of spatial clustering tasks. The final clustering results are obtained by combining membership information in different timestamps. This design overcomes one of the issues in defining the data space in traditional spatio-temporal clustering algorithms; the scaling of the temporal dimension can be highly influential to the

clustering result (Kisilevich et al., 2009). Additionally, our algorithm reconstructs clusters in parallel, and therefore mitigates the fire merging issue in the FSR algorithm. Other than that, since parameter tuning is not available in the FSR algorithm, a visualization tool is designed to enable us to choose near-optimal values of parameters.

In essence, our algorithm clusters hotspots into bushfires with arbitrary shape and size, which can help track the movement, coverage and intensity of every bushfire. Meanwhile, it provides an automatic process to detect new bushfire ignitions from the satellite hotspot data, which may be beneficial for urgent fire resource planning and deployment to remote areas in the future. Importantly, by applying this algorithm on satellite hotspot data from the past bushfire season, we are able to reconstruct fire events to enrich our knowledge of bushfire behaviour.

The general choice for bushfire modelling is the generalised additive model (**GAM**). For example, Read, Duff, and Taylor (2018) used **GAM** to model the risk of lightning-caused bushfires in Victoria. In this research, a spectrum of statistical models are examined as the classifier to predict the cause of the bushfire ignition, including **GAM**, as well as other computationally intensive models that are popular in the field of modern machine learning but rarely used in bushfire modelling. Our final model can be used to produce predictions of the cause of 2019-2020 Australia bushfires in Victoria. To our knowledge, there is not yet any academic research on modelling and predicting the cause of this bushfire season.

Overall, this research offers a complete workflow for bushfire analysis and monitoring, starting with the detection of bushfire ignitions from satellite hotspot data, through the production of predictions of the cause of these ignitions. We believe this workflow can be adopted into future bushfire investigation, particularly to establish the cause of bushfires.

## 1.4 Scope of the thesis

This thesis focuses on analysing 2019-2020 bushfires in Victoria with open-source remote sensing data. The remainder of the thesis is structured as follows. In chapter 2, we review the existing literature in the relevant fields. In chapter 3, we outline the data source and data integration, and provide a brief data exploratory analysis of the compiled datasets.

In chapter 4, we describe the clustering algorithm and present the clustering results. In chapter 5, we outline the strategy of model building and discuss the results. Finally, in chapter 6, we give a brief conclusion of the research and discuss limitations of our work and possible future extensions.



# Chapter 2

## Review of literature

### 2.1 Spatio-temporal clustering

To detect time and location of the bushfire ignitions from the satellite hotspot data, hotspots need to be grouped into meaningful clusters. This can be achieved by using spatio-temporal clustering.

Spatio-temporal data contain at least three dimensions, which are spatial location (e.g. latitude and longitude) and time. In the survey of spatio-temporal clustering algorithms, Kisilevich et al. (2009) summarised that there are five types of point-wise spatio-temporal data, including spatio-temporal events, geo-referenced variable, geo-referenced time series, moving points and trajectories. Table 2.1 shows that the classification of data is based on their temporal and spatial dimensions.

Meanwhile in satellite hotspots clustering, the data type is even more complicated. Each moving object possesses a spatial extension with arbitrary size and shape (Kisilevich et al., 2009).

**Table 2.1:** A classification of spatio-temporal point-wise data types (Kisilevich et al., 2009)

	Single snapshot	Updated snapshot	Time series
Fixed location	Spatio-temporal events	Geo-referenced variable	Geo-referenced time series
Dynamic location		Moving points	Trajectories

Before performing the clustering algorithm on spatio-temporal data, deciding how to represent the temporal information in a high-dimensional space is challenging (Kisilevich et al., 2009). This is because the end result of the clustering algorithm can be highly sensitive to the temporal resolution. Hence, scaling of the temporal dimension varies by application.

In the late '90s, researchers in the University of Munich proposed an influential density-based algorithm, Density Based Spatial Clustering of Applications with Noise (**DBSCAN**), for discovering clusters (Ester et al., 1996). Ester et al. (1996) mentioned three difficulties of clustering algorithms which motivated their work: (1) requirements of domain knowledge to determine the hyperparameters, (2) arbitrary shape of clusters and (3) computational efficiency.

**DBSCAN** defines a maximum radius  $\epsilon$  to construct densities around each point, which are spheres in the case of 3-dimensional space. To overcome the impact of noise, at least a minimum number of points (**MinPts**) are contained by the sphere. Finally, intersected spheres are considered to belong to the same cluster.

The time complexity of **DBSCAN** is  $O(N \log(N))$ , and the memory complexity is only  $O(N)$ . By visualizing the density of distance to the k-th nearest neighbour from each point, optimal values of  $\epsilon$  and **MinPts** can be chosen.

Although **DBSCAN** is popular among many fields, it is not suitable for clustering hotspot data. First, it does not address the issue of defining the scaling of the temporal dimension. Second, the behaviour of bushfires is not taken into account. The **DBSCAN** algorithm assumes the clustering rules work in both directions of a timeline, which does not capture the reality that bushfires evolve over time in one direction.

A better algorithm for satellite hotspots clustering is Fire Spread Reconstruction (**FSR**) (Loboda and Csiszar, 2007). This algorithm starts from the earliest observed hotspot. It constructs a tree based on three rules: (1) the earliest observed hotspot is the root of the tree, (2) any node is within a 2.5km radius from its parent and (3) any node is observed no later than four days from its parent. When the tree is closed and there are still unassigned

hotspots, the algorithm continues at the earliest unassigned hotspot to construct a new tree. Finally, each tree is a cluster, and the earliest hotspot is defined as the ignition point.

**FSR** works well in reconstructing bushfire spread, but it constructs clusters sequentially. If two fires start at a different locations but have overlapping coverage, **FSR** considers them to be a single fire when they meet each other. In this situation, the coverage of one of the fires increases dramatically during a short period of time, which does not correctly reflect the real speed of the bushfire. Besides, Loboda and Csiszar (2007) did not provide details of parameter tuning in their research. Instead, they briefly delivered the reason for choosing four days as the time threshold, which was because they wanted to tolerate missing observations due to cloud and smoke cover.

**DBSCAN** is mature and efficient with a parameter tuning tool, but is not suitable for temporal data like satellite hotspot data. **FSR** is designed for satellite hotspot data but lacks detailed consideration of parameter tuning and fire interactions. Inspired by these two algorithms, our work designs a spatio-temporal algorithm that can efficiently and robustly cluster hotspots with respect to the temporal behaviour of bushfires. Meanwhile, we provide a tool to help make parameter choices.

## 2.2 Bushfire modelling

Existing research on bushfire modelling can be divided into two main categories: simulation and analytical modelling.

In simulation modelling, Keane et al. (2004) developed landscape fire succession models (LFSMs) to simulate spatial fire behaviour, including fire ignition and fire spread, that accounted for fire and vegetation dynamics. Bradstock et al. (2012) used a model called **FIRESCAPE** which involved simulating fire behaviours with fuel and weather conditions. Simulation methods are cost- and time-effective ways to model bushfires, but ignition cause is not considered (Clarke et al., 2019). These methods seldom address the ignition methods that we are interested in.

Analytical modelling is another common way to build bushfires models. A commonly used modelling framework for analysing bushfire ignition is the generalised additive model

(**GAM**). Bates, McCaw, and Dowdy (2018) used this model with a Gaussian distribution to predict the number of lightning ignitions in Warren Region, Western Australia. Some studies implemented **GAM** with a binomial distribution to predict ignition probabilities in New South Wales, Australian Capital Territory and Victoria (Read, Duff, and Taylor, 2018; Zhang, Lim, and Sharples, 2017). Mixed-effect models are an alternative to incorporate spatial dependence and weather factors. Duff, Cawson, and Harris (2018) showed that by treating drought factor and intercept as random effects, and forest type as a grouping variable, a mixed-effect model with only three variables achieved 70% mean accuracy to predict fire occurrence in Southern Australia. Simpler parametric models have also been widely used, including multiple linear regression, negative binomial regression and generalised logistic regression (Cheney et al., 2012; Plucinski et al., 2014; Collins, Price, and Penman, 2015). Instead of modelling bushfires, some researchers conducted statistical tests and exploratory methods to assess hypotheses about bushfires (Miller et al., 2017; Dowdy, Fromm, and McCarthy, 2017).

Common covariates for bushfire ignition analysis are weather conditions, vegetation types, topographic information and the environmental impact of human activities. In addition, various fire danger indices have been used in modelling. Some studies chose to use index variables developed by McArthur, such as the Forest Fire Danger Index (Clarke et al., 2019; Read, Duff, and Taylor, 2018), while others chose to use indices developed by the Canadian Forestry Service, such as Canadian Fire Weather Index and Drought Code (Plucinski et al., 2014). We question the usefulness of these indices for bushfire ignition analysis since they are mostly extracted from weather and vegetation information. Comparing the paper written by Zumbrunnen et al. (2012) to the authors' previous work (2011) on the same topic, the fire weather index was replaced with temperature and precipitation for climate proxies, as the latter were available for the entire study period. Similarly, fire weather indices are not be considered in our research given that we have access to climate and vegetation data.

Throughout review of the existing literature, we have not found any research that attempts to model the cause of the bushfires in Australia, nevermind doing so with hotspot data. Therefore, we build a model by incorporating parts of the modelling framework and

---

covariate choices mentioned above, and we use hotspot data to predict the cause of bushfires in Victoria during the 2019-2020 season.

Although numerous studies for ignition analysis have applied semi-parametric and parametric methods, little use of machine learning models has been made. From a predictive modelling perspective, exploiting modern algorithms to obtain better prediction performance is vital. Hence, these approaches are considered for this research.

Most of the existing work on bushfire ignition focuses on fewer than three states in Australia and models are rarely applied to all of Australia. This is mainly due to the lack of coherence in bushfire datasets provided by different states in Australia. For the same reason, we only focus on bushfires in Victoria in this research.



# **Chapter 3**

## **Data**

### **3.1 Data Sources**

One of the focuses of this work is to utilise open data, and collates these datasets to provide a data fusion with which to tackle the research questions. The motivating data source is satellite hotspots, which is different from what have been analysed previously in the literature. This data is collated with weather records, fuel layer, location of roads, fire stations and recreation sites, and causes of historical fire ignitions. The spatial and temporal details of each dataset are provided in Table 3.1. In addition, the original data types of each dataset are provided in Table 3.2. In both tables, there is an Index column indicates the dataset usage, which will be further discussed in section 3.1.8.

#### **3.1.1 Satellite hotspot data**

To track bushfires in Australia remotely with high temporal and spatial resolution, we use hotspot data taken from the Himawari-8 satellite (P-Tree System, 2020). This hotspot data is available on the Japan Aerospace Exploration Agency FTP site. We download only the data from October 2019 to March 2020, and details on how to download this data are provided by Williamson (2020). The dataset contains records of 1,989,572 hotspots for these six months in the full disk of 140 °east longitude.

### 3.1.2 Climate data

To better understand bushfires, we collect climate data from Bureau of Meteorology (**BOM**) and Commonwealth Scientific and Industrial Research Organisation (**CSIRO**). Weather properties including maximum temperature, minimum temperature, rainfall and solar exposure are retrieved via an open-source R package `bomrang` (Sparks et al., 2020), which is a data client of **BOM**. Historical weather records of 885 weather stations across Australia from January 1863 to April 2020 are downloaded. However, the number of publicly available climate attributes on **BOM** are limited. In order to maintain the coherent reproducible workflows, we download near-surface wind speed grids across Australia from January 1975 to December 2018 from **CSIRO** (McVicar, 2011) and the station-based wind speed from January 2017 to August 2020 from the Australia Automated Surface Observing System (**ASOS**) network mirrored by Iowa State University (Iowa State University, 2020). There are only a limited number of **ASOS** stations in Victoria and they are mostly located around Melbourne. Therefore, even though **ASOS** data is also available from 2000-2018, **CSIRO** data is used.

### 3.1.3 Road map

Since bushfires can be impacted by anthropogenic factors, we use the road map from the comprehensive open-source **OpenStreetMap** (OpenStreetMap contributors, 2020) to represent the reachability of bushfire ignition locations. The road map is one of the layers of the full archive, and consists of 1,797,217 roads belonging to 27 different road classes in Australia.

### 3.1.4 Fuel layer

Vegetation information is obtained from a 2018 nationwide forest dataset compiled by Australian Bureau of Agricultural and Resource Economics and Sciences (2018). This is the fifth and the latest national State of the Forest Report (**SOFR**). Previous national **SOFRs** were published in 1998-2013 and superseded.

### 3.1.5 Fire stations

Data of Country Fire Authority (**CFA**) fire stations are retrieved from Department of Environment, Land, Water & Planning (2020[a]). It contains 52,716 Victorian topographic features, including rivers, water bodies, transport, facilities and fire stations.

### 3.1.6 Recreation sites

Camping activities may be associated with accidental human-caused bushfires. Therefore, we download Victorian recreation sites from Department of Environment, Land, Water & Planning (2020[b]). The dataset contains 417 camping locations in Victoria.

### 3.1.7 Fire origins

We use the Victorian Department of Environment, Land, Water and Planning (**DELWP**) Fire Origins dataset to obtain historical bushfire ignition locations and causes (2019). This dataset provides each first reported location of the fire recorded by crews rather than the exact origin of the fire, which can be considered as a great approximation of ignition points.

### 3.1.8 Summary of dataset usage

In Table 3.1 and Table 3.2, column Index indicates different usage for datasets in this research. 1 is the satellite hotspot data, 2 is the wind speed data retrieved from **ASOS**, 3 are supplementary datasets and 4 is the historical bushfire ignition data.

1 is used in spatio-temporal clustering algorithm to identify ignition location and time. 3 combined with 4 are used as training, validation and test set to build a bushfire ignition classifier. 2 and 3 combined with the clustering results are used as predictors to predict the cause of bushfire ignitions in Victoria during the 2019-2020 season.

The summary of dataset usage and research workflow is provided in Figure 3.1.



**Figure 3.1:** Summary of dataset usage and research workflow

**Table 3.1:** Raw data information

Index	Data set name	Spatial Resolution	Temporal resolution	Time
1	Himawari-8 satellite hotspots data	0.02° ≈ 2km	Per 10 minutes	2019-2020
2	Australia ASOS - wind speed	mph	Hourly	2017-2020
3	Bureau of Meteorology climate data		Daily	1863-2020
3	CSIRO - near-surface wind speed	2° ≈ 200km	Daily	1975-2018
3	Openstreetmap - road map	2m		2020
3	Forest of Australia - fuel layer	100m		2018
3	Victorian CFA fire stations	20m		2020
3	Victorian recreation sites	10m		2020
4	Vicotiran fire origins	100m	Daily	1972-2018

**Table 3.2:** Raw data types

Index	Dataset name	Type	Files	Total size (MB)
1	Himawari-8 satellite hotspots data	Comma separated values file	6	908
2	Australia ASOS - wind speed	Comma separated values file	1	25
3	Bureau of Meteorology climate data	Comma separated values file	1	740
3	CSIRO - near-surface wind speed	Binary raster file	16070	384
3	Openstreetmap - road map	Shapefile - LINESTRING	1	2078
3	Forest of Australia - fuel layer	Amiga disk file	1	125
3	Victorian CFA fire stations	Shapefile - POINT	1	30
3	Victorian recreation sites	Shapefile - POINT	1	2.5
4	Vicotiran fire origins	Shapefile - POINT	1	22

## 3.2 Data processing for historical bushfire ignitions

The programming language we use to perform data manipulation for this research is R (R Core Team, 2020). Victorian historical fire origins, recreation sites, CFA fire stations and the road map are geospatial data in the shapefile format, which are processed using the tools in `sf` (Pebesma, 2018). Fuel layer and near-surface wind speed are also geospatial data but are stored in the grid format. We use tools in `raster` (Hijmans, 2020) to manipulate the data. Other datasets including satellite hotspot data and BOM climate data are stored in csv format, which can be handled using the package `tidyverse` (Wickham et al., 2019).

We only keep historical fire origins since 2000 due to the limitation of other supplementary datasets. We drop all cases with unknown cause from the dataset under the assumption that the major causes distributed in a same way in cases with unknown cause. Given the dataset contains fires other than bushfires, we drop non-bushfire cases based on their cause. Fires caused by reasons including "WASTE DISPOSAL", "BURNING BUILDING" and "BURNING VEHICLE" are not considered as bushfires. Causes of fire are then recategorized into 5 classes, including lightning, accident, arson, burning off and others. We end up having very few cases in others (0.9%) which motivates us to drop it.

BOM weather records including maximum temperature, minimum temperature, solar exposure and rainfall are station-based data. Weather stations are distributed unevenly across Victoria and are often under maintenance. Therefore, we match every historical fire origin to its 10 nearest weather stations, and then extracted the nearest non-missing daily weather metrics. Differently, CSIRO wind speed data provides us with a full grid of values in Victoria. Therefore, we project historical fire origins onto their corresponding grid entries to extract the daily average wind speed. Finally, weather metrics in the past 720 days for every historical fire origin are summarised into several numeric variables which can be found in the Appendix.

Besides, CFA stations locations are filtered from the raw data by matching the FEATSUBTYP field equal to "fire station". Natural logarithm of the distance to the nearest recreation site, fire station and road are computed for every historical fire origin.

Vegetation information included vegetation type, height and cover are stored in the grid format which is extracted in the same way as the [CSIRO](#) wind speed data.

We also truncate hotspot data by using a recommended threshold of fire power ([Williamson, 2020](#)), which is over 100 (irradiance over 100 watts per square metre), to reduce noise from the background, and then select the hotspots within the boundary of Victoria. The further steps of data processing for hotspot data including clustering and joining with other covariates will be introduced in chapter [4](#).

### **3.3 Compiled data**

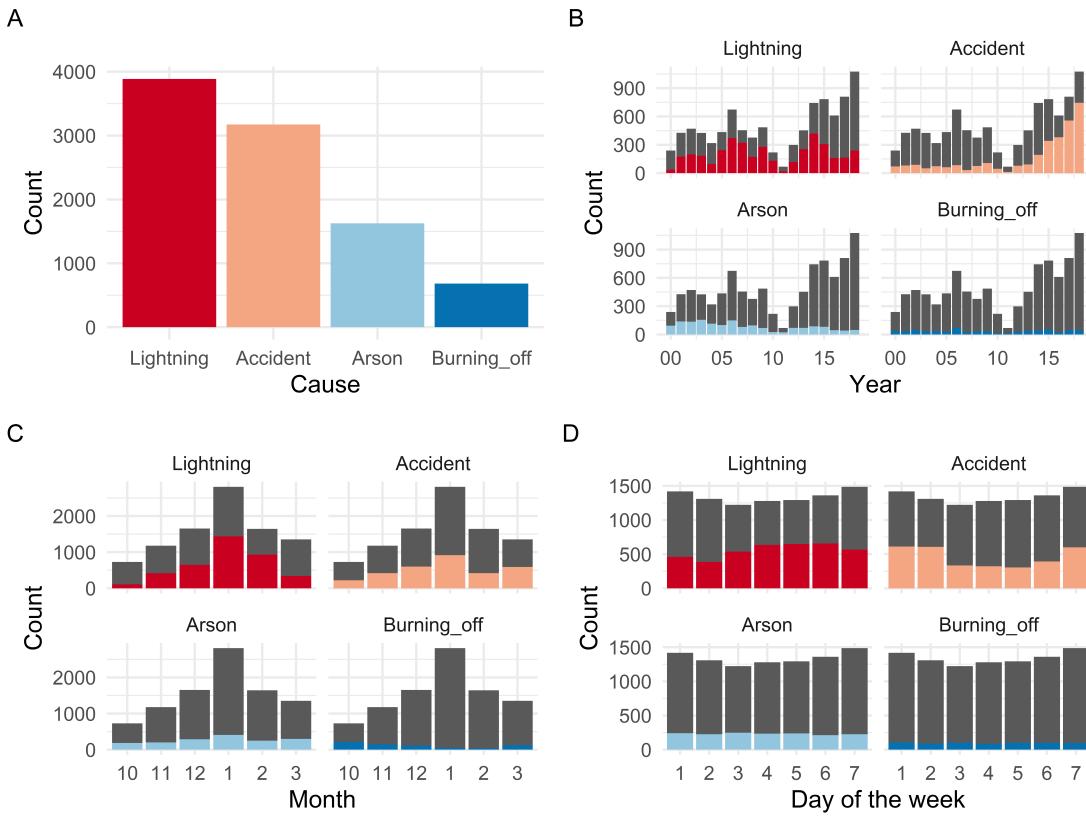
The end result of the data cleaning are two different datasets. One is the compiled dataset used for model fitting which contains 9369 observations and 55 fields, including a field representing the cause of the bushfire ignition. The other is the satellite hotspot data for clustering, which contains 75,936 observations and four fields: unique identifier, longitude, latitude and time.

Covariates for fitting the ignition classifier include month, day, day of the week, longitude, latitude, forest type, forest crown cover, forest height, rainfall in the past 2 years, wind speed in the past 2 years, temperature in the past 2 years, solar exposure in the past 2 years, as well as natural logarithm of proximities to the nearest fire station, recreation site and road. Additional details about the covariate information can be found in the Appendix.

### **3.4 Exploratory data analysis of historical bushfire ignitions**

By performing the exploratory data analysis on the historical bushfire ignitions, we explore the relationship between covariates and causes of ignition.

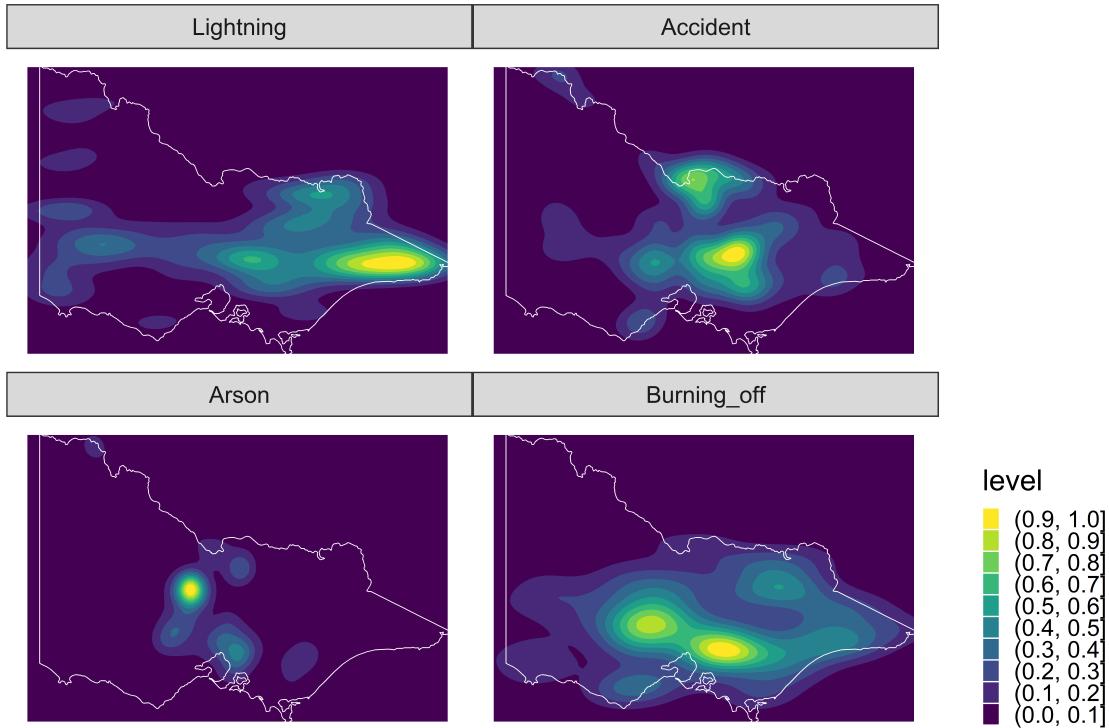
From January 2000 to December 2018, Victoria recorded 3886, 3173, 1627 and 683 bushfire ignitions caused by lightning, accident, arson and burning off, respectively. The histogram of ignition cause is given in (A) of Figure [3.2](#). We compare the observed distribution of causes of bushfires to that proposed by Beale and Jones ([2011](#)). Ignitions caused by



**Figure 3.2:** Histogram (A) and bar plots of four different causes of historical fire ignitions in Victoria by year (B), month (C) and day of the week (D). (A) Lightning and accident were the two main sources of bushfire ignitions, which took up 41% and 34% respectively. Proportion of ignitions caused by arson was around 17%. 7% of ignitions were reported as planned burns. (B) Accident-caused ignitions increased significantly since 2012. Meanwhile, there were only few cases in 2011, which was because 2011-2012 bushfire season mainly affected Western Australia instead of Victoria. (C) January was the most serious month in the bushfire season throughout these years. Besides, lightning-caused ignitions usually occurred in the hottest months. (D) Ignitions were almost equally likely to occur on every day of the week. However, accident-caused ignitions were more often on Sunday, Monday and Tuesday.

arson were significantly less than the data provided in their review (17% vs. 50%), while lightning-caused ignitions taking a greater percentage (41% vs. 6%).

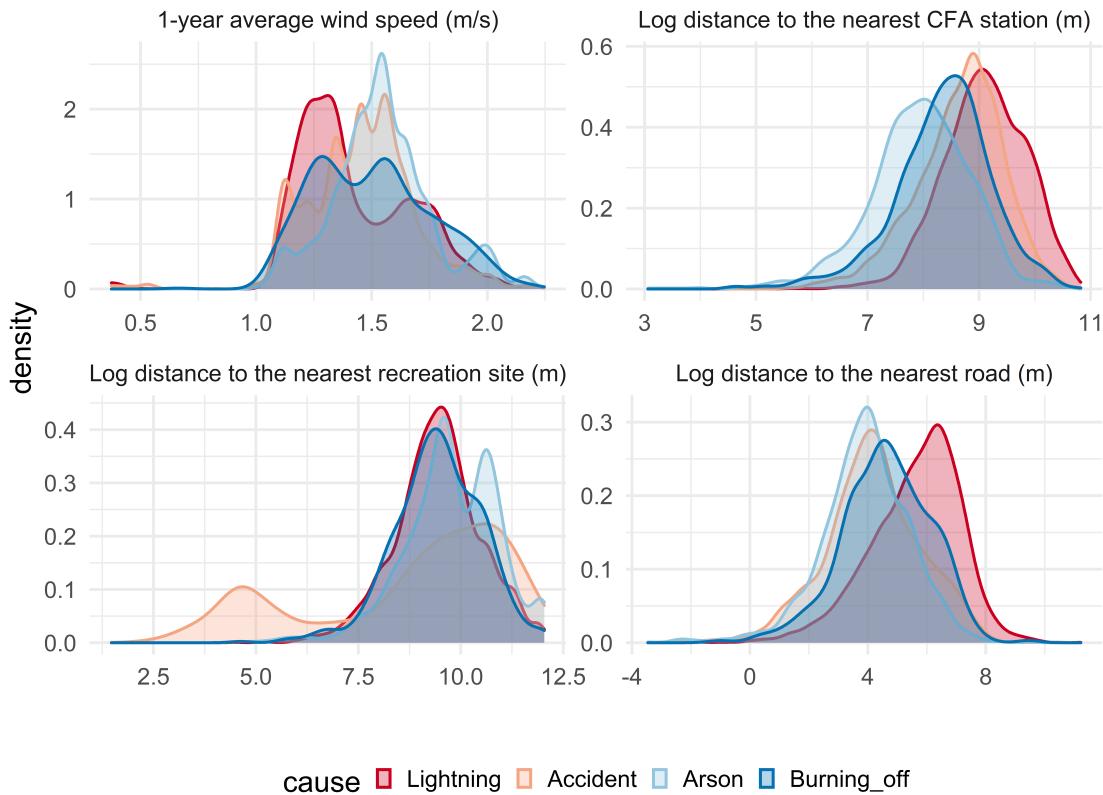
There was an abnormal increasing trend of accident-caused ignition over the recent years which is shown in (B) of Figure 3.2. In addition, from (C) of Figure 3.2, we find that lightning-caused ignitions were most likely occurred in January and February, which were the hottest months in Victoria. It indicates that lightning-caused ignitions were related to temperature. Moreover, according to (D) of Figure 3.2, people were careless in managing fire risk on Sunday and in the first two days of the week.



**Figure 3.3:** 2D density plot of historical bushfire ignitions. The density plot shows arson and accident-caused ignitions were different from each other. Accident-caused ignitions were denser. In addition, lightning-caused ignitions were less likely to occur in Melbourne metropolitan area.

In Figure 3.3, lightning-caused ignitions were more likely to occur in the mountain area of Victoria and human-caused ignitions were closer to the urban area. It suggests that spatial pattern might be useful in ignition method classification.

According to the density plot of 1-year average wind speed shown in the upper left of Figure 3.4, the slower yearly wind speed was an indicator of lightning-caused ignitions in contrast to ignitions caused by other sources. In the upper right of the same figure, we find that lightning-caused ignitions were far from the nearest fire station and arson-caused ignitions were closer to the nearest fire station. It was potentially because the population density around the fire station was larger. Besides, lightning-caused ignitions were less reachable shown in the bottom right of the Figure 3.4. Furthermore, in the bottom left of the figure, accident-caused ignitions were significantly close to recreation sites, which suggests that camp fire was a source of bushfire ignitions.



**Figure 3.4:** Density plot of 1-year average wind speed (Upper left), log distance to the nearest CFA station (Upper right), log distance to the nearest recreation site (Bottom left) and log distance to the nearest road (Bottom right). (Upper left) Wind speed was one of the key factor to classify different types of bushfire ignitions. Lightning-caused ignitions were more likely to occur in windless years. (Upper right) Arson-caused ignitions were closer to fire stations. In contrast, lightning-caused ignitions were far from the fire stations since they were often occurred in the mountain area. (Bottom left) Accident-caused ignitions were very likely to be camp fire. (Bottom right) Lightning-caused ignitions were significantly less reachable, which could cause difficulties to fight them.

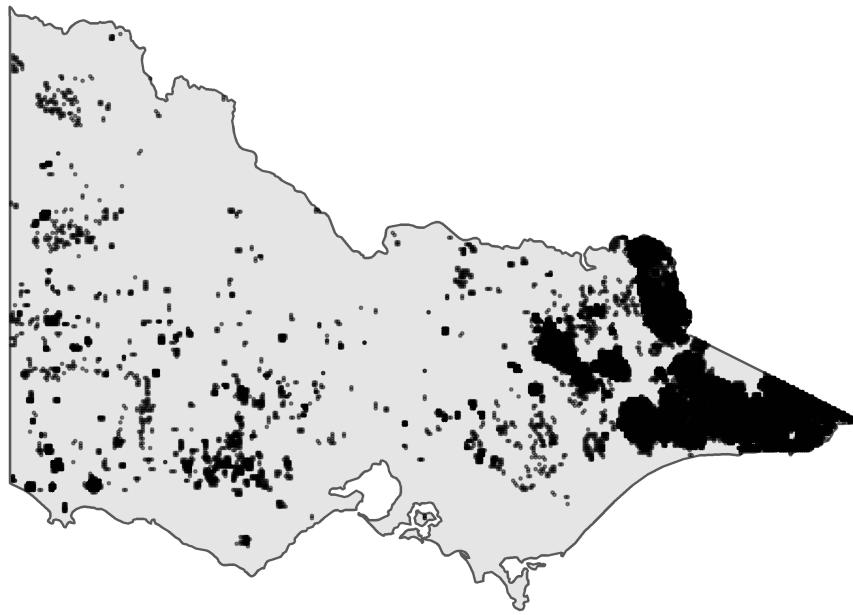


## **Chapter 4**

# **Detecting bushfire ignitions from hotspot data**

### **4.1 Overview of the hotspot data**

The hotspot data contains 75936 observations (shown in Figure 4.1). There is a large concentration of hotspots in the east of Victoria which points to the large scale effect of the fires. Generally, though hotspots are dispersed throughout the state, in many separated small clumps. To determine the locations where the bushfires started, these isolated hotspots can be easily labelled with a cluster membership, and the first time that a hotspot appears in the region would be considered the ignition point and time. Zooming in to the denser hotspot regions sees some concentrated clusters, but is generally much harder to tease apart into unique fires. Spatio-temporal clustering will be developed which uses the first appearance of a hotspot, and spatial distance from other hotspots to group the hotspots into clusters and identify the ignition point and time.



**Figure 4.1:** Distribution of 75936 satellite hotspot data in Victoria from October 2019 to March 2020. Most of the hotspots are observed in the east of Victoria. Hotspot data are difficult to use in ignition analysis.

## 4.2 Outline of the algorithm

Some clustering algorithms for spatio-temporal data exist but require modifications to work for this problem. There are six conditions that the clustering algorithm needs to satisfy, (1) it does not require any scaling of the temporal dimension, (2) it iteratively runs in the positive direction of the timeline to mimic the bushfire dynamics, (3) hotspots at later timestamps have no impact on the clustering result of hotspots observed earlier, (4) it keeps bushfires separate even they are connected at later timestamps, (5) it is a time-efficient and memory-efficient algorithm that is capable for the personal computer and (6) all its parameters can be tuned via diagnostic tools. These form the basis of our own algorithm for hotspots clustering.

The input for the algorithm is the hotspot data with 4 attributes: unique identifier, longitude, latitude, and time. The output is a list of labels describing the corresponding clusters

of hotspots. There are four steps: (1) slice the temporal dimension, (2) cluster hotspots spatially, (3) broadcast and update the membership label and (4) compute ignition locations. These are described in details in the following sections.

#### 4.2.1 Slice the temporal dimension

This step uses the parameter named *ActiveTime*, which controls the size of the interval. A natural interpretation of this parameter is the time a fire can stay smouldering but undetectable by satellite before flaring up again. In Loboda and Csiszar's work (2007), a similar parameter is used as the tolerance time for the missed observations due to cloud cover.

Given a fixed value of *ActiveTime* and the length of the time frame  $T$ , the algorithm will slice the timeline into several intervals:

$$S_t = [\max(1, t - \text{ActiveTime}), t], t = 1, 2, \dots, T$$

, where  $T$  and  $t$  have the same unit as *ActiveTime*. For example, if we have 48 hours of data and we set *ActiveTime* = 24 hours, the algorithm will produce 48 intervals,  $S_1 = [1, 1], S_2 = [1, 2], \dots, S_{25} = [1, 25], S_{26} = [2, 26], \dots, S_{47} = [23, 47], S_{48} = [24, 48]$ .

#### 4.2.2 Cluster hotspots spatially

The spatial clustering is computed within a time interval. There is one parameter, *AdjDist*, which represents the potential distance a fire can spread with respect to the temporal resolution of the data. For example, if *AdjDist* = 3km and the temporal resolution of the data is 10-minute, then the potential speed of the bushfire is  $3\text{km}/10\text{min} = 18\text{km/h}$ . Similarly, in Ester et al. (1996), a radius-based parameter  $\epsilon$  is defined to connect neighbouring points.

Given a fixed value of *AdjDist* and the interval  $S_t$ , conceptually the algorithm will:

- (a) Connect each pair of hotspots  $h_i$  and  $h_j$  if the proximity from  $h_i$  to  $h_j$  is less or equal to *AdjDist*, where  $h_i$  is the  $i$ th hotspot in the interval  $S_t$  and  $h_j$  is the  $j$ th hotspot in the interval  $S_t$ .

- (b) A connected component is called  $M_{t,n}$ ,  $1 \leq n \leq N$ , where  $N$  is the total number of components in the interval  $S_t$ . Hotspots in the same component  $M_{t,p}$  will be assigned with the same unique membership label, where  $M_{t,p}$  is the  $p$ th component in  $S_t$ .

The practical implementation of the clustering algorithm involves using a modified version of the breadth-first search algorithm to reduce the memory complexity of the clustering algorithm from  $O(N^2)$  to  $O(N)$ . Details information about this can be found in the Appendix.

#### 4.2.3 Broadcast and update the membership label

With current clustering results, the next step is to broadcast the clustering results from earlier intervals to update the membership label iteratively.

This step starts from the second interval  $S_2$  till the last interval  $S_T$ . Given the interval  $S_t$ , the algorithm will:

- (a) If a hotspot  $h$  belongs to both  $S_{t-1}$  and  $S_t$ , it will succeed its membership label from  $S_{t-1}$ . We call these hotspots  $H_s = \{h_{s,1}, h_{s,2}, \dots\}$ .
- (b) If a hotspot  $h$  belongs to  $S_t$ , but does not belong to  $S_{t-1}$ , we call these hotspots  $H_c = \{h_{c,1}, h_{c,2}, \dots\}$ . If a hotspot  $h_i \in H_c$  shares the same component  $M_{t,p}$  with hotspots  $H_j \in H_s$ , it will succeed the membership label from the nearest hotspot  $h_{nearest} \in H_j$ .

Overall, substep (a) broadcasts the membership labels from previous clustering results by linking identical hotspots in both  $S_{t-1}$  and  $S_t$ . Substep (b) spreads the membership labels according to the clustering result in  $S_t$ . The algorithm combines the knowledge in different intervals to produce the overall clustering result. By far, every hotspot has its membership label.

#### 4.2.4 Compute ignition locations

The earliest observed hotspot of a cluster is used as the ignition point. If there are multiple earliest hotspots within a cluster, the centroid of these points will be computed and used as the ignition point.

#### 4.2.5 Satisfying the six conditions

In terms of condition (1), the algorithm doesn't depend on the scaling of the temporal dimension, but instead uses the parameter *ActiveTime* to control the time tolerance. Condition (2) and (3) are embedded in the algorithm. The two substeps for label updating in step 4 is designed to meet condition (4). For condition (5), the memory complexity of this algorithm is  $O(N)$ , and the time complexity is  $O(N^2)$ . Although the time complexity of this algorithm is not as great as **DBSCAN**, it is still acceptable for the personal computer. Justification that condition (7) is satisfied is provided in the Appendix.

### 4.3 Clustering results

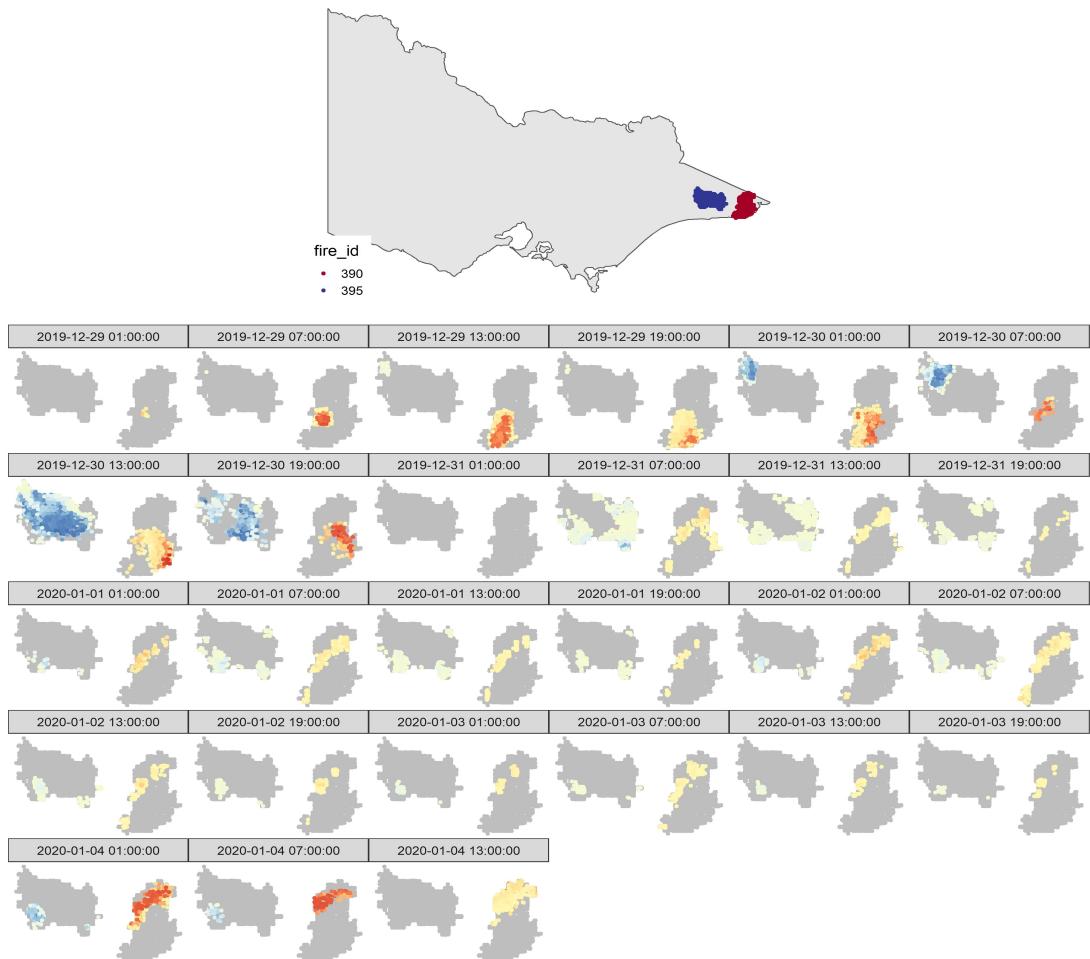
The clustering results by applying the algorithm on the hotspot data in Victoria is shown in Figure 4.2. From the distribution of ignition points (Figure 4.2), we do not observe too many overlapped points. This is consistent with the fact that the environment around the ignition point is often severely damaged and is less likely to be ignited again. Comparing to the hotspot data (Figure 4.1), ignition points can provide us with direct information of ignition locations and ignition time. This is crucial for our research on the cause of bushfires.

By using the algorithm, we can also reconstruct the bushfire dynamics to study the bushfire behaviour. Figure 4.3 shows two fires and their lifetime movements as produced by the algorithm. From this plot, we can diagnose several features of the algorithm. First, the algorithm can provide the ignition, movement, coverage and intensity of the bushfire. Second, the parameter *AcitiveTime* is working in the algorithm. At 01:00 am 31 December 2019, there are missing values but the algorithm continues reconstructing the bushfire. Finally, bushfires reconstructed by this algorithm can have arbitrary size and shape and



**Figure 4.2:** Distribution of 1022 ignitions in Victoria from October 2019 to March 2020 produced by our spatio-temporal clustering algorithm. Most of the ignitions are distinct spatially.

multiple branches far away from each other. This feature is essential for the correctness of the algorithm.



**Figure 4.3:** The snapshots of hotspots show the bushfire behaviour of bushfire “390” and bushfire “395”, which is two clusters from the results of the clustering algorithm. Fire “390” started at 01:00 am 29 December 2019 and fire “395” started at 07:00 am 29 December 2019. The grey shadow behind the hotspots is the coverage of the bushfire. The darker the colour of the hotspot, the higher power it contains. Missing hotspots data at 01:00 am 31 December 2019 suggests that there was smoke above that area, but it does not trouble the algorithm. Moreover, the split of hotspots within a bushfire shows that the algorithm can tackle bushfire with fronts in different directions.

## 4.4 Data integration for ignition points in 2019-2020

We discuss the data integration for historical fire origins in the data processing section, but we haven't outline the data integration for the clustering results. The only difference is the source of the wind speed data, other covariates such as vegetation factors, temperature and rainfall follow the same procedure to join with the ignition points as provided in the data processing section. In terms of the wind speed, we use the daily **ASOS** wind speed data from the nearest station for each ignition points. We then convert the wind speed data from mph to m/s. The end result of this step is a dataset with same covariates as the dataset used for model fitting.

# **Chapter 5**

## **Classification of ignition causes**

### **5.1 Model description**

We use a random forest model to classify historical bushfire ignitions. A random forest (Breiman, 2001) is an ensemble learning method for building tree-based prediction models. It is perhaps one of the most regularly used machine learning models in various fields (Boulesteix et al., 2012; Heung et al., 2016). It combines the predictions of many decision trees constructed using the bootstrapping the sample and randomly selecting variables/features at each tree node, takes the majority vote as the forest prediction. A random forest model guards against overfitting of the training set by only using out-of-bag predictions from each tree.

Other candidate models we examine in this research are the multinomial logit model (Berkson, 1944), the generalised additive model (GAM) multinomial logistic regression (Yee and Wild, 1996) and the XGBoost (Chen and Guestrin, 2016).

A multinomial logit model (Berkson, 1944) is a generalization of the logistic regression to multi-class problems, which is commonly used as the baseline model in predictive modelling.

A generalised additive model (Hastie and Tibshirani, 1990) is a generalised linear model with additive smooths terms in the link function and GAM multinomial logistic regression (Yee and Wild, 1996) is its extension to multi-class problems. Generalised additive model

is relatively popular in the field of bushfire ignition analysis. Some examples are Bates, McCaw, and Dowdy (2018), Read, Duff, and Taylor (2018) and Zhang, Lim, and Sharples (2017).

A XGBoost (Chen and Guestrin, 2016) is an open-source distributed gradient boosting library. It provides a parallel tree boosting to solve complex regression and classification problems efficiently. Gradient boosting (Breiman, 1998) is an important technique in machine learning, which belongs to the class of boosting algorithms. It is a method to build a strong learner, which often referred to as an ensemble model, by aggregating a set of weak learners iteratively. Numerous competitions, for example, the Higgs boson machine learning challenge (Adam-Boudarios et al., 2015) and the Global energy forecasting competition 2012 (Hong, Pinson, and Fan, 2014), have shown that XGBoost is one of the dominant methods in building prediction models on structured data.

In this research, the model building process includes feature selection, hyperparameter tuning and candidate model selection. These are described in details in the following sections.

The multinomial logit model, generalised additive model, random forest and XGboost are available in package `nnet` (Venables and Ripley, 2002), `mgcv` (Wood, 2011), `randomForest` (Liaw and Wiener, 2002) and `xgboost` (Chen et al., 2020) respectively. Besides, package `lime` (Pedersen and Benesty, 2019) is used to perform feature selection and package `caret` (Kuhn, 2020) is used to control the training, hyperparameter tuning and candidate model selection.

In terms of the train-test split, we randomly select 80% of total data as the training set, and the rest 20% data is test set. The total number of training samples is 7497 and the total number of test samples is 1872.

## 5.2 Feature selection

In feature selection, a reasonable principle is to select the most important features. Concerning the variable importance, Strobl et al. (2007) in their research has shown that the global variable importance, particularly random forest variable importance, can be bias

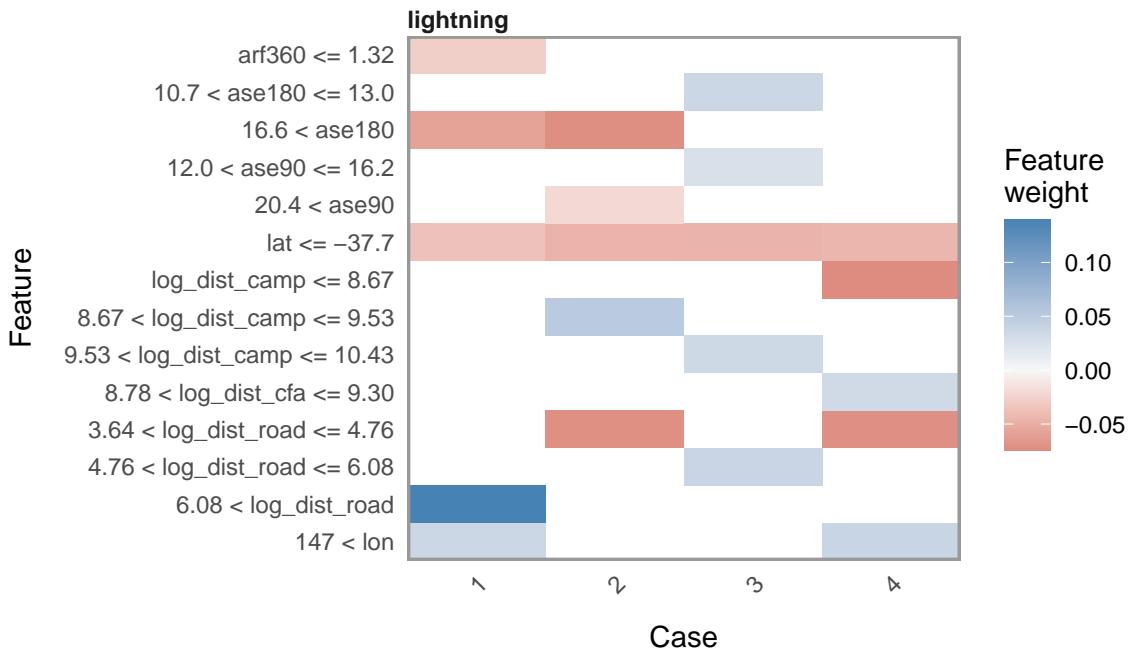
and misleading. Alternatively, unlike the global variable importance provided by many other packages, the package `lime` (Pedersen and Benesty, 2019) provides the local variable importance under the assumption that machine learning models are linear at the local scope.

Given an observation, `lime` randomly samples data points around the predictors, and obtains their predictions by passing them into the machine learning model. It then fits a simple model, typically a ridge regression (Hoerl and Kennard, 1970) or a lasso regression (Tibshirani, 1996), on these data points. Due to the characteristic of the lasso regression, it is possible to select the most important variables based on the regularization path. By repeating this process for large enough observations, variables being frequently selected are the most important variables. Figure 5.1 is an example of the result produced by `lime`.

The strategy used in this research to perform feature selection is first fitting and tuning a full model with all covariates using 3-fold cross-validation grid searching controlled by `caret` (Kuhn, 2020), then passing in 100 observations for `lime` to find the top 10 most important variables. The final set of variables selected for each candidate model is given in Table 5.2. Table 5.1 gives the detail information about each variable. From the result, we find the characteristics of different candidate models in ignition method classification. Simpler models, such as the multinomial logistic regression model and the GAM multinomial logistic regression model are preferred to use climate covariates, while the random forest model and the XGBoost model rely on ignition location and anthropogenic covariates.

**Table 5.1:** Detail information of variables involved in the feature selection process.

Covariate name	description	Units
lon	Longitude	degrees
lat	Latitude	degrees
FOR_TYPE	Forest type. Eg. Acacia, Callitris, Casuarina, etc.	
COVER	Forest crown cover	
HEIGHT	Forest height class	
arf360	Average rainfall in the past 360 days	mm
ase90	Average global solar exposure in the past 90 days	MJ/m <sup>2</sup>
ase180	Average global solar exposure in the past 180 days	MJ/m <sup>2</sup>
ase720	Average global solar exposure in the past 720 days	MJ/m <sup>2</sup>
amaxt90	Average maximum temperature in the past 90 days	Celsius degree
amaxt180	Average maximum temperature in the past 180 days	Celsius degree
amaxt720	Average maximum temperature in the past 720 days	Celsius degree
amint180	Average minimum temperature in the past 180 days	Celsius degree
ws	Average wind speed on that day	m/s
aws_m12	Average wind speed in last 12 months	m/s
aws_m24	Average wind speed in last 24 months	m/s
log_dist_cfa	Natural logarithm of the distance to the nearest CFA station	m
log_dist_camp	Natural logarithm of the distance to the nearest recreation site	m
log_dist_road	Natural logarithm of the distance to the nearest road	m

**Figure 5.1:** An example of selecting the most important variables for the random forest model with respect to the predicted probability of the bushfire ignited by lightning in 4 cases using lime. The feature weight is the weighted importance within a case. In this example, the common feature that influences all observations is latitude (lat).

**Table 5.2:** The top 10 most important variables for each candidate model ranked in descending order. Variables on the top are more important. The differences in choices of variables across candidate models can be observed. Random forest and XGBoost exploit the location variables and anthropogenic variables. Vegetation factors are most influential in multinomial logistic regression. Solar exposure and wind speed are most important in GAM multinomial logistic regression.

Multinomial logistic regression	GAM multinomial logistic regression	Random forest	XGBoost
FOR_TYPE	ase180	log_dist_camp	log_dist_road
HEIGHT	ase90	log_dist_road	log_dist_camp
ase90	aws_m12	log_dist_cfa	log_dist_cfa
amaxt180	log_dist_camp	lon	lat
amaxt90	amaxt180	ase180	lon
ase180	aws_m24	lat	ws
log_dist_cfa	log_dist_road	aws_m24	ase180
COVER	amaxt90	aws_m12	ase720
amint180	lon	arf360	amaxt720
amaxt720	amaxt720	ase90	ase90

### 5.3 Hyperparameter tuning and candidate model selection

The hyperparameter tuning for each candidate model is performed by using 3-fold cross-validation grid searching controlled by package `caret` (Kuhn, 2020). We set up a grid of potential hyperparameters and evaluate their performance cell by cell. The grid, the definition of every hyperparameter and the optimal hyperparameters is given in the Appendix.

The final step of the model building process is candidate model selection. Model performance is compared by using both overall prediction accuracy and multi-class AUC. Multi-class AUC is defined by Hand and Till (Hand and Till, 2001) and it is available in package `pROC` (Robin et al., 2011). This metric generalises the commonly used AUC into multiple class classification problems by averaging pairwise comparison of classes.

### 5.4 Results

After performing feature selection and parameter tuning, we find that random forest outperforms all other candidate models in both overall prediction accuracy and multi-class AUC. Thus, we choose the random forest model as our final model. Model performance

**Table 5.3:** Performance of the candidate models. Random forest model is the best in terms of accuracy and multi-class AUC.

Model	Accuracy	Multi-class AUC
Multinomial logistic regression	0.5272	0.7424
GAM multinomial logistic regression	0.6779	0.8233
Random forest	0.7495	0.8795
XGBoost	0.7388	0.8752

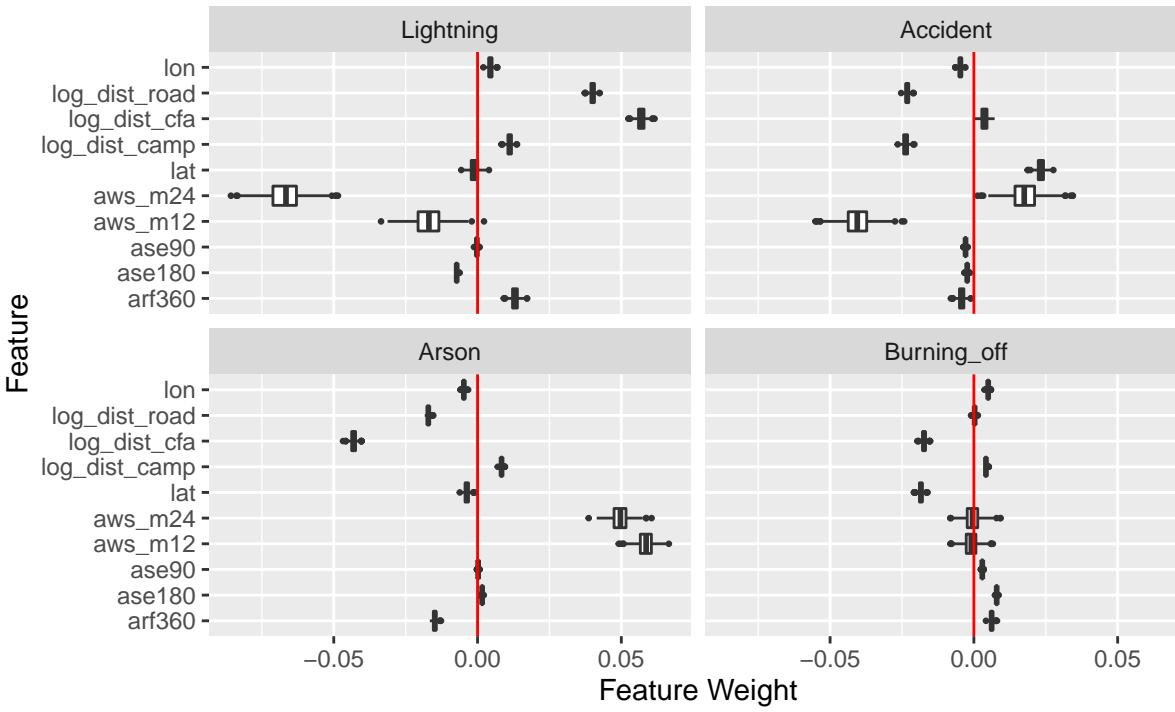
**Table 5.4:** Confusion matrix of random forest model. The overall accuracy is 0.7495.

	Lightning	Accident	Arson	Burning_off	Total
Prediction:lightning	703 (90.5%)	77 (12.1%)	50 (15.4%)	44 (32.4%)	874
Prediction:accident	51 (6.6%)	494 (77.9%)	89 (27.4%)	38 (27.9%)	672
Prediction:arson	18 (2.3%)	55 (8.7%)	175 (53.8%)	22 (16.2%)	270
Prediction:buring_off	5 (0.6%)	8 (1.3%)	11 (3.4%)	32 (23.5%)	56
Total	777	634	325	136	1872

is given in Table 5.3. More details about the model performance can be found in the Appendix.

The overall accuracy of our model is 74.95%. The confusion matrix is shown in Table 5.4. It suggests that lightning-caused and accident-caused ignitions can be easily classified from other causes. 77.9% of accident-caused and 90.5% of lightning-caused ignitions are correctly recognised by the model, which is a reliable result. Meanwhile, the model is not very confident with arson (53.8%) and burning off (23.5%). An error rate map for model diagnostic can be found in the Appendix.

The weighted variable contribution to the probability of different causes produced by `lime` is the scaled coefficients obtained from the lasso regression at the local scope. Figure 5.2 shows proximity to the nearest CFA station and proximity to the nearest road have a high positive impact on the probability of lightning-caused bushfire, while 2-year average wind speed has a high negative impact on the probability. Patterns in arson are almost the opposite of what has been shown in lightning. Latitude and 2-year average wind speed have some positive impact on the probability of accident-caused bushfire, while average wind speed in past 12 months, proximity to the nearest road and proximity to the nearest camping site have negative to the probability. Variable contribution to the probability of



**Figure 5.2:** Variable contribution to the probability of different causes. Variable has a positive weight means it has a positive impact on the probability. The same rule applies to negative weights. The magnitude shows the impact strength, which can be seen as the marginal effect at the local scope.

planned burn is relatively small, and the proximity to the nearest CFA station and latitude contribute negatively to the probability.

For future fire investigation, if a bushfire starts at a remote area in a windless year, it is very likely to be lightning-ignited bushfire. In contrast, if the bushfire is very close to the CFA station and starts in a windy year, it is possible to be arson. Moreover, the accident-caused bushfire usually starts near the recreation site and road in a windless year after a windy year.

## 5.5 Predicting ignition cause for 2019-2020 season

A fitted random forest model, along with covariate data in 2019-2020 bushfire season, can be used to produce the prediction of the cause of the bushfires in Victoria during the 2019-2020 season. Figure 5.3 shows the prediction produced by the final model. And Table 5.5 summarizes the prediction. According to the prediction, most majority of the bushfires in 2019-2020 season are caused by lightning. However, there are 138 bushfires caused by

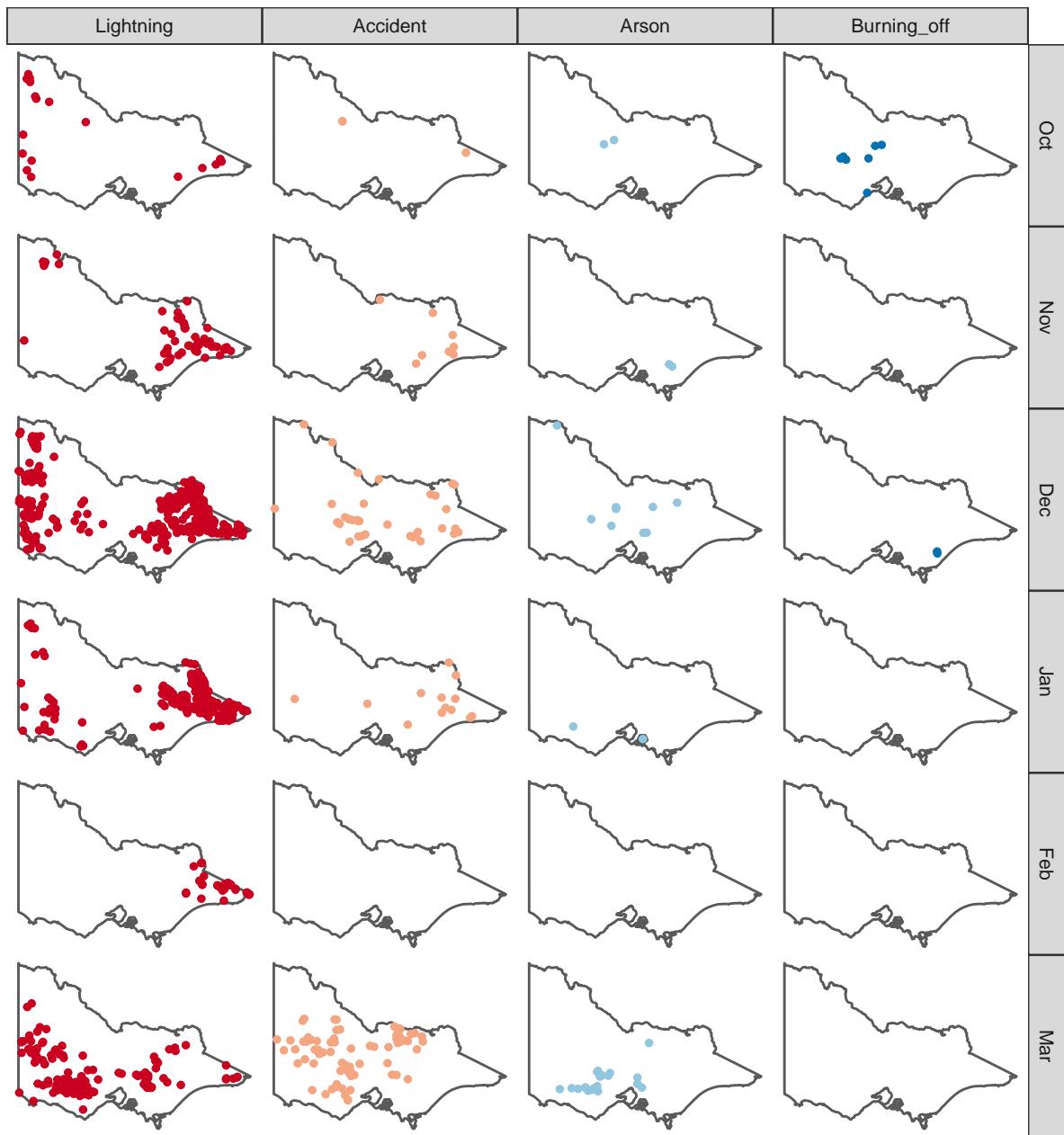
**Table 5.5:** A summary of the predicted causes of 2019-2020 Australia bushfires. Our model predicts 82% of the bushfire were lightning, 14% were accident, and only 4% and 1% were arson and planned burns respectively.

Cause	Oct	Nov	Dec	Jan	Feb	Mar	Total
Lightning	19	57	315	266	32	149	838 (0.82%)
Accident	3	8	34	13	0	80	138 (0.14%)
Arson	2	2	10	2	0	21	37 (0.04%)
Burning_off	7	0	2	0	0	0	9 (0.01%)

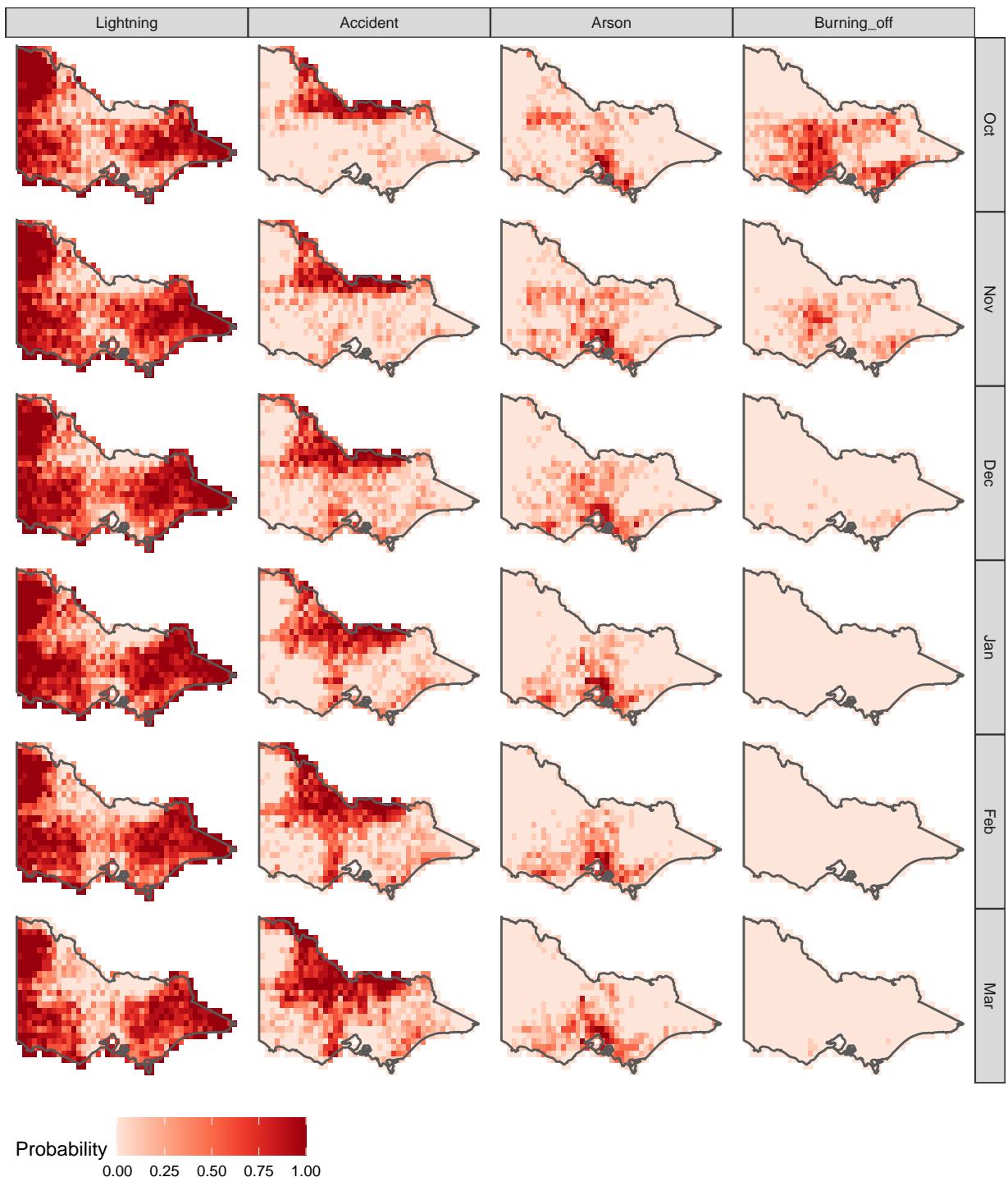
accidents which take up 14% of the total fires. Majority of the accident-caused bushfires are ignited in March. Besides, 37 bushfires are caused by arsonists. There was a noticeable bushfire at French island in January, which caused serious damage to the koala habitat. Our model predicts its cause is arson. Very few planned burns are predicted after October 2019 which suggests the correctness of our model. It is because fire restrictions normally start in October.

Furthermore, we provide a map with 0.2 degree spatial resolution for quick decision making, which is given in Figure 5.4. If the investigator observes a new ignition between December and March and the weather condition is similar to the 2019-2020 bushfire season, prediction of the cause can be made immediately by checking Figure 5.4.

Other than that, Figure 5.4 also reveals the temporal and spatial characteristics of different causes. Probability of lightning-caused bushfire is almost time and spatial invariant from October to March. Bushfires start in the east and the west of Victoria are likely to be predicted by our model as lightning-ignited bushfires. The probability of accident-caused bushfire evolves by time. In October, only the north of Victoria has a high probability of accident-caused bushfire, but the high probability region spreads to the south over time and reaches its peak in March. Probability of arson-caused bushfire excluding the Melbourne region decreases as time goes from October to March. Planned burns only occur in October and November.



**Figure 5.3:** Prediction of the cause of bushfire ignitions in Victoria during the 2019-2020 season produced by the final model. Majority of the bushfires are caused by lightning. Most of the Planned burns occurred before November. Bushfires caused by arson mainly occur in December 2019 and March 2020. A large proportion of bushfires in March 2020 are predicted to be accident-caused.



**Figure 5.4:** A map with 0.2 degree spatial resolution for quick decision making of the cause of the bushfire during a bushfire season. This map is based on the assumption that the long-run weather condition of the new ignition is similar to 2019-2020 Australia bushfire season. Users need to match the location of the observed ignition and the observed date with the map to obtain the prediction for each cause. The darker the region the higher the probability.

## 5.6 Fire risk map

Our predictive model can produce the probability of the bushfire ignited by a certain type of sources  $P(M|S, \mathcal{F})$ , but it does not provide us with the risk of the bushfire ignited by that source  $P(M, S|\mathcal{F})$ , where  $M$  is the method of the bushfire ignition,  $M \in \{\text{lightning, arson, accident, burning off}\}$ ,  $S$  is the bushfire is ignited and  $\mathcal{F}$  is the supplementary information of the bushfire ignition, such as location, time, weather conditions, distribution of vegetation.

Knowing  $P(M|S, \mathcal{F})$  is useful for bushfire investigation but is not particularly helpful for bushfire management. In general, decision makers want to know the overall risk  $P(S|\mathcal{F})$  and the risk of the bushfire ignited by a certain method  $P(M, S|\mathcal{F})$ .

A possible method to obtain  $P(M, S|\mathcal{F})$  is by using the decomposition of the conditional probability:

$$P(M, S|\mathcal{F}) = P(M|S, \mathcal{F})P(S|\mathcal{F})$$

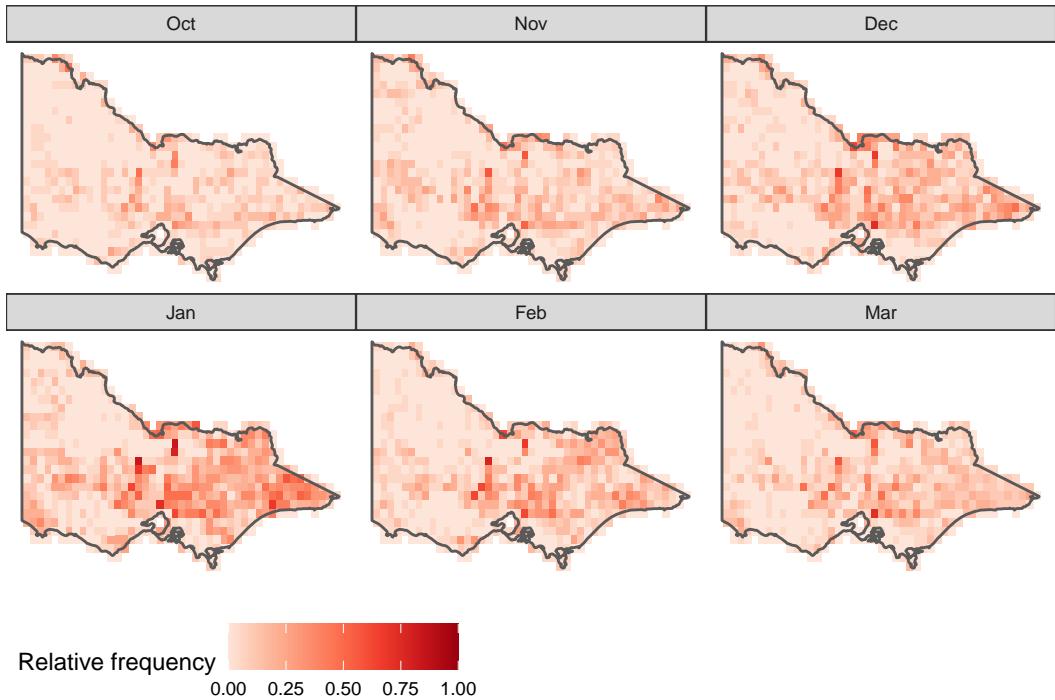
We will demonstrate how this works by using a simple method to estimate  $P(S|\mathcal{F})$  and eventually yield  $P(M, S|\mathcal{F})$  in the rest of the section.

If we only consider month  $MON$  and location  $L$  as  $\mathcal{F}$ , a reasonable estimator of the probability of at least one bushfire ignited in a region  $l$  in a certain month  $mon$  is the relative frequency calculated from historical records:

$$P(S|MON = mon, L = l) = \frac{\sum_{i=1}^N I(\text{bushfire ignition occurred in month } mon, \text{year } y_i)}{N}$$

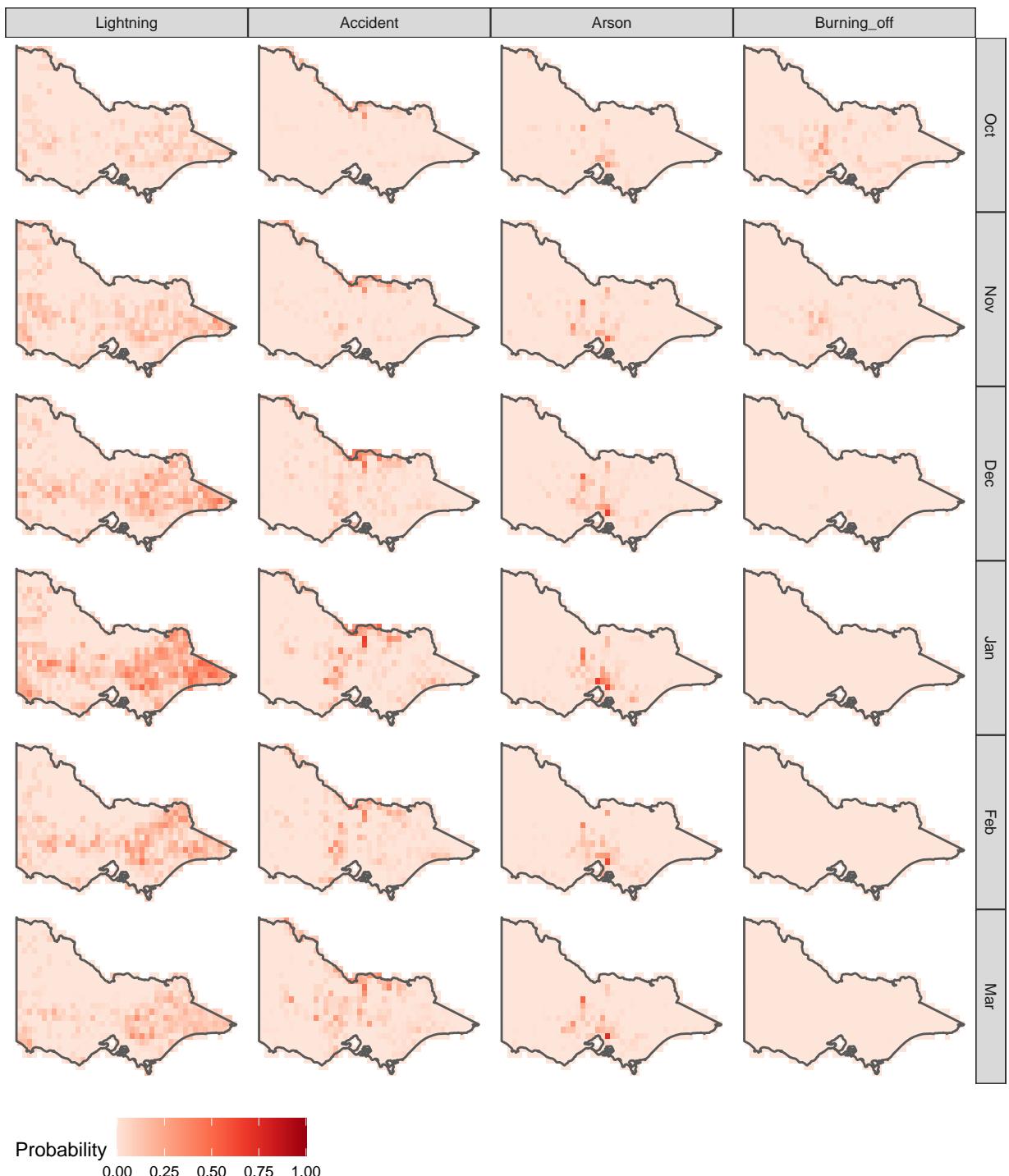
, where  $N$  is the number of years in the historical records and  $y_i$  is the  $i$  th year of the historical records.

Figure 5.5 shows the estimate of  $P(S|MON = mon, L = l)$  using 20 years historical records. If we assume the weather conditions are similar to 2019-2020 bushfire season,  $P(M|S, MON = mon, L = l)$  can be produced by our random forest model and the joint



**Figure 5.5:** A map of the relative frequency of the bushfire ignition occurred in a region in the past 20 years. It is an estimate of the probability of at least one bushfire ignition occur in a given region and a given month.

probability  $P(L, S | \mathcal{F})$  can then be calculated. Figure 5.6 shows the final outcome. From the result, we notice that January is the most dangerous month in terms of lightning-caused bushfire and the east of Victoria is the high risk area. The risk of arson-caused bushfire is concentrated in a relatively small region near Melbourne.



**Figure 5.6:** The joint probability of bushfire ignited by different methods in different months. The east of Victoria has the highest risk of lightning-caused bushfire in January. Arson risk is concentrated around Melbourne.



# **Chapter 6**

## **Discussion and conclusion**

In conclusion, we have developed a spatio-temporal clustering algorithm for detecting bushfire ignitions from satellite hotspot data, which provides a possible way to automate or at least supplement bushfire ignition data beyond field observation. We fit a random forest model for classifying the cause of bushfire ignitions in Victoria, which demonstrates good predictive power (79% overall accuracy and 0.8795 mult-class AUC). Wind speed, latitude, proximity to the nearest fire station, proximity to the nearest recreation site and proximity to the nearest road have most influence to the predicted probability. The prediction of the cause of bushfire ignitions in Victoria during the 2019-2020 season produced by the predictive model suggests the main cause is lightning (82%). Availability of satellite hotspot data and supplementary open data make it possible to deploy the clustering algorithm and the predictive model used in this research, which could benefit bushfire investigation of past and future events and real-time bushfire monitoring.

### **6.1 Policy Implications**

Our prediction suggests that the main cause of the 2019-2020 Australia bushfires is lightning. According to the clustering results, these bushfires are mostly start in remote areas, like East Gippsland. Our fire risk map gives similar result that indicates east of Victoria in January is at high risk of lightning-caused bushfire. Strategies may need to be taken by

Forest Fire Management Victoria and the Country Fire Authority to minimise lightning-ignited bushfire risk in these areas. Meanwhile, modern technology for monitoring and fighting bushfires like satellite sensors, community sensors and drones could be deployed in these areas to overcome the difficulties of reaching the ignition point.

Besides, according to our prediction, the number of accident-caused bushfires rises significantly in March 2020. We suspect they are correlated to certain types of human activities or weather conditions. Bushfire investigators in Country Fire Authority may need to reveal the motivations behind this phenomenon to reduce this controllable impact on bushfire ignition.

Finally, future bushfire investigation could be performed more concisely by utilising satellite hotspot data and modern machine learning framework. We have shown this is achievable and we suggest the government could spend some resources on developing a set of reliable and accurate models to predict the cause of the bushfire ignition remotely by fusing the available data, the knowledge of climate scientists and the experience of investigators. This could overcome the workforce shortage and reduce the countless time spent on the investigation.

## 6.2 Limitation and possible extensions

### 6.2.1 Data

In this research, vegetation dynamics has not been taken into account in modelling. Instead, a static map of vegetation information is used. This could lead to ineffective use of vegetation factors. A possible future direction will be incorporating vegetation data collected in multiple years.

In the prediction of 2019-2020 season, **ASOS** wind speed data is used as covariate. **ASOS** network only has a limited number of stations in Victoria, so the weather conditions integrated to estimated bushfire ignitions could be inaccurate. To improve the accuracy of the prediction, **BOM** wind speed data can be used in the future, but they are not publicly available.

### 6.2.2 Clustering algorithm

During the reconstruction of the bushfire dynamics, only satellite hotspot data is used in our clustering algorithm. Other influential factors such as distribution of vegetation, slope and weather conditions have not been taken into account. These factors will affect the speed of the bushfire and the direction of the spread. An extension that we plan to pursue is a more comprehensive and complicated algorithm that allows variable speed of bushfire and variable time tolerance during the clustering.

### 6.2.3 Modelling

Only 4 statistical models have been tested in this research. Other popular machine learning models such as artificial neural network and support vector machine may have a better performance. However, these modelling frameworks typically require significant amount of time in hyperparameter tuning. Besides, advanced modelling techniques like model fusion and dimensionality reduction can be applied to this problem in future work.

### 6.2.4 Fire risk map

We show the use of our predictive model in producing the fire risk map. However, a simple method is used to estimate the probability of at least one bushfire ignited in a given region and in a given month. None of the changes of the distribution of vegetation and differences in the weather conditions have been considered. In the future work, we will develop another model to provide accurate prediction of the risk of the bushfire ignition  $P(S|\mathcal{F})$ . Combining the outcomes of two models, we will eventually be able to produce the precise risk of bushfire ignited by different methods  $P(M, S|\mathcal{F})$ .

### 6.2.5 Shiny app

General public and officers from fire departments may not have time and capability in running the code of our predictive model. An interactive web-based shiny application could be a more accessible way for people to benefit from our research. It could provide elegant interactive plots for exploratory data analysis and real-time prediction for decision

making in bushfire management. A shiny app is under development by Liu (2020). The product will be published on the server of the Department of Econometrics and Business Statistics, Monash University recently. An early stage release can be found on <https://ebsmonash.shinyapps.io/VICfire/>.

## Appendix A

# Covariate information

**Table A.1:** Detail information of covariates used in modelling.

Covariate name	description	Units
month	Month	
day	Day	
dow	Day of the week. 1-7	
lon	Longitude	degrees
lat	Latitude	degrees
FOR_TYPE	Forest type. Eg. Acacia, Callitris, Casuarina, etc.	
COVER	Forest crown cover. 1-6. Check COVER look up table	
HEIGHT	Forest height class. 1-6. Check HEIGHT look up table	
rf	Rainfall on that day	mm
rf7	Average rainfall in the past 7 days	mm
arf14	Average rainfall in the past 14 days	mm
arf28	Average rainfall in the past 28 days	mm
arf60	Average rainfall in the past 60 days	mm
arf90	Average rainfall in the past 90 days	mm
arf180	Average rainfall in the past 180 days	mm
arf360	Average rainfall in the past 360 days	mm
arf720	Average rainfall in the past 720 days	mm
se	Global solar exposure on that day	MJ/m <sup>2</sup>
ase7	Average global solar exposure in past 7 days	MJ/m <sup>2</sup>
ase14	Average global solar exposure in past 14 days	MJ/m <sup>2</sup>

**Table A.1:** Detail information of covariates used in modelling. (continued)

Covariate name	description	Units
ase28	Average global solar exposure in past 28 days	MJ/m <sup>2</sup>
ase60	Average global solar exposure in past 60 days	MJ/m <sup>2</sup>
ase90	Average global solar exposure in past 90 days	MJ/m <sup>2</sup>
ase180	Average global solar exposure in past 180 days	MJ/m <sup>2</sup>
ase360	Average global solar exposure in past 360 days	MJ/m <sup>2</sup>
ase720	Average global solar exposure in past 720 days	MJ/m <sup>2</sup>
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
amaxt28	Average maximum temperature in the past 28 days	Celsius degree
amaxt60	Average maximum temperature in the past 60 days	Celsius degree
amaxt90	Average maximum temperature in the past 90 days	Celsius degree
amaxt180	Average maximum temperature in the past 180 days	Celsius degree
amaxt360	Average maximum temperature in the past 360 days	Celsius degree
amaxt720	Average maximum temperature in the past 720 days	Celsius degree
mint	Minimum temperature on that day	Celsius degree
amint7	Average minimum temperature in the past 7 days	Celsius degree
amint14	Average minimum temperature in the past 14 days	Celsius degree
amint28	Average minimum temperature in the past 28 days	Celsius degree
amint60	Average minimum temperature in the past 60 days	Celsius degree
amint90	Average minimum temperature in the past 90 days	Celsius degree
amint180	Average minimum temperature in the past 180 days	Celsius degree
amint360	Average minimum temperature in the past 360 days	Celsius degree
amint720	Average minimum temperature in the past 720 days	Celsius degree
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
aws_m3	Average wind speed in last 3 months	m/s
aws_m6	Average wind speed in last 6 months	m/s
aws_m12	Average wind speed in last 12 months	m/s
aws_m24	Average wind speed in last 24 months	m/s
log_dist_cfa	Natural logarithm of the distance to the nearest CFA station	m
log_dist_camp	Natural logarithm of the distance to the nearest recreation site	m

**Table A.1:** *Detail information of covariates used in modelling. (continued)*

Covariate name	description	Units
log_dist_road	Natural logarithm of the distance to the nearest road	m

**Table A.2:** *COVER look up table (Australian Bureau of Agricultural and Resource Economics and Sciences, 2018)*

COVER code	Forest crown cover	Description
1	20-50%	Woodland
2	50-80%	Open
3	>80%	Closed
4	n/a	Plantation
5	>=20%	Unknown
6	<20%	Non forest

**Table A.3:** *HEIGHT look up table (Australian Bureau of Agricultural and Resource Economics and Sciences, 2018)*

HEIGHT code	Forest height class	Description
1	2m-10m	Low
2	10m-30m	Medium
3	>30m	Tall
4	n/a	Plantation
5	>=2m	Unknown
6		Non forest



## **Appendix B**

# **Modified breadth-first search algorithm used in step 2 of the clustering algorithm**

This modified breadth-first search algorithm includes 6 substeps:

- (a) Append a randomly selected hotspot  $h_i$  to a empty list  $L$ , where  $h_i$  was the  $i$ th hotspot in the interval  $S_t$ .
- (b) Let pointer  $P$  point to the first element of the list  $L$ .
- (c) Visit every  $h_i \in S_t$  where  $h_i \notin L$ . Meanwhile, if  $geodesic(h_i, P) \leq AdjDist$ , append  $h_i$  to list  $L$ .
- (d) Move pointer  $P$  to the next element of the list  $L$ .
- (e) Repeat (a) to (d) till the pointer  $P$  reach to the end of the list  $L$ .
- (f) Assign a new membership to all hotspots  $h_i \in L$ . Repeat (a) to (f) for unassigned hotspots in interval  $S_t$ .

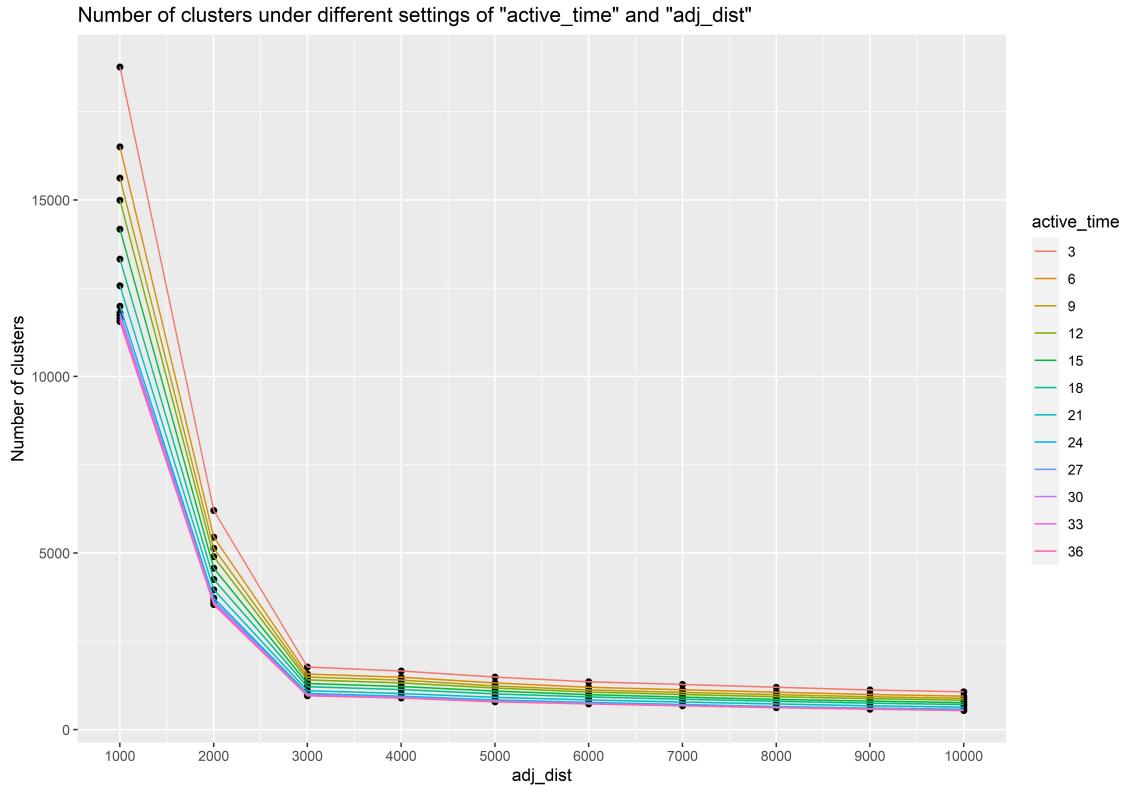


## Appendix C

# Effects of parameter choices in the clustering algorithm

The popularity of **DBSCAN** (Ester et al., 1996) is partially due to the parameter tuning tools it provides. In this section, we will introduce the parameter tuning tool for our algorithm.

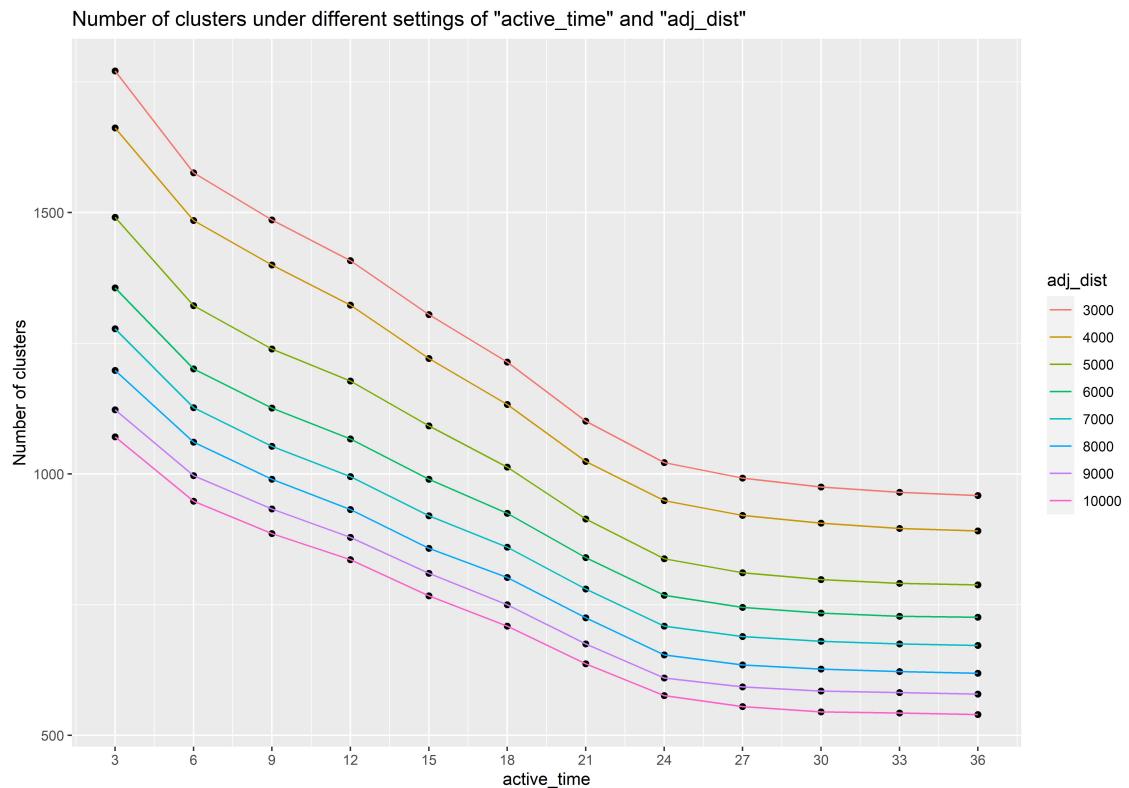
There are only two parameters used in the algorithm, which are *ActiveTime* and *AdjDist*. Increase the tolerance of undetectable time or the potential fire speed will usually reduce the total number of clusters. However, if there are large gaps between clusters spatially, increase the parameter *AdjDist* will not significantly reduce the number of clusters. Similarly, if there are large gaps between clusters temporally, the increase of *ActiveTime* will have limited impact on the number of clusters. In clustering algorithms, one of the metrics to measure the quality of the clustering results is the gap between clusters. Therefore, if we can find a point where the marginal effect of *ActiveTime* and *AdjDist* on the number of clusters is small, we may potentially obtain a reasonable choice of the parameters. Meanwhile, if the gaps are large enough, we may observe small marginal effect when *ActiveTime* and *AdjDist* are over certain values. Thus, it is insufficient to pick the optimal value just by asking for a small marginal effect. Instead, we are going to check the first-order derivative of the marginal effect. The motivation is if the first-order derivative of



**Figure C.1:** A visualization tool for parameter tuning in our algorithm. It works like a scree plot. We need to choose a point with a large second-order derivative. The reasonable choice of the parameter *AdjDist* is 3000km.

the marginal effect is large, it means we are crossing a line where a great proportion of noisy hotspots are not seen as individual clusters anymore.

Figure C.1 and Figure C.2 shows the effects of parameter choices on the number of clusters. It works like the scree plot (Cattell, 1966). The scree plot is originally used for finding statistically significant components in principal component analysis (**PCA**). In our application, we need to find the “elbow” of the graph to choose the value for our parameters. The “elbow” is an indication of the first-order derivative of the marginal effect is large. In Figure C.1, it’s very clear we need to choose 3000km for *AdjDist*. And in Figure C.2, we choose 24 hours for *ActiveTime*.



**Figure C.2:** A visualization tool for parameter tuning in our algorithm. It works like a scree plot. We need to choose a point with a large second-order derivative. The reasonable choice of the parameter ActiveTime is 24 hours.



## Appendix D

# Hyperparameter tuning for ignition classifiers

**Table D.1:** A grid of values tested in hyperparameter tuning for each candidate model.

Hyperparameter	Range
Multinomial logistic regression decay	$\{0.2h \mid h = 1, 2, \dots, 10\}$
Random forest mtry	$\{h \mid h=1, 2, \dots, 10\}$
XGBoost	
max_depth	$\{3, 5, 7, 9\}$
nrounds	$\{50h \mid h = 1, 2, \dots, 200\}$
eta	$\{0.3, 0.2, 0.1, 0.05, 0.025, 0.0125, 0.00625\}$
subsample	$\{0.05h \mid h = 10, 11, \dots, 18\}$
gamma	$\{0.2h \mid h = 0, 1, \dots, 5\}$
colsample_bytree	$\{0.05h \mid h = 10, 11, \dots, 18\}$
min_child_weight	$\{1, 3, 5, 7\}$

**Table D.2:** A description of each hyperparameter. The definition of the hyperparameters is referenced from the documentations of the package *nnet* (Venables and Ripley, 2002), package *randomForest* (Liaw and Wiener, 2002) and package *xgboost* (Chen et al., 2020).

Hyperparameter	Description
Multinomial logistic regression decay	Parameter for weight decay
Random forest mtry	Number of variables randomly sampled as candidates at each split
XGBoost	
max_depth	Maximum depth of a tree
nrounds	The number of rounds for boosting
eta	Step size shrinkage used in update to prevents overfitting
subsample	Subsample ratio of the training instances
gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree
colsample_bytree	subsample ratio of columns when constructing each tree
min_child_weight	Minimum sum of instance weight (hessian) needed in a child

**Table D.3:** The final result of the hyperparameter tuning.

Hyperparameter	Value
Multinomial logistic regression decay	0.2
Random forest mtry	1
XGBoost	
max_depth	5
nrounds	4800
eta	0.025
subsample	0.85
gamma	0.8
colsample_bytree	0.55
min_child_weight	1

## Appendix E

### Model performance

**Table E.1:** Confusion matrix of multinomial logit model. The overall accuracy was 0.5272.

	Lightning	Accident	Arson	Burning_off	Total
Prediction:Lightning	568 (73.1%)	259 (40.9%)	74 (22.8%)	33 (24.3%)	934
Prediction:Accident	182 (23.4%)	277 (43.7%)	120 (36.9%)	51 (37.5%)	630
Prediction:Arson	22 (2.8%)	77 (12.1%)	120 (36.9%)	30 (22.1%)	249
Prediction:Burning_off	5 (0.6%)	21 (3.3%)	11 (3.4%)	22 (16.2%)	59
Total	777	634	325	136	1872

**Table E.2:** Confusion matrix of GAM model. The overall accuracy was 0.6779.

	Lightning	Accident	Arson	Burning_off	Total
Prediction:Lightning	663 (85.3%)	114 (18%)	64 (19.7%)	45 (33.1%)	934
Prediction:Accident	74 (9.5%)	434 (68.5%)	106 (32.6%)	33 (24.3%)	630
Prediction:Arson	31 (4%)	72 (11.4%)	144 (44.3%)	30 (22.1%)	249
Prediction:Burning_off	9 (1.2%)	14 (2.2%)	11 (3.4%)	28 (20.6%)	59
Total	777	634	325	136	1872

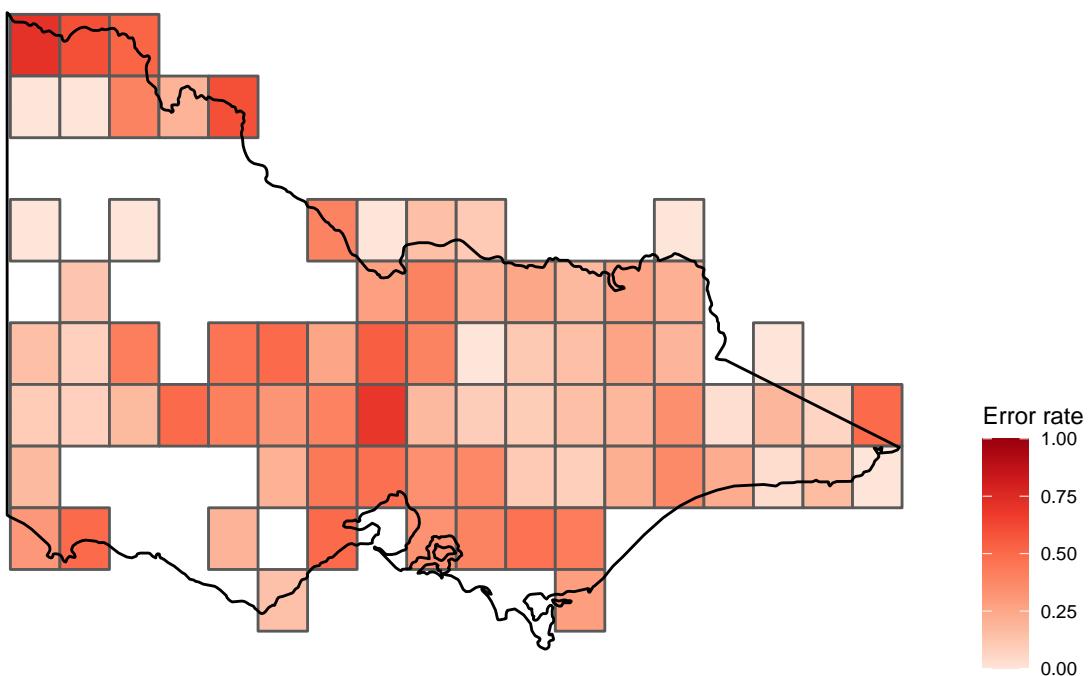
**Table E.3:** Confusion matrix of XGBoost model. The overall accuracy was 0.7388.

	Lightning	Accident	Arson	Burning_off	Total
Prediction:Lightning	695 (89.4%)	87 (13.7%)	42 (12.9%)	36 (26.5%)	934
Prediction:Accident	53 (6.8%)	465 (73.3%)	85 (26.2%)	38 (27.9%)	630
Prediction:Arson	22 (2.8%)	72 (11.4%)	183 (56.3%)	22 (16.2%)	249
Prediction:Burning_off	7 (0.9%)	10 (1.6%)	15 (4.6%)	40 (29.4%)	59
Total	777	634	325	136	1872

## **Appendix F**

### **Error rate map**

In order to explore the prediction performance of the final model on the test set, we plot a error rate map to reveal the spatial patterns. The plot is given in Figure F.1. From the error rate map, we can observe that our model correctly predicts most of the cases in the east of Victoria, which is the mountain area. However, it performs worse near the Melbourne region. Besides, our model does not fit well on the boundary of the north-west of Victoria.



**Figure F.1:** The spatial patterns of the error rate of the final model. We omit regions with less than 5 bushfires occurred. Our model makes very few mistakes in the east of Victoria but has a higher error rate near Melbourne.

## **Appendix G**

# **Supplementary material**

Supplementary materials include figures, codes and documentations can be found in the Github repository of this project <https://github.com/TengMCing/bushfire-paper>



# Bibliography

- Adam-Bourdarios, C, G Cowan, C Germain, I Guyon, B Kégl, and D Rousseau (2015). The Higgs boson machine learning challenge. In: *NIPS 2014 Workshop on High-energy Physics and Machine Learning*, pp.19–55.
- Arnold, JB (2019). *ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'*. R package version 4.2.0. <https://CRAN.R-project.org/package=ggthemes>.
- Australian Bureau of Agricultural and Resource Economics and Sciences (2018). *Forests of Australia*. <https://www.agriculture.gov.au/abares/forestsaustralia/forest-data-maps-and-tools/spatial-data/forest-cover> (visited on 05/21/2020).
- Bates, BC, L McCaw, and AJ Dowdy (2018). Exploratory analysis of lightning-ignited wildfires in the Warren Region, Western Australia. *Journal of environmental management* 225, 336–345.
- Beale, J and W Jones (2011). Preventing and reducing bushfire arson in Australia: A review of what is known. *Fire technology* 47(2), 507–518.
- Berkson, J (1944). Application of the logistic function to bio-assay. *Journal of the American statistical association* 39(227), 357–365.
- Boulesteix, AL, S Janitza, J Kruppa, and IR König (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2(6), 493–507.
- Bradstock, RA, GJ Cary, I Davies, DB Lindenmayer, OF Price, and RJ Williams (2012). Wildfires, fuel treatment and risk mitigation in Australian eucalypt forests: insights from landscape-scale simulation. *Journal of Environmental Management* 105, 66–75.

- Breiman, L (1998). Arcing Classifiers. *The Annals of Statistics* **26**(3), 801–824.
- Breiman, L (2001). Random forests. *Machine learning* **45**(1), 5–32.
- Cattell, RB (1966). The scree test for the number of factors. *Multivariate behavioral research* **1**(2), 245–276.
- Chen, T and C Guestrin (2016). Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp.785–794.
- Chen, T, T He, M Benesty, V Khotilovich, Y Tang, H Cho, K Chen, R Mitchell, I Cano, T Zhou, M Li, J Xie, M Lin, Y Geng, and Y Li (2020). *xgboost: Extreme Gradient Boosting*. R package version 1.1.1.1. <https://CRAN.R-project.org/package=xgboost>.
- Cheney, NP, JS Gould, WL McCaw, and WR Anderson (2012). Predicting fire behaviour in dry eucalypt forest in southern Australia. *Forest Ecology and Management* **280**, 120–131.
- Clarke, H, R Gibson, B Cirulis, RA Bradstock, and TD Penman (2019). Developing and testing models of the drivers of anthropogenic and lightning-caused wildfire ignitions in south-eastern Australia. *Journal of environmental management* **235**, 34–41.
- Collins, KM, OF Price, and TD Penman (2015). Spatial patterns of wildfire ignitions in south-eastern Australia. *International Journal of Wildland Fire* **24**(8), 1098–1108.
- Costa-Luis, CO da (2019). ‘tqdm’: A Fast, Extensible Progress Meter for Python and CLI. *Journal of Open Source Software* **4**(37), 1277.
- Council, Climate (2019). This is not normal: climate change and escalating bushfire risk. *Climate Council Briefing Paper* **12**.
- Csárdi, G and R FitzJohn (2019). *progress: Terminal Progress Bars*. R package version 1.2.2. <https://CRAN.R-project.org/package=progress>.
- Department of Environment, Land, Water & Planning (2019). *Fire Origins - Current and Historical*. <https://discover.data.vic.gov.au/dataset/fire-origins-current-and-historical> (visited on 05/21/2020).
- Department of Environment, Land, Water & Planning (2020[a]). *CFA - Fire Station*. <https://discover.data.vic.gov.au/dataset/cfa-fire-station-vmfeat-geomark-point> (visited on 05/21/2020).
- Department of Environment, Land, Water & Planning (2020[b]). *Recreation Sites*. <https://discover.data.vic.gov.au/dataset/recreation-sites> (visited on 05/21/2020).

- Dowdy, AJ, MD Fromm, and N McCarthy (2017). Pyrocumulonimbus lightning and fire ignition on Black Saturday in southeast Australia. *Journal of Geophysical Research: Atmospheres* **122**(14), 7342–7354.
- Duff, TJ, JG Cawson, and S Harris (2018). Dryness thresholds for fire occurrence vary by forest type along an aridity gradient: evidence from Southern Australia. *Landscape Ecology* **33**(8), 1369–1383.
- Ester, M, HP Kriegel, J Sander, X Xu, et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Kdd*. Vol. 96. 34, pp.226–231.
- Graham, T and T Keller (2020). Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight. *The Conversation* **10**.
- Grolemund, G and H Wickham (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software* **40**(3), 1–25.
- Hand, DJ and RJ Till (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning* **45**(2), 171–186.
- Harris, CR, KJ Millman, SJ van der Walt, R Gommers, P Virtanen, D Cournapeau, E Wieser, J Taylor, S Berg, NJ Smith, R Kern, M Picus, S Hoyer, MH van Kerkwijk, M Brett, A Haldane, JF del Río, M Wiebe, P Peterson, P Gérard-Marchant, K Sheppard, T Reddy, W Weckesser, H Abbasi, C Gohlke, and TE Oliphant (2020). Array programming with NumPy. *Nature* **585**(7825), 357–362.
- Hastie, TJ and RJ Tibshirani (1990). *Generalized additive models*. Vol. 43. CRC press.
- Heung, B, HC Ho, J Zhang, A Knudby, CE Bulmer, and MG Schmidt (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma* **265**, 62–77.
- Hijmans, RJ (2020). *raster: Geographic Data Analysis and Modeling*. R package version 3.3-13. <https://CRAN.R-project.org/package=raster>.
- Hoerl, AE and RW Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67.
- Hong, T, P Pinson, and S Fan (2014). *Global energy forecasting competition 2012*.
- Hyndman, RJ (2020). *Monash Honours Thesis Rmarkdown Template*. <https://github.com/robjhyndman/MonashHonoursThesis>.

## BIBLIOGRAPHY

---

- Iowa State University (2020). *ASOS-AWOS-METAR Data Download*. [https://mesonet.agron.iastate.edu/request/download.phtml?network=AU\\_ASOS](https://mesonet.agron.iastate.edu/request/download.phtml?network=AU_ASOS).
- Kaplan, J (2020). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables*. R package version 1.6.2. <https://CRAN.R-project.org/package=fastDummies>.
- Keane, RE, GJ Cary, ID Davies, MD Flannigan, RH Gardner, S Lavorel, JM Lenihan, C Li, and TS Rupp (2004). A classification of landscape fire succession models: spatial simulations of fire and vegetation dynamics. *Ecological modelling* **179**(1), 3–27.
- Kisilevich, S, F Mansmann, M Nanni, and S Rinzivillo (2009). “Spatio-temporal clustering”. In: *Data mining and knowledge discovery handbook*. Springer, pp.855–874.
- Knaus, C (8, 2020). Police contradict claims spread online exaggerating arson’s role in Australian bushfires. *The Guardian*. (Visited on 10/06/2020).
- Kuhn, M (2020). *caret: Classification and Regression Training*. R package version 6.0-86. <https://CRAN.R-project.org/package=caret>.
- Liaw, A and M Wiener (2002). Classification and Regression by randomForest. *R News* **2**(3), 18–22.
- Lisa Richards, Nigel Brew and L Smith (2020). *2019-20 Australian bushfires-frequently asked questions: a quick guide*. Parliament of Australia. [https://www.aph.gov.au/About\\_Parliament/Parliamentary\\_Departments/Parliamentary\\_Library/pubs/rp/rp1920/Quick\\_Guides/AustralianBushfires](https://www.aph.gov.au/About_Parliament/Parliamentary_Departments/Parliamentary_Library/pubs/rp/rp1920/Quick_Guides/AustralianBushfires).
- Liu, C (2020). *VicFire shiny app*. <https://github.com/timtam3/Bushfire/tree/master/VICfire>.
- Loboda, T and I Csiszar (2007). Reconstruction of fire spread within wildland fire events in Northern Eurasia from the MODIS active fire product. *Global and Planetary Change* **56**(3-4), 258–273.
- McAneney, J, K Chen, and A Pitman (2009). 100-years of Australian bushfire property losses: is the risk significant and is it increasing? *Journal of environmental management* **90**(8), 2819–2822.
- McVicar, T (2011). *Near-Surface Wind Speed. v10. CSIRO. Data Collection*. <https://doi.org/10.25919/5c5106acbc02>.

## BIBLIOGRAPHY

---

- Miller, C, M Plucinski, A Sullivan, A Stephenson, C Huston, K Charman, M Prakash, and S Dunstall (2017). Electrically caused wildfires in Victoria, Australia are over-represented when fire danger is elevated. *Landscape and Urban Planning* **167**, 267–274.
- Müller, K (2017). *here: A Simpler Way to Find Your Files*. R package version 0.1. <https://CRAN.R-project.org/package=here>.
- OpenStreetMap contributors (2020). *Planet dump retrieved from https://planet.osm.org*. <https://www.openstreetmap.org>.
- P-Tree System (2020). *JAXA Himawari Monitor - User's Guide*. <https://www.eorc.jaxa.jp/ptree/userguide.html> (visited on 05/21/2020).
- Padgham, M and MD Sumner (2020). *geodist: Fast, Dependency-Free Geodesic Distance Calculations*. R package version 0.0.4. <https://CRAN.R-project.org/package=geodist>.
- Pebesma, E (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal* **10**(1), 439–446.
- Pedersen, TL and M Benesty (2019). *lime: Local Interpretable Model-Agnostic Explanations*. R package version 0.5.1. <https://CRAN.R-project.org/package=lime>.
- Plucinski, M, W McCaw, J Gould, and B Wotton (2014). Predicting the number of daily human-caused bushfires to assist suppression planning in south-west Western Australia. *International Journal of Wildland Fire* **23**(4), 520–531.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. <https://www.R-project.org/>.
- Read, N, TJ Duff, and PG Taylor (2018). A lightning-caused wildfire ignition forecasting model for operational use. *Agricultural and Forest Meteorology* **253**, 233–246.
- Robin, X, N Turck, A Hainard, N Tiberti, F Lisacek, JC Sanchez, and M Müller (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77.
- RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA. <http://www.rstudio.com/>.
- Schloerke, B, D Cook, J Larmarange, F Briatte, M Marbach, E Thoen, A Elberg, and J Crowley (2020). *GGally: Extension to 'ggplot2'*. R package version 2.0.0. <https://CRAN.R-project.org/package=GGally>.

- Schowengerdt, RA (2006). *Remote sensing: models and methods for image processing*. Elsevier.
- Chap. 1, pp. 1–44.
- South, A (2017). *rnaturrearth: World Map Data from Natural Earth*. R package version 0.1.0.  
<https://CRAN.R-project.org/package=rnaturrearth>.
- Sparks, AH, J Carroll, J Goldie, D Marchiori, P Melloy, M Padgham, H Parsonage, and K Pembleton (2020). *bomrang: Australian Government Bureau of Meteorology (BOM) Data Client*. R package version 0.7.0. <https://CRAN.R-project.org/package=bomrang>.
- Strobl, C, AL Boulesteix, A Zeileis, and T Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics* **8**(1), 25.
- Tibshirani, R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288.
- Tierney, N, D Cook, M McBain, and C Fay (2020). *naniar: Data Structures, Summaries, and Visualisations for Missing Data*. R package version 0.5.2. <https://CRAN.R-project.org/package=naniar>.
- Van Rossum, G and FL Drake (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
- Venables, WN and BD Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. <http://www.stats.ox.ac.uk/pub/MASS4>.
- Wickham, H (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software* **40**(1), 1–29.
- Wickham, H, M Averick, J Bryan, W Chang, LD McGowan, R François, G Grolemund, A Hayes, L Henry, J Hester, M Kuhn, TL Pedersen, E Miller, SM Bache, K Müller, J Ooms, D Robinson, DP Seidel, V Spinu, K Takahashi, D Vaughan, C Wilke, K Woo, and H Yutani (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**(43), 1686.
- Wilke, CO (2020). *ggridges: Ridgeline Plots in 'ggplot2'*. R package version 0.5.2. <https://CRAN.R-project.org/package=ggridges>.
- Williamson, G (2020). Example code to generate animation frames of Himawari-8 hotspots.  
<https://gist.github.com/ozjimbob/80254988922140fec4c06e3a43d069a6> (visited on 05/21/2020).

## BIBLIOGRAPHY

---

- Wood, SN (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* **73**(1), 3–36.
- Xie, Y (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.20. <https://github.com/rstudio/bookdown>.
- Yee, TW and C Wild (1996). Vector generalized additive models. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(3), 481–493.
- Zhang, Y, S Lim, and JJ Sharples (2017). Wildfire occurrence patterns in ecoregions of New South Wales and Australian Capital Territory, Australia. *Natural Hazards* **87**(1), 415–435.
- Zumbrunnen, T, P Menéndez, H Bugmann, M Conedera, U Gimmi, and M Bürgi (2012). Human impacts on fire occurrence: a case study of hundred years of forest fires in a dry alpine valley in Switzerland. *Regional Environmental Change* **12**(4), 935–949.
- Zumbrunnen, T, GB Pezzatti, P Menéndez, H Bugmann, M Bürgi, and M Conedera (2011). Weather and human impacts on forest fires: 100 years of fire history in two climatic regions of Switzerland. *Forest Ecology and Management* **261**(12), 2188–2199.