Using Remote Sensing Data to Understand Fire Ignition during the 2019-2020 Australia 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

May 2020

Contents

Chapter 1

Statement of the topic

Along with the extreme heatwave in Australia 2019-2020, one of the most devastating bushfire season in history had been witnessed. Lighting strikes and arson were been discussed among pubilc as the main cause of this disaster. This research will explore the methods of fire ignition during 2019-2020 Australia bushfires season and provide a model to predict the fire risk of neighbourhoods. Hotspots data from the JAXA's Himawari-8 satellite and weather data from the Australian Bureau of Meteorology will be used in ignition identification and bushfire danger estimation. In addition, an interactive web application embeds with research outcomes will be built for data visualization purpose.

1.1 Motivation

In Australia, numerous occasions of bushfires cause losses of properties and life every year. Recorded since Hobart bushfires in 1967, the insurance claimed for building losses was greater than $A10 million (McAneney, Chen & Pitman, 2009) McAneney, Chen and Pitman (2009) show in their research that the average number of buildings had been destroyed by bushfire per year is 84, accounting for 20% of total building losses in hazard events.

The Australian bushfire season in 2019-2020, compared with other major bushfires in history, had a severer impact on environment and properties. According to the Parliament of Australia report (2020) on this devastating bushfire season, 3094 houses had been destroyed during this crisis and burnt land was over 17M hectares. These two figures are

the highest in history. Fortunately, only 33 people including firefighters died comparing to 173 in Black Saturday and 47 in Ash Wednesday (Parliament of Australia, 2020).

Therefore, understanding what caused the 2019-2020 Australian bushfires is extremely important which may provide us with information for future legislation and fire risk management.

Another motivation for us to conduct this research is the difficulty of finding the cause of bushfire in Australia. Beale and Jones (2009) state in their research that only 58.9% of the cause of fires is known. In the known cases, the percentage of deliberate and suspicious ignitions is around 50%. Besides, 35% of known cases are caused by accidents and 6% of known cases are caused by nature, for example, lightning. Provide probability information on the cause of bushfires may help investigators in practice.

1.2 Research aim and questions

This research aims to answer the following questions:

1. How to find the ignition methods during 2019-2020 Australia bushfire season? How to identify fire from hotspots data? How to classify the cause of ignition without classes labels? Will likelihood of ignitions increase near the campping site and road? Will likelihood of ignitions increase near lightning strikes?
2. How to model fire risk of neighbourhoods? Are tree-based models better than Generalized linear model and Generalized additive model in fire risk prediction? How to take into account the temporal patterns and spatial patterns in fire risk modelling? How to apply regularization and add noise in fire risk modelling?

1.3 Research plan

This reasearch will start from data collection. Various datasets will be collected for this research, and the primary datasets will be hotspots data, weather data and map data.

To understand the ignition of bushfires, a customized clustering algorithm will be developed to convert hotpots data into fire history, which will contain the starting time

and coordinates of each fire. This algorithm will mainly involve simulating fire growth, deciding fire boundaries controlled by tolerance and assigning hotspots data to the most probable cluster. After the clustering result being obtained, fire history will be visualized to diagnostic the performance of the algorithm. It will be done by comparing the behaviour of the same fire under different sets of hyperparameters.

Exploratory data analysis of fire history and its relative factors, like weather condition, distance to the nearest road and distance to the nearest recreation site will be performed. Prior knowledge and featuring engineering will be needed to fully understand the relationship. We expect to discover relationships between the ignition of fire with these factors, which can help us identify the cause of bushfires later on.

In order to examine the sources of fire ignition, different strategies will be used depending on the outcome in the previous section. If the findings from the analysis are strong and directly related to potential sources of fire ignition, hypothesis tests will be conducted to examine the pattern. If the evidence is weak, we will consider developing another clustering algorithm on fire history. This algorithm will be designed to maximize the distance between bushfire started with different causes in a high dimensional space. A probability model then can be built on top of it, which can provide a probabilistic answer for the cause of bushfires during 2019-2020 bushfire season.

Models for predicting fire risk of neighbourhoods will be built using raw hotspots data instead of the fire history because the hotspots data can be considered generated from a partially observable Markov process, and the underlying state is the development of the bushfire. From low complexity models like logistic regression to high complexity models like random forest will be tested.

For sharing our research outcome, a shiny app will be built and hosted online. In addition, both static and dynamic visualization tools will be considered using. Due to the nature of Spatio-temporal data, which has at least 3-dimensional features, static map view without faceting can only provide limited information. Meanwhile, faceting map view with time will be limited by the size of caravans. Animation based map view is computationally expensive and distracting though it provides more information. Better ways for visualizing Spatio-temporal data will be explored during the development. The potential product will

be an interactive map view with triggers to transform data and manipulate the aesthetics specifications.

1.4 Scope of research

<left blank>

Chapter 2

Literature Review

The literature review contains two parts. The first part is the historical and recent researches in Australian bushfires modelling. The second part is the existing tools of spatiotemporal data visualization and analysis.

2.1 Bushfires modelling

Existing researches in bushfires modelling can be divided into two main categories, one is simulation modelling, another is analytical modelling.

In simulation modelling, Keane et al. (2004) have attempted to use landscape fire succession models (LFSMs) to model fire behaviour including fire ignition and fire spread. They are a group of spatial simulation models which taking into account fire and vegetation dynamics. Similarly, Bradstock et al. (2012) used FIRESCAPE model which mainly involved simulating fire behaviours with fuel and weather conditions. Simulation methods are cost-effective and time-effective in modelling bushfires (Clarke et al., 2019). However, Clarke et al. (2019) also stated in their research that ignition likelihood is not well discussed and considered in these models. Besides, these methods seldomly address ignition types of bushfires which we are interested in. Therefore, simulation modelling will not be considered in our bushfire ignition research.

Alternatively, analytical modelling is a more popular way to build bushfires models. In analytical modelling, the general framework for analysing bushfires ignitions is generalised additive model (GAM). Bates, McCaw and Dowdy (2018) used it for predicting the number of lightning ignitions. Besides, some studies include a logit link to extend the model for bushfires ignition likelihood prediction (Read, Duff and Taylor, 2018; Zhang, Lim and Sharples, 2017). Moreover, mixed-effects had been considered for spatial and weather factors (Duff, Cawson and Harris, 2018). Simpler models were also been used in this field, like multiple linear regression, negative binomial regression and generalised logistic regression (Cheney, Gould, McCaw and Anderson, 2012; Plucinski, McCaw, Gould and Wotton, 2014; Collins, Price and Penman, 2015). Particularly, instead of using a model, some researches performed statistical testing and exploratory data analysis to test certain hypothesises of bushfires (Miller et al., 2017; Dowdy, Fromm and McCarthy, 2017).

Common covariates for ignitions analysis are weather conditions, vegetation types, topographic information and anthropogenic variables. Moreover, various indexes had been used in modelling. Some researches choose to use indexes variables developed by McArthur such as the Forest Fire Danger Index (Clarke et al., 2019; Read et al., 2018), while others choose to use indexes developed by Canadian Forestry Service such as Canadian Fire Weather Index and Drought Code (Plucinski et al., 2014). We doubt that these indexes are irrelevant with fire ignition prediction because they are extracted from weather and vegetation information. However, it may help improve our model performance because it can be viewed as features generated from feature engineering. We will test if these metrics are significant in our research.

Although numerous studies for ignitions analysis have applied semiparametric and parametric methods, little analytics attention has been paid to more complex model such as tree-based model, support vector machine and artificial neural network. These tools are well developed in machine learning, which will be considered in this research.

Most of the existing works of bushfire ignition analysis focus on a certain area such as south-eastern Australia (Clarke et al., 2019), or a certain state such as Victoria (Read, Duff & Taylor, 2018). In this research, we will primarily focus on Victoria but eventually extend the results throughout Australia. In addition, little research has been done in analysing the

cause of 2019-2020 bushfires. Due to unknown causes of 2019-2020 bushfires, clustering, which is a rarely considered method in this field, will be used to identify ignition types.

Chapter 3

Data collection and processing

There are numerous open source data sets available that are collated to provide the data to address the research questions. The main data resource that differs from that used in the literature is the satellite data that records hotspots. The next sections describe how the data is accessed, and and pre-processed for later analysis.

3.1 Sources

Table 3.1 summarises the data sources.

3.1.1 Hotspots

Hotspot data is downloaded from the JAXA's Himawari-8 satellite. This satellite is positioned in geostationary orbit at 140 degrees east longitude, and the revisit period is 10-minute. Its management system - JAXA's P-Tree system, provides WildFire observation product with 2km spatial resolution. (P-Tree System, 2020). Details on how to download this data are provided by [ozjim post].

3.1.2 Weather

Weather data were collected from the Australian Bureau of Meteorology, by using an R package - Bomrang (Adam, Mark, Hugh & Keith, 2020). Due to the limitation of APIs provided by the package, we crawled data from BOM's website for extra information.

Table 3.1: *Data information*

Data set name	Spatial Resolution	Temporal resolution	Time
Hotspots data - JAXA's Himawari-8 satellite	$0.02^\circ \approx 2km$	Per 10 minutes	2015-2020
Weather data - Australian Bureau of Meteorology		Daily	2019-2020
Map - OpenStreetMap	2m		2020
Fuel layer - Australian Bureau of Agriculture and Resource Economics and Sciences	100m		2018
Victorian CFA fire stations - Department of Environment, Land, Water & Planning	20m		2020
Victorian Recreation sites - Department of Environment, Land, Water & Planning	10m		2020
Fire Origins - Department of Environment, Land, Water & Planning	100m		1972-2019

3.1.3 Fuel layer

To characterise the fuel we used forest of Australia (2018) from the Australian Bureau of Agriculture and Resource Economics and Sciences. It is a fuel layer contains the vegetation information across Australia.

3.2 Pre-processing

3.2.1 Data types

The hotspot and weather data is csv format, which can be processed generally with the `tidyverse` tools (Wickham et al., 2020) in R. Fuel layer and other map data are presented as geospatial objects, which can be processed using the tools in the `sf` (Pebesma, 2020).

3.2.2 hotspots clustering

Hotpots are point data obtained by light detection from the satellite. They are snapshots of bushfires every 10 minutes. However, just like it is hard to identify a single plant from the top view of a grassland, information about individual bushfire can not be directly derived from the raw hotspot data. In order to understand the ignition time and ignition place of each bushfire, we need to separate hotspots into clusters and trace the growth of each cluster. To preprocess the hotspots data, we selected the observations in Australia from the full disk. Meanwhile, hotspots with irradiance under 100 watt per square metre will be deleted. We restricted our study to hotspots that have significant firepower. An hour

id has been assigned to each observation range from 1 to T, represents the relative time the hotspot being observed. On top of the tidy hotspots data, a clustering algorithm was developed to identify fire clusters. Details about the algorithm can be found in table 3.2.

By using this algorithm, we assigned each hotspot a cluster membership which we called *fire_id*. Meanwhile, we recorded the characteristics of each cluster including its centroid, starting time, ending time and movement. The inspiration behind this algorithm is the behaviour of real world bushfire which can be summarised as two hyperparameters, the distance of spread in each hour and the lifetime of fire since last observed, represented by r_0 and t_0 respectively. These two hyperparameters have been used to determine if a new hotspot belongs to an existing or new cluster.

Table 3.2: *A clustering algorithm for hotspots*

Algorithm 1 Hotspots clutering

input: Hotspots dataset H : (Hour_id$^{(n)}$, Coordinates$^{(n)}$), n = 1, 2, ... N
An empty dataset F : (Fire_id$^{(m)}$, Coordinates$^{(m)}$, Active$^{(m)}$), m = 1, 2, ...
An empty vector K $\in \mathbb{N}_1^n$
A distance hyperparameter $r_0 \in \mathbb{R}^+$
A time hyperparameter $t_0 \in \mathbb{N}^+$

output: A vector K $\in \mathbb{N}_1^n$ contains memberships of hotspots
A dataset F contains fire clusters information including memberships, latest centroids and time from last updated

1 : select subset $H_c \in$ H where Hour_id == 1
2 : calculate distance matrix D for Coordinates in H_c
3 : assign 1 to a zero adjacency matrix A for where D $\leq r_0$ // hotspots with relative distance less or equal to r_0 will be considered belong to the same cluster
4 : create undirected unweighted graph G from A
5 : record memberships of G to K
6 : record clusters classes to Fire_id and record clusters centroids to Coordinates of F
7 : set Active in F to t_0 // Active clusters are fire being observed in the last t_0 hour
8 : **for** hour = 2, ... T **do**
9 : let Active -1 and select subset $F_c \in$ F where Active ≥ 0
10 : select subset $H_c \in$ H where Hour_id == hour
11 : append Coordinates from F_c to H_c
12 : repeat step 2 - 4
12 : **for** $h_i =$ each hotspot in H_c **do**
13 : **if** h_i share the same membership as one of active clusers in F_c **then**
14 : copy the corresponding Fire_id of the nearest active cluser to K
15 : **else** copy the membership from G to K
16 : **end if**
17 : **end for**
18 : update F for clusters involed in current timestamp and reset corresponding Active to t_0
19 : **end for**

3.3 Compiled data

The end result of the data pre-processing is a set of data tables, that are related by indexes. Figure 3.1 shows these tables and the indexes which allow information to be compared across tables. Technically, this figure is called a conceptual entity relationship diagram. It will be useful for describing the data model to be used in both modelling and web interface for communicating fire risk.

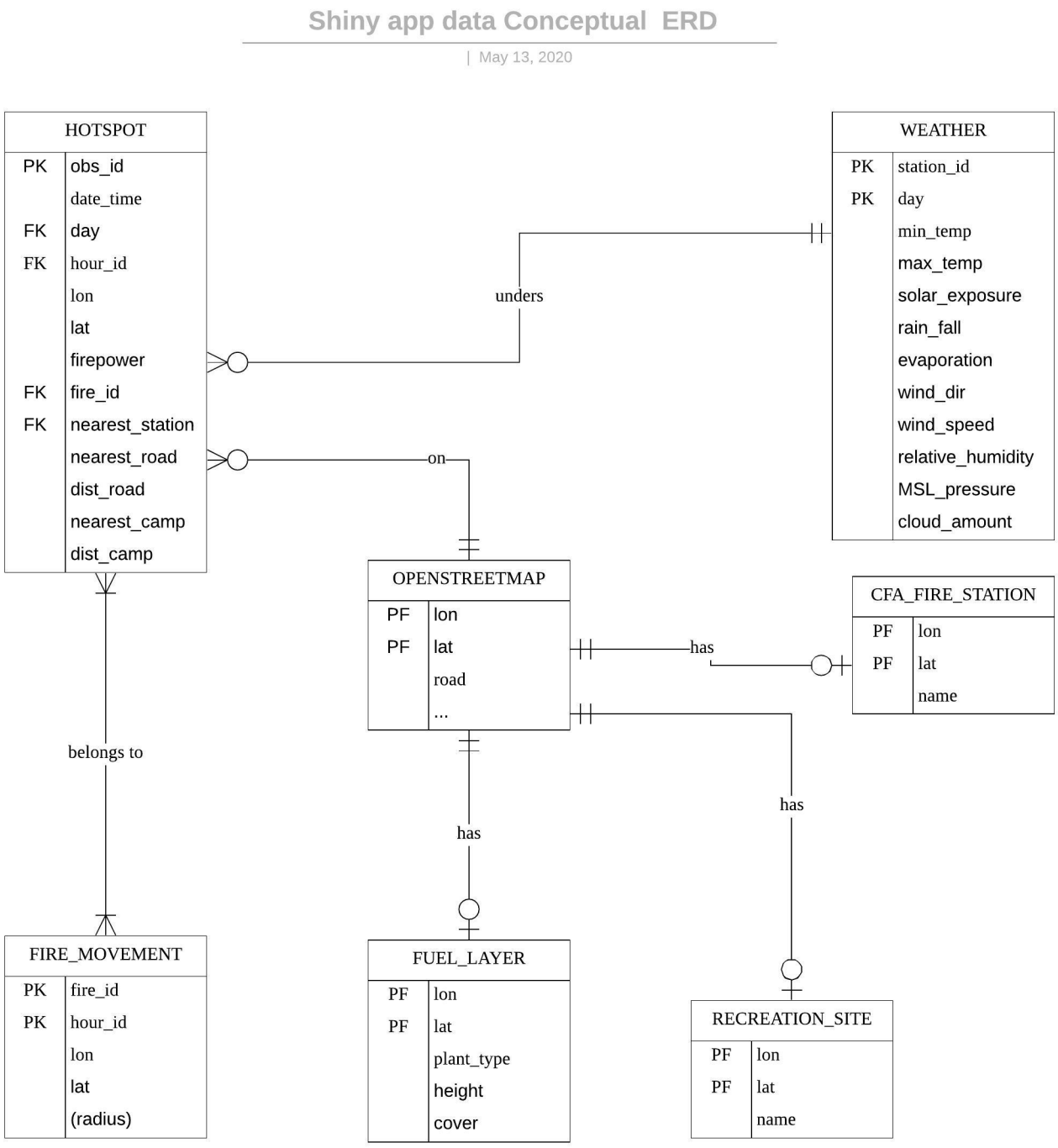

Figure 3.1: *Entity relationship diagram illustrating the relational tables of the compiled data. Tables correspond to processed hotspot data, weather, local facilities like CFA sites. This data structure is useful for the data modeling and web app development.*

Chapter 4

Exploratory data analysis

Chapter 5

Modelling

<plan>

5.1 Predicting ignition method

<plan>

5.2 Modelling fire risk

<plan>

Chapter 6

Timeline

The research plan for this semseter can be found in Table 6.1. Future research plan can be found in Table 6.2.

Table 6.1: *Research plan till week 9*

Timeline	Tasks
Week 2	Geographic data background reading
Week 3	Collect Remote sensing data (JAXA himawari-8 satellite) and explore BOM weather data APIs (Bomrang)
Week 4	Collect Road Map (OpenStreetMap) and read articles in SpatioTemporal data visualization and modelling
Week 5	Develop clustering algorithm for remote sensing data
Week 6	Test diferent hyperparameters for clusetring
Week 7	Exploratory data analysis on fire clusters and data integration
Week 8	Feature planning for the shiny app
week 9	Write research proposal and prepare the first presentation

Table 6.2: *Research plan since June*

Timeline	Tasks
June - July	Modelling fire ignition and fire risk
August	Consolidate findings and create mockups of the shiny app
September	Develop the shiny app and perform different levels of testing
October	Write thesis and prepare the second presentation

Chapter 7

Supplementary materials

Chapter 8

Biblography

Chapter 9

Words from last week

9.1 Data

9.2 Methodology

Both supervised and unsupervised learning will be implemented to reach the research aims.

To understand the ignition of bushfires, a customized clustering algorithm will be developed to convert hotpots data into fire history, which will contain the starting time and coordinates of each fire. This algorithm will mainly involve simulating fire growth, deciding fire boundaries controlled by tolerance and assigning hotspots data to the most probable cluster. After the clustering result being obtained, fire history will be visualized to diagnostic the performance of the algorithm. It will be done by comparing the behaviour of the same fire under different sets of hyperparameters.

Exploratory data analysis of fire history and its relative factors, like weather condition, distance to the nearest road and distance to the nearest recreation site will be performed. Prior knowledge and featuring engineering will be needed to fully understand the relationship. We expect to discover relationships between the ignition of fire with these factors, which can help us identify the cause of bushfires later on.

In order to examine the sources of fire ignition, different strategies will be used depending on the outcome in the previous section. If the findings from the analysis are strong and directly related to potential sources of fire ignition, hypothesis tests will be conducted to examine the pattern. If the evidence is weak, we will consider developing another clustering algorithm on fire history. This algorithm will be designed to maximize the distance between bushfire started with different causes in a high dimensional space. A probability model then can be built on top of it, which can provide a probabilistic answer for the source of bushfire ignition during 2019-2020 bushfire season.

Models for predicting fire risk of neighbourhoods will be built using raw hotspots data instead of the fire history because the hotspots data can be considered generated from a partially observable Markov decision process, and the underlying state is the development of the bushfire. From low complexity models like logistic regression to high complexity models like deep neural network will be tested.

For sharing our research outcome, a shiny app will be built and hosted online. In addition, both static and dynamic visualization tools will be considered using. Due to the nature of Spatio-temporal data, which has at least 3-dimensional features, static map view without faceting can only provide limited information. Meanwhile, faceting map view with time will be limited by the size of caravans. Animation based map view is computationally expensive and distracting though it provides more information. Better ways for visualizing Spatio-temporal data will be explored during the development. The potential product will be an interactive map view with triggers to transform data and manipulate the aesthetics specifications.

9.3 Preliminary Results

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.