Development of Bushfire Risk Prediction Web App in Victoria using Open Data

Brenwin Ang & Helen Evangelina

12/11/2021

Contents

Introduction and Motivation

Bushfires is a common and natural phenomemon that occurs frequently in many places around the world. Victoria, Australia however, is one of the most fire-prone regions in the world - given its fire conducive weather and fuel conditions. It is an intrinsic and inevitable part of Australia's environment which has significantly shaped our landscape (Australian Government Geoscience Australia 2021). Survey studies by ANU shown that about 80% affects Australians in some way whether through direct damage or indirectly through anxiety or smoke (Biddle et al. (2020)).

There have been catastrophic bushfire events in the past including Black Saturday 2009 in Victoria, Ash Wednesday 1983 in Victoria and South Australia, the 2006 December bushfires, as well as the recent 2019-2020 bushfires. (Victoria (2021)) Canada, USA, Turkey, Greece, Italy and Russia faced a devastating bushfire season in the summer of 2021 since it was the hottest July ever recorded, raising concern for Australia since it is moving into summer. Especially since the 2019-2020 bushfire in Australia known colloquially as Black Summer resulted to more damages on property and the environment compared to other bushfire events in history, with 3094 houses destroyed and over 17M hectares of land burned (Richards, Brew, and Smith 2020).

Even though bushfires cannot be avoided naturally, their consequences can be minimised. Thus, it is important to understand and predict the risk of bushfire as understanding the risk of fire of certain areas would help in developing strategies for mitigating the risks. Effective bushfire risk management pivots around providing relevant and timely data to both the emergency personnel and public. For instance, a prediction model can tell us whether an area has a high risk of bushfire, so resources can be focused on areas with high fire risk.

This report condenses the research work on bushfire risk modelling done by Brenwin Ang and Helen Evangelina as part of their internship for ETC5543 Business Analytics Creative Activity coursework. This work focus its window of study in Victoria. This is an extension to the work done by Di Cook, Emily Dodwell and Patrick Li on hotspot clustering algorithm. It is based on Patrick Li's thesis titled "Using Remote Sensing Data to Understand Fire Ignitions in Victoria during the 2019-2020 Australian Bushfire Season" (Li 2020). Weather is considered a determinant for forest fires. Therefore, environmental variables such as temperature, relative humidity, wind speed, radiation, drought index, etc will be used in the modelling as predictors.

Generally, fire risk is defined as the likelihood of a fire occurring, multiplied by the severity of the fire (Merton Council 2013). In the case of this report, fire risk is defined as likelihood of fire ignition occurring. The overall objective of this work is to develop a `Shiny` web app to monitor potential fire ignitions, incorporating fire risk model developed to monitor the 2021-2022 Victorian bushfire season based on environmental variables in where users would be able to toggle the input values of the environmental variables which would be incorporated the risk predictions. The resulting app seeks to enable easy access to bushfire information, raise awareness to bushfires and provide data information to make informed decisions to better adapt to the many impacts of climate change. The corresponding github repository including Patrick's thesis can be found here.

Table 1: data source, variables, format and temporal resolution used in this study

data_source
SILO (https://www.longpaddock.qld.gov.au/silo/)
ERA5 Reanalysis data (https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels-monthly-means?t
BoM's AWRA-L (http://www.bom.gov.au/water/landscape/#/sm/Actual/month/-26.32/132.54/3/Point/Separate/-15.6/
Department of Environment, Land, Water and Planning (DELWP) (https://discover.data.vic.gov.au/dataset/forest-types-

Data Sources and Integration

Following data analysis is based off the map of Victoria that is divided into 20x20 cells. Each cell is roughly $0.451ř \times 0.257ř$ (or $50km \times 28km$). Note that only cells within Victoria were considered. The partitioning the study area into equally-sized marked grid cells greatly facilitates the analysis- allows comparisons between cells, point-based fire ignitions to be viewed not just as individual points but as a spatially varying density throughout the study area. The cell cize can be determined through exploratory data analysis (Cardille and Ventura (2001)). From (other) gridded data or vector data, it also allows numeric and categorical variables to be co-registered to each cell. A coarse grid cell was deemed adequate to spread out the spatial variation or likelihood of fire ignitions while conserving landscape-level attributes. Nevertheless, this is easily tunable.

The study also only focuses on bushfire seasons of October through to March given most fire occur during these period. Data dates from 2016 to 2021. The period of data used presents a trade-off- the data must be long enough to capture the temporal variability but not too long that it ignores the changes in the data such as climate change or human-induced spatial ignition patterns. (Parisien et al. (2005))

Fire Ignitions Data

Fire ignitions are represented as spatial points on the map. There were two sources of bushfire ignitions data.

Firstly, satellite hotspot data from Himawari-8 satellite taken from the Japan Aerospace Exporation Agency FTP site. Himawari-8 data is a remote sensing data collected by remote sensors carried by a satellite. Upside of this data is that it provides a high temporal and spatial resolution. It is able to capture bushfires starting from remote areas very difficult to access by foot. Additionally, it has a 10-minute time resolution. Fire ignitions were detected in real-time using a clustering algorithm produced previously. (more information on how this is done can be found in Patrick's thesis on github).

Secondly, historical First Responder's data, this can be thought of as fires reported upon sight. That is, based on where the First Responder saw the fire. This is sourced from the Department of Environment, Land, Water and Planning (DELWP). Most of these were reports from volunteers manning fire toweres erected around Victoria. (See appendix for map of fire towers around Victoria)

Climate and Landscape Variables

For a bushfire to ignite, it needs an ignition source and conducive weather and landscape conditions. Fire ignitions can be classified into 4 categories- lightning, arson, accident and planned burning off (back burning which escaped containment line). Previous analysis suggests more than 80% of fires can be attributed to lightning. Cook (2020).

In this analysis, 11 covariates are considered to identify their effect on bushfire ignitions. These variables are: maximum temperature (`max_temp`) ($řC$), relative humidity (`rh`) (%), solar radiation (`radiation`) (MJ/m^2), derived FAO56 short crop evapotranspiration rate (`et_short_crop`) (mm), daily rainfall (`daily_rain`) (mm), leaf area index in high vegetation in $m^2\ m^{-2}$ (`lai_hv`) ($m^2\ m^{-2}$), leaf area index in low vegetation (`lai_lv`) ($m^2\ m^{-2}$), 10m wind speed (`WS10`) ($m\ s^{-1}$) and forest type in Victoria (`vic_forest`). Information about

the data sources of each of the variables presented in Table 1 above. *More information regarding the data sources can be found in the Appendix.

A subset of these variables namely `max_temp`, `rh`, `WS10` and `et_short_crop` (proxy for meteorological drought index) are used in McArthur Forest Fire Danger Index (FFDI), which is used widely around Australia to measure the degree of danger of fire in Australian forests. Dowdy et al. (2009) A plethora of studies have investigated the causes of widespread bushfires. However, to date, most of the existing literature focuses mainly on bushfire spread rather than ignitions.

As seen in the table, many of these data sources provide too fine a spatial grid resolution than required for our structure. Therefore, these are regridded to match the 20×20 gridded Victoria map. For numeric variables (all except `vic_forest`), the values of the finer cells within our larger cells are averaged. Categorical variable (`vic_forest`) are encoded into indicator variables, for example, any grid cell containing a forest is set to a value of 1, and 0 otherwise.

Additionally, 2 months lag of each of the variable (except `vic_forest` since it is periodic data) was included in the modelling. Studies have shown that not just current conditions but ongoing conditions such as heatwaves or drought creates favourable for bushfire ignitions. Deb et al. (2020)

In the process, we automated the workflow of extracting the variables from the various sources- giving access to a wide range of potential predictors.

Himawari-8 hotspot data processing

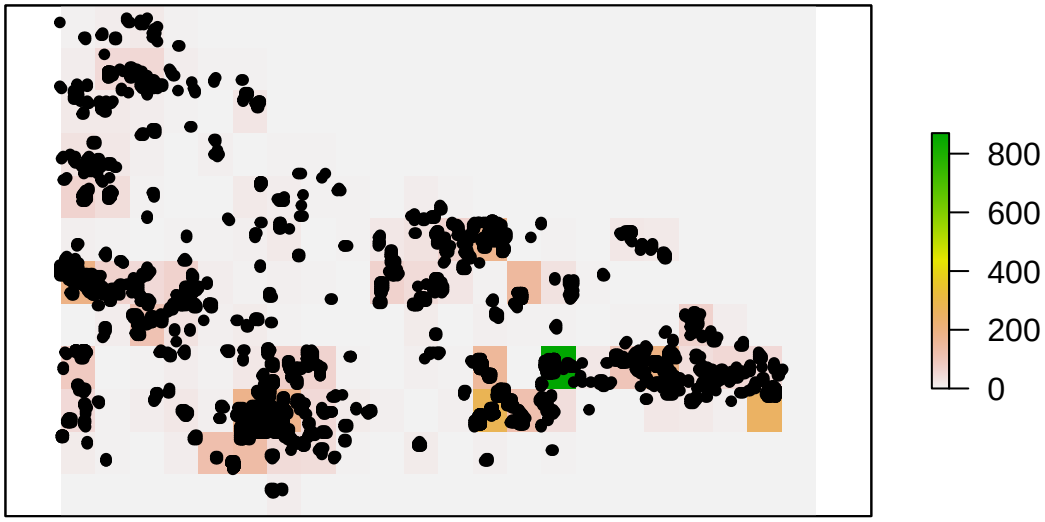
The hotspot data from Himawari-8 (P-Tree System 2020) firstly needs to be selected to those within the boundary of Victoria. To reduce noise, it is then filtered based on the fire power with a recommended threshold of over 100 (irradiance over 100 watts per square metre) (Williamson 2020).

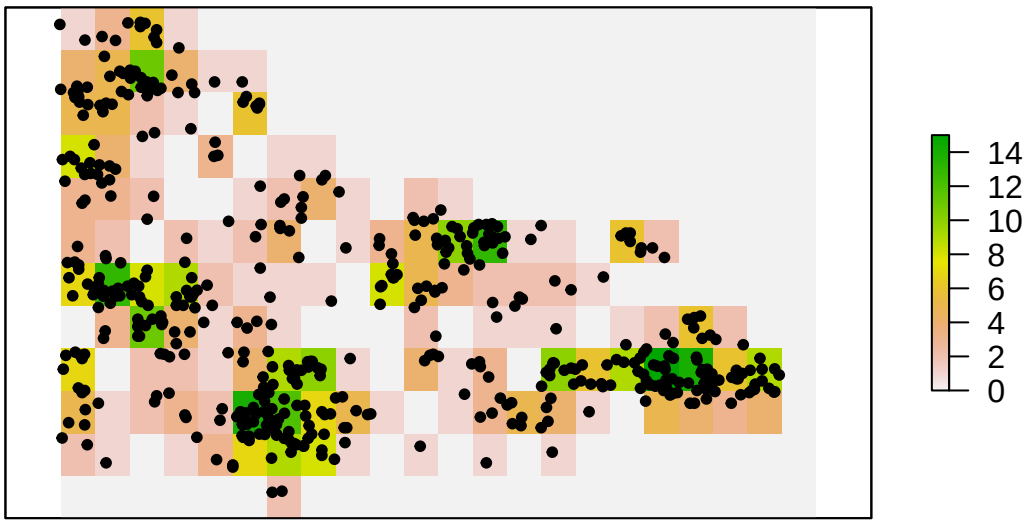
The hotspot data from Himawari-8 needs to be grouped into clusters because some hotspots are branches of an existing bushfire. Therefore, we use spatio-temporal clustering to group the hotspots into clusters. Clustering algorithm used in this project is based on Weihao Li's thesis (citation & reference), inspired by two existing clustering algorithm, Density Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996) and Fire Spread Reconstruction (FSR) (Loboda and Csiszar 2007).

The issue with using DBSCAN for clustering hotspot data is its algorithm assumes the clustering rules work in both directions of a timeline, which does not capture the reality that bushfires evolve over time in one direction. It is not suitable for temporal data. FSR reconstructs bushfire spread well, however it is constructing the clusters sequentially. This means that FSR will consider two fires to be a single fire if they commence at different locations but they overlap. Because of this, FSR does not reflect the real speed of bushfire correctly. Other limitation of FSR is that it lacks detailed consideration of parameter tuning. Weihao Li's clustering algorithm is inspired by these two algorithms, and it can efficiently and robustly cluster hotspot to consider the temporal behavior of bushfires.

There are four steps in the clustering algorithm. The first step is to slice the temporal dimension based on a parameter named ActiveTime. Next, the hotspots are clustered spatially by using a parameter called AjdDist, which reflects the potential distance a fire can spread with respect to the temporal resolution of the data. Hotspots in the same component will be assigned a unique membership id. The third step is to broadcast the clustering results and update the membership label. Lastly, the ignition locations are computed with the earliest observed hotspot indicates the ignition point. If there are several earliest hotspots, the centroid of these points is used.

The clustering algorithm slices the data by its temporal dimension and splits the spatio-temporal clustering tasks into thousands spatial clustering tasks. The result of the clustering is a dataset consisting of four variables – unique identifier, longitude, latitude, and time. The final hotspot data used for this project consists of 2,917 observations from 2016 to 2021. The maps below illustrate the spatial distribution of ignition points on the raw hotspot data and on the clustered hotspot data.


800
600
400
200
0


14
12
10
8
6
4
2
0

Model Building

To prevent overlaps of fire ignitions, only one of the historical or satellite fire ignition data is chosen for the purpose of modelling. Historical fire data might not represent the accurate locations of the bushfire ignitions. Fire might start in a remote location which might be hard to visually access or monitor. Therefore, to choose which fire data to use for our modelling, a comparison of the historical fire data and the satellite data is done. The satellite data used here is the clustered satellite data.

Comparison of Hotspot and Historical Fire Data

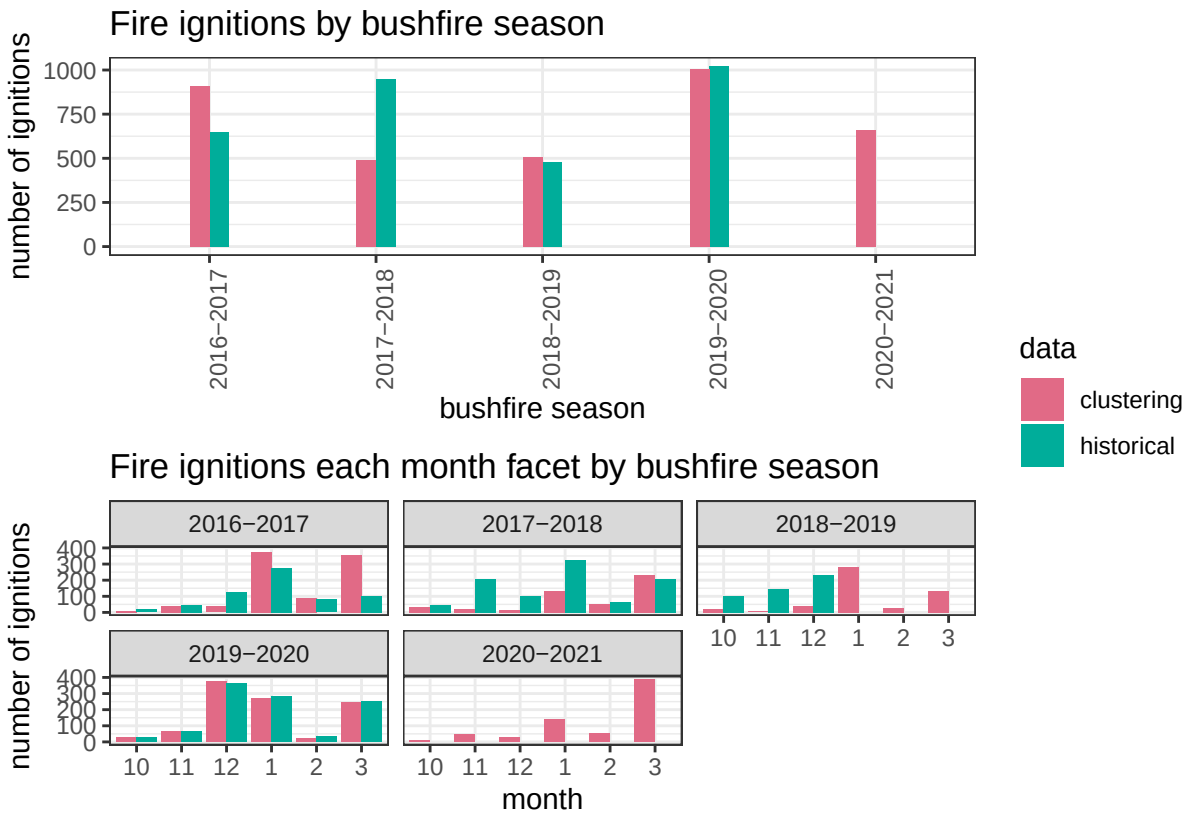


Figure 1: The comparison of the number of fire ignitions between hotspot data and historical data against bushfire seasons(top) and month(bottom).

Number of ignitions in each data set is plotted against bushfire season (top) and month (bottom) in Figure 1 above. The lack of agreeableness both the datasets is apparent. In particular, there were considerably more bushfire ignitions in 2016-2017 bushfire season in the clustered data especially in the months January to March. Meanwhile in 2017-2018, there were more observations in the historical data except for March. This suggests that the number of fires differs on the historical data and hotspot data. One obvious insight here is that the number of observations is not consistent monthly, indicating that month is an important factor in determining the risk of fire.

The differences in the spatial distributions of ignitions in both the data sets were done by computing the difference in number of ignitions (points) contained within in cell i.e. historical ignitions - satellite ignitions per cell. From figure 2, there is obvious difference in how the fire ignitions are scattered around the map.

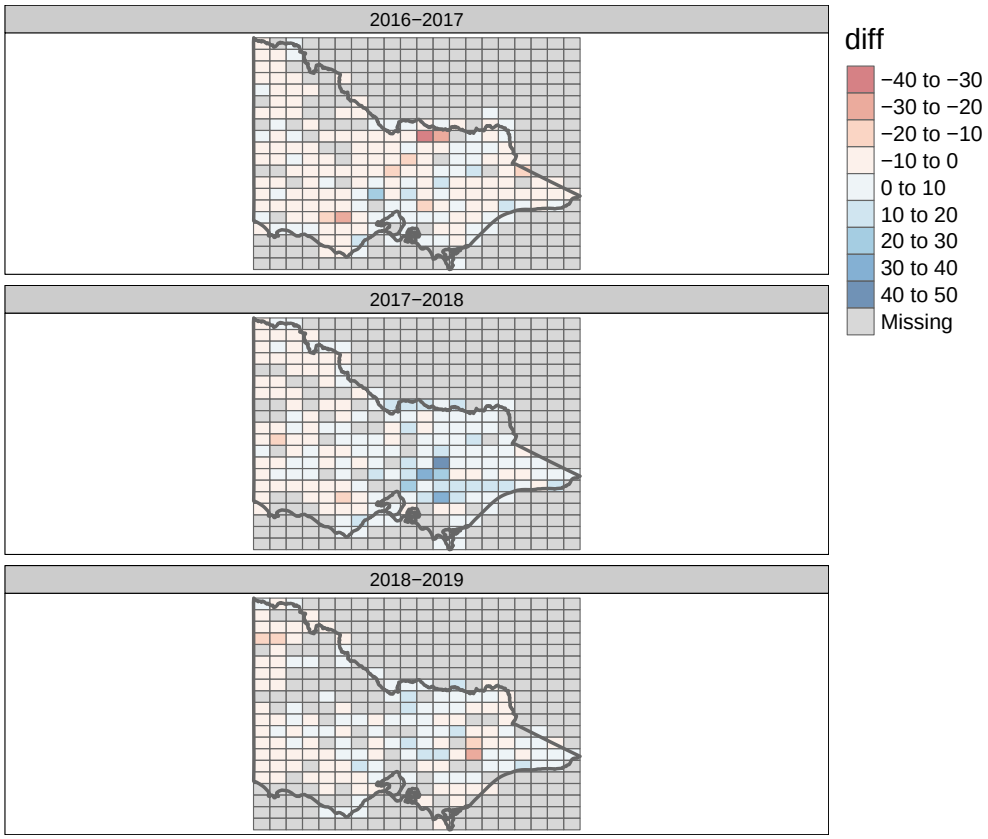


Figure 2: The spatial representation of the differences in the historical data and the satellite data. Differences are computed by substracting the number of fires in the historical data with the number of fire in the satellite data.

Historical data is more concentrated in the map while satellite data tend to be more scattered. There is also strikingly more observations near Melbourne CBD for historical data sets, indicating that there are more fires in areas where there are more people. Meanwhile, the satellite data shows that there are so little fires around the city center which might be due to the filtering of firepower. There are a lot of fires in remote areas which are not captured by the historical data, which might be due to these remote fires not being able to be monitored by people directly. Satellite data, on the other hand, can capture these fires well.

In conclusion, satellite or hotspot data provides more accurate locations and time of bushfire ignitions as satellite data captures fires in remote areas not visually accessible by people. There are a few hypothesis pertaining to the inaccuracies of the historical data. Firstly, it is difficult to distinguish the number of fires one observes with the naked eye. For the same reason, it is less accurate and more subjective to gauge the location of a fire as compared to satellite technology. Equally important is the issues regarding the consistency of fire tower manning at each station. Therefore, the satellite ignition data is used.

Overview

The most common choice for bushfire risk modelling is the generalised additive model (GAM), which is used by Bates, McCaw, and Dowdy (2018) to predict the number of lightning ignitions in Western Australia. Simpler parametric models have also been used to predict the risk of bushfire, including multiple linear regression and generalized logistic regression. The predictors used for these models are environmental variables like weather variables and vegetation types. However, little of the models use hotspot data to predict the risk of bushfire.

For the purpose of modelling, the data is split into training and test set. The training set consists of those data for 2016-2020, while the test set consists of the data for 2020-2021. Splitting data to training and test set is important in model building to check the accuracy of the model and prevent over-fitting or under-fitting. A model might perform well in the training set, but not perform well once it is applied to the test set. The accuracy of our model can be tested using the test set. In this case, the test set consists of the most recent year because the 2021-2022 is more likely to be similar to 2020-2021. The reason different bushfire period is chosen as the test set is because we want to see if our model can provide a good prediction for different periods. We explored some models with our final model being random forest model.

Random Forest Model

Random forest (Breiman 2001) is a supervised learning technique that is constructing hundreds of regression trees by using ensemble learning, combined to a robust prediction model. It consists of a large number of decision trees, which are combined by taking the mean of the predicted y values as the prediction of all trees. A random forest model is built from a bootstrap sample of the data. Each tree at each parent node is constructed from p randomly selected predictors (Matsuki, Kuperman, and Dyke 2016). The trees run in parallel without any interactions amongst them (Bakshi 2020). Random forest is well suited for non-linear relationships, which is suitable for our case as the relationship between environmental variables and the fire risk is non-linear. In this case, we are using random forest for regression problem in where the splitting in each node is based on minimising RSS.

An issue with modelling the risk of bushfire based on environmental variable is the collinearity between variables. For example, relative humidity changes when temperature changes. An evidence of multicollinearity in our data can be seen in 3. Random forest model can handle multicollinearity problem well as it offers protection against the impact of collinearity between predictors (Matsuki, Kuperman, and Dyke 2016). The reason for this is because random forest only considers a subset of predictors at each split, resulting to decorrelated trees. Even though multicollinearity is not a problem for the random forest algorithm itself, it might be a problem for the variable importance. Feature importance might be affected by multicollinearity as the importance of variables with high collinearity will be offset by each other.

Based on the previous analysis, month is an important variable in determining the risk of fire as the risk of bushfire differs monthly. Month is treated as a factor variable in the model. A random forest model with all

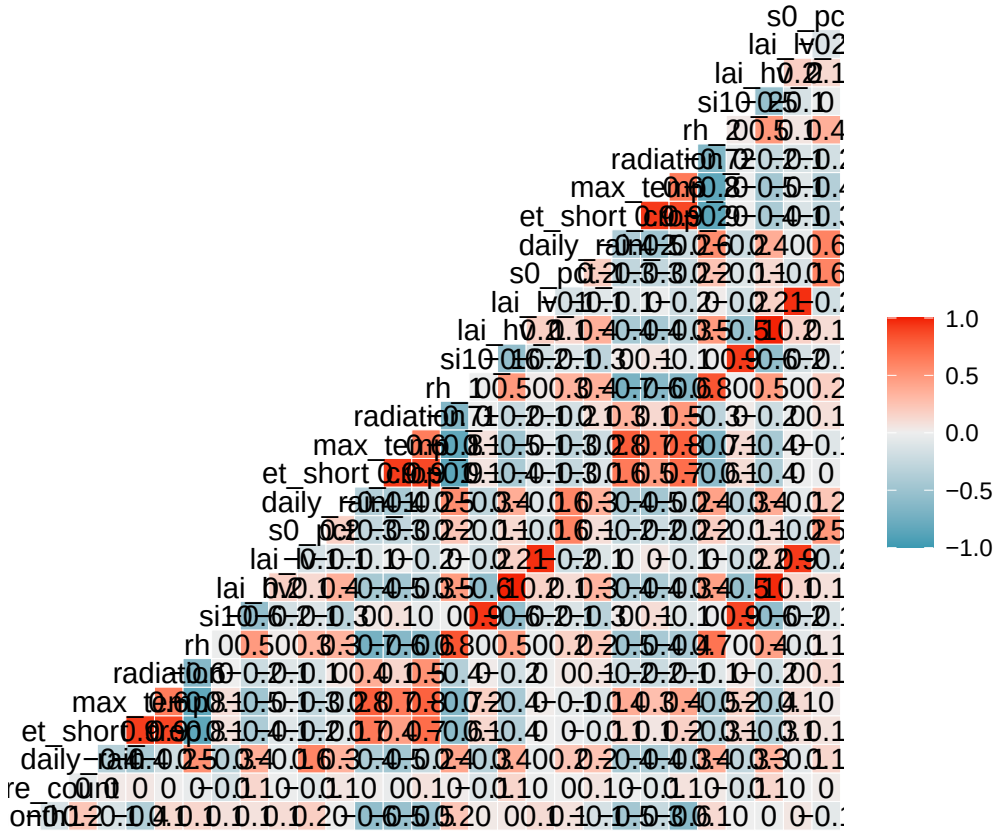


Figure 3: The figure shows the correlation matrix of the variables in our data. There are some variables in our data with high correlations, such as radiation has a really high correlation with et_short_crop (0.9).

Table 2: Summary statistics of the random forest model on test set.

RMSE	MAE	R-Square	MSE
1.480753	0.6660982	0.3960729	2.19263

Table 3: Summary statistics of the random forest model on test set.

RMSE	MAE	R-Square	MSE
1.547397	0.8506499	0.2186106	2.394438

the variables in the dataset as the explanatory variables is fitted. Our first random forest model is a model fitted using `ranger` (Wright and Ziegler 2017) without specifying the number of trees (ntree) and the number of variables used in each tree (mtry). The result is a random forest model by using 500 ntree and 5 mtry. This model has a mean of squared residuals of 2.1878 and R-squared of 0.396 as seen on table 2. Test MSE resulted from this model is 2.379 with an R-squared of 0.23, indicating that random forest is better than a linear regression model. Summary statistics of the test set on the random forest model is on table 3.

Features Selection

We explored if only including the important variables would make the model better. Variable importance on a random forest with multicollinearity in the data could not really be trusted, therefore `lime` (Pedersen and Benesty 2021) is used to understand which variables contribute most to the prediction of bushfire risk. Local Interpretable Model-agnostic Explanation (LIME) can explain the predictions of a regression problem in an interpretable manner, which is done by learning an interpretable model locally around the prediction (Adyatama 2020). The data is split into data with high fire counts and data with low fire counts. Then, 4 observations are sampled from each data, which will tehn be passed to the `explain` function.

Figure 4 illustrates the predictors which explain most of the fire risk prediction for the low fire counts with the x-axis showing the relative magnitude and direction of each predictors. While figure 5 shows those for the high fire counts. It can be seen that forest and month are really important in both cases, followed by surface soil moisture (`s0_pct`), wind speed (`si10`), and `radiation.` The top 15 features are mostly the same for all the observations. These important features are used to fit a random forest model, however this new model has a higher mean of squared residuals. Hence, we decided to proceed with the random forest model with all of the variables.

Parameters tuning

To increase the accuracy of the model, paramaters tuning is done by using the `tuneRanger` package (Probst, Wright, and Boulesteix 2018) to find the best values of parameters for the model. Lowest error is achieved when mtry is equal to 8 and ntree equal to 500. Another random forest is then fitted with these parameters and all of the predictors. This results to a 1.212 mean of squared error and R-squared of 0.399, which is slightly better than the previous model. Applying the model to the test set results to a slightly higher MSE (2.39) and a slightly lower R-squared (0.21), indicating that the tuned model is slightly worse than the previous model. This might be due to the random forest model to overfit the training data, hence why it does not perform better on the test set. Summary statistics of the tuned model on the test set can be seen on table 4 Even though the tuned model performs better on the training set, it performs worse on the test set. This indicates that the tuned model might over-fit the training set, therefore we choose to go with the un-tuned model.

To evaluate the model more, the residuals are calculated and plotted against the predicted values. There are some high residuals, indicating that the model is not predicting some observations well. Figure 6 shows the representation of the actual monthly fire counts and predicted fire counts on the training set. One clear

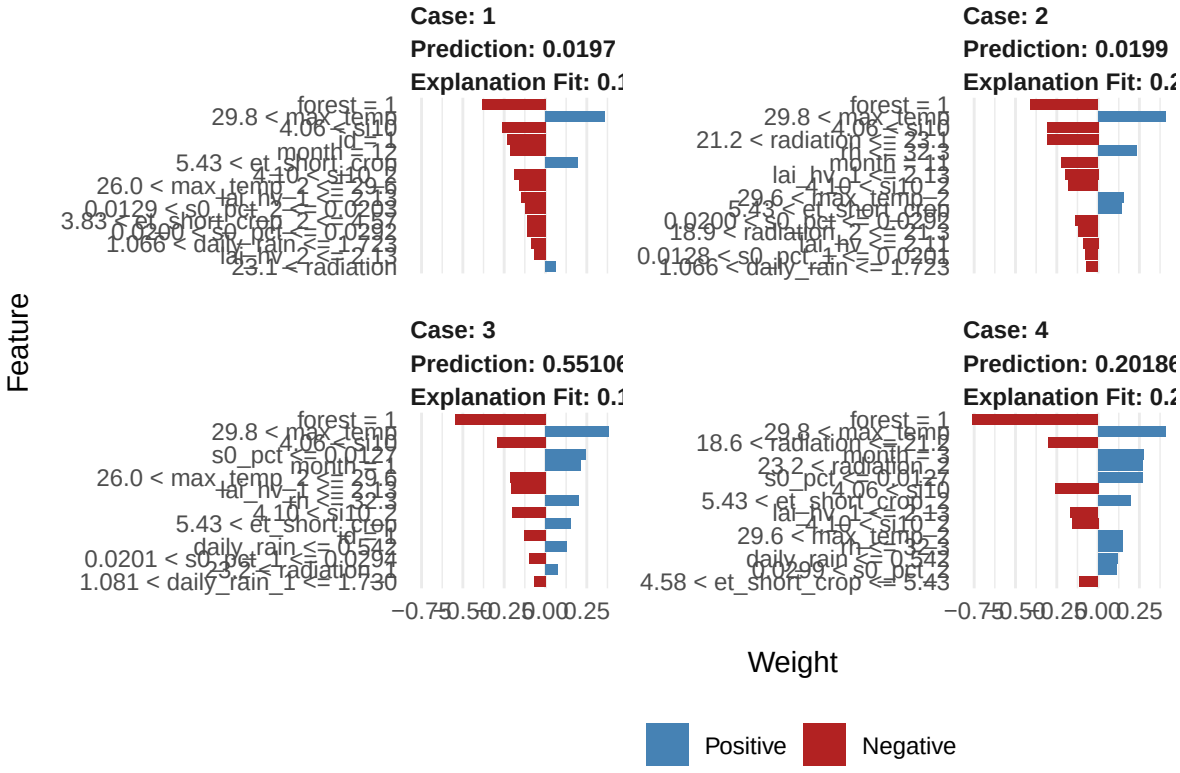


Figure 4: The figure illustrates the variables which are deemed as important in predicting the fir risk for the low fire counts.

Table 4: Summary statistics of the random forest model on test set.

RMSE	MAE	R-Square	MSE
1.56121	0.8865385	0.1972493	2.437377

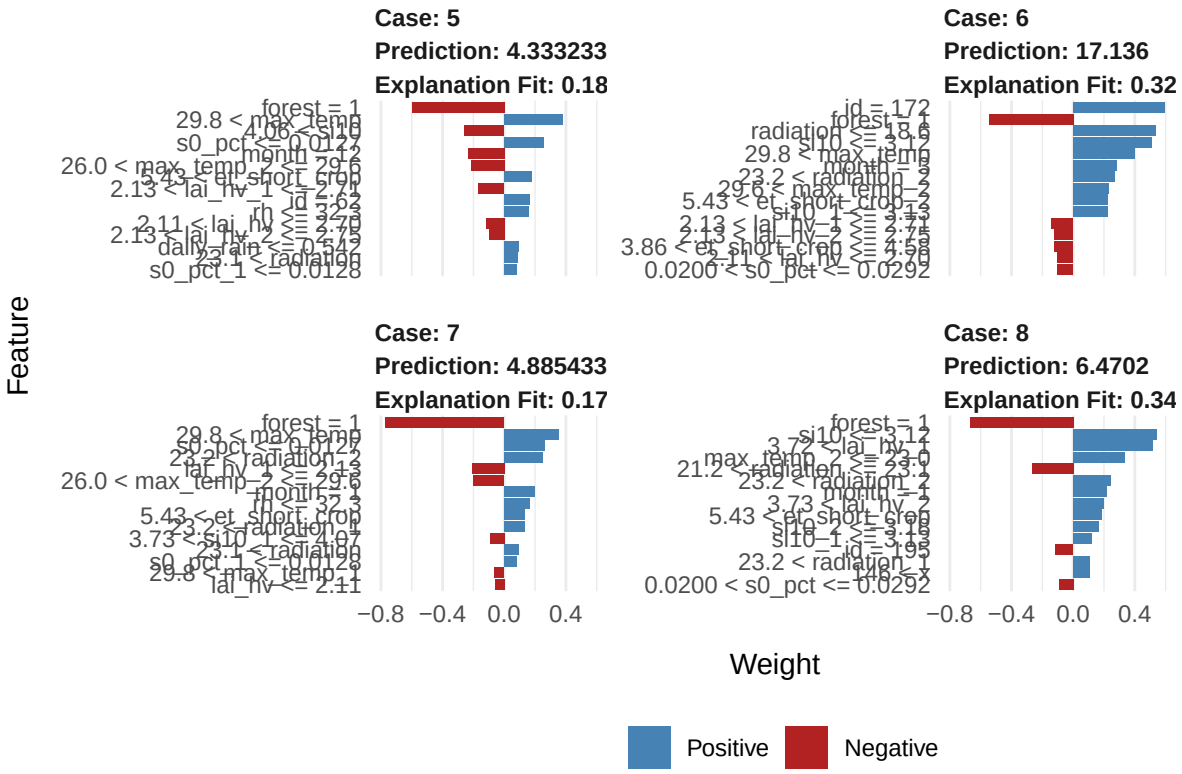


Figure 5: The figure illustrates the variables which are deemed as important in predicting the fir risk for the high fire counts.

thing to notice here is that our model tends to under-predict the fire counts, especially in March. This is due to the many zeros in our response variable which might bring down the prediction. Figure 7 shows the comparison for the test set. The predictions for the test set do not really capture the observed fire counts, which might be due to the different situations in 2020-2021 and the different number of fires every bushfire period. 2020-2021 has a higher number of fires, especially in January and March, which needs to be further explored in future works. Looking at the proportions of fire instead of counts might be better because looking at the pattern on the map, the model predicts those cells with higher observed fire counts to have higher predicted fire counts compared to others.

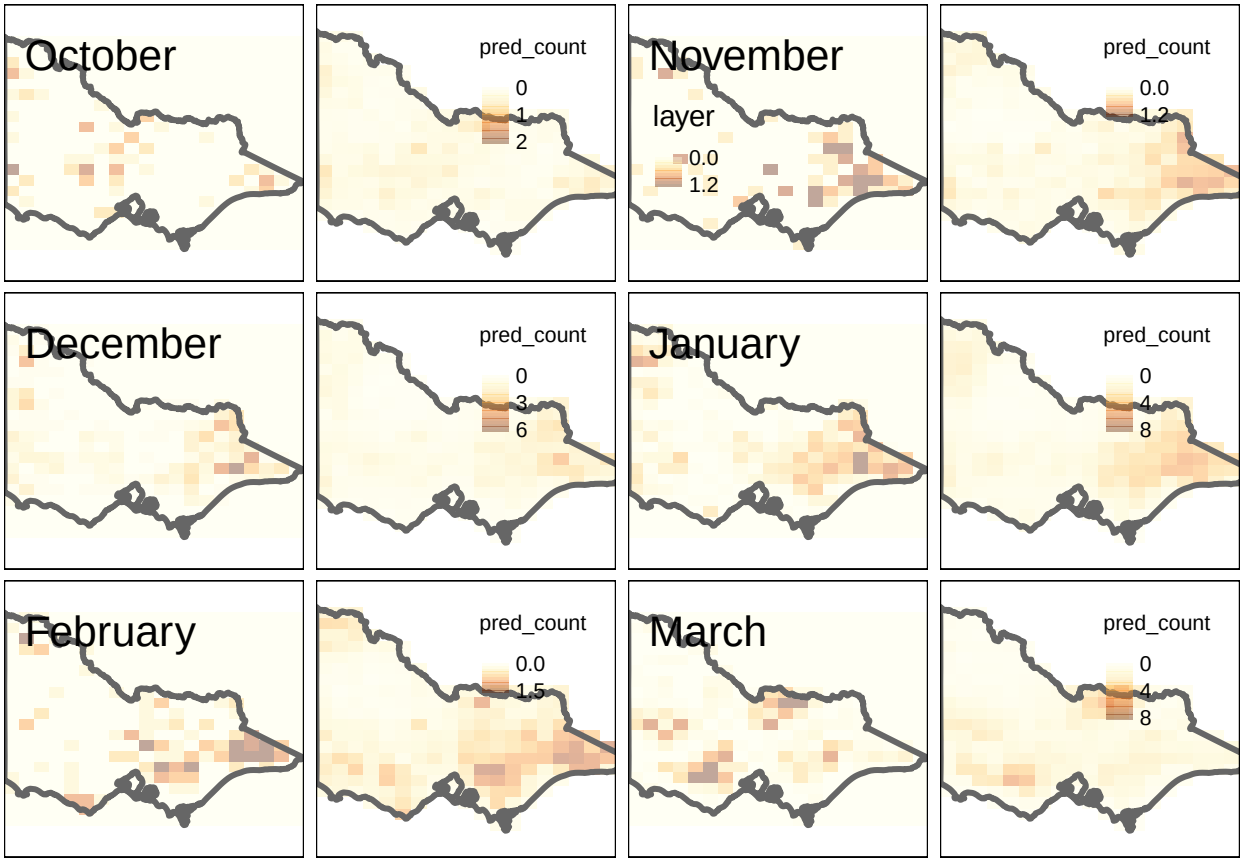


Figure 6: A map of the monthly fire counts of the actual (left) vs the predicted values (right) resulted from the random forest model for the training set.

After exploring some modeling alternatives, random forest is chosen as the final model. The first reason is due to the non-linear relationship between the weather variables and fire counts, a linear model would not be sufficient. Random forest is great for working with non-linear relationships between the response and explanatory variables. Secondly, random forest can handle multicollinearity problem well. Other than that, random forest works well on large datasets and standardising the data is not required as it uses a rule-based approach. For these reasons, random forest outperforms other modelling techniques. A limitation of random forest model is it cannot work with previously unseen data because the random forest regressor is unable to discover trends is not able to discover trends for it to extrapolate outside the training data.

Another limitation of the model is that there are a lot of zeros in the data, which might result to the under-prediction of fire counts. There are some high residuals which might be due to the huge amount of zeros in the data, which might affect the predictions for those with similar weather conditions but with a high fire counts. Additionally, the random forest uses variables which represent weather conditions, which might not account for accident-based fires. We might also be missing some variables that are useful for predicting the

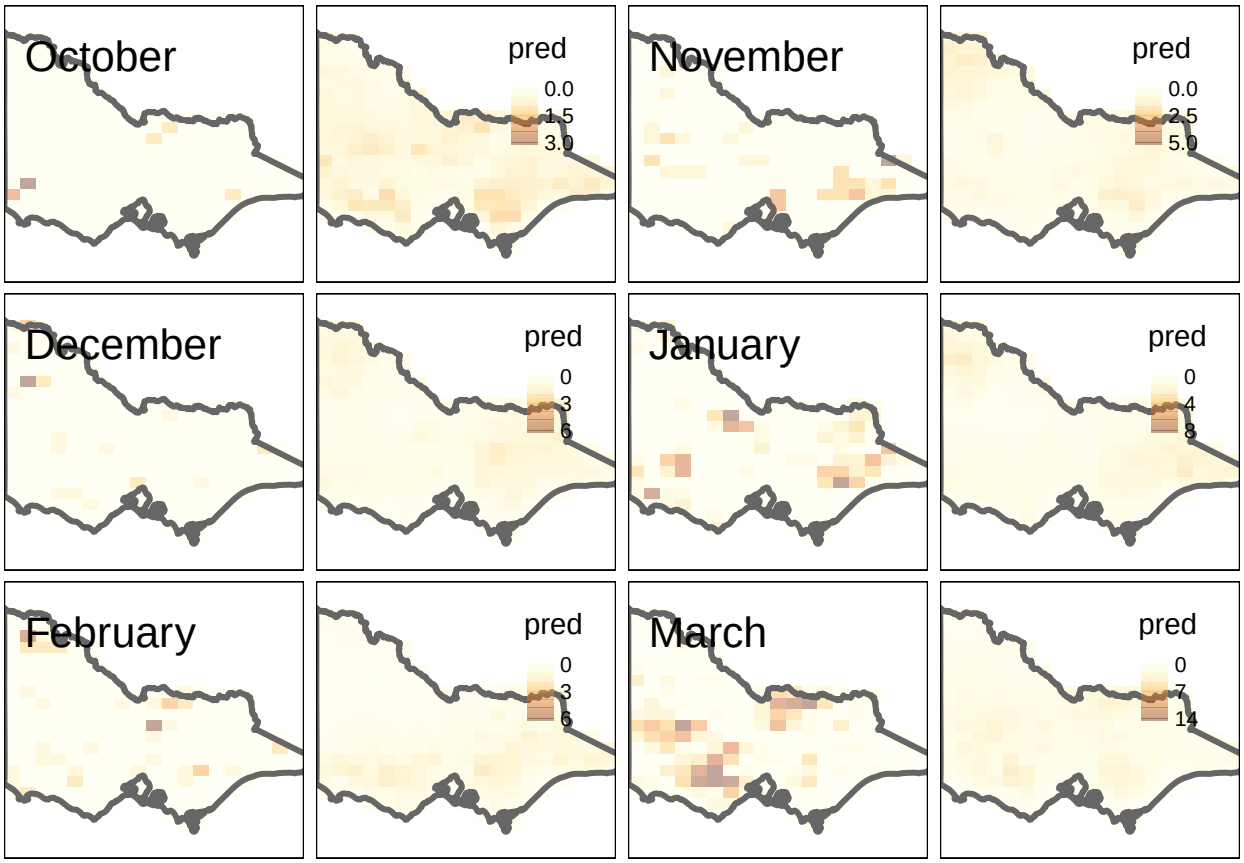


Figure 7: A map of the monthly fire counts of the actual (left) vs the predicted values (right) resulted from the random forest model for the test set.

Table 5: Variables with largest Variance Inflation Factors(VIF)

variable	VIF
et_short_crop_1	188.94
et_short_crop_2	181.49
et_short_crop	165.06
lai_hv_1	112.72
max_temp_1	68.31
max_temp_2	67.70

risk of fire, like distance to road. This could be a consideration for our future work to include more variables.

Choosing the Random Forests and Model Alternatives

A lasso regression was attempted to parse out important variables. Depending on the penalty (λ) term imposed on the least square optimisation problem has the effect of shrinking (unimportant) variables to zero/very small coefficient. While lasso regression can be a good regularisation technique to sieve out important variables, it is inadequate for modelling in our scenario. This is mainly due to the structure of our variables. The climate variables are highly correlated as one might expect. For example, with high `daily_rain` one can expect a lower `max_temp` or a variable with its lag variables are highly correlated. These relations exist across many of our variables thus indicating high multicollinearity

To demonstrate, we fit a simple linear model and compute the Variance Inflation Factor (VIF). The 6 variables with the highest VIF is shown in 5. A general rule of thumb is that if VIF > 10 then multicollinearity is high Perlato (2020). Lasso's objective function provides unstable solutions in prescence of collinear features or features with very similar information. Schreiber-Gregory (2018)

Therefore, due to the inherent interplay between the variables, lasso regression is not used for modelling. For the same reason, generalised linear models(GLMs) attempted such as Poisson regression were ruled out.

Since ensemble learning with random forest process can deal with these issues, it was deemed suitable in our modelling scenario. Random forests had been used in several studies to model bushfire susceptibility. (Gigović et al. (2019))

Shiny Web application (User-Interface)

The outputs of data processing and bushfire predictions are put together in an app. This app is an extension to an previously made `RShiny` app (publicly available at Monash University's Department of Econometrics and Business Statistics website. The previous app allow users to visualise bushfires in Victoria at a period of time of their choosing. Outputs are based on the clustering results. Also, a model predicts the cause of bushfire; given a bushfire ignition. Our complementary goal (as per Research grant) is to develop risk model to visualise and monitor potential fire ignitions and track fires from satellite hotspots, in real time, for 2021-2022 Victorian bushfire season.

`RShiny` allows for elegant interactive exploratory plots and updates in real-time for swift decision making. It is also highly customisable to enhance users' interface. Furthermore, it supplies convenience function to assemble the html website in `R` and is easy to deploy.

The current project builds on the "Fire Risk" Tab. On top of the historical bushfire information and bushfire `cause` prediction, this complementary addition serves to inform users with historical information risk and up-to-date predictions around bushfire risk. We maintained a similar design for the structure of the app-keeping separate tabs for historical information and predictions. This is to cater for different users who might be using the app. Fire personnels might be interested in the weather and landscape variables creating the conducive fire environment. Plots and data are made downloadable. While users would be interested in bushfire predictions on the go while planning a trip.

The current state of the historical bushfire risk Information tab is shown above. in Figure 8. User interactivity include choosing the bushfire seasons to overlay ignition points. User can also choose the months to compute number of bushfires from. Upon clicking a grid cell, the data table and plots are produced. The table provides information of the different weather and landscape variables. Plots show the historically the number of bushfires and values of each variable (max temperature in this case) which might be of interest to users. These are made downloadable for interested users.

Additionally, a rough sketch of bushfire fire risk map in the app is shown above (Figure 9. It incorporates the random forest model to make predictions of the bushfire risk in a particular cell in real time. Users will be able to toggle the various variables and make forecasts under different scenarios. A detail is that not all cells shown will increase the same amount but instead at a relative rate (by %) to account for the fact that different places have different weather conditions.

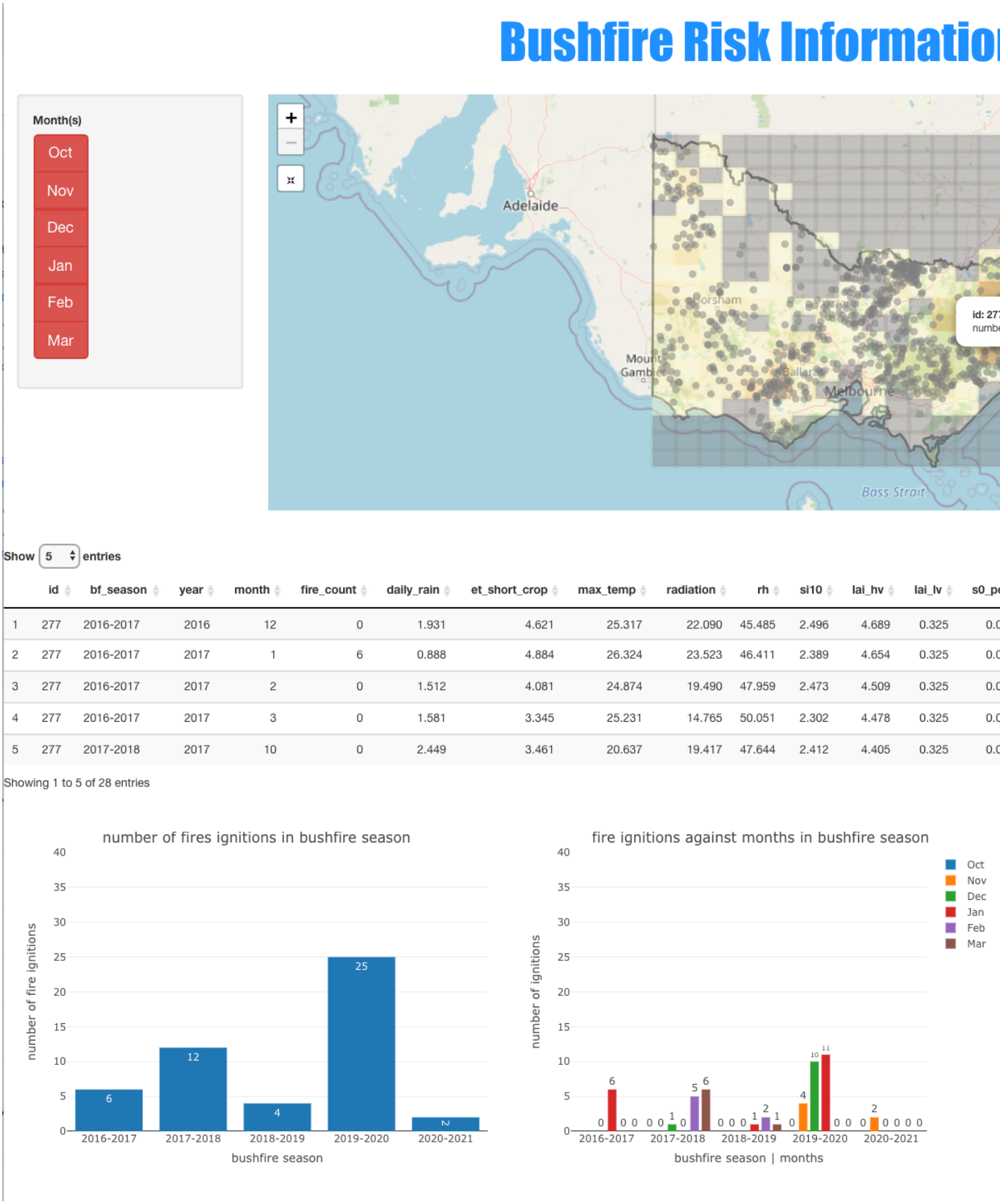


Figure 8: Bushfire Risk Information tab

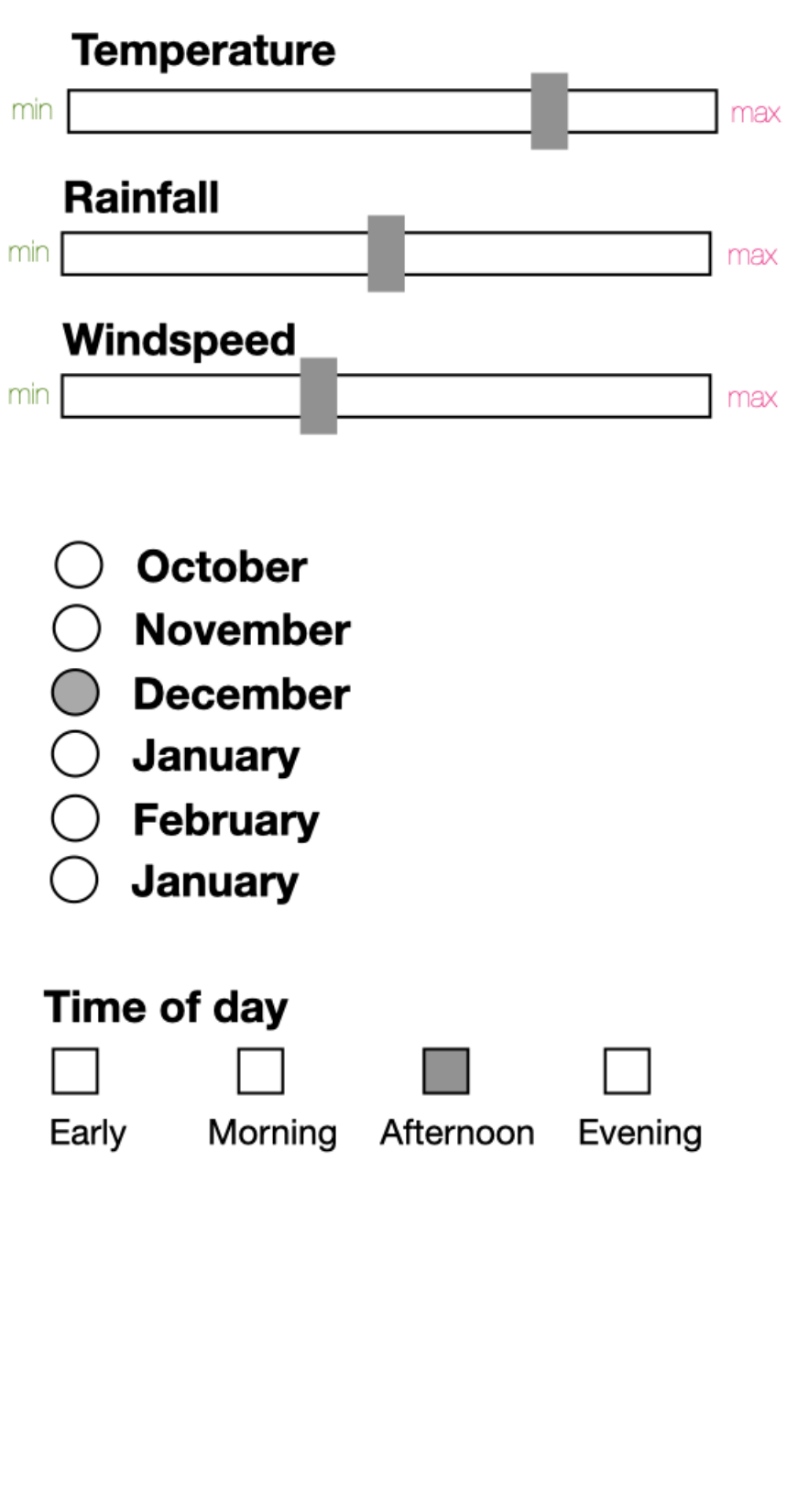


Figure 9: rough plan for bushfire risk map

Summary and future works

To assess bushfire risk in Victoria, A Shiny web application is developed to help monitor potential fire ignitions using the satellite data. For predicting the risk of fire, different modelling alternatives were explored. This will be used as a prototype to predict the bushfire risk 2021-2022 bushfire season based on weather and landscape conditions. The best model was deemed to be a random forest model which appropriately incorporates various weather and climate variables. Random forest model outperforms other modelling techniques because random forest is great to capture the non-linear relationship between the response and explanatory variables. Furthermore, random forest handles the multicollinearity structure in the data well. The model gives a training mean squared error of 1.216 and R-squared of 0.396. However, applying the model to the test set results to a slightly lower R-squared and higher MSE. The reason for this might be because the model tends to over-fit the training set, resulting to less accurate predictions for the test set. Additionally, our model tends to under-predict due to a lot of zeros in our data. It is also important to note that the response (fire ignitions) has been erratic going from one year to another. Even though random forest does not provide a perfect prediction, it provides a pretty accurate predictions, especially if looking at the relative counts in proportion-wise. The model predicts cell which have higher observed fires to have higher predicted fires in comparison to other cells.

Even though more than 80% of fires occur by lightning, another limitation of our model is it uses variables which represent weather conditions, which might not accurately account for accident-based fires. Some variables that might be useful for predicting the risk of fire, like distance to road are still attempted to be factored in. This could be a consideration for our future work to include more variables. The bushfire risk tab in the application that incorporates the predictions is being worked on.

Appendix

Data Sources

`SILO`

`SILO` is a database of Australian climate data from 1889 to current hosted by Queensland Department of Environment and Science (DES). It provides daily metereological datasets for a range of climate variables. The datasets are constructed from observational data obtained from the Bureau of Meteorology (BoM) and other suppliers. *more information on how data is constructed here.

`ERA5 renalysis data`

ERA5 is the fifth generation European Centre for Medium-Range Weather Forecasts (ECMWF) reanlysis for global climate and weather. It also supplies data from 1979 onwards for a whole range of climate and weather data.

Reanalysis combines model data and observations across the world into a globally complete and consistent dataset using the laws of physics. *link

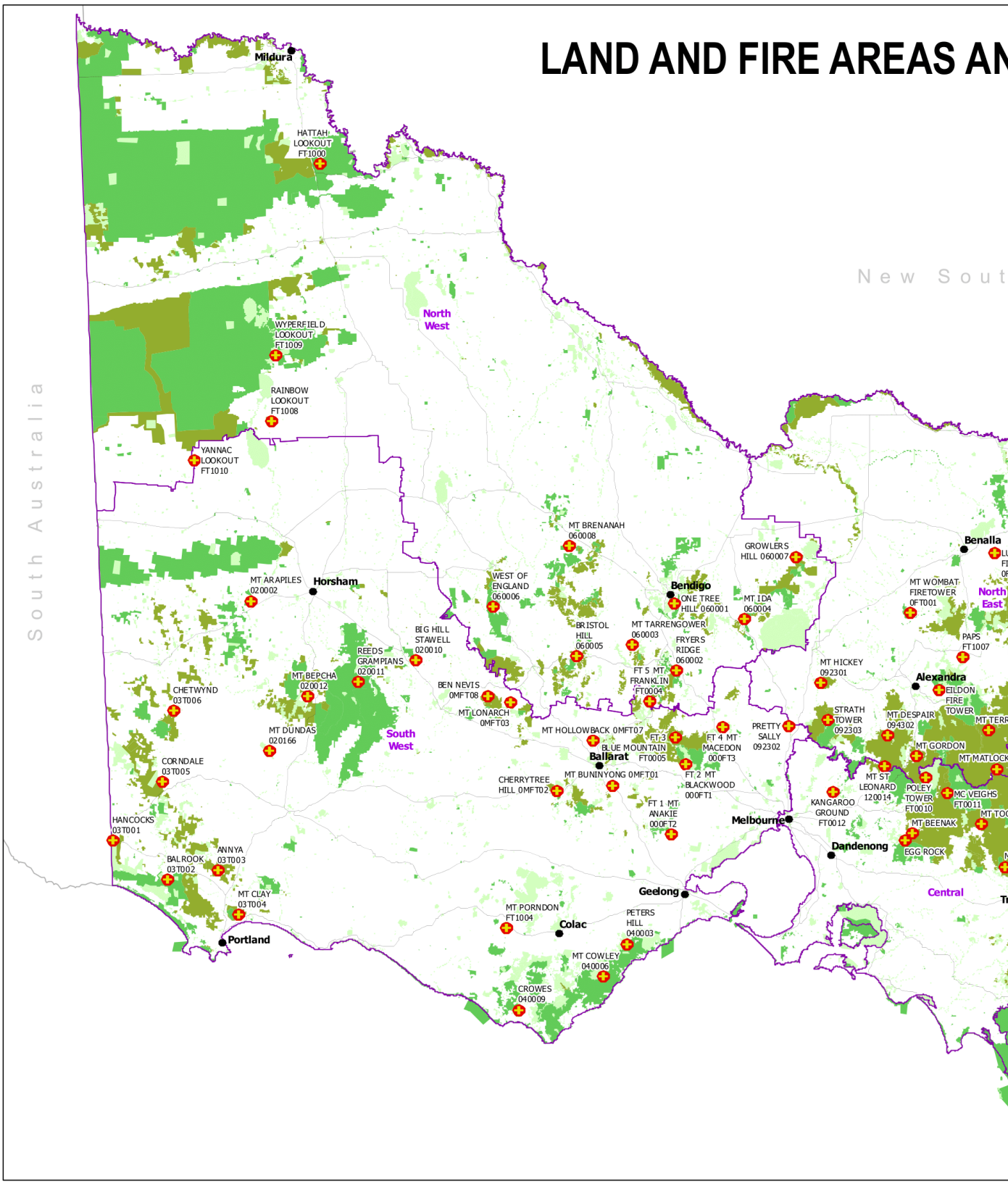
`Victoria forest data`

Victoria forest vector data delineates Victorian forest types and its attributes referenced through a forest type code. Data can be downloaded from dicover.data.vic.gov.au.

`Surface soil moisture`

BoM's Australian Water Resources Landscale model provides an array of estimates depicting Austrlia's water balance including surface soil moisture, evapotranspiration among others.

Fire towers in Victoria



Source

Adyatama, Arga. 2020. *Interpreting Black Box Regression Model with Lime*. https://algotech.netlify.app/blog/interpreting-black-box-regression-model-with-lime/.

Australian Government Geoscience Australia. 2021. *Bushfire*. https://www.ga.gov.au/scientific-topics/community-safety/bushfire.

Bakshi, Chaya. 2020. *Random Forest Regression*. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84.

Biddle, Nicholas, Ben Edwards, Diane Herz, and Toni Makkai. 2020. "Nearly 80." https://theconversation.com/nearly-80-of-australians-affected-in-some-way-by-the-bushfires-new-survey-shows-131672.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45: 5–32.

Cardille, Jeffrey A, and Stephen J Ventura. 2001. "Occurrence of Wildfire in the Northern Great Lakes Region: Effects of Land Cover and Land Ownership Assessed at Multiple Scales." *International Journal of Wildland Fire* 10 (2): 145–54.

Cook, Diane. 2020. "Open Data Shows Lightning, Not Arson, Was the Likely Cause of Most Victorian Bushfires Last Summer." https://theconversation.com/open-data-shows-lightning-not-arson-was-the-likely-cause-of-most-victorian-bushfires-last-summer-151912.

Deb, Proloy, Hamid Moradkhani, Peyman Abbaszadeh, Anthony S. Kiem, Johanna Engström, David Keellings, and Ashish Sharma. 2020. "Causes of the Widespread 2019–2020 Australian Bushfire Season." *Earth's Future* 8 (11): e2020EF001671. https://doi.org/https://doi.org/10.1029/2020EF001671.

Dowdy, Andrew J, Graham A Mills, Klara Finkele, and William de Groot. 2009. "Australian Fire Weather as Represented by the McArthur Forest Fire Danger Index and the Canadian Forest Fire Weather Index." *Centre for Australian Weather and Climate Research Tech. Rep* 10: 91.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and others. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Kdd*, 96:226–31. 34.

Gigović, Ljubomir, Hamid Reza Pourghasemi, Siniša Drobnjak, and Shibiao Bai. 2019. "Testing a New Ensemble Model Based on SVM and Random Forest in Forest Fire Susceptibility Assessment and Its Mapping in Serbia's Tara National Park." *Forests* 10 (5): 408.

Li, Weihao. 2020. *Using Remote Sensing Data to Understand Fire Ignitions in Victoria During the 2019-2020 Australian Bushfire Season*. url.

Loboda, TV, and IA 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): 258–73.

Matsuki, Kazunaga, Victor Kuperman, and Julie A. Van Dyke. 2016. "The Random Forests Statistical Technique: An Examination of Its Value for the Study of Reading." *Scientific Studies of Reading* 20 (1): 20–33. https://doi.org/10.1080/10888438.2015.1107073.

Merton Council. 2013. *Fire Safety Risk Assessment*. https://www.merton.gov.uk/assets/Documents/www2/fire_safety_risk_assessment_-_june_2013.pdf.

Parisien, Marc-André, VG Kafka, KG Hirsch, JB Todd, SG Lavoie, PD Maczek, and others. 2005. "Mapping Wildfire Susceptibility with the BURN-P3 Simulation Model."

Pedersen, Thomas Lin, and Michaël Benesty. 2021. *Lime: Local Interpretable Model-Agnostic Explanations*. https://CRAN.R-project.org/package=lime.

Perlato, Andrea. 2020. "Deal Multicollinearity with LASSO Regression." https://www.andreaperlato.com/mlpost/deal-multicollinearity-with-lasso-regression/.

Probst, Philipp, Marvin Wright, and Anne-Laure Boulesteix. 2018. "Hyperparameters and Tuning Strategies for Random Forest." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. https://doi.org/10.1002/widm.1301.

P-Tree System. 2020. "JAXA Himawari Monitor - User's Guide." https://www.eorc.jaxa.jp/ptree/userguide.html.

Richards, Lisa, Nigel Brew, and L Smith. 2020. *2019-20 Australian Bushfires-Frequently Asked Questions: A Quick Guide*.

Schreiber-Gregory, DN. 2018. "Regulation Techniques for Multicollinearity: Lasso, Ridge, and Elastic Nets." In *SAS Conference Proceedings: Western Users of SAS Software 2018*, 1–23.

Victoria, Forest Fire Management. 2021. "Past Bushfires a Chronology of Major Bushfires in Victoria from 2013 Back to 1851." https://www.ffm.vic.gov.au/history-and-incidents/past-bushfires.

Williamson, Grant. 2020. "Example code to generate animation frames of Himawari-8 hotspots." https://gist.github.com/ozjimbob/80254988922140fec4c06e3a43d069a6.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1): 1–17. https://doi.org/10.18637/jss.v077.i01.