NERC Hackathon 1 - Bath Air Squad summary

A Bayesian framework for analysing
air pollution effect on COVID-19 infection rate

Daniel Burrows[1], Piotr Morawiecki[1], Laura Oporto[1], and Yang Zhou[1]

[1] Department of Mathematical Sciences, University of Bath

June 5, 2020

Abstract

Since its breakout in China at the beginning of 2020, the COVID-19 virus has spread violently across the world, with various regions and countries suffering different levels of impact as a result of the pandemic. The increasing quantity of subsequent COVID-19 related data enables a better understanding of the dynamics of the viral infection to be formed. This study presents a spatio-temporal Bayesian framework in R that analyses the impact of air pollution and social factors on the infection rate in the United States and Italy. The effects of different pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3) and the Air Quality Index (AQI) are studied in different time scales to assess their impact on COVID-19 incidence and potentially inform future environmental policy as countries ease their lockdown restrictions.

Keywords: Bayesian modelling, INLA, COVID-19, air pollution

1 Background

In an effort to combat the COVID-19 crisis, numerous research collaborations have arisen that seek to contribute towards a collective response in some way or another. One branch of this response focuses on understanding how air pollution and its impact on the respiratory system could potentially make humans more vulnerable to the virus.

Based on previous literature [1], [2], we observed a causality between long-term exposure to common outdoor air pollutants such as NO_2, $PM_{2.5}$, O_3 and different levels of damage to the human respiratory and immune systems, all of which can make a person more vulnerable to a virus infection [3]. Additionally, other studies such as [3] and [1] mention that different chemicals can have various impact results on people within different age groups.

Additionally, exposure to air pollutants can develop dangerous background diseases that are major contributors to COVID-19 fatality. For example, from data collected from Italian records, there are five top background diseases that caused the deaths. Many well established studies have shown that those diseases all have a strong correlation to exposure to NO_2 [4].

Although COVID-19 has caused a huge number of fatalities, the total fatality numbers globally have decreased by a significant percentage, including all air pollution related diseases [2]. Therefore it brings us the question: how much does air pollution contribute towards the infection or even the severity of COVID-19? Aside from the effect of the easy-transmission nature of this virus and the convenient transportation systems, how much danger are we putting ourselves in because of pollution?

1.1 Problem Description

Based on our literature review and the epidemic situations to date, we have selected two of the most severely influenced countries to study: Italy and the United States. We want to understand how air pollutants (particularly $PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, and O_3) impact the possibility of being infected by the disease. The data we have selected is focused on outdoor pollution because more systematic data-sets are available. Parameters and data used in our research are measured at county level for the US and region level for Italy. Caution needs to be taken when measuring air pollution and its impact. We need to make sure the selection of area units are both minimised within the area exposure and maximised between the area exposure variability [5]. County level is chosen for US because it distinguishes the meaningful boundaries between urban, suburban and rural areas. Additional relevant public data is available. The regional choice of level for Italy is primarily limited because of the lack of availability of data.

The complexity of the impact of air pollution on viral infection is partially related to social structure and can also

be influenced by the weather conditions and geo-spatial features of a certain region. We have included several of the social factors such as personal income and population density. A full list of variables can be found in Table 1 and Table 2. We have included these parameters in an attempt to provide a clear and broad image of these influencing factors. Policy-making can also easily be adjusted based on the information in our model.

We develop a spatio-temporal Bayesian framework, which is capable of estimating the effect of different factors (including environmental, demographic and economic parameters) on the infection rate. We use the framework to obtain results based on data from the USA and Italy to demonstrate the short- and long-term effects of pollution and compare relative effects of different pollutants. Details of implementation can be found in Section 3 and in supplement materials.

1.2 Supplementary Materials

- Digital presentations on Dashboards: Dashboards of Italy and Dashboards of the US
- Data source are provided on Appendix on page 6
- Model Code is available on Github at BathAirSquad_NERCHackathonOne
- Comparison Model used is at PM-COVID [5].
- Summary presentation on Google Drive

2 Model Formulation

Bayesian models have been applied with considerable success in spatially modelling health data [6]. There are two main benefits of using Bayesian frameworks:

1. They provide probability distributions of model parameters, which unlike point value estimates, can be naturally used for hypothesis testing,
2. They provide a natural way for learning, integrating prior information/beliefs.

We denote the quantity of interest in region i and day t by a_t^i, and the average probability of infecting other people within a given time interval by p_t^i. In this study, a_t^i will denote either the number of COVID-19 infections, or the number of deaths resulting from the virus, from which the context will be clear. Moreover we assume that the amount of people infected with the virus in a given region is large enough so that we may assume that people moving between regions has a negligible effect on the number of people infected.

If the number of people infected by a given carrier is iid (independent and identically distributed) the number of people infected on the next day follows the Poisson distribution[6]:

$$a_{t+1}^i - a_t^i = \text{Poisson}(a_t^i p_t^i) \tag{1}$$

We expect the probability of infection to depend on several factors (e.g. population density, air density). To ensure p_t^i is positive we propose to model its value as:

$$\ln\left(p_t^i\right) = \theta_0 + \theta X_t^i + u^i + v_t + w_t^i \tag{2}$$

where θ_0 is intercept, θ is a vector of linear parameters, X_t^i is a vector of factors for a given region and day, u^i describes the spatial random effects, v_t the temporal random effects, and w_t^i the remaining random effects.

Given the data $\{a_t^i\}$ and factors $\{X_t^i\}$, and possibly prior information about the model parameters, we use Bayesian inference to find the posterior distribution of the model parameters θ_0, θ, u, v and w. Adopting a Bayesian approach not only provides an estimate of the effect of each parameter, but also furnishes credibility intervals that quantify the credibility of the resulting estimates.

The key modelling problem is to choose proper parameters. Using only air pollution can cause inaccuracies in the final results. Even if the correlation between air pollution and infection rate could be observed, it might not imply causation; e.g. if both are caused by a third factor. For example the high density population may both cause a given area to be highly polluted (e.g. because of car traffic) and cause a higher infection rate (because people might have more contact with each other). To avoid such a misleading correlation we include some third factors in the model as well.

Secondly, the parameters can be highly correlated, which causes their linear parameters to be highly sensitive to input data. For example as we observed high correlation between the concentration of NO, NO_2 and NO_x, only NO_2 is used to represent these factors (much more data is available for NO_2 than for all other NO_x).

Sometimes an overdispersion is observed when using Poisson model, i.e. the values have wider distribution than expected. In such cases a negative binomial distribution can be used, in which standard deviation is modelled using an independent constant. It is applied if the overdispersion appears, i.e. the values are more widely distributed than expected [7].

3 Framework implementation

The framework was implemented in R using the INLA (INtegrated Laplace Approximation) package. It provides approximations of a desired posterior distribution by numerical integration. Especially for multi-parameter models

the computation is much more efficient than the MCMC methods (such as Metropolis-Hastings algorithm).

The framework uses two data sets: the first with spatial data (for example population density for each region) and the second with spatio-temporal data (for example number of confirmed COVID-19 cases in given region on a given day). The current framework uses data sets for the United States for 3251 counties from the starting day on 22-02-2020. Our model provides the interface for the user to update the datasets through provided data preprocessing files (`Pollutant_data_preprocessing.R` and `INLA_dataset_preprocessing.R`). The former reads in air pollution monitor data from the US Environmental Protection Agency (EPA), then aggregates them and creates spatial visualisations. Its output is then read in by the second script. This script merges the output with other data sets obtained from US Census website, and CSSEGISSandData and PM_COVID GitHub repositories. These two scripts need to be performed **only** if the data sets need to be updated, otherwise one can directly use `INLA_Bayesian_framework`. The script structure is presented in Figure 1.

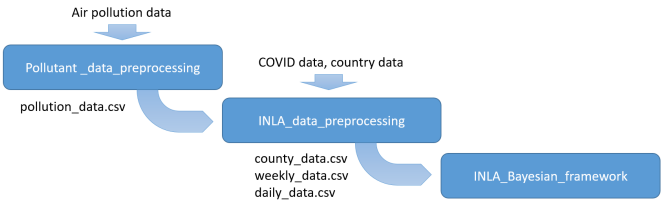

Figure 1: INLA framework file structure.

The code allows users to easily choose which variables to include in the model by choosing corresponding factor names in the `covariates` variable and providing the proper form of equation as variable `formula`. For the US counties-level model, five default modes are available. For example in model 1 we set:

```
covariates <- c("AQI_2weeks","popdensity")
formula <- new_cases ~ AQI_2weeks +
    popdensity + f(id, model = "iid") +
    f(day, model = "ar1") +
    f(rowId, model = "iid")
```

which means that the model uses the `AQI_2weeks` (average AQI from the last two weeks) and the `popdensity` (population density) parameters together with random effects corresponding to each county (`id`), each day (`day`) and remaining random effect different for each record (`rowId`).

The list of all parameters used for modelling US COVID-19 infection rates at county-level is attached in the Table 1 on Appendix A. Not all of the counties in the US have complete data in all of the parameters. Only those with covariates are used by the model. In particular, the EPA air pollution data of different pollutants is only available for a relatively small fraction of the total number of counties. Moreover the user can choose whether to use COVID-19 confirmed infection or deaths case data (by setting `useDeaths` to `FALSE` or `TRUE`, respectively). The former are available for longer days than the latter (especially early in the pandemic). However the former can be biased because they are limited by the number of tests each county was conducting per day. Data recording the number of deaths is usually considered to be a more reliable variable in epidemic models.

Users can also choose whether they prefer the daily or the weekly measure data (by setting `useWeekly` to `FALSE` or to `TRUE`, respectively). The daily data provides more accurate results. However, if the computation cost is a concern, the weekly data can be used to produce reliable results.

The third part that users can choose is whether to use the Poisson or the negative binomial distribution to describe number of new cases (see Section 2 for theoretical details).

4 Results

To demonstrate the potential of the developed framework, we conducted two experiments to investigate the effect of air pollution on the infection rate. In the first experiment we wanted to see whether and how the Air Quality Index (AQI) affects the infection rate. We expect this the effect to be two-fold. Firstly, the current pollution level may impact the people susceptible to virus infection. Secondly, there may be long term effects of living in polluted environment that decreases the immunity, e.g. a higher chance of developing asthma. To differentiate between these two scenarios we build four models with pollution averaged over different time intervals. For short-time effect we averaged the pollution over 2 weeks[1] (`AQI_2weeks`) and 2 months (`AQI_2months`) for the given data record. For long-time effect we took the average pollution in 2019 (`AQI_2019`) and 2016-2019 (`AQI_4years`). Additionally we included population density (`popdensity`) for the reasons discussed in 2.

The results obtained by this framework are presented in Figure 2. Here only figures for 2019 are presented because results for 2016 to 2018 are similar. The top two graphs shows the 95% credibility interval for the value of linear parameters corresponding to each of the covariates. On the rightmost graph we clearly see that linear parameter for `AQI_2019` is strongly positive, which shows that high average AQI (corresponding to high pollution) correlates positively with high infection rate. The effect of different values of AQI is additionally presented in the bottom graphs, where it is assumed that all other parameters (here only `popdensity`) stay constant level (equal to their mean value in the dataset). The continuous line represents the median of predictions, the dark area represent 50% credibility interval, light area 90% confidence interval and vertical dashed line the current mean AQI. For example the result

[1] Two weeks is the time during which the symptoms of COVID-19 develop.

shows that the AQI could most probably the most significant factor. If AQI could be decrease by 2%, the infection rate would decrease from originally 5% to 4%.

This correlation is however not significant for temporary (short-term) pollution exposure. Neither 2 weeks and 2 months did not correlate with the infection rate. This could mean that the impact of long-term consecutive exposure to high concentration of pollution is higher on assisting viral infection [1]. This does not mean that the short-term effects are negligible, as the confidence interval are very wide. Moreover the change of significance of population density indicates that there may be some factors not included in the model, which affects the results. Further analysis of additional factors and increasing data set size may increase the significance of obtained results.

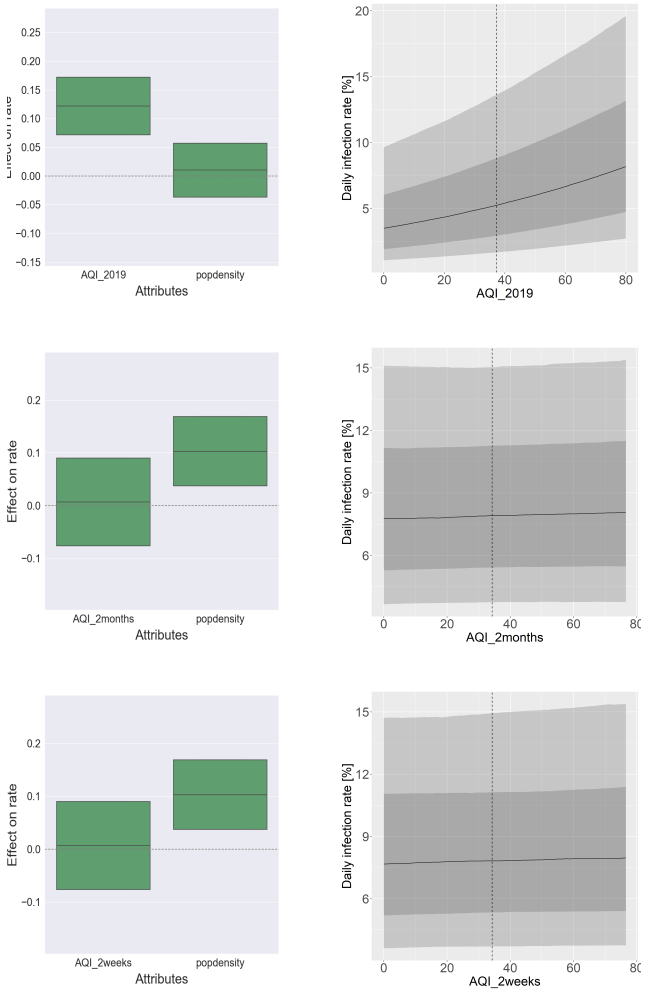

Figure 2: The effect of Air Quality Index (AQI) in different time scales: annual average (top row), two-month average (middle row) and two-week average (bottom row).

In the second experiment we analysed the combined effects of involving all those pollutants ($PM_{2.5}$, PM_{10}, NO_2, SO_2, CO, O_3). As input data we used the average value of measurements from monitor stations for 2019. As we can see in Figure 3, almost all pollutants seems to have a positive effect on the infection rate. However it cannot be confirmed with the 95% confidence interval (in all cases 2.5% quantile is below 0). The result can be improved either after updating the pollutant data (often monitoring data time series ended at February, March or April) or using a different source of air quality data (for example satellite images). The effect of ozone ozone (O_3), $PM_{2.5}$ and PM_{10} seems to be the highest and most significant[4] [1]. Unexpectedly the effect of CO is significantly negative, i.e. high concentration of CO corresponds to low infection rate. This observation is worth further exploration in models with different combinations of covariates, to exclude possible bias in the obtained results.

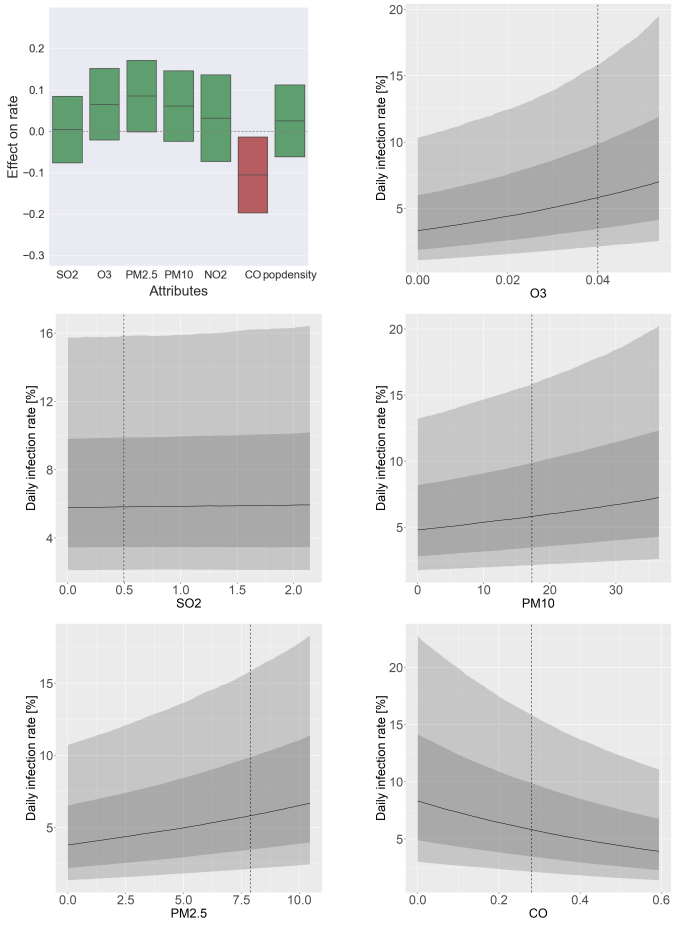

Figure 3: The effect of different pollutants on the infection rate.

5 Conclusions and Further Study

5.1 Conclusions

The framework we developed allows to examine the impact of air pollutants on the COVID-19 infection rate. It uses Bayesian inference, which allows to estimate the confidence intervals of parameters for a wide class of models. The

framework is flexible and allows user even with limited programming experience to easily build and analyse own models.

We successfully assembled data sets, which allow to conduct analysis for USA and Italy, but can be used of other countries as well. We used these to show that the air pollutant have significant long-term effect on infection rate. Moreover we estimate relative effects of six different pollutants. The ozone, $PM_{2.5}$ and PM_{10} seem to have highest effect, but this result need to be confirmed by further investigation.

Wide range of environmental, demographic, economic and state policy data, that we included in the framework may allow researchers to easily investigate much more factors potentially affecting the infection rate. These may be used to make potential suggestions to policy makers to protect human health for example by reducing pollution and save those associated unnecessary health expenditure

5.2 Limitations and Further Study

One extension of the current research is to combine the results of our model with an epidemic spreading model such as the SIR model, which uses the infection rate as an input model parameter. This would allow an estimation of the effect of air pollution on different epidemic scenarios and in turn potentially inform future environmental policies.

We used monitored air pollution data across counties in US and Italy. However, not every region provides in situ observations. One improvement we were planning to test is use the GEOS-Chem transport model to simulate a more accurate estimated pollution data. However due to the lock-down in UK, we could not run this simulation on our supercomputer to compare the result.

Another important model comparison that could be carried out is by running our model based on data from different countries and on different time scale. The impact of indoor pollution on COVID spread is also worth modelling, but such data as for now are not collected.

References

[1] Daniele Contini and Francesca Costabile. Does air pollution influence covid-19 outbreaks?, 2020.

[2] Frédéric Dutheil, Julien S Baker, and Valentin Navel. Covid-19 as a factor influencing air pollution? *Environmental Pollution (Barking, Essex: 1987)*, 2020.

[3] Jonathan Ciencewicki and Ilona Jaspers. Air pollution and respiratory viral infection. *Inhalation toxicology*, 19(14):1135–1146, 2007.

[4] Yaron Ogen. Assessing nitrogen dioxide (no2) levels as a contributing factor to the coronavirus (covid-19) fatality rate. *Science of The Total Environment*, page 138605, 2020.

[5] Xiao Wu, Rachel C Nethery, Benjamin M Sabath, Danielle Braun, and Francesca Dominici. Exposure to air pollution and covid-19 mortality in the united states. *medRxiv*, 2020.

[6] Moraga. *Geospatial health data: Modeling and visualization with R-INLA and shiny*. CRC Press, 2019.

[7] James G Booth, George Casella, Herwig Friedl, and James P Hobert. Negative binomial loglinear mixed models. *Statistical Modelling*, 3(3):179–191, 2003.

A Descriptions of input parameters

Here we listed all the variables and parameters we have used in our model.

parameter	description	year	source
id	county FIPS number	-	CSSEGISandData
state	state name	-	CSSEGISandData
county	county name	-	CSSEGISandData
population	county's population	2020	CSSEGISandData
area	county's area	2010	census.gov
AQI_2019	average AQI for 2019	2019	epa.gov
AQI_4years	average AQI for 2016-19	2016-19	epa.gov
SO2	concentration of SO2 [ppb]	2019	epa.gov
O3	concentration of O3 [ppm]	2019	epa.gov
PM2.5	concentration of PM2.5 [ug/m3]	2019	epa.gov
PM10	concentration of PM10 [um/m3]	2019	epa.gov
NO2	concentration of NO2 [ppb]	2019	epa.gov
NO	concentration of NO [ppb]	2019	epa.gov
NOx	concentration of NOx [ppb]	2019	epa.gov
CO	concentration of CO [ppb]	2019	epa.gov
incomePerCapita	the average personal income per capita	2007	census.gov
hospitalExpenditure	the local government expenditure on hospitals	2002	census.gov
healthExpenditure	the local government expenditure on health	2002	census.gov
poverty	the number of population live in poverty	2018	PM_COVID
popdensity	the population over the land area in km^2	2018	PM_COVID
medianhousevalue	the median value of owned houses measured in 1000$	2018	PM_COVID
pct_blk	the % of African-American minority	2018	PM_COVID
medhouseholdincome	the median value of household income	2018	PM_COVID
pct_owner_occ	the % of owner occupied housing	2018	PM_COVID
hispanic	the % of hispanic minority	2018	PM_COVID
education	the % with a lower than high school education	2018	PM_COVID
pct_asian	the % of Asian minority	2018	PM_COVID
pct_native	the % of native American	2018	PM_COVID
pct_white	the % of white American	2018	PM_COVID

Table 1: Parameters describing US counties (`county_data.csv`).

parameter	description	year	source
state	state name	-	CSSEGISandData
id	county FIPS number	-	CSSEGISandData
day	day id (days passed since 2020-01-22)	-	CSSEGISandData
previous_deaths	cumulative deaths in the last time step	2020	CSSEGISandData
new_deaths	new deaths in the previous time interval	2020	CSSEGISandData
previous_cases	cumulative confirmed cases in the previous time step	2020	CSSEGISandData
new_cases	new cases confirmed in the previous time interval	2020	CSSEGISandData
stateOfEmergency	state of emergency is declared (TRUE/FALSE)	2020	PM_COVID
closedSchools	schools are closed (TRUE/FALSE)	2020	PM_COVID
closedDayCares	day cares are closed (TRUE/FALSE)	2020	PM_COVID
bannedVistitorsToNursingHomes	visits in nursing homes are banned (TRUE/FALSE)	2020	PM_COVID
stayAtHome	stay-at-home policy in place (TRUE/FALSE)	2020	PM_COVID
closedBusinesses	non-essential businesses are closed (TRUE/FALSE)	2020	PM_COVID
closedRestaurants	restaurants are closed except take out (TRUE/FALSE)	2020	PM_COVID
closedGyms	gyms are closed (TRUE/FALSE)	2020	PM_COVID
closedMovieTheatres	movie theatres are closed (TRUE/FALSE)	2020	PM_COVID
frozeEvictions	Froze evictions (TRUE/FALSE)	2020	PM_COVID
freezingUtilityShutOffs	Order freezing utility shut offs in place	2020	PM_COVID
frozeMortgagePayments	mortgage payments are frozen (TRUE/FALSE)	2020	PM_COVID
waivedOneWeekPeriod	1-week waiver for unemployment insurance	2020	PM_COVID
AQI_2weeks	AIQ averaged over last two weeks	2020	epa.gov
AQI_2months	AIQ averaged over last two weeks	2020	epa.gov

Table 2: Parameters describing daily and weekly county data (`daily_data.csv` and `weekly_data.csv`).

The Table 3 provides all the publicly available data sources from Italy we used in the research.

	Data	Description
COVID data	Italian Civil Protection Department (ICPD, 2020)	02. 24 - 06. 4, 2020
Pollution data	Daily pollutant levels(NO2,PM2.5,PM10), in $\mu g/m^3$	2016 - 2019
Area per region	National Institute of Statistics (Instituto Nazionale di Statistica)	2020
Population	Instituto Nazionale di Statistica	2019
Map shapefile	Github repository	2020
Income and health expenditure	Instituto Nazionale di Statistica	2020

Table 3: Data sources used for modelling the Italian region of COVID-19 outbreak.