Seeing through the smoke: Evaluating the respiratory health effects of the 2019-2020 Victorian bushfires

A thesis submitted for the degree of

Bachelor of Commerce (Honours)

by

Callum Shaw

28893786

Department of Econometrics and Business Statistics

Monash University

Australia

November 2020

Contents

Abstract

Respiratory illness from exposure to bushfire smoke was a significant driver of harm to public health during the 2019-2020 bushfire season. Unfortunately, it has been challenging for researchers to explore the effects of bushfire smoke precisely because of the sparsity of air quality measurement stations.

We developed a two-stage approach to estimate the health effects of bushfire related particulate matter between 11/11/2019 and 25/01/2020. First, satellite data was used in a linear mixed-effects model to estimate bushfire-related particulate matter concentrations across Victoria. Second, a varying-parameter negative binomial GLM was used to estimate the causal effects of bushfire smoke.

Model simulation suggests that bushfire smoke caused 580 (95% CI: 69 - 1092) emergency room presentations. Simulation results also indicate that it is mostly working-age adults in inner-city areas who were harmed by bushfire smoke. These results highlight a need to reform workplace safety laws to minimise work-related outdoor activity during future crises.

Acknowledgements

I would like to thank my supervisors Dianne Cook and Xueyan Zhao for their generous time and support across this year. It would not have been possible to complete this project without their help. I would also like to thank Stephen Duckett and William Mackey at the Grattan Institute for help accessing data and for their advice. Finally, I have relied heavily on financial support from Monash University and the Reserve Bank of Australia this year - I would like to acknowledge both organisations for their support.

This report uses open-source software developed by R Core Team (2020), Mackey (2020), Pebesma (2018), Henry (2020), Bengtsson (2020), Wang, Cook, and Hyndman (2020), Wickham et al. (2019), Grolemund and Wickham (2011), Bivand, Pebesma, and Gomez-Rubio (2013), Gräler, Pebesma, and Heuvelink (2016), and Csárdi and FitzJohn (2019). I thank the authors of these software.

Chapter 1

Introduction

1.1 Motivation and research aims

The 2019-2020 bushfire season was one of the worst on record. High temperatures across 2019 led vegetation to dry out, which significantly amplified the severity of fires across Australia. The direct impact of these bushfires was huge. Estimates suggest that the fires destroyed over 10,000 buildings, including 3,500 homes. Tragically, 34 people lost their lives due to the direct health effects of these fires.

While these figures are shocking, the indirect consequences of the fires were likely even more severe and substantial. In particular, there is a wide range of adverse health effects that can be caused by exposure to bushfire smoke. Bushfire smoke is associated with large increases in fine particulate matter (PM2.5)[1] which can travel deep into the lungs and cause a range of respiratory problems including asthma and chronic obstructive pulmonary disease. These health effects were, in aggregate, almost surely significantly more harmful than the direct health impact of the fires.

Policymakers need to be able to quantify the impact of bushfire smoke on public health to ensure that they can effectively respond to future bushfire crises. Critically, they must understand how the effects of bushfire smoke differ across demographic and geographic

[1] PM2.5 refers to particulate matter finer than 2.5 microns.

groups so that public safety policies can be designed more effectively and hospitals can be better prepared for spikes in emergency room visits.

Estimating the causal effects of bushfires in Victoria on respiratory health has been challenging, however. The distribution of Victorian air quality measurement stations is relatively sparse and uneven. While there is adequate coverage in Melbourne and Eastern Victoria, there is little to no coverage across the rest of the state (see Figure 1.1). This makes it difficult for researchers to assess the population's exposure to PM2.5 accurately.

Several other factors make modelling the relationship between bushfires and respiratory morbidity a challenge. First, hospital presentations and air quality are recorded at different locations. This means that it is necessary to develop some spatial matching, averaging or interpolation method to create a usable dataset. Second, there are sources of air pollution that are not related to bushfires - which makes it necessary to separate bushfire related air pollution from background pollution. Third, there are other variables such as humidity which can impact respiratory health. It is necessary to control for these variables. Finally, Victorian hospital data is a longitudinal overdispersed count dataset. Modelling this dataset in a way that controls for both unobserved effects and the heterogeneous effects of our independent variables is non-trivial.

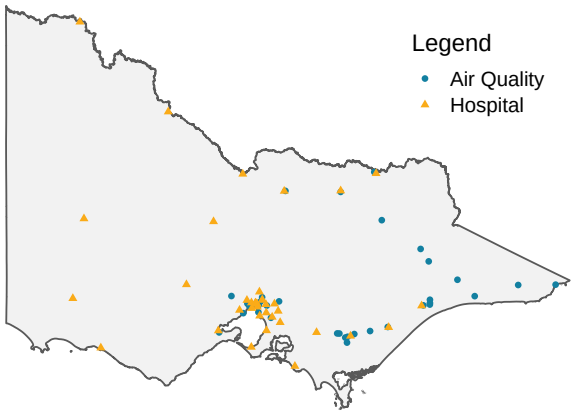

Figure 1.1: *Locations of air quality measurement stations and hospitals in Victoria. Coverage is only adequate in Melbourne and Eastern Victoria.*

The objectives of this thesis are:

1) To use both the satellite imaging data for Aerosol Optical Depth and the observed ground PM2.5 data to derive a complete high-resolution grid of PM2.5 data for the

state of Victoria for every day between November 11th 2019 and January 25th 2020, using a linear mixed-effect model with time-varying coefficients;

2) To develop an approach to further separate PM2.5 into a 'bushfire' and a 'background' component;

3) To estimate a varying-coefficient fixed-effect Negative Binomial model to quantify the demographically and spatially heterogeneous causal effects of bushfires on hospital respiratory emergency room presentations; and

4) To use the estimated model to simulate several policy-relevant scenarios to explore how the bushfires impacted different demographic and geographic groups.

Our modelling suggests that there were roughly 580 (95% CI: 69 - 1092) respiratory ED presentations between November 11th 2019 and January 25th 2020 that can be attributed to bushfire related PM2.5. We also find that adults aged 18-64 and hospitals in inner-city neighbourhoods were most impacted by bushfire related PM2.5. These findings suggest that working-age adults exhibited more risky behaviour during extreme smoke events. This is likely due to their increased need to travel for work and other activities.

Policymakers should consider reviewing workplace safety regulations. Potential policy changes could include encouraging office staff to work from home during extreme pollution events and closing high-risk workplaces such as construction and outdoor retail. Health officials should also assess the capacity of inner-city hospitals to receive a surge in respiratory presentations.

1.2 Literature review

We centred our review of the existing literature around three main areas: (a) identifying the health effects of bushfire smoke, (b) modelling the health effects of air pollution, and (c) addressing the limitations of our air quality measurement system.

1.2.1 Identifying the health effects of bushfire smoke

A meta-analysis of the physical health effects of non-occupational exposure to wildfire smoke by Liu et al. (2014) found that wildfire smoke was consistently associated with an increase in respiratory morbidity - with 43 out of 45 studies of respiratory diseases reporting a significant adverse association with wildfire smoke. The relationship was not as consistent for cardiovascular morbidity and non-accidental mortality, however - with only six out of fourteen and nine out of thirteen studies reporting a statistically significant adverse relationship. A similar review by Reid et al. (2016) found evidence that exposure to bushfire smoke was associated with an increased risk of respiratory infections and all-cause mortality.

A more detailed meta-analysis by Dennekamp and Abramson (2011) found that most of the harm from bushfire smoke-related particulate matter appears to come from increased morbidity of asthma and chronic obstructive pulmonary disease (COPD). The authors noted that the literature which decomposed particulate matter into 'bushfire' and 'background' components was relatively sparse and that more work in this area could help policymakers better respond to future crises.

1.2.2 Modelling the health effects of air pollution

Poisson GAMs

Studies of the health effects of air pollution tend to rely heavily on Poisson Generalised Additive Models (GAMs) (Simpson et al. (2000); Tham et al. (2009); Morgan et al. (2010)). These models are attractive because they (a) treat the dependent variable as 'count' data, and (b) can account for more complex relationships between the dependent, independent and confounding variables. These models typically include a vector of confounding variables (e.g. temperature, humidity) and a smoothed function of time (to control for unobserved confounding variables) (Peng and Dominici (2008)). Schwartz (1994) suggested that LOESS represent this smoothed function. Other authors have suggested using smoothing splines, penalised splines, and natural splines (Dominici, McDermott, and Hastie (2004); Ramsay, Burnett, and Krewski (2003); Touloumi et al. (2004)).

Poisson GAMs may be suitable for long term studies of the health effects of anthropogenic air pollution, but they are likely inappropriate for studying the health effects of bushfires. Bushfires are likely to lead to sizeable short term spikes in emergency room presentations, which will violate the strict equidispersion conditions of the Poisson distribution. Replacing the Poisson distribution with the more flexible negative binomial distribution is likely to be necessary for our application.

Semi-parametric methods such as smoothing splines are generally used to account for annual seasonality, whereas our study only uses data from a single summer. We should therefore be able to develop a fully parametric model structure.

Hierarchical modelling

Another notable feature of these studies is the hierarchical structure of the data. Most studies are multi-site, which means that the data are indexed both by time and by site/location. It is highly likely that the effects of air pollution and other meteorological variables vary by site. For example, hospital presentations at a children's hospital may be impacted differently by a change in air quality than a rural hospital that caters mainly to adults.

To deal with this, it is not uncommon for papers to adopt model structures that allow the relationship between PM2.5 and health outcomes to vary by region or hospital. In studies with a long time period, this is typically achieved by estimating a separate model for each hospital/region (Ostro et al. (2007)). These structures may be inappropriate in studies with a shorter time period; however, so a fixed effects or random effects specification may be preferable.

Other considerations

The models we have described above do not distinguish between different sources of pollution, which makes it difficult to isolate the effects of bushfire smoke. Morgan et al. (2010) handles this problem by decomposing particulate matter into 'bushfire PM10' and 'background PM10'.[2] On 'bushfire' days (when PM10 is above the 99th percentile), they define background PM10 as the 30-day moving average of PM10, and bushfire PM10

[2]This paper uses PM10 rather than PM2.5 data.

as the difference between the moving-average and the observed value. On other days, background PM10 and overall PM10 are the same.

These models also assume that the user has access to air quality measurement data at the same location where individuals are exposed to air quality. In practice, we only have access to a small set of point air quality estimates at measurement stations across a region. Fitting a model requires the user to predict particulate matter concentrations at some location close to a hospital.

1.2.3 Addressing the limitations of air quality measurement systems

Gaps in the coverage of air quality measurement stations have made it difficult to accurately measure the population's exposure to fine particulate matter (Rappold et al. (2011)). To address this problem, researchers have developed a series of modelling approaches that translate satellite-derived measurements (namely Aerosol Optical Depth) into estimates of PM2.5. Aerosol Optical Depth is a satellite-derived estimate of the amount of light that is absorbed by fine particulate matter in the earth's atmosphere. This makes it an excellent proxy for PM2.5 since higher levels of light absorption tend to correlate with higher levels of particulate matter.

Chudnovsky et al. (2012) uses a mixed-effect model with day-specific intercepts and slopes to estimate the concentration of PM2.5 in New England, USA. They find that this approach effectively accounts for time-varying heterogeneity and leads to a good model fit. Similar studies in India (Unnithan and Gnanappazham (2020)) and China (Zhang et al. (2019)) have found that including day-specific random effects significantly improved model fit.

1.2.4 Literature summary

While many researchers have made important contributions the environmental epidemiological literature by developing methods to either (a) estimate the effects of particulate matter on respiratory health outcomes or (b) produce higher resolution PM2.5 estimates using satellite data, these two kinds of methods have rarely been applied together.[3] In

[3]See Kloog et al. (2012) for an example where methods from both fields are used.

particular, there have been almost no papers that apply these methods together to assess the impacts of bushfires in Australia.

1.3 Research outline and contributions

The remaining components of our thesis are structured as follows:

In Chapter 2, we develop a method to address the inadequate spatial coverage of air quality measurement stations across Victoria. To do this, we collect Aerosol Optical Depth data from NASA's TERRA satellite network. We combine the AOD data with a range of other meteorological variables in a linear mixed effect model where ground-based PM2.5 is the dependent variable. This model allows for day-specific intercepts and slopes, which effectively captures many sources of time-variant unobserved heterogeneity. We then use this model to predict PM2.5 at a fine (10km grid) level across Victoria.

In Chapter 3, we model the relationship between bushfire related PM2.5 and daily respiratory ED presentations using a varying-parameter negative binomial generalised linear model (GLM). Our structure allows the estimated intercepts and coefficients for estimated PM2.5 to vary by hospital, which allows us to account for differing patient risk profiles and hospital sizes. We also present the methods used to decompose PM2.5 into a background and bushfire component, alongside the methods we used to match satellite data to hospitals.

In Chapter 4, we outline the simulation-based methodology we used to explore how different regional and demographic groupings were impacted by bushfire related particulate matter. Our results show that bushfire related particulate matter most heavily impacted inner-city hospitals and working-age adults aged 18 to 64. We argue that these results point to a need to reform workplace safety regulations to reduce the amount of time spent outdoors when smoke levels are hazardous. Potential reforms could include temporarily closing outdoor retail spaces and construction sites while encouraging office staff to work from home. We also highlight some strategies that may help health departments better prepare for bushfire-related spikes in emergency department presentations.

This thesis makes several important contributions. First, it is the first paper that we know of to use satellite data to estimate particulate matter concentrations during the 2019-2020 bushfire season. Our novel approach has produced an air quality dataset that could be used in further studies of the crisis. Second, our paper develops a multi-level model that allows the effects of bushfire smoke to differ by heterogeneous age and regional groupings. In addition, our fixed effects structure helps account for many sources of unobserved heterogeneity - alleviating concerns of endogeneity or of correlation between hospital/age groups and our other regressors. This model is, therefore, an important tool that can be used to simulate policy scenarios and thus help better manage future crises. Finally, our simulation results illustrate that it is urban, inner-city hospitals and working-age adults that were impacted most by bushfire smoke. This is a novel finding that has several important implications for policymakers; however, it has not been reported widely in the media or the literature.

Chapter 2

Deriving PM2.5 estimates from satellite data

The sparse and uneven distribution of air quality measurement stations makes it challenging to estimate the health effects of bushfire smoke. We can address this problem by using satellite imagery to produce higher resolution estimates of PM2.5. To do this, we collect satellite-based estimates of Optical Depth (AOD), temperature, and dewpoint temperature. A linear mixed-effect model with time-varying parameters is then used to estimate a grid of average daily PM2.5 values across Victoria. All data are daily, and our sample period is the 11th of November 2019 to the 25th of January 2020.[1]

2.1 Data

2.1.1 Ground-based PM2.5 measurements

The Victorian Environmental Protection Agency (EPA) provided access to PM2.5 measurements at 33 air quality measurement stations across Victoria. We were provided data at the hourly level; however, we aggregated it to obtain the 24-hour mean PM2.5 concentration for all 33 stations. The data we were provided with is from January 2008 to February 2020,

[1] We chose this period because of data availability.

however much of it is missing. We have a complete set of ground measurements from the 11th of November 2019 to the 25th of February 2020.

While most of the literature used untransformed PM2.5, we used log 24-hour mean PM2.5. Bushfires tend to drive a dramatic increase in the number of extreme air quality events. The presence of these extreme events can create numerical instabilities which make it challenging to fit the mixed-effects models. The untransformed series also has very high variance, skewness and kurtosis. Using a log-transformation helps stabilise our model fitting and enables us to make more reasonable assumptions about the distribution of our dependent variable and our model's residuals, but also allows us to maintain a relatively easily interpretable set of results.

2.1.2 Other meteorological products

The European Centre for Medium-Range Weather Forecasts (ECMWF) produces global hourly estimates of a range of meteorological variables, including temperature, wind speed, and humidity through the ERA5 dataset. The ECWMF provides this data at a 30km grid scale.

We collect estimates of the following meteorological variables:

- 2m temperature: Temperature of the air at 2m above the surface.
- 2m dewpoint temperature: A measurement of humidity, dewpoint temperature is the temperature that the air must be cooled to produce condensation at 2m above the surface.

2.1.3 Aerosol optical depth

Aerosol Optical Depth (AOD) measures the proportion of light that is absorbed by particulate matter in the atmosphere. Higher absorption rates tend to correlate with higher concentrations of particulate matter, which makes AOD a good proxy for PM2.5.

The AOD data we collected for Victoria came from the Moderate Resolution Imaging Spectroradiometer (MODIS) Terra and Aqua satellites. We collected the MCD19A2 product

from NASA's Level 1 and Atmosphere Archive and Distribution System Distributed Active Archive Center (LAADS DAAC). We obtained data at a 1km gridded scale at a 0.47 μm wavelength. The number of observations provided per day depends on the orbit of the satellite system.

There are three major problems we faced with the AOD data - the inconsistent time between measurements, long computation times, and missing data. To create consistent periods, we averaged across all observations on a given day to produce a grid of 'average' 47 μm readings across Victoria (which will correspond to the 24-hour exposure to PM2.5). The grid was upscaled from 1km to 10km, which reduced the amount of data we needed to process and thus reduced computation times.

2.1.4 A model for AOD interpolation

The missing data problem was slightly more challenging to manage. Poor visibility often means that satellite networks cannot observe AOD levels which creates large gaps in AOD coverage. We addressed this issue by using spatial interpolation to fill in the missing data. The strong spatial dependence of the AOD series meant that spatial interpolation was preferable to temporal interpolation in this case. This strong spatial dependence can be seen in Figure 2.1, which plots the spatial distribution of AOD on days with relatively good AOD coverage.

We chose inverse distance weighting (IDW) as the interpolation method.[2] IDW was chosen over more sophisticated spatial/spatio-temporal interpolation methods (such as kriging) because it is computationally cheap, which is of critical importance given the amount of data we need to process. To implement this method, we assumed that the value of some missing value $AOD_{t,j}$ at grid location j on day t could be described as the following weighted sum of its 100 closest neighbours:

$$AOD_{t,j} = \frac{1}{100} \sum_{i=1}^{100} \frac{w(AOD_{t,i}) \times AOD_{t,i}}{w(AOD_{t,i})}$$

[2]See Appendix A for a more detailed discussion of interpolation methods.

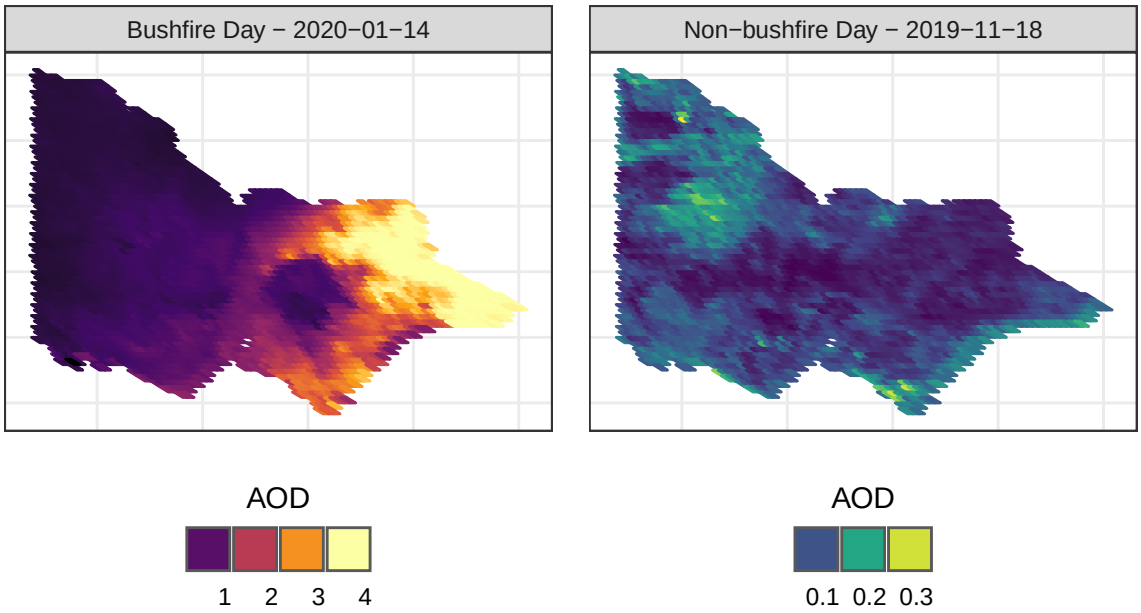

Figure 2.1: *Spatial distribution of AOD on 14/01/2020 and 18/11/2019. Different colour schemes are used to draw attention to the difference in AOD concentration across these two days. There is strong spatial dependence on both days of high bushfire activity and under normal conditions.*

Where the weight for a given location i is given by the inverse square of its distance from j, e.g.

$$w(AOD_{t,i}) = \frac{1}{d(i,j)^2}$$

and where $d(i,j)$ is the Euclidean distance between location i and location j.

2.1.5 Data integration

The data we collected for this stage of the model were all at different spatial scales - the PM2.5 data were points estimates, the AOD data was at a 10km grid level, and the meteorological data was at a 30km grid level. To merge the AOD with the meteorological dataset, we matched each 10km TERRA (AOD) grid with the nearest 30km ECMWF grid. This matching process allowed us to preserve the 10km resolution. We used this dataset to make our final PM2.5 predictions.

We then identified the cells of the grid that contained the air quality measurement stations and filtered our dataset to focus on the AOD, temperature and dewpoint temperature at

these locations. This filtered data was then joined to the PM2.5 data by location and day to create a dataset that we could use to fit our model.

2.2 Model

2.2.1 Model structure

Our estimated mixed-effects model contains day-specific slopes for Aerosol Optical Depth and day-specific intercepts. Previous work (Chudnovsky et al. (2012), Zhang et al. (2019)) has shown that this model structure can effectively control for unobserved time-varying heterogeneity that can impact the AOD-PM2.5 relationship. We also allow log(AOD) to interact with temperature and dewpoint temperature, which can help account for some of the effects of temperature and humidity on AOD measurement equipment.

Our estimated model is:

$$\log(PM_{t,j}) = \beta_{0,t} + \beta_{1,t}\log(AOD)_{t,j} + \beta_3 TEMP_{t,j} + \beta_4 DEWTEMP_{t,j} + \beta_5 TEMP_{t,j} \times \log(AOD)_{t,j} + \beta_6 DEWTEMP_{t,j} \times \log(AOD)_{t,j} + \varepsilon_{t,j}$$

where:

$$\beta_{0,t} = \gamma_0 + u_{0,t}$$

$$\beta_{1,t} = \gamma_1 + u_{1,t}$$

$$u_{1,t}, u_{2,t} \sim N_2(\mathbf{0}, \Sigma_0)$$

$$\varepsilon_{t,j} \sim N(0, \sigma^2)$$

Where $\log(PM_{t,j})$ is the natural logarithm of the PM2.5 measurement at location j at time t, $\log(AOD_{t,j})$ is the natural logarithm of AOD at time t and location j, and $TEMP_{t,j}$,$DEWTEMP_{t,j}$ are temperature and dewpoint temperature at location j and time t respectively. β_3, β_4, β_5 and β_6 are the fixed coefficients for temperature, dewpoint temperature, the interaction between log(AOD) and temperature, and the interaction between log(AOD) and dewpoint temperature respectively. $\varepsilon_{t,j}$ is an i.i.d normal error term.

$\beta_{0,t}$ and $\beta_{1,t}$ are the time-varying parameters for the intercept and log(AOD) respectively. They are composed of a fixed component (γ_0 & γ_1) and a random component ($u_{0,t}$ & $u_{1,t}$). These random components follow a bivariate normal distribution with a mean of zero and a variance covariance matrix Σ.

The model is fitted using restricted maximum likelihood via the `lmer` function in the **lme4** R package.

2.2.2 Model validation

The hierarchical structure of our model means that standard cross-validation and model assessment criteria are likely to be inappropriate. This is because we are seeking to use our model to predict PM2.5 levels where we have no observed data. Our model assessment criteria should account for this fact.

We use a Leave-One Group Out Cross Validation (LOGO CV) approach to solve this problem. This strategy is similar to leave one out cross-validation, but rather than leaving one observation out we leave one 'group' (in our case an air quality measurement station) out. A LOGO CV approach allows us to assess how accurately our model will be able to predict PM2.5 levels at unobserved locations. This will be more effective for our application since we will be using predicted PM2.5 at locations with no ground-measured PM2.5 in the later stages of our modelling process.

2.3 Results and model evaluation

2.3.1 Model fit

We report the estimated fixed effects parameters and standard error in Table 2.1. All fixed effects regressors appear to be highly statistically significant, except for the log(AOD) dewpoint temperature interaction term.

2.3.2 Model performance

The model's overall performance is decent - producing an in-sample R^2 of 0.76 and a cross-validated R^2 of 0.71 on the transformed data series. The model achieves an in-sample

Table 2.1: *Model Fixed Effects Estimates*

	Estimate	Std. Error	df	t value	Pr(>|t|)
Intercept	-42.8456	4.8920	219.8534	-8.7582	0.0000
Temperature	0.0729	0.0101	494.8664	7.1980	0.0000
Dewpoint Temperature	0.0830	0.0145	541.3884	5.7175	0.0000
log(AOD)	28.1187	4.9452	380.2574	5.6861	0.0000
log(AOD) * Temperature	-0.1171	0.0151	655.7602	-7.7760	0.0000
log(AOD) * Dewpoint Temperature	0.0198	0.0115	770.0429	1.7209	0.0857

mean absolute error of 0.4398 and a cross-validated mean absolute error of 0.4843 (see Table 2.2).

Table 2.2: *In-sample and cross-validated model performance statistics*

	MAE	RMSE	R-Squared
In-sample	0.4398	0.6402	0.7641
Cross-validated	0.4843	0.7122	0.7081

We note that the model tends to produce slight overestimates when actual PM2.5 is low and underestimates when actual PM2.5 is high. This can be seen in Figure 2.2, which plots the predicted and actual PM2.5 levels for both the cross-validated predictions and in-sample predictions.

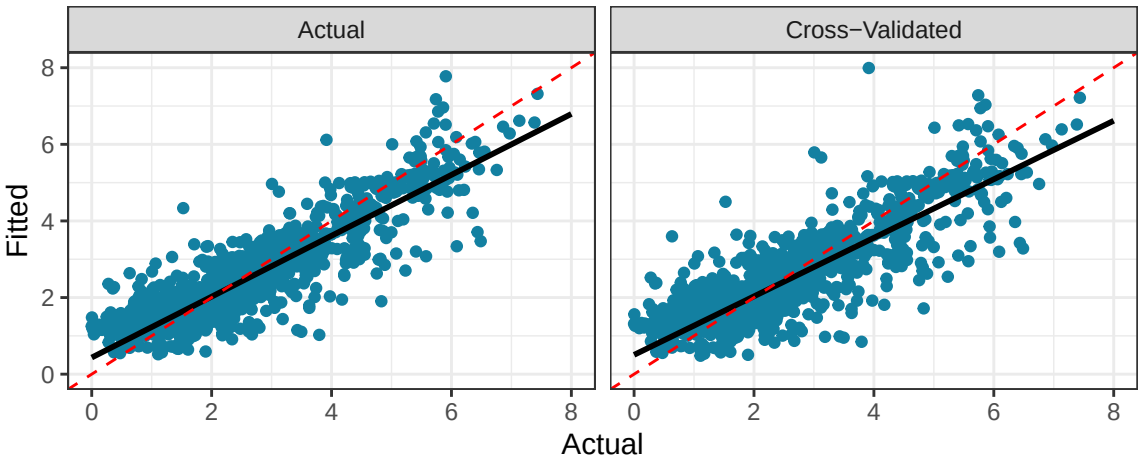

Figure 2.2: *Fitted log(PM2.5) vs actual log(PM2.5). The red line is a 1-1 line, and the black line is a linear regression fit. In and out of sample (cross-validated) model fit is good, but the model appears to have a tendency to produce underestimates when actual PM2.5 is high.*

Chapter 3

Estimating the heterogeneous health effects of bushfire smoke

Our satellite-derived estimates of PM2.5 provides a much higher resolution view of the spatial distribution of PM2.5 across Victoria than could reasonably be produced using ground-based measurements. These estimates allow us to more accurately estimate the population's exposure to bushfire-related particulate matter, which will facilitate a more precise analysis of the impact of bushfire smoke on different demographic groups. Producing this analysis requires a well-specified model that describes the relationship between bushfire related PM2.5 and health outcomes (respiratory emergency department presentations in our case).

We develop such a model by:

- Decomposing our estimated PM2.5 values into 'bushfire' and 'background' components; and
- Fitting a multi-level (fixed effects) negative binomial regression generalised linear model that describes the relationship between the daily count of respiratory ED presentations and bushfire-related PM2.5 (controlling for other environmental factors).

3.1 Data

3.1.1 Emergency room presentations

Emergency room presentation data was collected from the Victorian Department of Health via the Grattan Institute. The data set provides daily hospital-level presentation data for 39 Victorian hospitals with emergency rooms. The data is segmented into three age bands (under 18,18 to 64 and 65+) and into three main diagnosis groups (respiratory, injury and other).

There are a few features of this dataset that are important to consider. First, the distribution of respiratory ED presentations varies by hospital. The largest median count of presentations is 34, whereas the smallest median count of presentations (seen across five hospitals) is 1. These differences may be attributable to the size of the population the emergency department services, the size of the emergency department itself, or differences in the kinds of patients the department services.

The presence of these differences is difficult to account for because hospitals do not serve a fixed number of individuals, which makes it impossible to make per-capita adjustments. For example, a hospital in a tourism-dependent beach town may serve a smaller population in the winter, and a larger population in the summer. The seasonal influx of tourists would not be accounted for in census estimates of a region's population.[1]

Second, Victorian hospitals are not evenly distributed across Victoria. Instead, they are primarily clustered in dense urban areas such as Melbourne, Geelong and Bendigo (see Figure 3.1). Hospitals in densely populated regions will serve smaller geographic areas than populations in sparsely populated areas. This will make it difficult for us to 'allocate' estimated PM2.5 levels to hospitals.

Third, there are some hospitals that treat primarily children or provide primarily perinatal care (which means they mostly treat patients in the 18 to 64-year-old age group). There are a few cases in our sample where children/elderly adults present to perinatal hospitals

[1] We allow the parameters of the model to vary by hospital, which allows us to account for differences in the size of emergency departments, differences in patient demographics and any other time-invariant sources of heterogeneity.

and adults present to children's hospitals. These observations are treated as outliers and removed.[2]

3.1.2 Spatial matching of hospital and satellite data

Hospitals admissions are indexed at the hospital level (point spatial data); however, they may be a result of exposure to particulate matter or other meteorological variables from a wider catchment area. Defining these areas is challenging because hospitals in densely populated areas are likely to have a smaller catchment area. In contrast, rural hospitals are likely to have a wider catchment area. This would make using a fixed radius to form catchment areas inappropriate.

Instead, we allocate each reading in our estimated PM2.5 grid to its nearest hospital, which creates 35 'catchment areas' for Victorian hospitals. Four inner-city hospitals do not have a catchment area because they are not close enough to the centre of any grid. These four hospitals are allocated their closest grid. Figure 3.1 shows how this creates a set of regions which are relatively small in hospital-dense areas, and very wide in hospital-sparse areas. Our regressors are constructed by taking mean values for each catchment area.

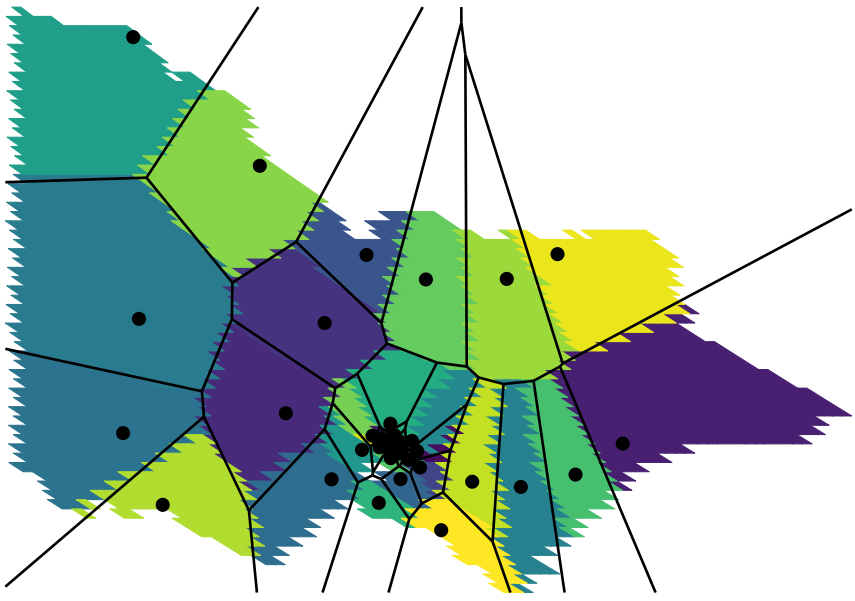

Figure 3.1: *Estimated hospital catchment zones for Victoria. Black points are hospital locations. Hospitals in densely populated regions have smaller catchment zones than hospitals in sparsely populated regions.*

[2] Additionally, there are some hospitals that are very close to children's hospitals and so rarely receive patients aged under 18. These observations are also removed.

This approach assumes that there is a constant population density across our constructed catchment areas and that individuals will always travel to the nearest hospital as the crow flies. These assumptions are clearly not strictly true, but they act as a reasonable enough approximation of the truth to allow us to make sensible inference about the impact of environmental factors on respiratory health.

3.2 Decomposing PM2.5 into a bushfire and a background component

Our PM2.5 estimates do not allow us to make causal inferences about the impact of bushfires on public health because fine particulate matter can come from many sources. Estimating the health effects of bushfire smoke will therefore require us to decompose our estimates into a 'bushfire' and a latent 'background' component which captures PM2.5 from standard urban and other environmental sources.

Morgan et al. (2010) outlines a method for decomposing PM10 into a 'background' and a 'bushfire' component using a simple 99th percentile threshold, which we described in Chapter 1. We apply a similar method. We assume that PM2.5 is, under normal conditions, produced by miscellaneous processes such as vehicle exhaust and industrial production. We then assume that bushfires introduce an additional source of pollution in the form of wood smoke and ash.

A simple threshold method is used to decompose PM2.5.[3] First, we calculate the 99th percentile of log(PM2.5) for November, December and January in Victoria using historical ground-based PM2.5 data from January 2013 to January 2019. These values are used as thresholds to identify periods and regions where bushfires are leading to hazardous air quality. We also calculate the median log(PM2.5) concentration on days where PM2.5 is below the threshold for the same set of months (again using ground readings). These values are used as 'background' PM2.5 on bushfire days. The estimated threshold and median values are shown in Table 3.1.

[3]We explored more advanced decomposition methods using satellite-derived bushfire hotspot data. The complex meteorological processes involved in this problem meant that this approach was not feasible.

Table 3.1: *99th percentile and historical median log(PM2.5) values*

Month	Threshold	Median (Exc. fire days)
January	2.8415	1.8833
November	2.5744	1.7017
December	2.5331	1.7991

Our estimated PM2.5 grid is then broken down by month and matched with a threshold and a median log(PM2.5) level. For each cell in our grid, on days where the estimated log(PM2.5) concentration exceed its given threshold, we set bushfire PM2.5 to the difference between the estimated PM2.5 level and the allocated historical median PM2.5 level. On normal days, we set a cell's background PM2.5 to the estimated PM2.5 level, and bushfire PM2.5 to zero.

This approach to PM2.5 decomposition assumes that bushfires are the only cause of extreme spikes in air quality. We believe this is a reasonable assumption for our sample period because there were no other major events that lead to short term deterioration of air quality on the same scale as bushfires. This assumption may not hold for longer studies, or for studies in other regions, however.

3.3 Estimating causal effects – a varying-coefficient negative binomial model

Our model structure is designed to model the age and region varying effects of bushfire smoke. To achieve this, we use fixed effect/dummy variables interaction terms.

Formally, let $Y_{t,i,k}$ be a variable which represents the count of respiratory ED presentations for patients in age band k at hospital i on day t. We assume that $Y_{t,i,k}$ follows a negative binomial distribution with a time, hospital and age varying rate parameter $\mu_{t,i,k}$ and a fixed shape parameter θ. This assumption allows us to model the mean count $\mu_{t,i,k}$ and introduces a random error to account for unobserved heterogeneity.

$$Y_{t,i,k}|\mu_{t,i,k},\theta \sim \mathcal{NB}(\mu_{t,i,k},\theta)$$

Now assume that $\mu_{t,i,k}$ can be described as:

$$\mu_{t,i,k} = \exp(\beta_{0,i,k} + \beta_{1,i,k} FIRELPM_{t,i} + \beta_{2,i,k} BGLPM_{t,i} + \beta_{3,k} DEWTEMP_{t,i} + \beta_{4,k} TEMP_{t,i} + \beta_5 WEEKEND_t)$$

where:

$$\beta_{0,i,k} = \gamma_0 + H_{0,i} + G_{0,k} + \delta_{0,i,k}$$
$$\beta_{1,i,k} = \gamma_1 + H_{1,i} + G_{1,k} + \delta_{1,i,k}$$
$$\beta_{2,i,k} = \gamma_2 + H_{2,i} + G_{2,k} + \delta_{2,i,k}$$
$$\beta_{3,k} = \gamma_3 + G_{3,k}$$
$$\beta_{4,k} = \gamma_4 + G_{4,k}$$

Where $FIRELPM_{t,i}$ is bushfire log(PM2.5), $BGLPM_{t,i}$ is background log(PM2.5), $TEMP_{t,i}$ is temperature, $DEWTEMP_{t,i}$ is dewpoint temperature, and $WEEKEND_t$ is a weekend dummy variable. The intercept and slopes $\beta_{0,i,k}$, $\beta_{1,i,k}$, and $\beta_{2,i,k}$ are composed of a fixed component γ, an age-specific component G_k, a hospital specific component H_i and a hospital & age interaction component $\delta_{i,k}$. $\beta_{3,k}$, and $\beta_{4,k}$ are composed a fixed component γ and an age-specific component G_k. The intercepts are estimated by including interacted age & hospital dummies, and the slopes are estimated by interacting hospital and age dummy variables with our regressors. The weekend dummy variable has a fixed slope β_5.

The model is fitted using the `stats` R package.

3.3.1 Fixed effects structure

We chose a fixed-effects rather than a random-effects or a mixed-effects specification.[4] A random effects specification was considered, but was rejected because random effects specifications impose the assumption that the unobserved and observed effects for a given hospital and age group are uncorrelated. This is an assumption that may not hold, and thus may introduce endogeneity and make inference difficult. We have a full sample of emergency rooms in Victoria, which allows us to estimate fixed effects in a relatively

[4]We also explored Bayesian hierarchical models. Our initial results were very similar to our fixed effects results. We decided to proceed using fixed effects because it allows us to avoid having to manage the specification of prior distributions.

straightforward manner and avoid much of the complication associated with a random-effects specification. We also have a long enough sample period to estimate the parameters in our model accurately.

An alternative approach would have been to estimate individual models for each hospital and age group. We did not take this approach because the smaller sample sizes would make it harder for our model to identify the relevant relationships in the model accurately.

3.4 Results and model evaluation

3.4.1 Model fit

Distributional assumptions

We assume that the daily count of respiratory ED presentations follows a negative binomial distribution. An alternative option would be to allow respiratory ED presentations to follow a Poisson distribution. Negative binomial distributions allow for overdispersion, which can make them significantly more flexible than a Poisson distribution (which assumes equidispersion) but can also make model fitting more challenging through the addition of a new parameter. Adopting the more parsimonious Poisson distribution may be sensible if the equidispersion condition holds.

We test for overdispersion by fitting a Poisson GLM with the same mean count specification as our primary negative binomial model. We then apply the overdispersion test set out in Cameron and Trivedi (1990) by formulating:

$$VAR[Y_{t,i,k}] = \mu_{t,i,k} + \alpha * \mu_{t,i,k}^2$$

We then test if $\alpha > 0$. Under the null hypothesis, $\alpha = 0$ and a Poisson model is appropriate. Under the alternative hypothesis, $\alpha > 0$ and a negative binomial specification is appropriate. Our testing yielded a p-value of $8.5 * 10^{-7}$, so we conclude that our negative binomial specification is appropriate.

Model parameters

The estimated model contains over 300 parameters, which makes it infeasible for us to display the individual values and significance of all parameters in the model. Instead, we show the distribution of parameters in Figure C.1 in Appendix C.[5]

We also show the joint significance of all the main components of our model (bushfire PM2.5, background PM2.5, temperature and dewpoint temperature) in Table 3.2. These joint significance tests are conducted as a series of likelihood ratio tests. They test a null model (for example, a model where we include bushfire PM2.5 as a single regressor with no age or hospital variant slope) against the alternative 'full' model we described above. This enables us to establish the significance of the terms in our model as individual regressors and as regressors with age/hospital variant slopes.

Table 3.2: *Likelihood ratio tests of model parameters. Each row represents the likelihood ratio results associated with testing the restricted model against the unrestricted model described in this Chapter.*

Restricted Model	df	LR stat.	Pr(Chi)
Bushfire PM2.5			
No bushfire PM2.5	110	255.991	0.000
Bushfire PM2.5 included as a main effect	109	221.942	0.000
Bushfire PM2.5 with age-variant slopes	107	147.147	0.006
Background PM2.5			
No background PM2.5	110	146.835	0.011
Background PM2.5 included as a main effect	109	142.290	0.018
Background PM2.5 with age-variant slopes	107	131.954	0.051
Temperature			
No temperature	3	4.081	0.253
Temperature included as a main effect	2	1.500	0.472
Dewpoint Temperature			
No dewpoint temperature	3	29.433	0.000
Dewpoint temperature included as a main effect	2	15.548	0.000

These results suggest that bushfire particulate matter has a meaningful impact on respiratory health and that this impact varies by age and by hospital. This has important public policy implications that we address in Chapter 4.

[5]The figure highlights that the estimated parameters vary substantially across hospital and age group. There are several outliers in the plot which correspond to hospitals that tend to specialise in treating one age group.

The results also show that dewpoint temperature is statistically significant - both as an individual regressor and with variable slopes. Background PM2.5 appears to be significant when we include it with no fixed effect slopes, but there is some question as to the significance of its hospital and age specific slopes. Temperature is not significant.

3.4.2 Effect of bushfire PM2.5

The most interesting set of parameters is the fixed effect slope on PM2.5. In Figure 3.2, we plot the distribution of the slope of PM2.5 for all 39 hospitals in our sample across all 3 age brackets. The slope (calculated as $\beta_{1,i,k} = \gamma_1 + H_{1,i} + G_{1,k} + \delta_{1,i,k}$) can be interpreted in a relatively straightforward manner - a 1 per cent increase in PM2.5 will be expected to lead to a $\beta_{1,i,k}$ per cent increase in respiratory ED presentations.

Figure 3.2 shows that the effect of bushfire-related particulate matter varies by age group. It appears that working-age adults (18-64) are impacted most heavily by a given change in particulate matter, followed by older adults (65 and over). Children do not appear to be affected across our sample period. These results are surprising, as children and elderly individuals generally have less robust respiratory health than adults.[6] However, there is a range of behavioural factors that may drive these results, which are discussed in Chapter 4.

3.4.3 Model performance

Our estimated model's fitted values appear to have a similar distribution to the observed data (see figure 3.3); however, there is some tendency for the model to underestimate the frequency of zeros and other low values.[7]

Similar trends can be seen in Figure 3.4, which shows the fitted vs actual count of daily respiratory ED presentations. The overall fit seems good, but it is clear that the model produces slight overestimates for small values and slight underestimates for large values.

[6] There are some large outliers in the 0 to 17 age group. These outliers are generally small rural hospitals or specialist hospitals that do not see a large number of patients.

[7] We explored using hurdle and zero-inflated negative binomial GLMs to solve this problem, but they did not meaningfully increase the number of zeros predicted. Hurdle/zero-inflated models were also extremely unstable due to incidental parameter problems.

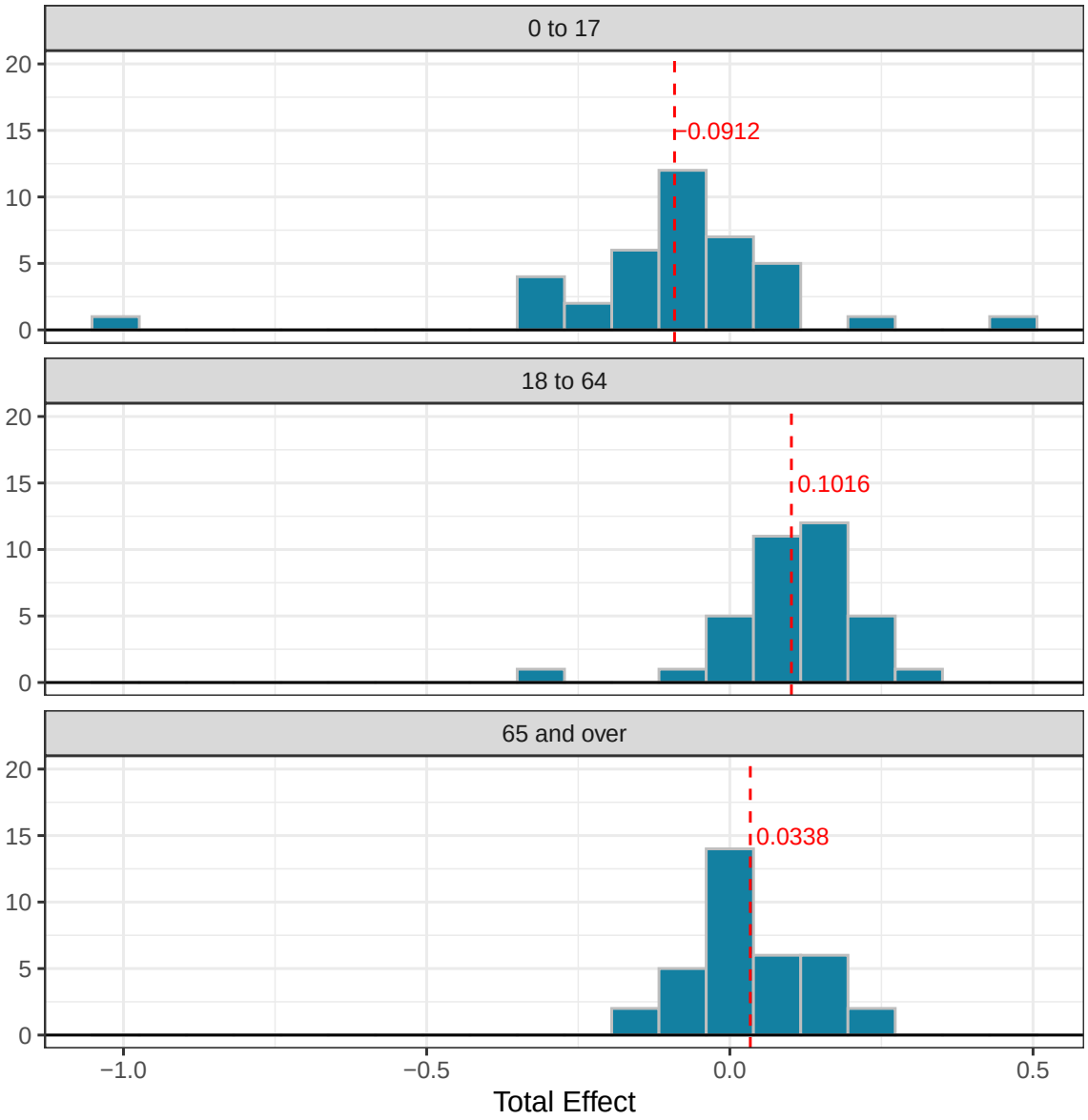

Figure 3.2: *Distribution of hospital-specific total effects of bushfire PM2.5, by age bracket. The average total effect for all hospitals is shown in red. The average total effect is highest for 18 to 64 year olds.*

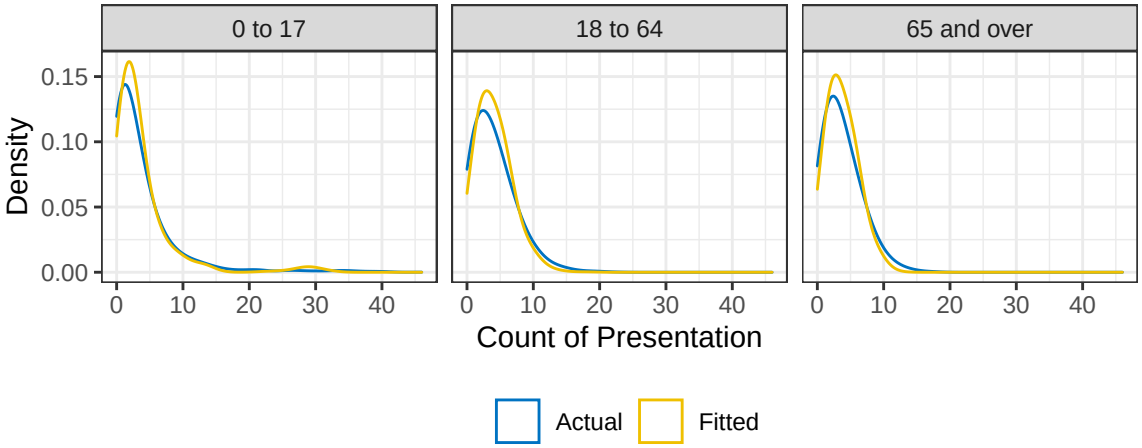

Figure 3.3: *Fitted (yellow) vs actual (blue) distribution of respiratory ED presentations in Victoria between 11/11/2019-25/01/2020, by age bracket. The two distributions are very similar, however there appear to be less very small counts in the fitted distribution.*

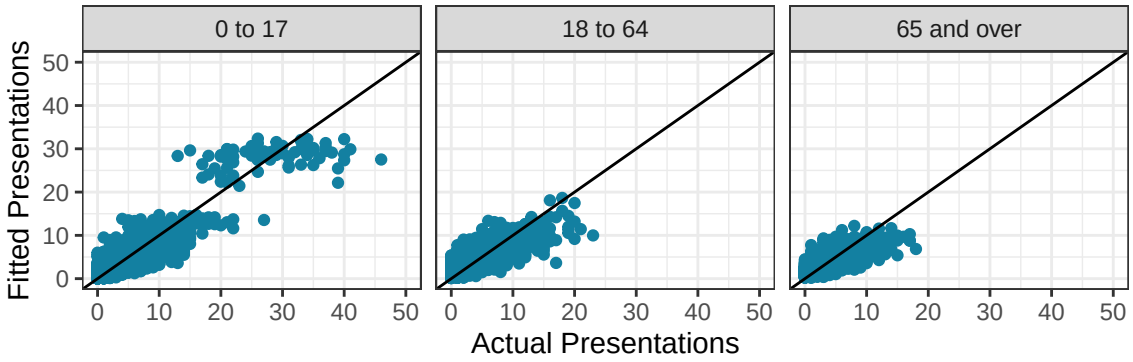

Figure 3.4: *Fitted vs actual scatterplot of respiratory ED presentations in Victoria between 11/11/2019-25/01/2020, by age bracket. The fitted distribution tends to perform worse for low counts.*

Chapter 4

Model simulation and policy analysis

4.1 Simulation design

We adopt a simulation-based approach to explore the relationship between bushfire-related PM2.5 and respiratory ED presentations in more detail. Specifically, we simulate from our fitted negative binomial model using (a) our observed a dataset (the main distribution) and (b) a counterfactual dataset where bushfire related particulate matter is set to zero, but all other regressors retain the same values as the observed dataset (the counterfactual distribution). We simulate 10,000 draws from the main distribution and the counterfactual distribution.

Formally, we construct two distributions:

$$Y_1|\mu_1, \theta \sim \mathcal{NB}(\mu_1, \theta)$$
$$Y_0|\mu_0, \theta \sim \mathcal{NB}(\mu_0, \theta)$$

Where Y_1 is a vector of length 8436 that contains the daily simulated count of ED presentations, by age band, under the observed dataset. Y_0 is a vector of the same length that contains the data simulated under the counterfactual dataset. μ_1 is the predicted mean count for each observation under the observed dataset and μ_0 is the predicted mean count for each observation under the counterfactual dataset. θ is a scale parameter that is constant across both distributions. Comparing the differences between the counts under the observed data and the counterfactual data will allow us to estimate the number of "excess" emergency department presentations associated with bushfire related fine particulate matter. Adopting this simulation-based approach allows us to construct a distribution of these 'excess' presentations - enabling us to quantify the uncertainty associated with these estimates. These distributions can then be broken down by age group and by hospital, which enables us to explore how bushfires impact different communities.

We estimate the number of excess presentations for each hospital, each day, by subtracting the values simulated from our counterfactual distribution from the values simulated from our main distribution. More formally, the estimated number of excess presentations for a given hospital i, on day t, for age bracket k, can be calculated as:

$$EXCESS_{i,t,k} = Y_{1,t,i,k} - Y_{0,t,i,k}$$

Our approach is preferable to comparing the average marginal effect across groups because it accounts for both the treatment effect and the treatment level of bushfire-related PM2.5. There is substantial regional variation in particulate matter concentrations which will be necessary for policymakers to consider when planning for future crises.

Simulation results suggest that there were 580 (95% CI: 69 - 1092) emergency room presentations in our sample period that can be attributed to bushfire related particulate matter. This interval is relatively wide, which reflects the noise in the PM2.5 and ED presentation data.

4.2 Policy implications

4.2.1 The most severe days account for the majority of excess presentations

The majority of excess presentations occurred in early January (see Figure 4.1) and December. There are almost no excess presentations in November. Most of the excess presentations occur on a limited number of days - with the five worst days accounting 297 out of the 580 excess presentations. These findings suggest that bushfire related particulate matter tends to cause harm acutely, in short bursts when air quality conditions are very poor.

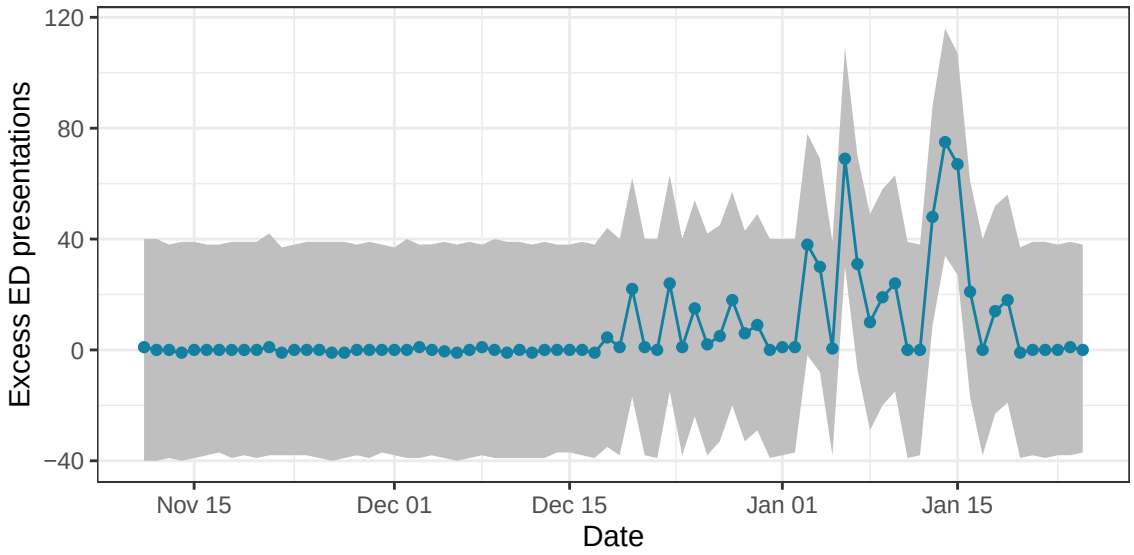

Figure 4.1: *Distribution of excess presentations by date. The blue dotted line represents the daily median excess presentation count, and the grey bands represent the 10th and 90th quantile. The series is very stable until January. The five worst days account for 297 out of 580 excess presentations.*

More proactive public safety/healthcare policy responses may be justified during these periods because these policies would only need to be enforced for short periods. An aggressive, short term set of strategies may enable policymakers to substantially reduce the number of ED presentations associated with bushfire smoke.

4.2.2 Bushfire PM2.5 impacts working-age adults most heavily

Identifying the groups that were most heavily impacted by bushfire particulate matter is a challenging task because the age brackets and hospitals which index our data have different sizes. Comparing the effects of bushfire particulate matter by looking at differences in excess presentations is not helpful because we cannot disentangle the effects of age bracket/hospital size from the effects of exposure.

To address this issue, we constructed a relative excess presentations statistic which measured the percentage change in respiratory ED presentations associated with bushfire PM2.5. We produced this measure by calculating the percentage difference between the number of respiratory presentations implied by the main and the counterfactual distribution. The resulting figure is scale-invariant - allowing us to make direct comparisons across age groups and hospitals.

We show the distribution of relative excess presentations by age in Figure 4.2. This plot highlights that bushfire smoke tended to impact adults aged 18 to 64 most heavily - with bushfire PM2.5 appearing to drive up the count of respiratory ED presentations by about 5.96% (95% CI: 3.15% to 8.87%). Adults aged 65 and over appear to be impacted less - with bushfires responsible for 1.69% increase in presentations (95% CI: -1.16% to 4.63%). Bushfire PM2.5 does not appear to have impact children at all - with a median decrease of 2.30% (95% CI: -5.12% to 0.59%).

These findings are surprising because young children and elderly adults tend to have more vulnerable respiratory systems. However, they may be attributable to differences in the amount of risk working-age adults take. Working-age adults have a greater need to travel during summer due to work,[1] and some adults work outside full time in industries like construction, which means they will tend to be exposed to more particulate matter. This higher exposure may explain why they are more likely to present at emergency rooms.

[1] Universities and schools were closed due for summer holidays for most of our sample period.

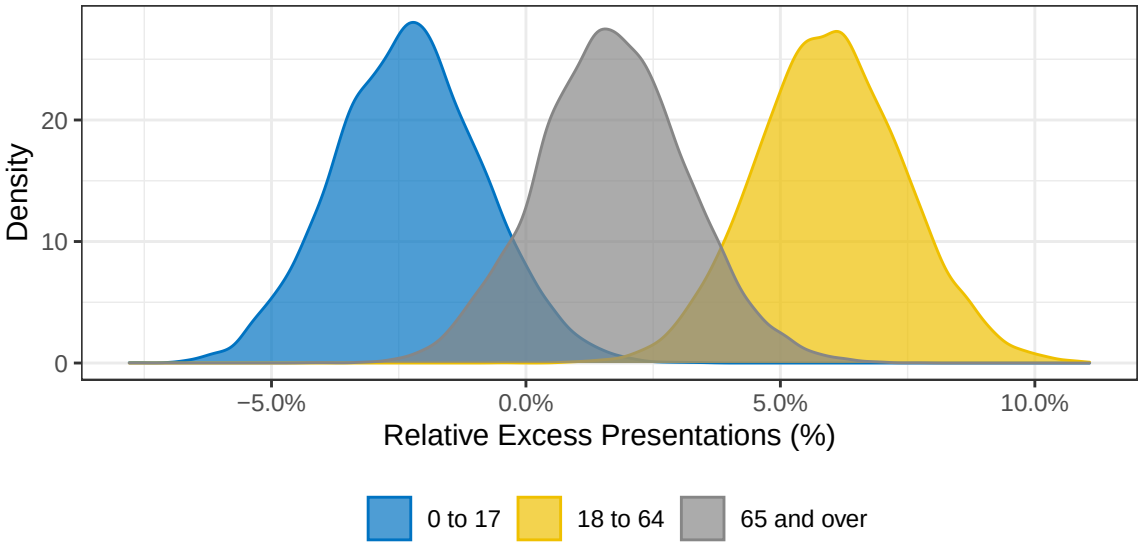

Figure 4.2: *Distribution of simulated relative excess presentations by age. 18 to 64 year olds were clearly impacted most heavily by bushfire related particulate matter.*

4.2.3 Hospitals near the CBD were impacted most heavily

Figure 4.3 highlights how bushfire PM2.5 has a much larger impact (in percentage terms) on 18 to 64 year-olds in inner-city hospitals than adult 18 to 64 year-olds in suburban or rural hospitals. Bushfire related PM2.5 increased inner-city respiratory ED presentations for adults aged 18 to 64 by about 12.6% (95% CI: 4.98% to 20.7%). 18 to 64 year-olds tend to travel in and around the CBD for work, so the substantial spike in presentations around the CBD suggests that work-related travel is a significant risk factor.

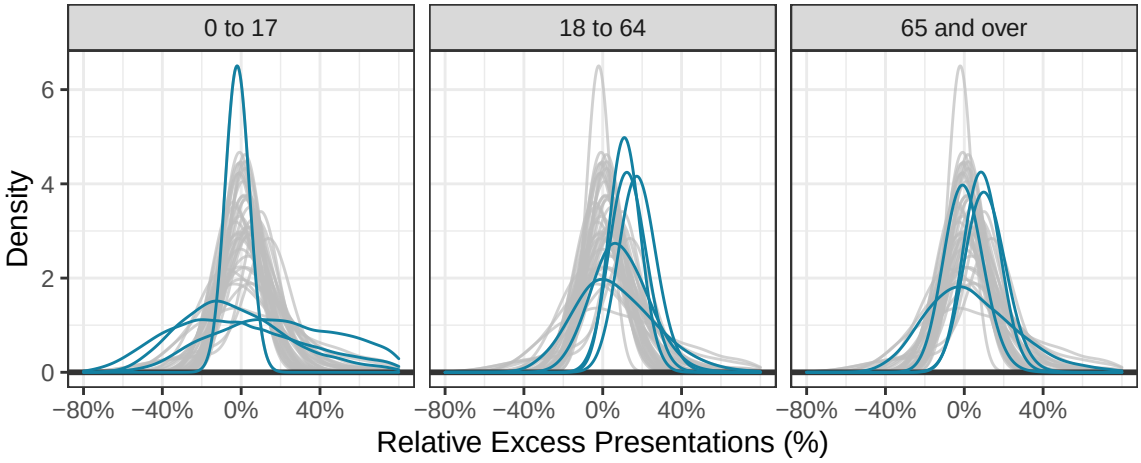

Figure 4.3: *Distribution of simulated relative excess presentations by age and hospitals. Inner city hospitals distributions are shown in blue, other hospitals are shown in grey. Inner city hospitals appear to have seen a larger spike in presentations from those aged over 18 than those aged under 18.*

4.2.4 Policy suggestions

Our results suggest that policymakers should:

1) Prepare for a spike in emergency room presentations around the CBD when air quality conditions deteriorate; and

2) Introduce policies that allow employees to work from home (where possible), and shut non-essential industry (e.g. construction) where workers will be outdoors when air quality levels are hazardous.

The first recommendation will ensure that hospitals have sufficient medical equipment and staff to handle a future surge in the number of patients presenting with respiratory problems. This will ensure that hospitals can provide all patients with the highest quality medical care, which will prevent a spike in adverse outcomes for presenting patients (e.g. further complications/mortality) because hospitals are overburdened.

To achieve this, health departments should conduct a review of their capacity to handle a surge of patients presenting with respiratory problems. If there is a shortage of specialised staff, it may be sensible to transfer pulmonologists[2] from suburban hospitals to inner-city hospitals on days where forecasted air quality is hazardous. If there is a shortage of equipment in inner-city hospitals,[3] then it may also be wise to move some specialised equipment from suburban to inner-city suburbs.

The second recommendation will help minimise the number of people that need to travel for work, which will drive down the amount of particulate matter those aged 18 to 64 will be exposed to. Lower exposure should help decrease the number of working-age adults who get admitted to emergency rooms across this period. The idea of allowing staff to work from home when air quality is hazardous is not unique or new (it was proposed by a number of unions during the crisis); however, our findings imply that there are serious risks associated with travelling and being outdoors that could be mitigated through better workplace safety regulations.

[2]Pulmonologists are doctors who specialise in treating diseases of the lung

[3]This is less likely as the department acquired a large quantity of equipment to prepare for the COVID-19 pandemic.

Chapter 5

Conclusion

Bushfire smoke is a severe threat to public health because it can cause a wide range of respiratory illnesses such a COPD[1] and exacerbate pre-existing conditions like asthma. However, researchers have historically struggled to accurately model the relationship between bushfire-related fine particulate matter and respiratory public health outcomes. This makes it challenging for policymakers to accurately assess which demographic groups and which regions are most at risk from bushfire smoke, which makes effective policy design and disaster preparation harder.

The most significant reason why this problem is so difficult is that air quality stations are sparsely and unevenly distributed across the state - making any assessment of the population's exposure difficult. There are other challenges, however, including the complex structure of hospital data, spatial differences in data, and the difficulties associated with isolating the impact of bushfires on air quality.

In this thesis, we developed a two-stage modelling approach that allowed us to overcome these limitations and effectively model the relationship between bushfire-related PM2.5 and Victorian respiratory ED presentations from 11/11/2019 to 25/01/2020. In short, our approach involved:

[1] COPD stands for chronic obstructive pulmonary disease.

1) Using satellite-based proxies for particulate matter, along with other meteorological variables, to estimate a high-resolution grid of PM2.5 across Victoria using a linear mixed-effect model with time-varying parameters.

2) Applying a simple threshold method to isolate the effects of bushfires on PM2.5.

3) Developing a varying-parameter fixed effect negative binomial GLM which describes the relationship between bushfire related PM2.5 and the daily count of respiratory emergency presentations (by hospital and age) while controlling for unobserved age and hospital-specific heterogeneity, temperature and dewpoint temperature.

4) Completing a series of simulation exercises to show how bushfire smoke impacted different demographic groups and regions during the 2019-2020 bushfire season.

The results of these simulations highlight how working-age adults and inner-city hospitals were impacted most heavily by bushfire smoke across the summer. This finding has important implications for the way that public health departments prepare for future crises, and underscores the need for better workplace safety policies during extreme air quality events.

This thesis makes important contributions to (a) the public discussion around how we should respond to future bushfire crises and (b) the broader environmental epidemiology literature. From a public discourse perspective, we have highlighted how higher risk-taking among working-age adults appears to have led to a spike in adverse respiratory health outcomes and suggested a series of policy approaches that could help minimise this spike in future crises. Our contributions to the epidemiological/public health literature are also substantial, as we have developed an approach that would be effective for modelling other large wildfires and bushfires across the globe. In particular, our high-resolution PM2.5 dataset could be of great value to future studies of the 2019/2020 bushfire season in Victoria.

There are several areas where further research could be valuable. It would be valuable to replicate this analysis with case-level, rather than aggregated hospital admissions or emergency room presentation data. This would have several distinct benefits:

1) Case-level data would allow us to better control for patient risk profiles and to identify how the fires impacted different demographic groups more clearly.

2) Case-level data would enable us to aggregate our data by home location.

3) Case-level data would allow us to calculate other aggregate measures of interest (such as excess admissions, excess deaths).

It would also be useful to replicate this study with data from New South Wales and Queensland. Air quality in Sydney was persistently high for an extended period in 2019/2020, so it would be interesting to see how fires around NSW impacted respiratory ED presentations. A nationwide study would enable us to compare our work with other papers in the literature and to add more value to the national policy debate.

Bibliography

Bengtsson, H (2020). *matrixStats: Functions that apply to rows and columns of matrices (and to vectors)*. R package version 0.56.0. https://CRAN.R-project.org/package=matrixStats.

Bivand, RS, E Pebesma, and V Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition*. Springer, NY. https://asdar-book.org/.

Cameron, A and P Trivedi (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics* **46**(3), 347–364.

Chudnovsky, AA, HJ Lee, A Kostinski, T Kotlov, and P Koutrakis (2012). Prediction of daily fine particulate matter concentrations using aerosol optical depth retrievals from the Geostationary Operational Environmental Satellite (GOES). *Journal of the Air & Waste Management Association* **62**(9), 1022–1031.

Csárdi, G and R FitzJohn (2019). *progress: Terminal progress bars*. R package version 1.2.2. https://CRAN.R-project.org/package=progress.

Dennekamp, M and MJ Abramson (2011). The effects of bushfire smoke on respiratory health. *Respirology* **16**(2), 198–209.

Dominici, F, A McDermott, and TJ Hastie (2004). *Improved semiparametric time series models of air pollution and mortality*.

Gräler, B, E Pebesma, and G Heuvelink (2016). Spatio-temporal interpolation using gstat. *The R Journal* **8** (1), 204–218.

Grolemund, G and H Wickham (2011). Dates and times made easy with lubridate. *Journal of Statistical Software* **40**(3), 1–25.

Henry, L (2020). *dtplyr: Data table back-end for 'dplyr'*. R package version 1.0.1. https://CRAN.R-project.org/package=dtplyr.

Kloog, I, BA Coull, A Zanobetti, P Koutrakis, and JD Schwartz (2012). Acute and chronic effects of particles on hospital admissions in New-England. *PLoS ONE* **7**(4). Ed. by MB Gravenor, e34664.

Liu, JC, G Pereira, SA Uhl, MA Bravo, and ML Bell (2014). A systematic review of the physical health impacts from non-occupational exposure to wildfire smoke.

Mackey, W (2020). *absmaps: Download and use maps data from the Australia Bureau of Statistics*. R package version 0.2.1.

Morgan, G, V Sheppeard, B Khalaj, A Ayyar, D Lincoln, B Jalaludin, J Beard, S Corbett, and T Lumley (2010). Effects of bushfire smoke on daily mortality and hospital admissions in Sydney, Australia. *Epidemiology* **21**(1), 47–55.

Ostro, B, WY Feng, R Broadwin, S Green, and M Lipsett (2007). The effects of components of fine particulate air pollution on mortality in California: Results from CALFINE. *Environmental Health Perspectives* **115**(1), 13–19.

Pebesma, E (2018). Simple features for R: standardized support for spatial vector data. *The R Journal* **10**(1), 439–446.

Peng, RD and F Dominici (2008). *Statistical Methods for Environmental Epidemiology with R*. Springer New York.

R Core Team (2020). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Ramsay, TO, RT Burnett, and D Krewski (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology* **14**(1), 18–23.

Rappold, AG, SL Stone, WE Cascio, LM Neas, VJ Kilaru, MS Carraway, JJ Szykman, A Ising, WE Cleve, JT Meredith, H Vaughan-Batten, L Deyneka, and RB Devlin (2011). Peat bog wildfire smoke exposure in rural North Carolina is associated with cardiopulmonary emergency department visits assessed through syndromic surveillance. *Environmental Health Perspectives* **119**(10), 1415–1420.

Reid, CE, M Brauer, FH Johnston, M Jerrett, JR Balmes, and CT Elliott (2016). *Critical review of health impacts of wildfire smoke exposure*. https://ehp.niehs.nih.gov/wp-content/uploads/124/9/ehp.1409277.alt.pdf.

Schwartz, J (1994). Nonparametric smoothing in the analysis of air pollution and respiratory illness. *Canadian Journal of Statistics* **22**(4), 471–487.

Simpson, R, L Denison, A Petroeschevsky, L Thalib, and G Williams (2000). Effects of ambient particle pollution on daily mortality in Melbourne, 1991-1996. *Journal of Exposure Analysis and Environmental Epidemiology* **10**(5), 488–496.

Tham, R, B Erbas, M Akram, M Dennekamp, and MJ Abramson (2009). The impact of smoke on respiratory hospital outcomes during the 2002-2003 bushfire season, Victoria, Australia. *Respirology* **14**(1), 69–75.

Touloumi, G, R Atkinson, AL Tertre, E Samoli, J Schwartz, C Schindler, JM Vonk, G Rossi, M Saez, D Rabszenko, and K Katsouyanni (2004). Analysis of health outcome time series data in epidemiological studies. *Environmetrics* **15**(2), 101–117.

Unnithan, SL and L Gnanappazham (2020). Spatiotemporal mixed effects modeling for the estimation of PM2.5 from MODIS AOD over the Indian subcontinent. *GIScience and Remote Sensing* **57**(2), 159–173.

Wang, E, D Cook, and RJ Hyndman (2020). A new tidy data structure to support exploration and modeling of temporal data. *Journal of Computational and Graphical Statistics*.

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.

Zhang, K, G de Leeuw, Z Yang, X Chen, X Su, and J Jiao (2019). Estimating spatio-temporal variations of PM2.5 concentrations using VIIRS-Derived AOD in the Guanzhong Basin, China. *Remote Sensing* **11**(22).

Appendix A

Imputation of AOD data

Our AOD data is an effective proxy for PM2.5, however ~29% of the data is missing due to equipment problems and poor visibility.

There were two possible approaches we could have taken to fill in the missing aerosol optical depth data. First, we could have carried forward the last observed value for a given grid (time series imputation). Second, we could have used some spatial interpolation method.

Spatial interpolation was preferrable for two reasons. First, spatial interpolation gives 'smoother' estimates of AOD that more accurately reflect the way that aerosol particles disperse through space. Second, time series interpolation can carry forward extreme AOD readings which would cause our PM2.5 estimates to be biased.

In Figure A.1 we plot the observed AOD values on 22/01/2020 - one of the days with the worst AOD coverage. In Figure A.2 and Figure A.3 we plot the imputed values using both inverse distance weighting (our chosen spatial interpolation method) and time series imputation. Both issues described above can be clearly seen

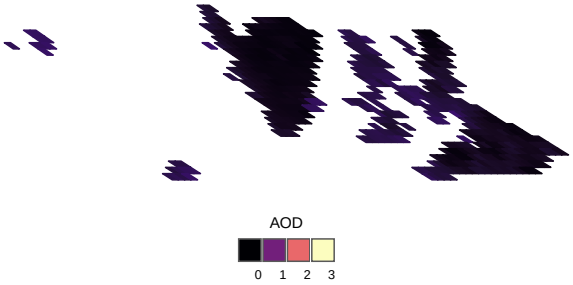

Figure A.1: *AOD coverage on 22/01/2020. No interpolation methods have been applied. The missing data problem is severe on this day.*

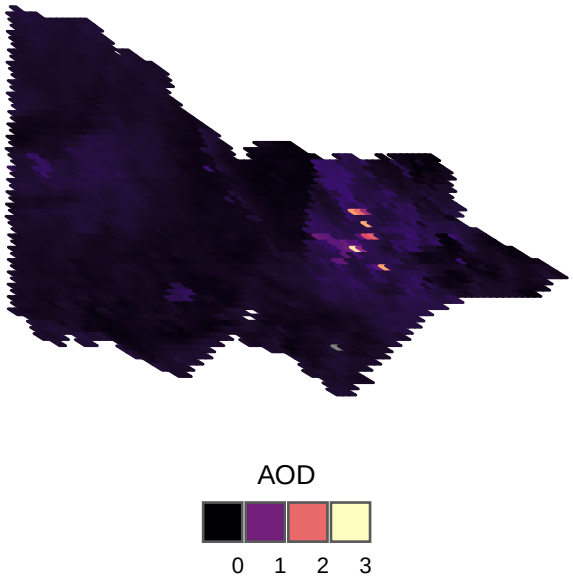

Figure A.2: *AOD coverage on 22/01/2020 using time series interpolation methods. Time series interpolation carries forward previous extreme values and does not capture spatial dependence.*

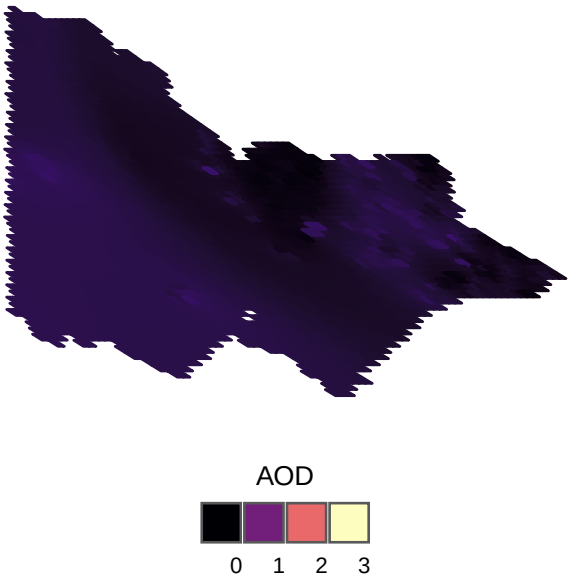

Figure A.3: *AOD coverage on 22/01/2020 using spatial interpolation methods. Extreme values are not carried forward and spatial dependence is captured.*

Appendix B

Data

B.1 Data description

Table B.1: *Data Description*

Variable	Description	Units
PM2.5	Concentration of particulate matter with diameter 2.5 micrometers or smaller. Measured at ground based stations.	Micrograms per cubic meter of air
AOD	Aerosol Optical Depth. A measure of solar extinction due to dust and other haze. Particulate matter absorbs light in the atmosphere, so higher AOD readings correlate with worse air quality.	AOD does not have a meaningful unit
DEWTEMP	Depoint temperature. The temperature to which the air must be cooled to become saturated with water vapour. A measure of humidity.	Degrees kelvin

Variable	Description	Units
TEMP	Temperature.	Degrees kelvin
FIRELPM	Bushfire-related log(PM2.5). Calculated using a threshold method described in Chapter 3.	Micrograms per cubic meter of air
BGLPM	Background log(PM2.5). Calculated using a threshold method described in Chapter 3.	Micrograms per cubic meter of air
ED Pres	Number of patients in a given age group presenting to a hospital emergency room.	Number of patients

B.2 Summary statistics

Table B.2: *Summary statistics - AOD-PM2.5 data*

	Mean	Std. Dev	Median	Min	Max	Range	Skewness	Exc. Kurt
PM2.5	32.74	102.50	6.81	0.02	1694.56	1694.54	8.22	95.72
AOD	0.49	0.66	0.32	0.03	4.00	3.97	3.56	13.36
DEWTEMP	283.27	3.09	282.98	274.66	291.85	17.19	0.23	-0.56
TEMP	290.76	4.62	289.89	280.75	307.76	27.01	0.62	-0.04
Log(PM2.5)	2.23	1.32	1.92	-3.87	7.44	11.31	0.91	1.84
Log(AOD)	-1.19	0.90	-1.13	-3.59	1.39	4.98	0.39	0.76

Table B.3: *Summary statistics - PM2.5-Hospital data*

	Mean	Std. Dev	Median	Min	Max	Range	Skewness	Exc. Kurt
Resp.ED Pres.	3.86	4.05	3.00	0.00	46.00	46.00	3.16	18.20
FIRELPM	0.42	0.82	0.00	0.00	5.05	5.05	2.10	3.95
BGLPM	1.72	0.38	1.80	0.42	2.83	2.41	-0.31	0.04
DEWTEMP.	282.31	3.25	282.17	270.90	290.57	19.67	0.09	-0.25
TEMP	291.28	4.92	290.47	280.32	309.28	28.96	0.69	0.22

Appendix C

Hospital-specific effects

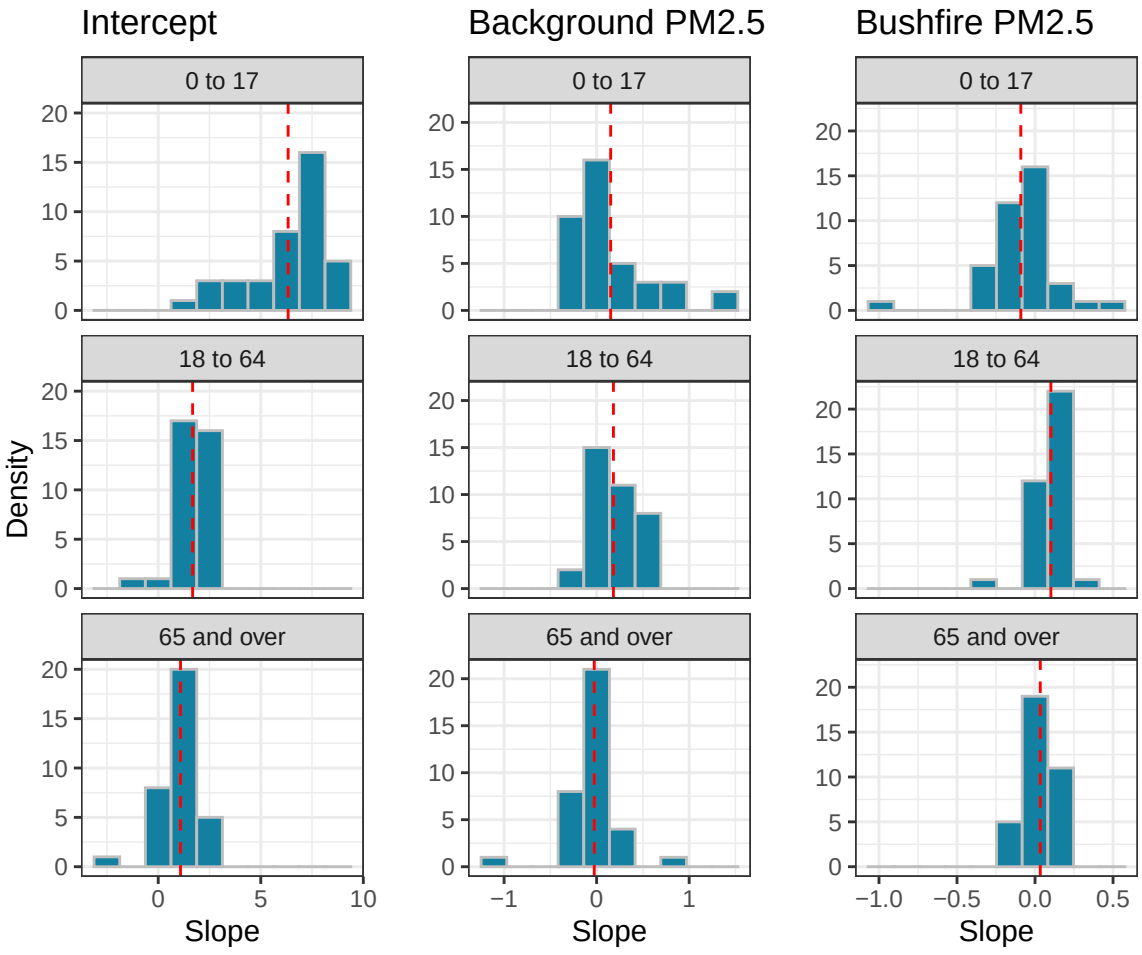

Figure C.1: *Distribution of hospital-specific model effects by age. Parameters seem roughly normally distributed for the '18 to 64' and '65 and over' demographics, but the '0 to 17' parameters appear to have a higher variance and more skewed distributions.*

Appendix D

Replication

This thesis uses confidential medical data. The full code and data for replication are available upon request and consultation with the appropriate parties. Any materials shared for the purposes of replication must not be used for any purposes except replication.

A version of our code has been made publicly available at an alternative github repo: https://github.com/anonymous345-star/4860.