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

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` (Grolemund and Wickham, 2011), `rnaturalearth 0.1.0` (South, 2017), `here 0.1` (Müller, 2017), `raster 3.3.13` (Hijmans, 2020), `bomrang 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^{th} November 2020

Abstract

The 2019-2020 Australian bushfire season was a devastating crisis. Debate about whether lightning or arson was the cause of the bushfires was fierce on social media. We developed a spatio-temporal clustering algorithm to detect the time and locations of bushfire ignitions in Victoria during the 2019-2020 season from the new available Himawari-8 hotspot data. The historical bushfire origins data over 2000-2018 was used to train a random forest model to classify the cause. The final model showed good predictive ability with 74.95% overall accuracy, and 90.5% accuracy in lightning-caused bushfires. The model predicted the major cause of estimated 2019-2020 bushfire ignitions in Victoria was lightning (82%), with arson responsible for just 3.62% of the total fires. The clustering algorithm and the predictive model developed in this research could help fire authorities to more readily monitor and investigate bushfire risk and causes using satellite hotpot data.

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, the insurance claims for building losses have exceeded 10 million Australia dollars (McAneney, Chen, and Pitman, 2009). The 2019-2020 Australia bushfires, compared with other major bushfires in history, had a more devastating impact on the environment and properties. 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 are the highest in history. Fortunately, fewer lives were lost: 33 (including firefighters) compared to 173 in the Black Saturday fires and 47 in Ash Wednesday fires (Lisa Richards, Nigel Brew and Smith, 2020).

Discussion about the cause of this crisis was a focal point on social media at the very beginning of the bushfire season. A group of people 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, which was one of the biggest climate organizations in Australia and was made up of climate scientists and experts, claimed that climate change not only worsened the bushfire season but also increased our cost of fighting fires (Council, Climate, 2019). Meanwhile, in the first few weeks of 2020, another group of

people brought a completely different argument onto the table. They claimed the cause of this bushfire season was arson instead of climate change. Besides, they attempted to replace the *#ClimateEmergency* hashtag with the *#ArsonEmergency* hashtag (Graham and Keller, 2020). Although police (Knaus, 2020) contradicted this controversial claim immediately, the spread didn't slow down at all. A research did 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 the accounts were suspicious with automated and inauthentic behaviour. They believed that the accounts were very likely ran by disinformation campaigns. However, there was no direct evidence to support their beliefs, particularly without knowledge of the owners of the accounts.

By far, we have 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 only known for 58.9% of fires in history. Among the known cases, about 13% are due to deliberate ignitions, 35% by accidents and 6% by natural, such as lightning. There are considerable variations 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 most majority of bushfires are ignited by lightning, fuel management such as planned burns may need to be delivered more often on a larger scale. In contrast, if arson plays an important role in the ignitions, improvements in regulations, enforcement, and research on environmental criminology may be in demand.

1.2 Remote sensing data

Remote sensing data are collected by remote sensors carried by a satellite or an aircraft. These sensors can detect the reflected energy from the earth and record it as signals (Schowengerdt, 2006). The signal will then be sent back to the Mission Control Center

on the ground to further process to high-quality products such as high-resolution colour image and sea surface temperature map.

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 provide a possible solution to this, particularly satellite hotspot data, which may be useful in detecting ignitions and movements of bushfires. New availability of hotspot data taken from Himawari-8 satellite (P-Tree System, 2020) has 10-minute time resolution that could be used to detect fires almost in real-time.

Knowing the precise ignition location and ignition time will allow us to reconstruct detailed information at the time, such as weather condition and the 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 locations of ignitions for 2019-2020 Australia bushfires in Victoria from satellite hotspot data, and (ii) build a predictive model to classify the potential methods of the bushfire ignition.

Clustering algorithm is an unsupervised learning algorithm to cluster 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 is to determine whether a hotspot is a new ignition point or a branch of an existing bushfire. This requires the algorithm runs recursively in a temporal manner with referencing 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 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 of defining the data space in traditional spatio-temporal clustering algorithms, that the scaling of the temporal dimension can be highly influential to the clustering result (Kisilevich et al., 2009). Besides, our algorithm reconstructs clusters in parallel that 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 allow us 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 of detecting new bushfire ignition from the satellite hotspot data, which may be beneficial for urgent fire resources planning and deployment in remote areas in the future. Importantly, by applying this algorithm on satellite hotspot data from the past bushfire seasons, we will be 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 explored as the classifier to predict the cause of the bushfire ignition, including **GAM**, and 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 the prediction 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.

In overall, this research offers a complete workflow for bushfire analysis and monitoring, starts from detecting bushfire ignitions from satellite hotspot data, and ends with producing prediction of the cause of the ignitions. We believe this workflow can be adopted into the future bushfire investigation, particularly for establishing the cause of bushfires.

1.4 Scope of the thesis

This thesis will focus on analysing 2019-2020 bushfires in Victoria with open-source remote sensing data. The remainder of the thesis will be structured as follows. In chapter 2, we will review the existing literature in the related field. In chapter 3, we will outline the data source and data integration, and provide a brief data exploratory analysis of the compiled datasets. In chapter 4, we will describe the clustering algorithm and present the clustering results. In chapter 5, we will outline the strategy of model building and discuss the results we found. In chapter 6, we will discuss the limitation of our work and possible future extension. Finally, in chapter 7, we will give a brief conclusion of the research.

Chapter 2

Review of literature

2.1 Spatio-temporal clustering

To detect time and locations 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 3 dimensions which are spatial location and time. Kisilevich et al. (2009) in their survey of spatio-temporal clustering algorithms summarised that there were 5 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 was based on the temporal dimension and the spatial dimension.

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 locatoin	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). It is because the end result of the clustering algorithm can be highly sensitive to the temporal resolution. Hence, scaling of the temporal dimension is vary 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 3 difficulties in clustering algorithm which motivated their work: (1) Requirements of domain knowledge to determine the hyperparameters, (2) Arbitrary shape of clusters and (3) Computation efficiency.

In the algorithm, they defined a maximum radius ϵ to construct densities around each point, which were spheres in the case of 3-dimensional space. Besides, to overcome the impact of noise, at least minimum number of points (`MinPts`) would be contained by the sphere. Finally, intersected spheres would be considered belonging to the same cluster.

The time complexity of **DBSCAN** is $O(Nlog(N))$, and the memory complexity is only $O(N)$. They have also shown that via visualizing the density of distance to the k-th nearest neighbour from each point, optimal ϵ and `MinPts` could be chosen.

Although **DBSCAN** is popular among many fields, it is not suitable for clustering hotspot data. First, it doesn't address the issue of defining the scaling of the temporal dimension. Second, the behaviour of bushfire has not been taken into account. The **DBSCAN** algorithm assumes the clustering rules work for both directions of the timeline, omitting the fact that bushfire evolves by time.

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 the 2.5km radius from its parent and (3) any node is observed no later than 4 days from its parent. When the tree is closed and there are still unassigned hotspots, the algorithm will continue at the earliest unassigned hotspot to construct a new

tree. Finally, each tree is a cluster, and the earliest hotspot will be defined as the ignition point.

FSR works well in reconstructing bushfire spread, but it constructs clusters sequentially. If two fires starts at a different location but had overlapping coverage, **FSR** will consider they are a single fire when their fronts meet with each other. In this situation, one of the fires will be smaller in scale and shorter in time, which can not correctly reflect the real state of the bushfire. Besides, Loboda and Csiszar (2007) didn't provide the detail of parameter tuning in their research. Instead, they briefly delivered the reason for choosing 4 days as the time threshold. It was because they wanted to tolerant missed observation 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 of detail consideration. Inspired by these two algorithms, our work is to design a spatio-temporal algorithm that can efficiently and robustly cluster hotspots with respect to the temporal behaviour of bushfire. Meanwhile, we need to 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, and accounting 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. The general framework for analysing bushfires ignition is a generalised additive model (**GAM**). Bates,

McCaw, and Dowdy (2018) used it with a Gaussian distribution for predicting the number of lightning ignitions in Warren Region, Western Australia. Some studies implemented **GAM** with a binomial distribution to predict ignition probability in New South Wales, Australian Capital Territory and Victoria (Read, Duff, and Taylor, 2018; Zhang, Lim, and Sharples, 2017). Mixed-effect model is an alternative to incorporate spatial dependence and weather factors. Duff, Cawson, and Harris (2018) have shown that by treating slope of drought factor and intercept vary by forest type, mixed-effect model with only 2 variables could achieve 70% mean accuracy for predicting fire occurrence in Southern Australia. Besides, 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 bushfire, some research conducted statistical tests and exploratory methods to assess hypotheses about bushfire. (Miller et al., 2017; Dowdy, Fromm, and McCarthy, 2017).

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

Throughout the review of the existing literature, we didn't find any research attempted to model the cause of the bushfires in Australia, and neither of them chose to apply hotspot data. Therefore, we intend to build a model by incorporating part of the modelling framework and covariate choices mentioned above, and use hotspot data to produce the prediction of the cause of 2019-2020 Australian bushfire season.

Although numerous studies for ignitions 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 less than 3 states in Australia. Model is seldomly applied across Australia. This is mainly due to the lack of coherence in bushfire datasets provided by different states in Australia. For the same reason, we will also only focus on bushfires in Victoria in this research.

Chapter 3

Data

3.1 Data Sources

A focus of this work is to utilise open-source data, and collates these data sets 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 indicate 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 used hotspot data taken from the Himawari-8 satellite (P-Tree System, 2020). The hotspot data was available on the Japan Aerospace Exploration Agency FTP site. Besides, only the data from October 2019 to March 2020 was downloaded. Details on how to download this data were provided by Williamson (2020). It contained records of 1989572 hotspots for 6 months in the full disk of 140 °east longitude.

3.1.2 Climate data

To better understand bushfires, we collected 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 were retrieved via an open-source R package `bomrang` (Sparks et al., 2020), which was a data client of **BOM**. Historical weather records of 885 weather stations across Australia from January 1863 to April 2020 were downloaded. However, the number of publicly available climate attributes on **BOM** were limited. To maintain the coherent reproducible workflows, we decided to download near-surface wind speed grids across Australia from January 1975 to December 2018 from **CSIRO** (McVicar, 2011) and the weather 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 were only a limited number of **ASOS** stations in Victoria and they mostly based around Melbourne. Even though **ASOS** data was also available from 2000 to 2018, **CSIRO** data was used because it had a higher spatial resolution.

3.1.3 Road map

Since bushfires could be impacted by anthropogenic factors, we used road map from a comprehensive open-source map **Openstreetmap** (OpenStreetMap contributors, 2020) to represent the reachability of ignition locations. The road map was one of the layers of the full archive, which consisted of 1797217 roads belonging to 27 different road classes in Australia.

3.1.4 Fuel layer

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

3.1.5 Fire stations

Data of Country Fire Authority (**CFA**) fire stations were retrieved from Department of Environment, Land, Water & Planning (2020[a]). It contained 52716 Victorian topographic features including rivers, water bodies, transport, facilities and fire stations.

3.1.6 Recreation sites

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

3.1.7 Fire origins

We used the Victorian Department of Environment, Land, Water and Planning (**DELWP**) Fire Origins dataset to obtain historical bushfire locations and ignition causes (2019). This dataset provided first reported location of fires recorded by crews rather than origins of a fire, which could 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 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 bushfires ignition classifier. `2` and `3` combined with the clustering results are used as predictors to predict the ignition causes of 2019-2020 bushfires.

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

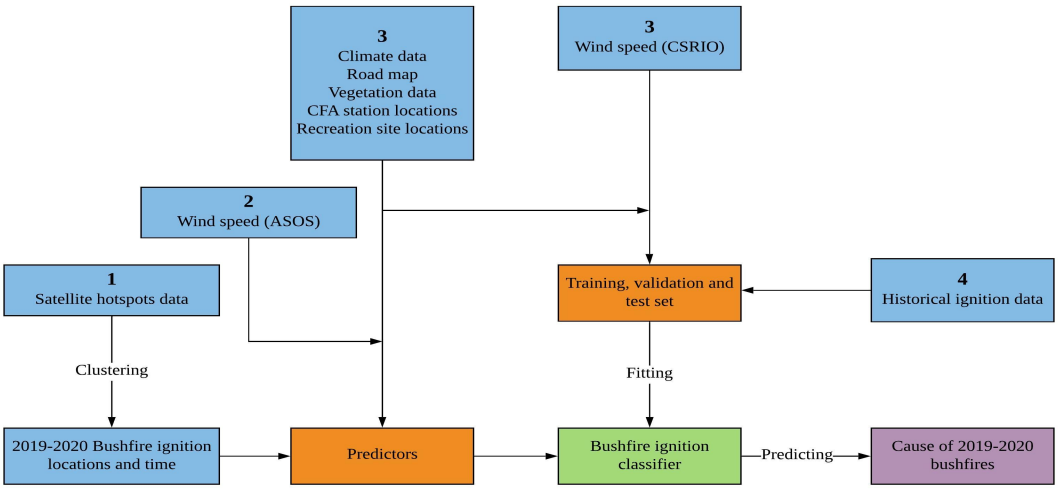

Figure 3.1: *Summary of dataset usage and research workflow*

Table 3.1: *Raw data information*

Index	Data set name	Spatial Resolution	Temporal resolution	Time
1	Himawari-8 satellite hotspots data	$0.02° \approx 2km$	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	$2° \approx 200km$	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 used to perform data manipulation in this research was `R` (R Core Team, 2020). Victorian historical fire origins, recreation sites, **CFA** fire stations and road map were geospatial data in the shapefile format, which could be processed using the tools in `sf` (Pebesma, 2018). Fuel layer and near-surface wind speed were also geospatial data but stored in the grid format. We used tools in `raster` (Hijmans, 2020) to manipulate the data. Other datasets including satellite hotspots data and **BOM** climate data were stored in csv format which could be handled using the package `tidyverse` (Wickham et al., 2019).

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

BOM weather records including maximum temperature, minimum temperature, solar exposure and rainfall were station-based data. Weather stations distributed unevenly across Victoria and were often under maintenance. Therefore, we matched each historical fire origin to its 10 nearest weather stations, and then extracted the nearest non-missing daily weather metrics. Differently, **CSIRO** wind speed data provided us with a full grid of values in Victoria. Therefore, we projected each historical fire origin onto its corresponding gird entry to extract the daily average wind speed. Finally, weather metrics in the past 720 days for every historical fire origin were summarised into several numeric variables which are provided in Table A.1.

Besides, **CFA** stations locations were 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 were computed for every historical fire origin.

Vegetation information included vegetation type, height and cover are stored in the grid format which was extracted in the same way as the **CSIRO** wind speed data.

We also truncated hotspot data by using a recommended threshold of fire power (Williamson, 2020), which was over 100 (irradiance over 100 watts per square metre), to reduce noise from the background, and then selectd the hotspots only inside 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 results of the data cleaning were two different datasets. One was the compiled dataset used for model fitting which contained 9369 observations and 55 fields, including a field represented the cause of the ignition. The other one was the satellite hotspot data that were ready for clustering, which contained 75936 observations and 4 fields. These 4 fields were unique identifier, longitude, latitude and time.

Covariates for fitting the ignition classifier included month, day and day of the week; longitude and latitude; forest type, forest crown cover and forest height; rainfall, wind speed, temperature and solar exposure in the past 2 years; Natural logarithm of proximity 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 explored the relationship between covariates and causes of ignition.

From January 2000 to December 2018, Victoria recorded 3886, 3173, 1627 and 683 bushfires ignitions caused by lightning, accident, arson and burning off respectively. The histogram of ignition cause is given in the upper left of Figure 3.2. The proportion of each cause was against the review of bushfires proposed by Beale and Jones (2011). Ignitions caused by

arson were significantly less than the data provided in their review, while lightning-caused ignitions taking a greater percentage.

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

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 suggested that spatial pattern might be useful in ignition classification.

According to the density plot of 1-year average wind speed shown in the bottom 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 left of the same figure, we found 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 upper right of the Figure 3.4. Moreover, in the bottom right of the figure, accident-caused ignitions were significantly close to camping sites, which suggests that camp fire was a source of bushfire ignitions.

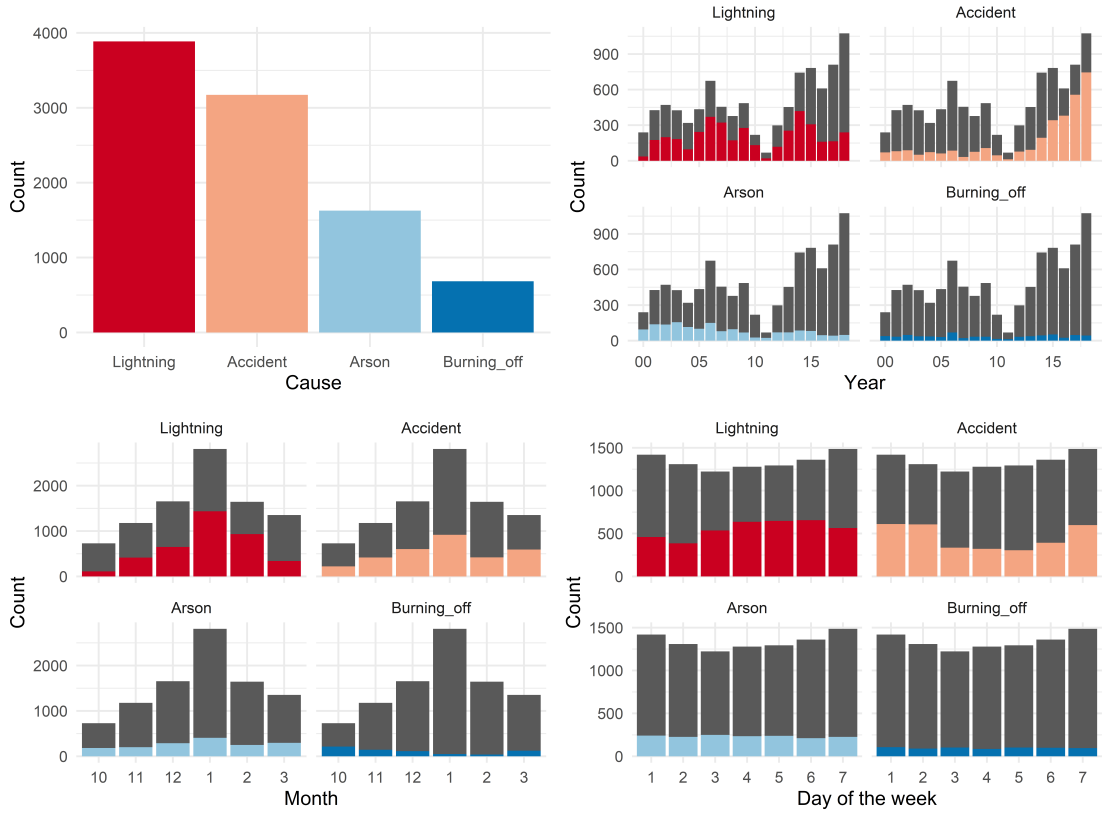

Figure 3.2: *Histogram (Upper left) and bar plots of four different casues of historical fire ignitions in Victoria by year (Upper right), month (Bottom left) and day of the week (Bottom right). (Upper left) Lightning and accident were the two main sources of bushfire ignitions, which took up 41% and 34% respectively. Proportion of ignitions caused by arson was around 17%. 7% of ignitions were reported as planned burns. (Upper right) 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. (Bottom left) January was the most serious month in the bushfire season throughout these years. Besides, lightning-caused ignitions usually occurred in the hottest months. (bottom right) 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.*

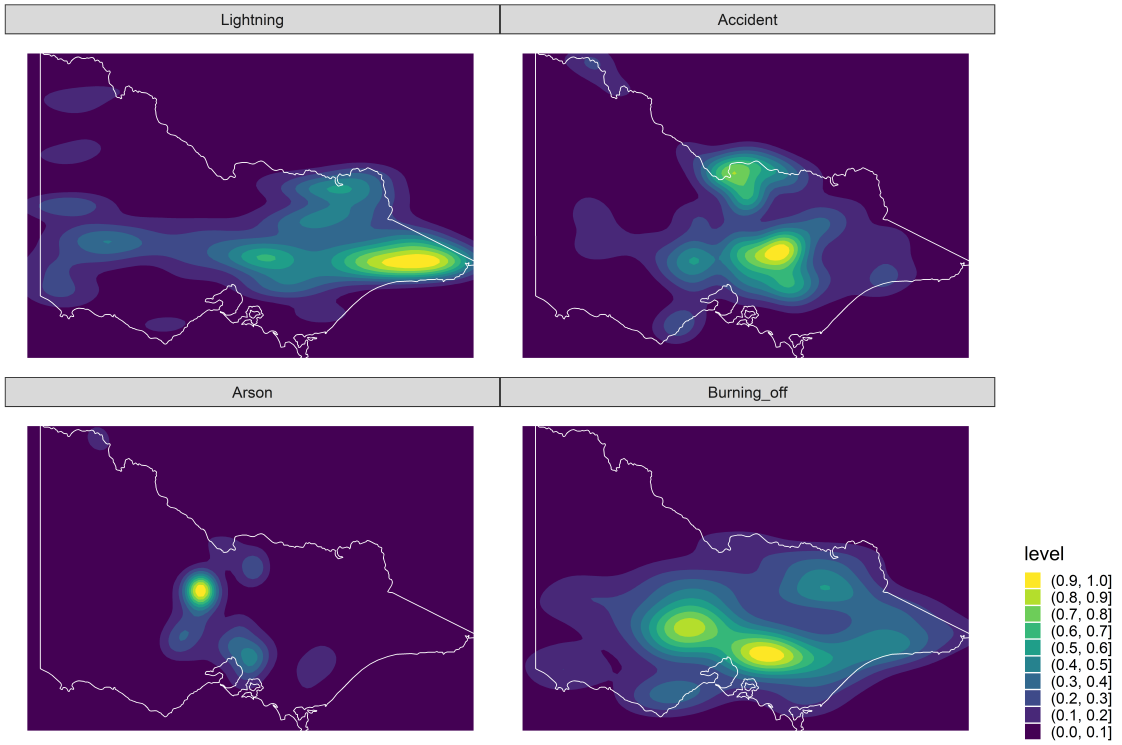

Figure 3.3: *2D density plot of historical bushfire ignitions. The density plot shows arson and accident-caused ignitions were different from each other. Accident-caused ignitions were denser. In addition, lightning-caused ignitions were less likely to occur in Melbourne metropolitan area.*

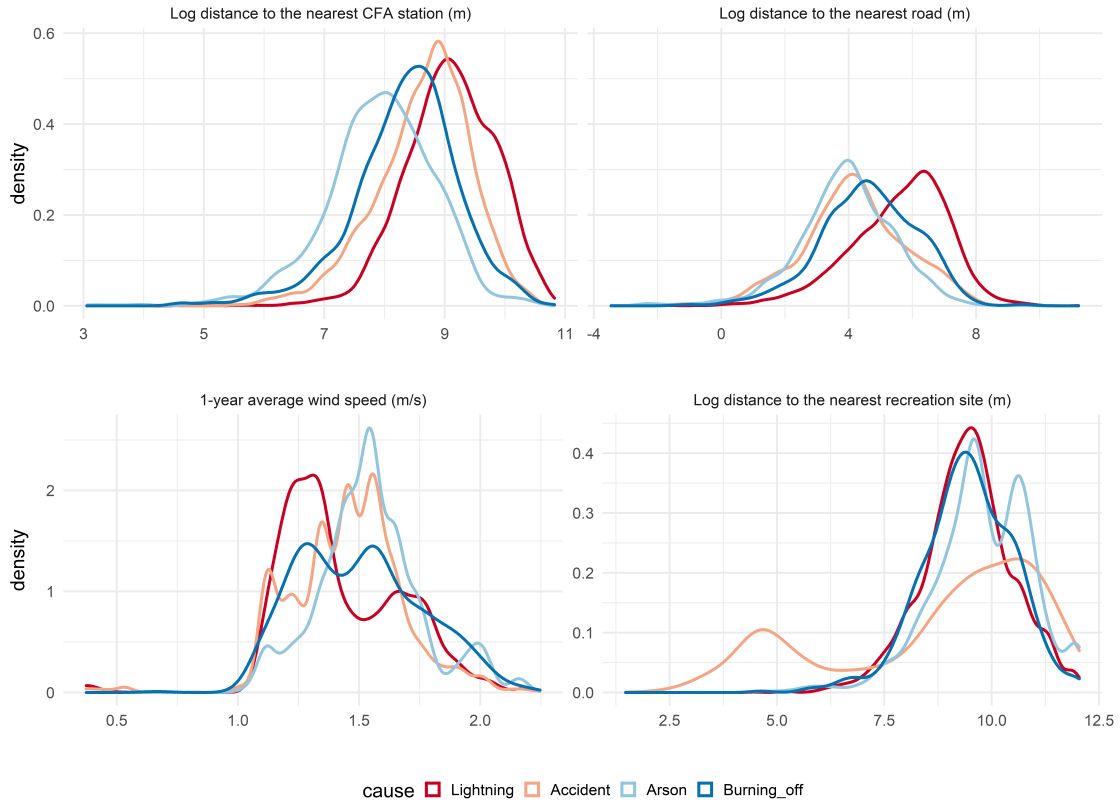

Figure 3.4: *Density plot of log distance to the nearest CFA station (Upper left), log distance to the nearest road (Upper right), log distance to the nearest recreation site (Bottom right) and 1-year average wind speed (Bottom left). (Upper left) Arson-caused ignitions were closer to fire stations.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

Chapter 4

Detecting bushfire ignitions from hotspot data

4.1 Overview of the hotspot data

The hotspot data contained 75936 observations which are shown in Figure 4.1. Generally, a group of similar hotspots in a spatio-temporal data space can be recognized as a series of bushfire dynamics. One of the characteristics of this data was bushfires in small scale were easier to be distinguished from the others because there are clear gaps, but some clusters in the east of the Victoria were close to each other spatially and temporally, which introduced extra difficulties of clustering. In essence, hotspots need to be clustered into different groups to identify the ignition locations.

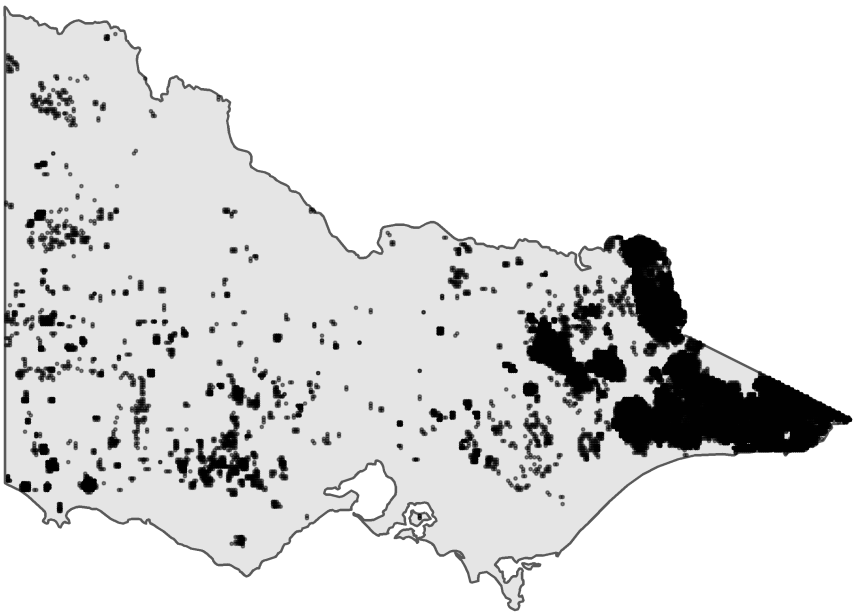

Figure 4.1: *Distribution of 75936 satellite hotspots data in Victoria from October 2019 to March 2020. Most of the hotspots are observed in the east of Victoria. Hotspots data are difficult to use in ignition analysis.*

4.2 Outline of the algorithm

We have discussed the existing clustering algorithm for spatio-temporal data in the literature review. Meanwhile, we point out the modification we need to consider in the development of our algorithm. To robustly and efficiently cluster satellite hotspots data, we summarised 6 conditions that the clustering algorithm needed to satisfy, (1) it does not require any scaling of the temporal dimension, (2) it iteratively runs at 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 keep 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.

With this in mind, we developed our own algorithm for hotspots clustering while meeting these conditions. The input of the algorithm is the hotspot data with 4 attributes, which

are the unique identifier, longitude, latitude, and time. The output are a list of labels describing the corresponding clusters of hotspots.

This algorithm can be divided into four steps: (1) Slice the temporal dimension, (2) Cluster hotspots locally, (3) Broadcast and update the membership label and (4) Compute ignition locations.

We will discuss these four steps sequentially in the following sections.

4.2.1 Slice the temporal dimension

To illustrate this step, we need to introduce the first parameter used in our algorithm, *ActiveTime*. *ActiveTime* is a parameter which controls the size of the interval. A nature interpretation of this parameter is the time a fire can stay smouldering but undetectable by satellite before flaring up again. In Loboda and Csiszar (2007)'s work, a similar parameter is used as the tolerance time for the missed observations due to cloud cover.

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

$$S_t = [max(1, t - ActiveTime), t], t = 1, 2, ..., T$$

, 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, $S_1 = [1, 1], S_2 = [1, 2], ..., S_{25} = [1, 25], S_{26} = [2, 26], ..., S_{47} = [23, 47], S_{48} = [24, 48]$.

4.2.2 Cluster hotspots locally

In the previous step, the algorithm breaks the temporal dimension. Thus, the local clustering for each interval only needs to address the hotspots spatially. This step involves the second parameter used in this algorithm, *AdjDist*. It 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 $3km/10min = 18km/h$. Similarly, in Ester et al. (1996)'s work, a radius-based parameter ϵ is defined to connect neighbouring points.

Given a certain value of $AdjDist$ and the interval $\boldsymbol{S}_t$, 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 ith hotspot in the interval $\boldsymbol{S}_t$ and h_j is the jth hotspot in the interval $\boldsymbol{S}_t$.

(b) A connected component is called $\boldsymbol{M}_{t,n}$, where n is less than the total number of components in the interval $\boldsymbol{S}_t$. Hotspots in the same component $\boldsymbol{M}_{t,p}$ will be assigned with the same unique membership label, where $\boldsymbol{M}_{t,p}$ is the pth component in $\boldsymbol{S}_t$.

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

4.2.3 Broadcast and update the membership label

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

This step started from the second interval $\boldsymbol{S}_2$ till the interval $\boldsymbol{S}_T$. Given the interval $\boldsymbol{S}_t$, the algorithm will:

(a) If a hotspot h belongs to both $\boldsymbol{S}_{t-1}$ and $\boldsymbol{S}_t$, it will succeed its membership label from $\boldsymbol{S}_{t-1}$. We call these hotspots $H_s = \{h_{s,1}, h_{s,2}, ...\}$.

(b) If a hotspot h belongs to $\boldsymbol{S}_t$, but does not belong to $\boldsymbol{S}_{t-1}$, we will call these hotspots $H_c = \{h_{c,1}, h_{c,2}, ...\}$. If a hotspot in H_c shares the same component $\boldsymbol{M}_{t,p}$ with any hotspot in H_s, it will succeed the membership label from the nearest hotspot in H_s.

In overall, substep (a) broadcasts the membership labels from previous clustering results by linking identical hotspots in both $\boldsymbol{S}_{t-1}$ and $\boldsymbol{S}_t$. Substep (b) spreads the membership labels according to the clustering result in $\boldsymbol{S}_t$. 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 Discussion of the algorithm

This section will reference to the conditions we discuss in the section 4.2. In terms of condition (1), this algorithm didn't depend on the scaling of the temporal dimension, instead, it uses the parameter *ActiveTime* to control the time frame. 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. Condition (7) will be justified in Appendix.

4.3 Results

The clustering results by applying our algorithm on the hotspots 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 hotspots 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 movement 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.

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. Combined with (Figure 4.1), It suggests the number of ignition has a positive relationship with the severity of the bushfire in an area.*

Second, the parameter *AcitveTime* is working in the algorithm. At 01:00 am 31 December 2019, there are missing values but the algorithm continues reconstructing the bushfire. Finally, the bushfire reconstructed by this algorithm can have arbitrary size and shape and multiple branches far away from each other. This feature is essential for the correctness of the algorithm.

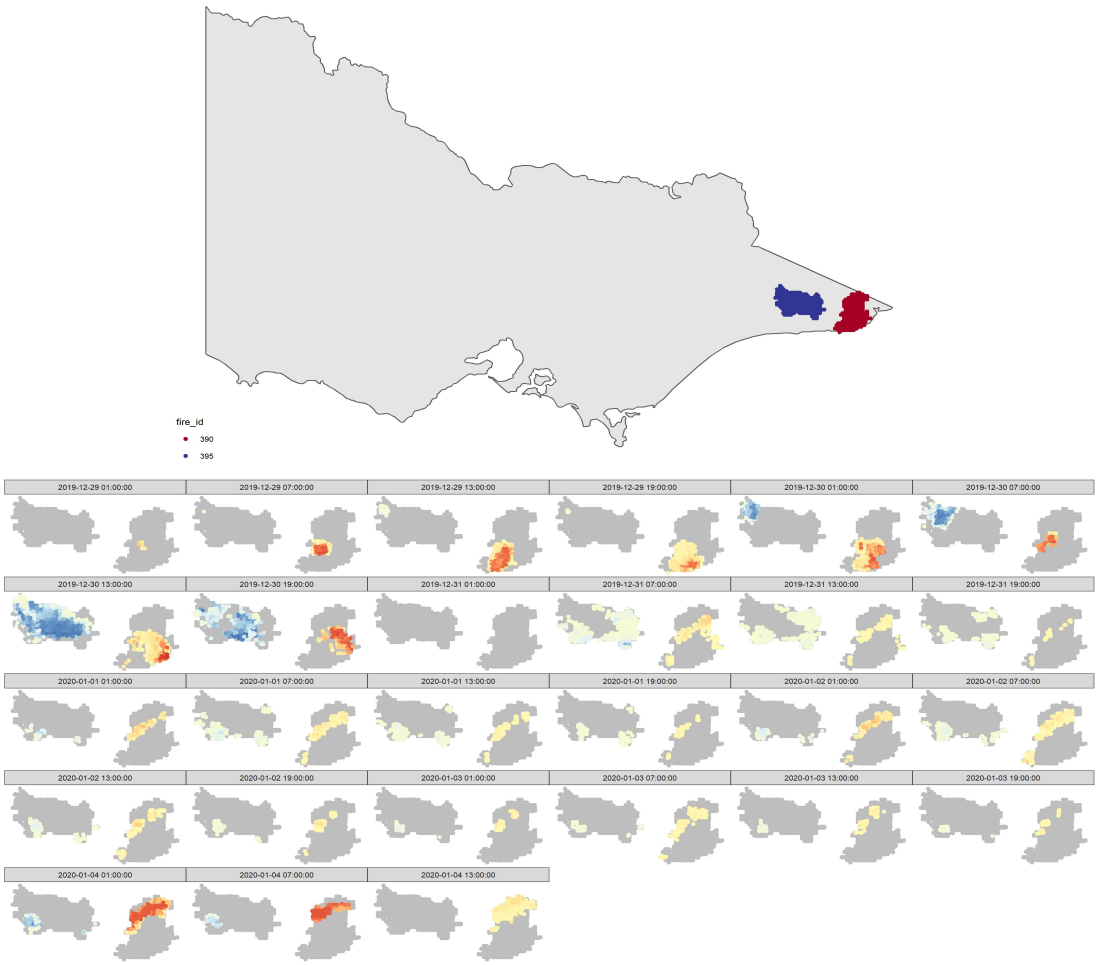

Figure 4.3: *The snapshots of hotspots show the bushfire behaviour of bushfire "390" and bushfire "395", which is two clusters from the results of the clustering algorithm. Fire "390" started at 01:00 am 29 December 2019 and fire "395" started at 07:00 am 29 Devember 2019. The grey shadow behinds the 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 differnet 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 will be the source of the wind speed data, other covariates such as vegetation factors, temperature and rainfall will 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. The wind speed will then be converted from mph to m/s. The end result of this step is a dataset with same covariates as the dataset used for modelling fitting.

Chapter 5

Classification of ignition causes

5.1 Model description

We use a random forest model to classify historical bushfire ignitions. Random forest (Breiman, 2001) is an ensemble learning method for building tree-based prediction models. It is perhaps one of the most regularly used black-box machine learning models in various fields (Boulesteix et al., 2012; Heung et al., 2016). It generates a certain number of decision trees by using the bootstrap aggregating technique and take the majority vote in the case of decision trees as the prediction. In contrast to the decision tree, random forest model can potentially reduce overfitting of the training set and have a built-in variable selection mechanism.

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

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

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. GAM 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).

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 week learners iteratively. Numerous competitions, for example, the Higgs boson machine learning challenge (Adam-Bourdarios 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.

The multinomial logit model, generalised additive model, random forest and XGboost are available in package `nnet` (**nnet**), `mgcv` (**mgcv**), `randomForest` (Liaw and Wiener, 2002) and `xgboost` (**xgboost**) respectively. Besides, package `lime` (**lime**) is used to perform feature selection and package `caret` (Kuhn, 2020) is used to control the training, hyperparameter tuning and candidate model selection process.

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. Within the 1872 test samples, we further randomly select 100 samples for feature selection. Thus, the number of remaining samples in the test set is 1772.

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. Unlike the global variable importance provided by many other packages,

the package `lime` (**lime**) provides the local variable importance under the assumption that machine learning models are linear at the local scope.

Given an observation, `lime` will randomly sample data points around the predictors, and obtain their predictions by passing them into the black-box model. It will then fit a simple model, typically a ridge regression (Hoerl and Kennard, 1970) or a lasso regression (Tibshirani, 1996), using 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 strategic 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 5.1. 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: *The top 10 most important variables for each candidate model ranked in descending order. Variables on the top are more important. The difference 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

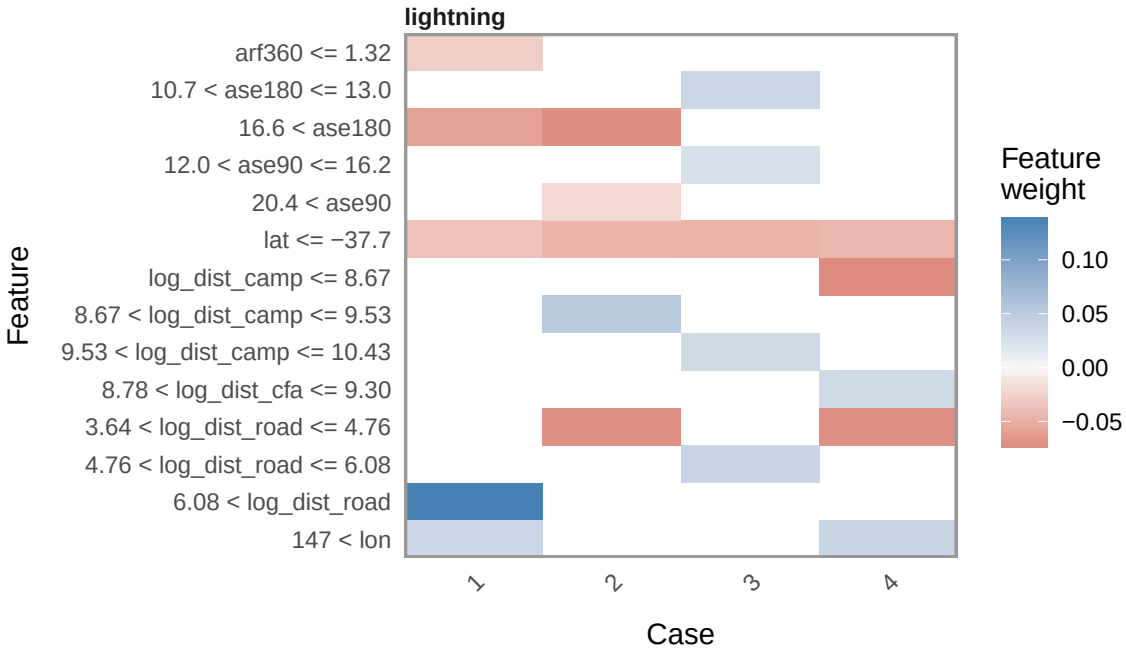

Figure 5.1: *An example of selecting the most important variables for the random forest model with respect to the predicted probability of the bushfire ignited by lightning in 4 cases using* `lime`*. The feature weight is the weighted importance within a case. In this example, the common feature that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

5.3 Hyperparameter tuning and candidate model selection

The hyperparameter tuning for each candidate model was done 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 for each hyperparameter and the optimal hyperparameters is given in the Appendix.

The final step of the modelling is candidate model selection. Model performance is compared by using both 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` (**proc**). 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 is better than all other candidate models in both prediction accuracy and multi-class AUC. Thus, we choose the random forest model as our final model. Model performance is given in Table 5.2. More details about the model performance can be found in the Appendix.

The overall accuracy of our model is 74.95%. The confusion matrix of the ignition classifier is shown in Table 5.3. It suggests that lightning-caused and accident-caused ignitions can be easily classified from other causes. Meanwhile, the model is not very confident with arson and burning off. 77.9% of accident-caused and 90.5% of lightning-caused ignitions are correctly recognised by the model, which is a reliable result.

Based on the prediction performance of the final model on the test set, we plot a map for error rate to reveal the spatial patterns. The plot is given in Figure 5.2. From the plot, we can observe that our model correctly predicts most of the cases in the mountain area, which is the east of Victoria. However, it performs worse in the Melbourne region. Besides, our model doesn't fit well on the boundary of the north-west of Victoria.

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

Table 5.2: *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 5.3: *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

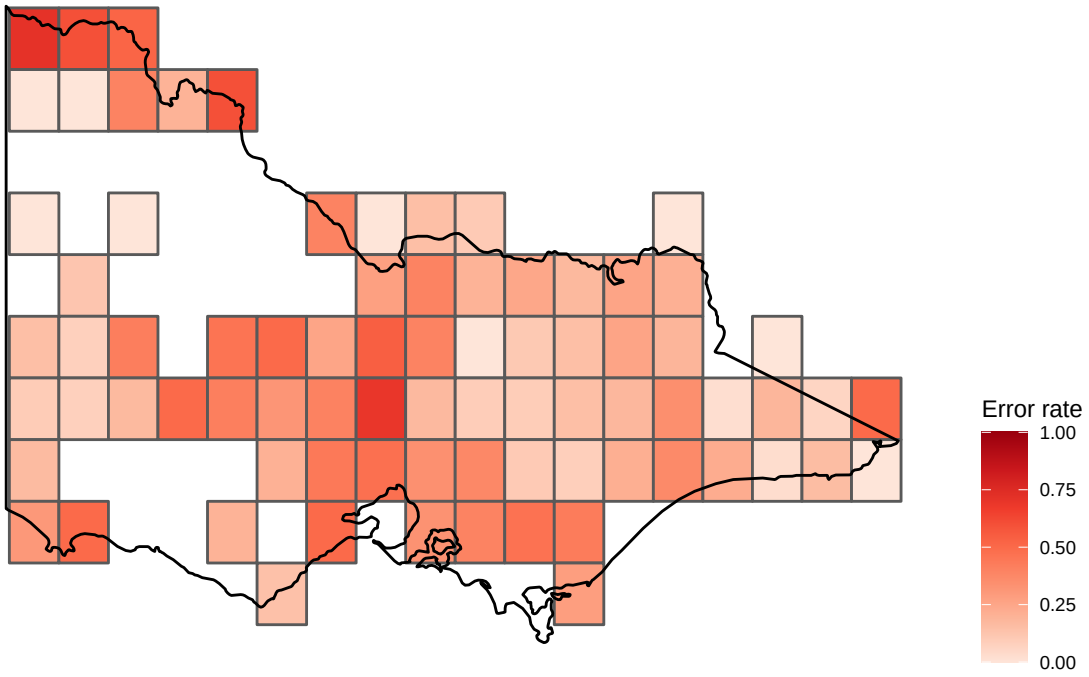

Figure 5.2: *The spatial patterns of the error rate of the final model. We omit regions with less than 5 bushfires occurred. Our model makes very few mistakes in the east of Victoria but has a higher error rate near Melbourne.*

5.3 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 average wind speed in the last 2 years has a high negative impact on the probability. Patterns in arson are almost the opposite of what has been shown in lightning. Latitude and average wind speed in past 24 months 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 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's 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.

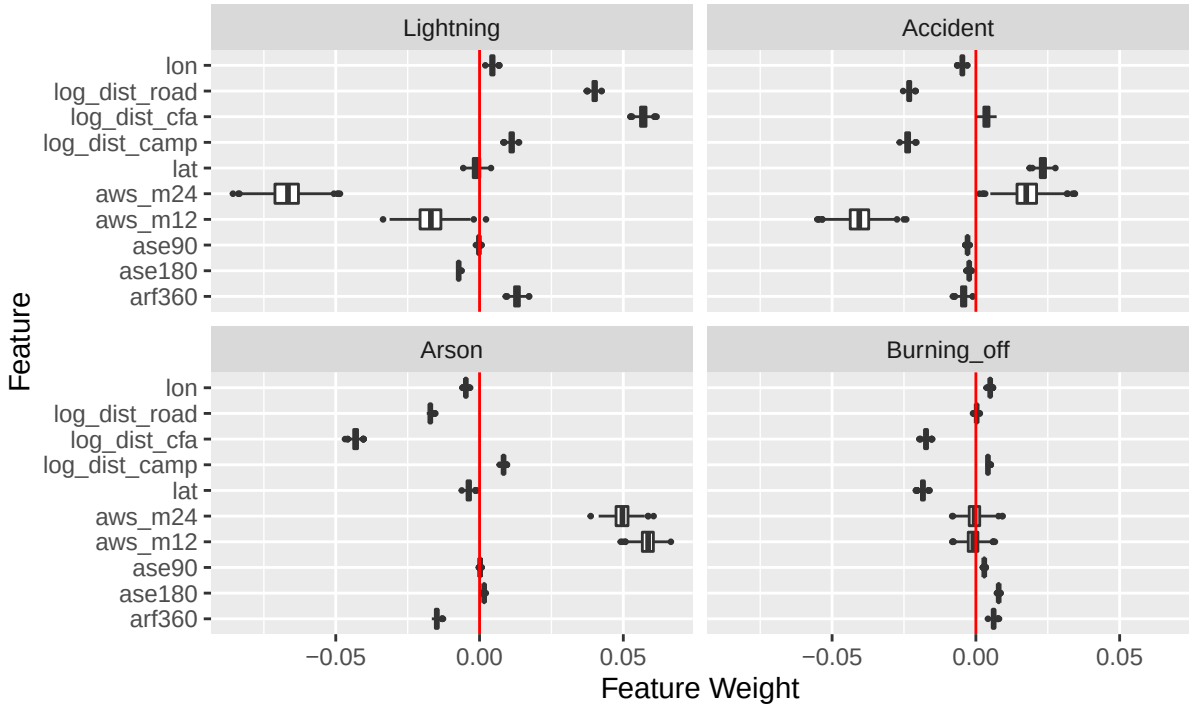

Figure 5.3: *Variable contribution to the probability of differenct causes. Variable has a positive weight means it has a postive impact on the probability. The same rule applies to negative weights. The magnitude shows the impact strength, which can be seen as the average marginal effect at the local scope.*

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 2019-2020 Australia bushfires. Figure 5.4 shows the prediction produced by the final model. And Table 5.4 summarize the prediction. According to the prediction, the most majority of the bushfires in 2019-2020 season were caused by lightning. However, there were 138 bushfires caused by accidents which took up 14% of the total fires. Most of the accident-caused bushfires were ignited in March. Besides, 37 bushfires were 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 was arson. Very few planned burns were predicted after October 2019 which suggests the correctness of our model that is because fire restrictions usually start in October.

The prediction in Figure 5.4 is only an example of using our model to predict the cause of the bushfire. Actually, with the location of the ignition and the corresponding weather conditions, covariates needed in this model can be easily produced, which makes our model adaptable for future fire prediction.

Furthermore, we provide a map with 0.2 degree spatial resolution for quick decision making of the cause of the bushfire, which is given in Figure 5.5. 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.

Other than that, Figure 5.5 also reveals the temporal and spatial characteristics of different causes. Probability of lightning-caused bushfire is almost time and spatial invariant. Bushfires in the east and the west of Victoria will be predicted by our model as lightning-ignited bushfires, which is reasonable. Because those areas are mostly national parks, forests and mountain regions, which could be ignited by lightning. 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 high probability region spread to the south over time and reach its peak in March. Probability of arson-caused bushfire outside the Melbourne region will decrease as time goes from October to March. Planned burns will only occur in October and November. It is a very low probability that a bushfire is caused by burning off in the summer of Victoria.

Table 5.4: *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 an 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%)

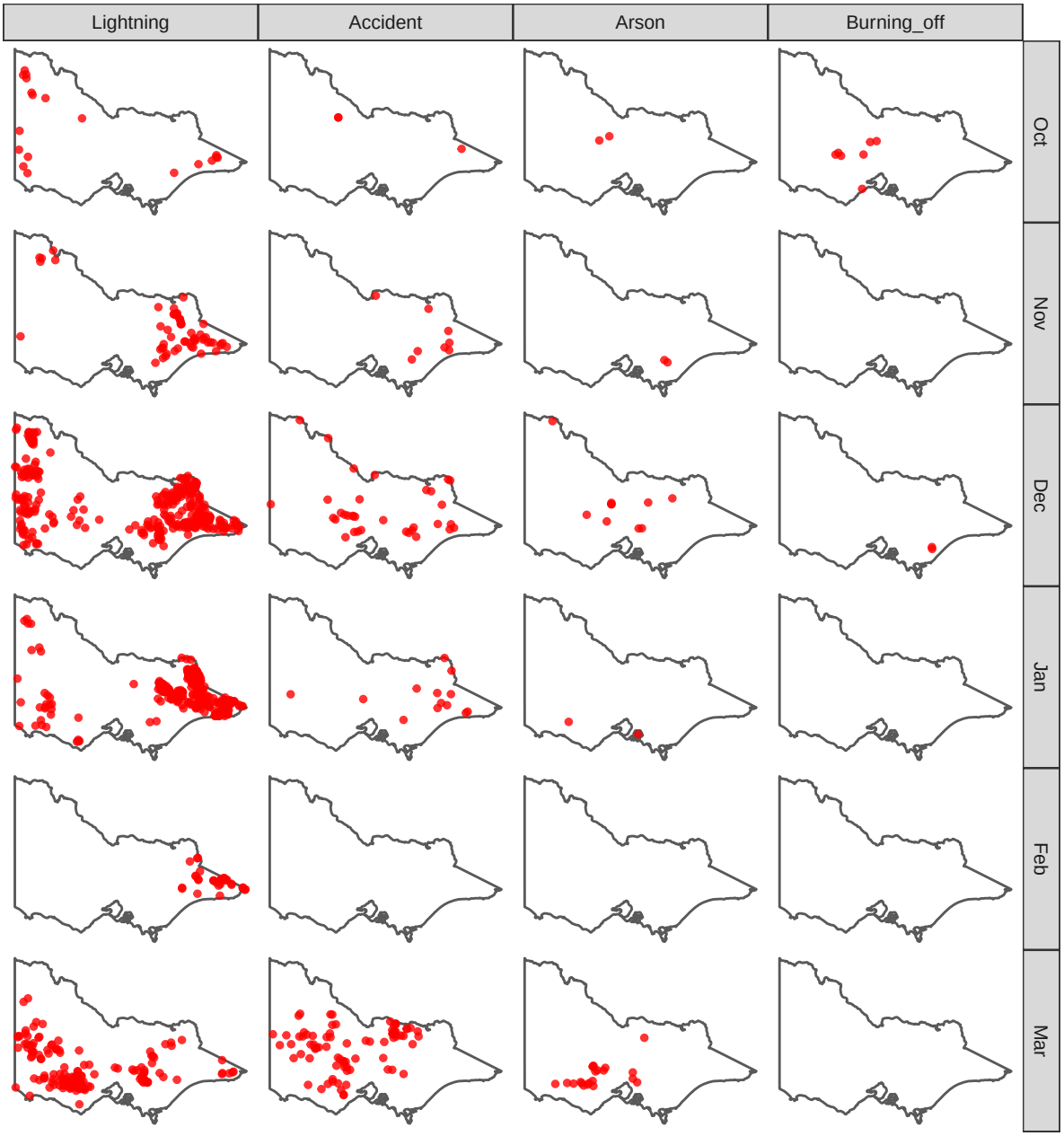

Figure 5.4: *Prediction of 2019-2020 bushfire ignition cause produced by our final model. Majority of the bushfires were caused by lightning. Most of the Planned burns occurred in October. There were two burning off in December 2019 predicted by the model which is abnormal. Arson were mainly in December 2019 and March 2020. The model predicts a lot of accident in March 2020.*

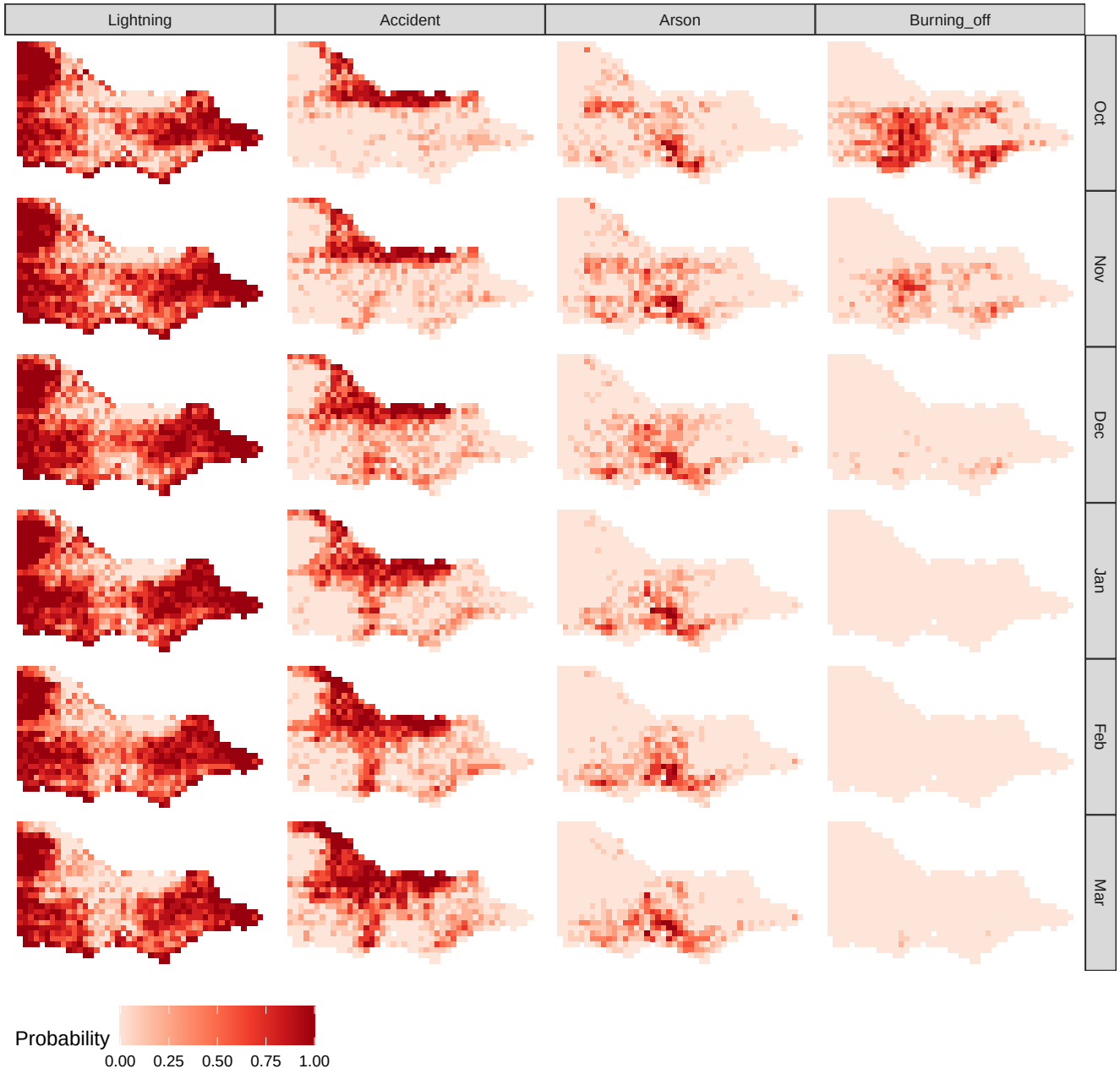

Figure 5.5: *A map with 0.2 degree spatial resolution for quick decision making of the cause of the bushfire during a bushfire season. This map is based on the assumption that the long-run weather condition of the new ignition is similar to 2019-2020 Australia bushfire season. Users need to match the 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.*

Chapter 6

Discussion

6.1 Policy Implications

Our prediction suggests that the main cause of the 2019-2020 Australia bushfires was lightning. And according to the clustering results, these fires were mostly started in remote areas, like East Gippsland. We suggest the Forest Fire Management Victoria and the Country Fire Authority to consider delivering planned burns on a larger scale in these regions in an earlier date ahead of the bushfire season. Smoke and pollution to the environment produced by the planned burns are unavoidable, but it is considerably better than the uncontrollable bushfires. We need to choose between them. 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. Budgets on bushfire prevention need to increase if possible, or at least, distribute heavier to these areas to employ more firefighters during the bushfire season.

Besides, according to our prediction, the number of accident-caused bushfires rose 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 hotspots 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 causes of the bushfire remotely by fusing the available data, the research 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

6.2.1 Data source

In this research, we didn't take into account the vegetation dynamics in modelling, instead, we only used a static map of vegetation information. This could lead to ineffective use of vegetation factors. However, we didn't find a similar public vegetation dataset in previous years. This could be mitigated in the future when there are more available data. Moreover, in the prediction, we used the ASOS wind speed as covariates. ASOS network only has a few stations near Melbourne so our weather conditions could be inaccurate. We knew there were accurate wind speed data on BOM, but they were not free to the public.

6.2.2 Clustering algorithm

Our clustering algorithm used only satellite hotspots data to reconstruct the bushfire dynamics. This didn't incorporate the vegetation factors, slope, smoke and weather conditions, which might affect the speed of the fire, the direction of the fire and the tolerance of the missing observations. This was already out of the scope of this research since it would take more time and effort but we already spent months to develop the current version.

6.2.3 Modelling

There were other machine learning models we didn't test on this task, like artificial neural network which might boost the performance. However, complex machine learning model

needed great effort in hyperparameters tuning which was possibly infeasible for our computing platform.

There were also other machine learning modelling techniques we didn't use in this research. The most impactful two would be feature engineering and ensemble modelling. We believed these methods could potentially increase our model performance, but we didn't apply them due to the time constraint. We might consider trying them in future work.

6.3 Future work

Our model can produce the probability of the bushfire ignited by a certain type of sources, but it doesn't provide us with the general risk of the bushfire ignited by that cause. Using the lightning-ignited bushfire as an example, it will give us

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

where $L = \textit{the bushfire ignited by lightning}$, $S = \textit{the bushfire ignited}$ and $\mathcal{F}$ is other information of the ignition, such as location, time, weather conditions, etc.

This probability is useful for bushfire investigation but is not particularly helpful for bushfire risk prediction. In general, decision makers wants to know the risk of the lightning-ignited bushfire $P(L, S|\mathcal{F})$ or the probability of the overall bushfirefire risk given certain conditions $P(S|\mathcal{F})$.

Our work can be extended by using the decomposition of the conditional probability:

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

We will demonstrate how this works by using the a naive method to estimate the $P(S|\mathcal{F})$ and eventually yield the $P(L, S|\mathcal{F})$.

A reasonable estimate of the probability of the bushfire starts in a region in a certain month is the fraction of the number of past December we observed bushfires on the total number

of past December. Figure 6.1 is the estimate of $P(S|\mathcal{F})$. As you can see, it produces a lot of missing values because we didn't observe bushfires in those regions in the past. The joint probability $P(L, S|\mathcal{F})$ can then be calculated by omitting other factors and simply using the prediction of 2019-2020 Australia bushfires. Figure 6.2 shows the final product. It can be directly used as the prediction of the risk of lightning-ignited bushfire.

Therefore, the potential future work will be developing a model to predict the overall risk $P(S|\mathcal{F})$ such that we can obtain the prediction of the risk of bushfire ignited by different causes.

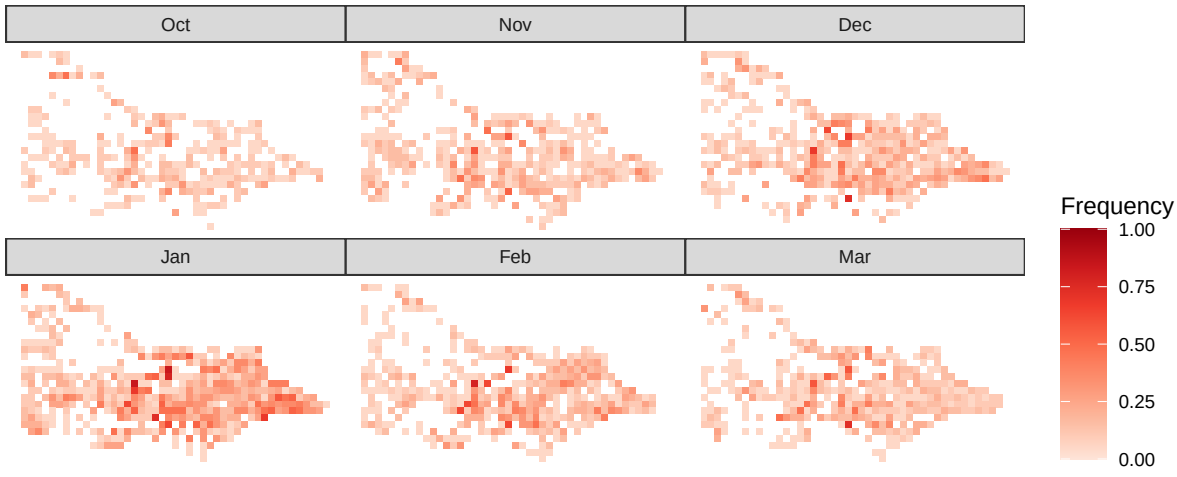

Figure 6.1: *A map of the frequency of bushfire occurred in a region. Some regions contain missing values because there was no observed bushfire.*

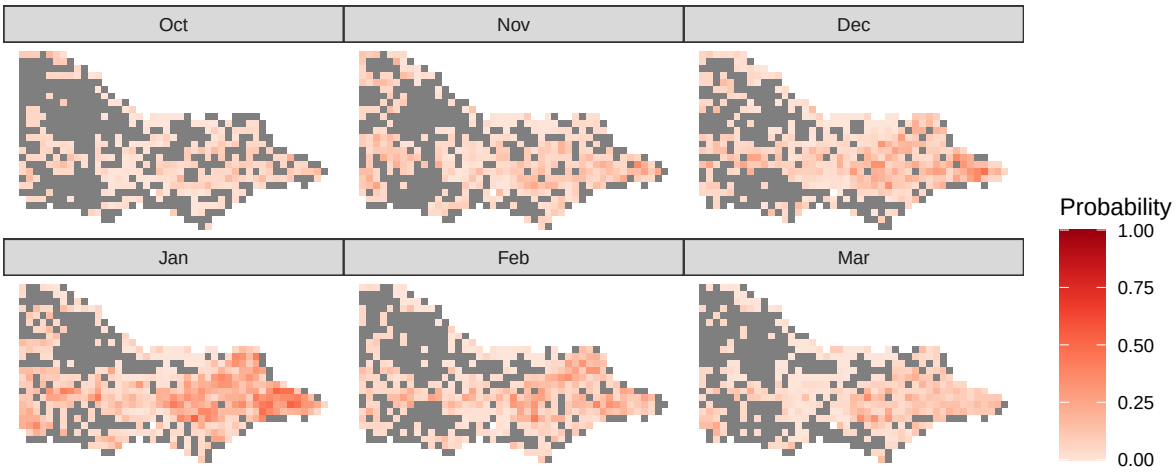

Figure 6.2: *The joint probability of lightning-ignited bushfire occur, which could be a future extension.*

Chapter 7

Conclusion

I will list the main points here (will be finished in the next 2 weeks). The conclusion will be brief given I have mentioned some of them in the contribution.

1. We predicted the cause of 2019-2020 bushfire, which showed the main cause.
2. The model can be used for any past or future ignition to predict cause, given the covariates are easily accessible
3. We developed an algorithm to reconstruct bushfire from hotspot data
4. Combine (1) and (3), it is a complete workflow to detect new ignition and predict the cause. This workflow can be smoothly adopted by others including CFA staff.

Appendix A

Covariate information

Table A.1: *Covariate information of the ignition classifier*

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 past 7 days	mm
arf14	Average rainfall in past 14 days	mm
arf28	Average rainfall in past 28 days	mm
arf60	Average rainfall in past 60 days	mm
arf90	Average rainfall in past 90 days	mm
arf180	Average rainfall in past 180 days	mm
arf360	Average rainfall in past 360 days	mm
arf720	Average rainfall in past 720 days	mm
se	Global solar exposure on that day	MJ/m^2
ase7	Average global solar exposure in past 7 days	MJ/m^2
ase14	Average global solar exposure in past 14 days	MJ/m^2

Table A.1: *Covariate information of the ignition classifier (continued)*

Covariate name	description	Units
ase28	Average global solar exposure in past 28 days	MJ/m^2
ase60	Average global solar exposure in past 60 days	MJ/m^2
ase90	Average global solar exposure in past 90 days	MJ/m^2
ase180	Average global solar exposure in past 180 days	MJ/m^2
ase360	Average global solar exposure in past 360 days	MJ/m^2
ase720	Average global solar exposure in past 720 days	MJ/m^2
maxt	Maximum temperature on that day	Celsius degree
amaxt7	Average maximum temperature in past 7 days	Celsius degree
amaxt14	Average maximum temperature in past 14 days	Celsius degree
amaxt28	Average maximum temperature in past 28 days	Celsius degree
amaxt60	Average maximum temperature in past 60 days	Celsius degree
amaxt90	Average maximum temperature in past 90 days	Celsius degree
amaxt180	Average maximum temperature in past 180 days	Celsius degree
amaxt360	Average maximum temperature in past 360 days	Celsius degree
amaxt720	Average maximum temperature in past 720 days	Celsius degree
mint	Minimum temperature on that day	Celsius degree
amint7	Average minimum temperature in past 7 days	Celsius degree
amint14	Average minimum temperature in past 14 days	Celsius degree
amint28	Average minimum temperature in past 28 days	Celsius degree
amint60	Average minimum temperature in past 60 days	Celsius degree
amint90	Average minimum temperature in past 90 days	Celsius degree
amint180	Average minimum temperature in past 180 days	Celsius degree
amint360	Average minimum temperature in past 360 days	Celsius degree
amint720	Average minimum temperature in 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: *Covariate information of the ignition classifier (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

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 with *ActiveTime* and *AdjDist* 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 the marginal effect

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

Figure B.1 and Figure B.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 B.1, it's very clear we need to choose 3000km for *AdjDist*. And in Figure B.2, we choose 24 hours for *ActiveTime*.

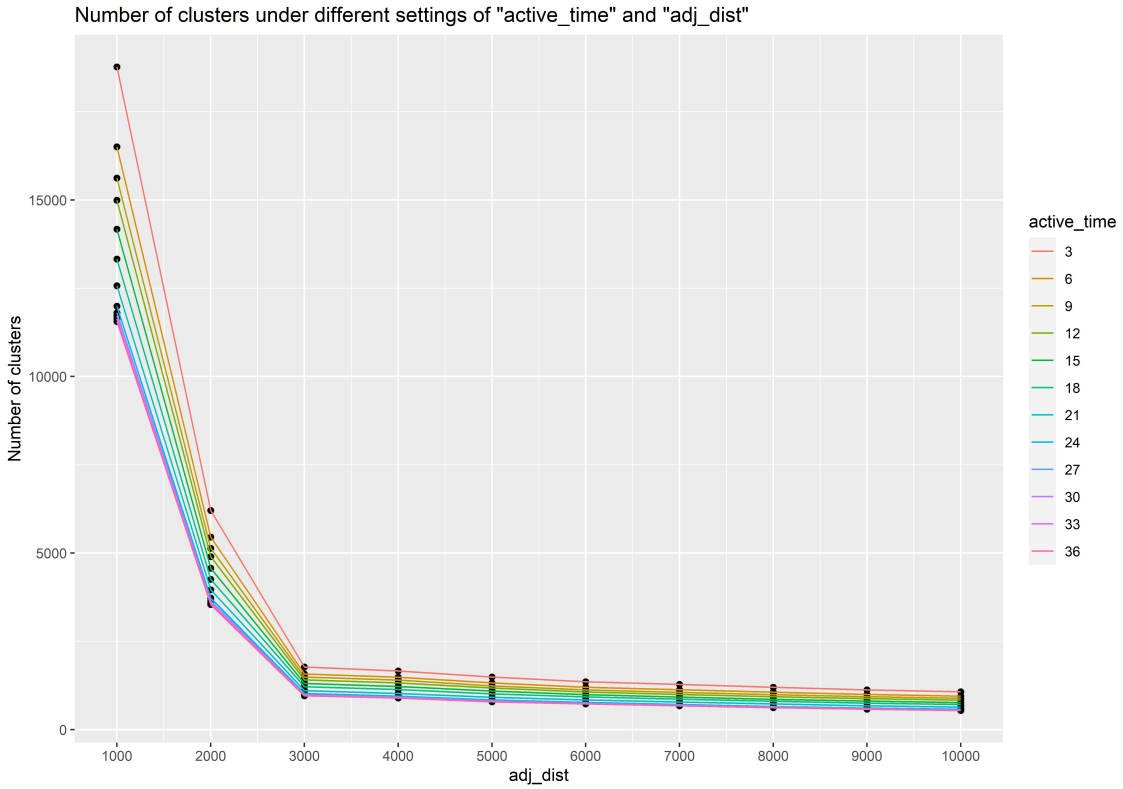

Figure B.1: *A visulization tool for parameter tuning in our algorithm. It works like a scree plot. We need to choose a point with a large second-order derivative. The reasonable choice of the parameter AdjDist is 3000km.*

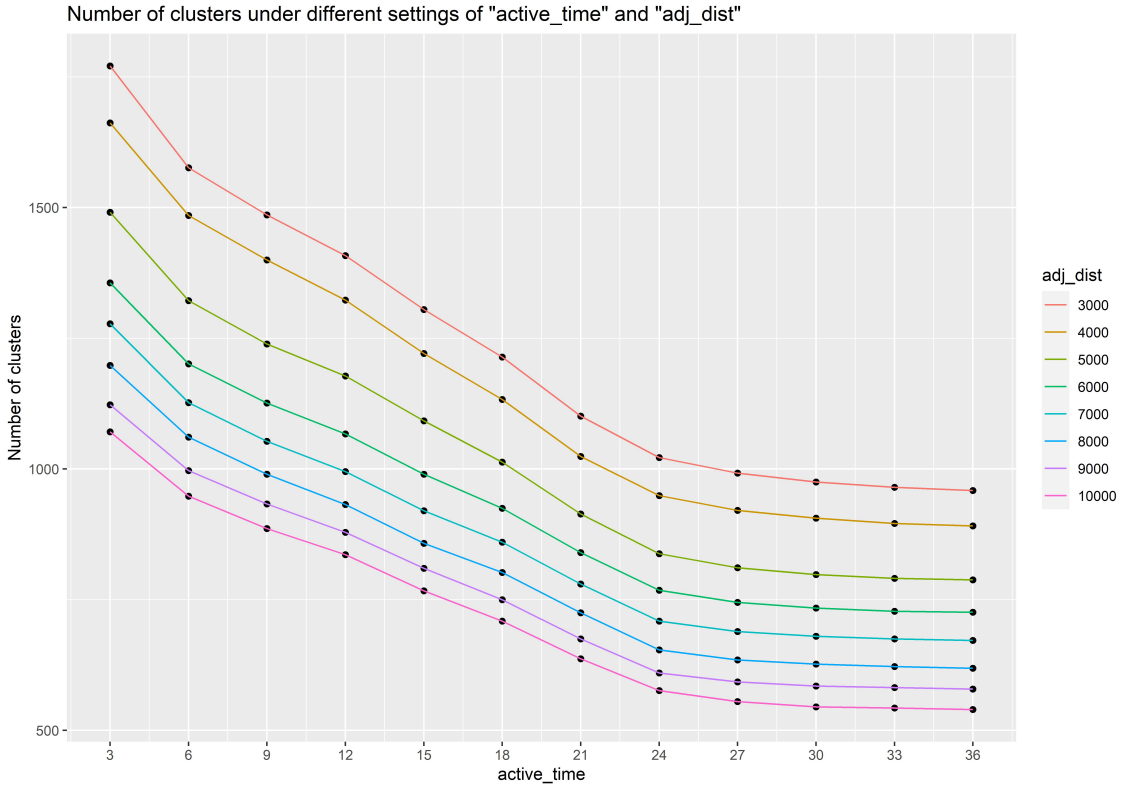

Figure B.2: *A visulization tool for parameter tuning in our algorithm. It works like a scree plot. We need to choose a point with a large second-order derivative. The reasonable choice of the parameter ActiveTime is 24 hours.*

Appendix C

Hyperparameter tuning for ignition classifiers

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

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

Table C.2: *A description of each hyperparameter. The definition of the hyperparameters is referenced from the documentations of the package* `nnet` *(***nnet***), package* `randomForest`*(Liaw and Wiener, 2002) and package* `xgboost` *(***xgboost***).*

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 C.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 D

Model performance

Table D.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 D.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 D.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 E

Supplementary material

Supplementary materials including 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 R'10, M Wiebe, P Peterson, P G'erard-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.

Iowa State University (2020). *ASOS-AWOS-METAR Data Download*. 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.

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/10.25919/5c5106acbcb02.

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). *rnaturalearth: World Map Data from Natural Earth*. R package version 0.1.0. https://CRAN.R-project.org/package=rnaturalearth.

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).

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.