Inference of seasonal forcings on reproduction numbers during epidemics

Fengyang Wang, Lu Lu, Ziyao Tian, Yi Xiao Zhang

Abstract

Amidst the current COVID-19 epidemic, the prediction of its future spread and the length of the epidemic becomes increasingly significant. We investigate the effects of temperature and humidity on the reproductive rate R of COVID-19, as warmer climates in the summer have historically resulted in diminished spread of influenza. Early epidemiology research in COVID-19 have used simple linear models such as ordinary least squares regression to estimate the relationship between R and climate. We estimate the reproductive rate of the virus using a novel Bayesian inference model with MCMC sampling to daily case data of various climate regions. While results are largely inconclusive, they provide us with a weak posterior update that R decreases with temperature and relative humidity.

Introduction

Exponential growth of the number of symptomatic cases is usually observed in the initial stages of infectious disease epidemics. Typically, the rate of this growth is decomposed into two quantities:

- The serial interval: time between onset of symptoms of two adjacent individuals in a transmission chain.
- The reproductive number, R: the expected number of new infections a single symptomatic individual will cause.

Due to inadequate public data to compute the serial interval, we use the estimated serial interval of 4.7 days with standard deviation 2.9 days obtained from Nishiura et al. (2020) [1]. The estimation of the latter quantity is important for policy, because infectious disease outbreaks will switch from an exponential growth regime to an exponential decay regime when the reproductive number falls below 1. The goal of mitigation or suppression policies is therefore to effect change on the reproductive number R. Many infectious diseases are believed to be seasonal, including most respiratory viruses, like the SARS-CoV-2 virus implicated in the 2019–2020 coronavirus pandemic. Wang et al. (2020) [3] have analysed how the weather of locations within China correlates with local point estimates of the reproductive number R. However, because their analysis is frequentist in nature, their estimates of the seasonal forcing do not come with any estimate of uncertainty.

We first produce a baseline model, which predicts the reproductive rate using the EpiEstim package, based on previous work by Thompson et al. (2019) [2], and applied to COVID-19 data as in Abbot et al. (2020) [4]. We then apply ordinary least squares regression against climate variables as well as other underlying regional factors, such as population density, GDP per capita, and latitude, similar to the work produced by Wang et al. (2020)[3]. We compare this against our novel Bayesian framework, and evaluate the significance of the

explanatory variables in each model.

Since Italy was one of the first countries affected by the epidemic, it has a longer history of case data before intervention and lockdowns took place. Furthermore, due to the many provinces of Italy having diverse climates, we have access to a larger range of temperature and humidity. Therefore, we sourced the Italian government DPC dataset[5], which contains time series incidence for each province since mid February. Using the temperature and humidity data from weather stations sourced from NOAA, we are able to estimate the climate for over 80 distinct provinces in the Italian COVID-19 dataset.

Methodology

Thompson et al. (2019) outline a Bayesian framework for estimation of $R(t)$ using symptomatic case counts, decomposed into local and imported cases. The EpiEstim R package implements ideas from this paper. A simplified summary of their approach, ignoring the imported cases:

1. Estimation of (discrete) distribution of serial interval from line-list data with intervals reflecting date of infection.
2. Estimation of time-varying R from time-bucketed symptomatic case counts.

We developed a method to estimate the strength of various seasonal forcings on R through MCMC. This involves both substantial deviations from the second part of the Thompson et al. model and also less less substantial adjustments for the nature of our data. Our method does not extend the first part of the Thompson et al. model.

To sanity-check the results of our method, we compare its conclusions against a point estimate for the seasonal forcings through linear regression with the EpiEstim estimates of R.

Incidence likelihood of the Thompson et al. model

In its estimation of time-varying R, the Thompson et al. model uses the following likelihood function for the number of new infections at time t:

$$I(t)|I(1),\cdots,I(t-1),\mathbf{w},R(t) \sim \text{Poisson}\left(R(t)\sum_{s=1}^{t} I(t-s)w_s\right)$$

where

- $I(t)$ is observed data containing the number of new infections at time t
- $R(t)$ is a parameter for the reproduction number at time t
- w_s denotes the proportion of serial intervals which take on discrete value s, i.e. $w_s = \text{P}(\text{SI} = s)$.

Because the Thompson et al. model uses a gamma-distributed prior for $R(t)$ (the conjugate prior for a Poisson likelihood) and because they chose a convenient family of serial interval distributions, the Bayesian posterior in the Thompson et al. model can be described by a closed form analytic expression.

Extension of Thompson et al. for R subject to forcings

Our Bayesian model extends the method of Thompson et al. to multiple locations, each with its own incidence time series $I_1(t),\cdots,I_n(t)$ such that

$$R_i(t) = \exp\left(\theta(t) + \lambda_i + \beta^T x_i(t) + \varepsilon_i(t)\right)$$

where:

- $x_i(t)$ is data of observed seasonal factors at time t and location i, with each feature standardized to mean 0 and variance 1
- β is a parameter describing the locally linear effect of each seasonal forcing
- $\theta(t)$ is a parameter for logarithm of country-wide R independent of modeled seasonal forcings
- λ_i is a parameter for local variation of $\log R$ independent of modeled seasonal forcings
- $\varepsilon_i(t)$ is a parameter for remaining fluctuation in $\log R$ not modeled above

Unlike the Thompson et al., our analysis fundamentally estimates the logarithm of R instead of R itself. This has some practical advantages: a linear model for R could predict negative values for sufficiently extreme climate conditions, but negative values for R do not make sense epidemiologically.

We maintain a very similar incidence likelihood function as the Thompson et al. model, but impose an additional correction for importations:

$$I_i(t)|P_i, x_i(t), I_i(1), \cdots, I_i(t-1), \mathbf{w}, \theta(t), \lambda_i, \beta, \varepsilon(t), \gamma \sim \text{Poisson}\left(\gamma P_i + R_i(t)\sum_{s=1}^{t} I_i(t-s)w_s\right)$$

where in addition to the parameters described (with $R_i(t)$ given by the equation above), the following additional data and parameters are introduced:

- P_i is data giving the population of location i
- γ is a parameter estimating the nationwide rate of imports per capita per day, on which we impose prior $\gamma \sim \text{Gamma}(2, 10^{-5})$

The only substantative change to this likelihood function, beyond the decomposition of $R_i(t)$ into the various parameters influencing it, is the addition of importation risk. The reason for this adjustment is that the dataset we are using does not distinguish between importations and local cases. Thus, the Thompson likelihood would be 0 upon detection of the first case, had we not modeled the importations separately.

The most important feature of our model is the parameter β which are the coefficients of the linear model of $\log R$. This feature is assumed to be global across all locations and times, and thus models inherent properties of the contagiousness of the outbreak, as opposed to local deviations in space or time.

To avoid the model ascribing deviations in R to the seasonal forcings without widespread evidence across locations and times, three other features of our model are the parameters $\theta(t)$, λ_i, and $\varepsilon_i(t)$. The first parameter, $\theta(t)$, describes the logarithm country-wide R before adjusting for location and climate and local idiosyncratic factors. That is, it can be thought of as a baseline R value affected simply by the national circumstances at the given time; such as national policies or human behavioral changes. This will encapsulate various unmodeled behaviours such as weekends or interventions. At time zero, our prior belief is that this R will match the consensus of international estimates for R_0, which vary widely but tend to cluster between 2 or 3 [TK: Citation needed?]. Because of the great uncertainty, there is significant variance in this assumption, so the prior is quite weak (note that the $\mu \pm 2\sigma$ values, when transformed exponentially, are below 1 and above 6 respectively).

$$\theta(0) \sim \text{Normal}(\log 2.5, 0.5^2)$$

The remaining days of θ are modeled using a simple autoregressive model:

$$\theta(t+1) \sim \text{Normal}(\theta(t), \theta^\sigma)$$

where θ^σ is a parameter describing the spread of daily drifts in θ. The prior on θ^σ is weakly informative:

$$\theta^\sigma \sim \text{Gamma}(2, 0.01)$$

The location-specific deviation from the country average in $\log R$ independent of modeled seasonal forcings is encoded in λ_i. The existence of this parameter helps the model explain the large variations of apparent R across different locations, some of which appear to be affected much more so than others. The remaining parameter $\varepsilon_i(t)$ encapsulates all other unexplained deviations across both time and location. All these parameters (and β) are given the following weakly informative prior:

$$\lambda_i, \varepsilon_i(t), \beta_k \sim \text{Normal}(0, 0.5)$$

Our technique is general and can be used for any epidemic and any combination of potential seasonal forcings.

As the model is much more complex than the one in the Thompson et al. paper, we are unable to derive an analytic expression to compute the posterior. Instead, we use `rstan` to run 4 chains of MCMC for 2000 iterations each.

Using OLS and EpiEstim to check results

The baseline model we build estimates the reproductive rate using a parametric SI model, where we specify the mean and standard deviation of the serial interval. Because we do not have access to line-list SI data for COVID-19, we are unable to effectively estimate the serial interval from data, thus we take the findings of Nishiura et al. (2020) [1], which estimated the mean serial interval to be 4.7 and the standard deviation to be 2.4. We use the model detailed in Wallinga and Teunis (2004) [8] for estimating R from bayesian inference on a fixed serial interval distribution and time series data, which can be easily implemented using the EpiEstim package produced by Cori et al. (2013) [7].

For each province, we take a 7 day rolling sum on the number of new cases each day, and feed the data into the parametric SI model, which is given distribution $w(\tau)$ From there, the model uses a likelihood based approach to estiamte R at each discrete timestep, computing the likelihood that a person at timestep i is infected by a person infected at timestep j, and summing the likelihood for each j across all its infectees to compute the R value at timestep j. Thus, for each province, we produce a daily time series estimate for R on each day. This method is outlined in detail in Wallinga and Teunis (2004)[8].

For each province, we take a 7 day rolling sum on the number of new cases each day, and feed the data into the parametric SI model, which is given distribution $w(\tau)$ based on the SI parameters. From there, the model uses a likelihood based approach to estiamte R at each discrete timestep, computing the likelihood that a

person at timestep i is infected by a person infected at timestep j, and summing the likelihood for each j across all its infectees to compute the R value at timestep j. Thus, for each province, we produce a daily time series estimate for R on each day. This method is outlined in detail in Wallinga and Teunis (2004)[8].

For each province, day pair, we have $\bar{x} = (x_1, x_2, x_3, x_4), y$ where:

- x_1: standardized GDP per capita
- x_2: standardized population density
- x_3: standardized air temperature
- x_4: standardized relative humidity
- y: logarithm of estimated R

We then perform ordinary least squares regression to find $\bar{\beta}$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Data

We sourced the following data:

- Italian case data aggregated per province from the Italian Civil Protection Department (DPC)
- 2017 demographic data for Italian provinces from Eurostat, via the `restatapi` package
- Weather data from NOAA's network of stations, via the `worldmet` package

A total of 102 provinces were included in the analysis. This is all provinces except those located in the region of Sardinia. The provinces in Sardinia were excluded because of a mismatch between the demographic data and the case data, due to a recent political reorganization in Sardinia.

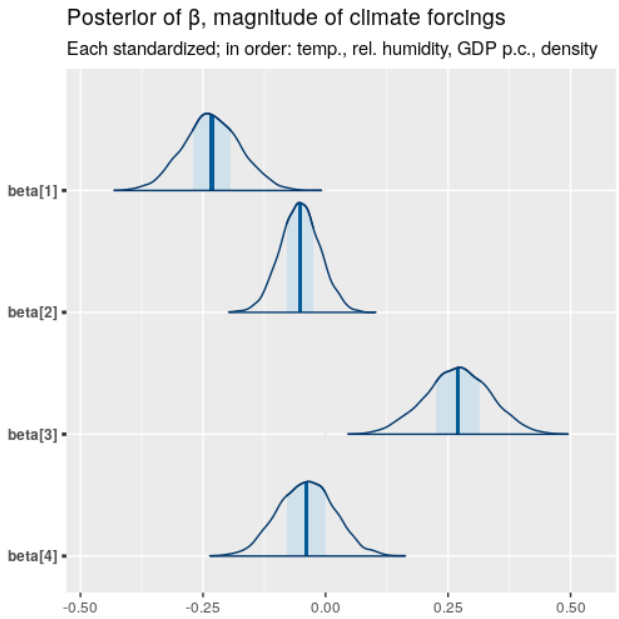
A total of 19 dates were included in the analysis, starting from the first date of February 24, 2020 present in the data and concluding on March 13, 2020. Although more dates are available, we felt that the dramatic interventions in Italy after March 13 reduced the relevance of seasonal forcings on transmission.

A total of 4 demographic and climatic features were included in the analysis: air temperature, relative humidity, GDP per capita, and population density. The demographic features are included as control so that the inferred coefficients for weather are not due to spurious correlations with demographics, as Northern Italy is generally denser, richer, and cooler on average than Southern Italy.

Results

Posterior Estimates

Applying the methods depicted from the methodology section, we obtain the following results.



Fitting our Bayesian model, which extends the method of Thompson et. al, values of beta were estimated for temperature, relative humidity, Gross Domestic Product (GDP) and population density, respectively. The mean estimate of β^T is

$$\beta^T = \begin{bmatrix} -0.232 & -0.054 & 0.263 & -0.036 \end{bmatrix}$$

in the overall model for R:

$$R_i(t) = \exp\left(\theta(t) + \lambda_i + \beta^T x_i(t) + \varepsilon_i(t)\right)$$

Note that each feature's data is standardized to mean 0 and variance 1.

The model indicates that higher temperatures are observed with lower R values, and higher GDP is observed with higher R values, while relative humidity and population density have close to no effect on R. Because of the confidence intervals, we are fairly confident of the directional correctness of this relationship for temperature and for GDP per capita.

In comparison, the EpiEstim and OLS model produced beta values of:

$$\beta^T = \begin{bmatrix} -0.051 & -0.045 & -0.081 & -0.051 \end{bmatrix}$$

for temperature, relative humidity, Gross Domestic Product (GDP) and population density, respectively, with

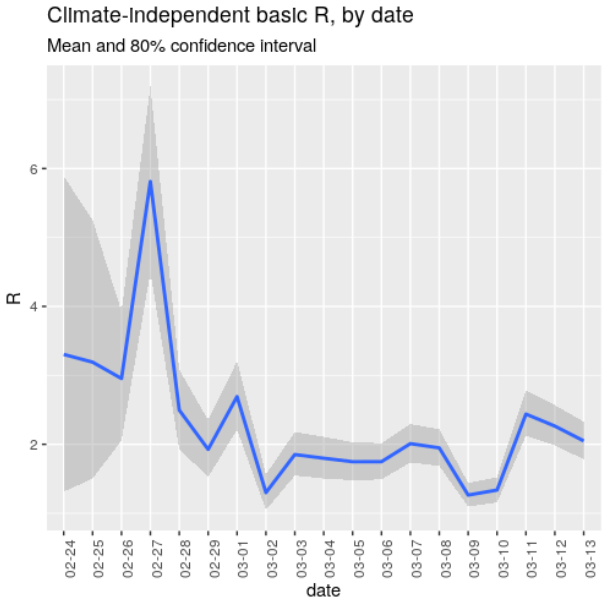
$$\beta_0 = 1.08802$$

in the overall model for R:

$$R_i(t) = \exp\left(\beta_0 + \beta^T x_i(t) + \varepsilon_i(t)\right)$$

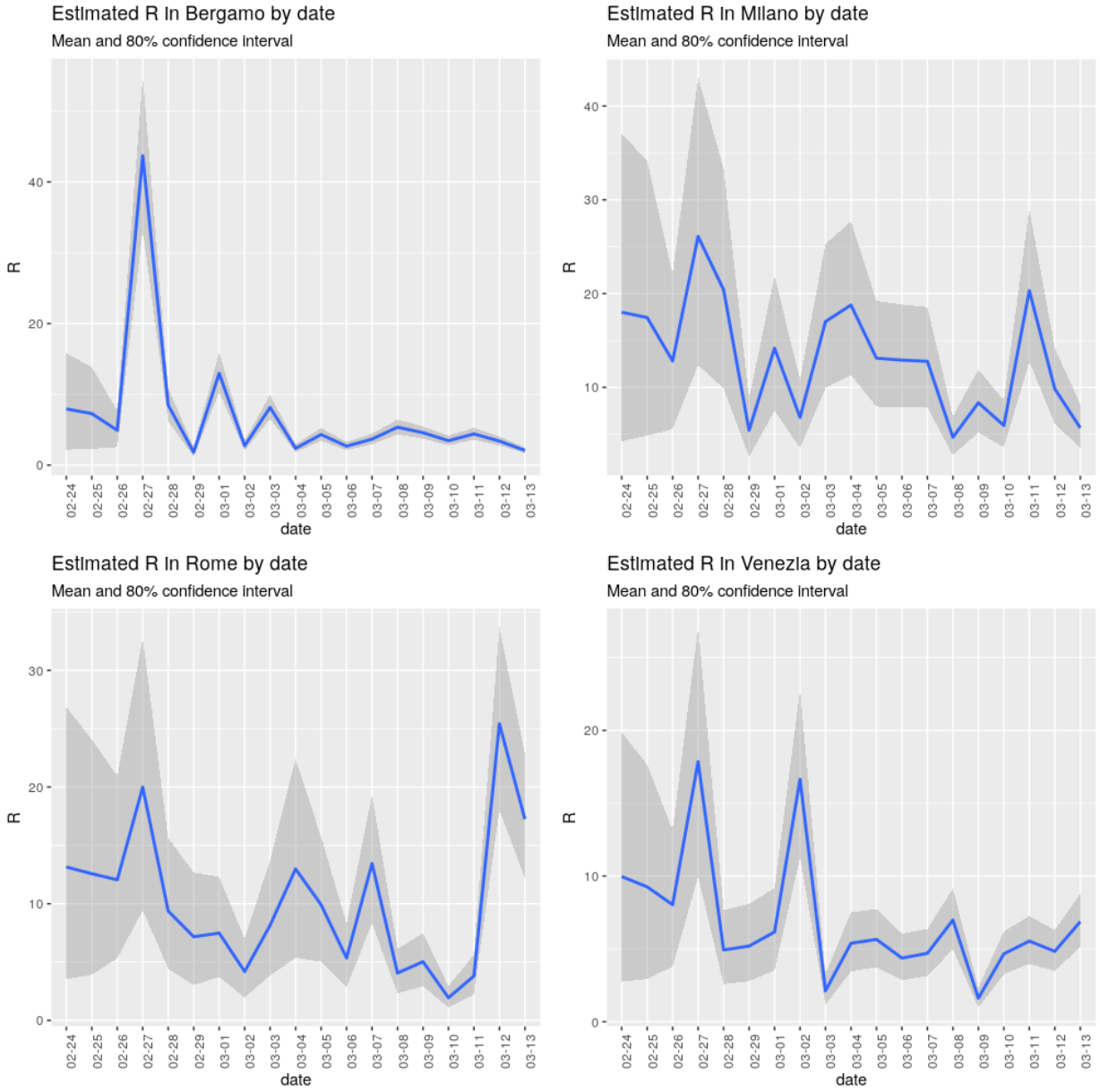
We note that the coefficient for GDP per capita has flipped signs from the Bayesian framework estimate, with the other coefficient remaining the same sign. The p-values for GDP per capita and density are 0.00042 and 0.019 respectively, which are both significant. However, the p-values for air temperature and relative humidity are 0.42 and 0.65 respectively, which are much less significant. The large discrepancy in values between coefficients between the two models is surprising, considering they are both based on a linear relationship between $\log(R)$ and the confounding varaibles. Further investigation would be required in the future to determine the cause.

Estimates of θ, transformed to climate-independent R by date



R is high in early phases, with an early spike on February 27, potentially due to increases in testing, which means the data (and hence the model) has underestimated the number of infectious cases compared to the ground truth. This result is aggregate across the entire nation and is net of any contribution to national weather changes.

Estimated R in Various Provinces



These plots show the estimated R values in each province computed through the samples from the MCMC. An important observation is that the confidence intervals for many provinces are very wide, because those provinces have relatively little data. However, one strength of our model is that no weighting on results from each province needs to be manually done. The data from larger provinces is automatically more important because the Poisson likelihood function returns smaller results as the rate parameter increases.

We have also analyzed the λ and ε parameters of the model, but because these are not as important, see the Appendix for plots and analysis.

In many cities, our estimated R_0 values are extremely high. This is because those cities reported a dramatic increase in cases, and our estimated R_0 values are indeed consistent with the data. However, we will discuss in the interpretation section the limits of our approach, and why we believe this may not match the ground

truth.

Interpretation

While the overall results are not surprising, some caution is important in the interpretation of these results. There are at least three sources of uncertainty not captured in the analysis:

- Uncertainty about the serial interval distribution (our method assumes the serial interval distribution is known)
- Uncertainty about timeliness of data (our model assumes cases are reported on the date of infection; in reality there is a significant delay, which the model does not attempt to adjust for)
- Uncertainty of the completeness of the data (our method assumes all cases are reported, which may not be true, as the high spikes in estimated R are unlikely to be ground truth; a more likely explanation is that early cases went undetected for a long time)
- Uncertainty about the accuracy and currency of data (our model assumes that temperature, relative humidity, GDP per capita, and population data are local and current, but in reality the demographic data is from 2017 and the weather data is measured by the closest station that may be in some cases not even located in the province)

Hence, it is likely that the uncertainty in posterior estimations of β by the model is understated. While the model provides some evidence that temperature is effective at reducing spread, we do not believe this evidence to be as strong as claimed.

Discussion

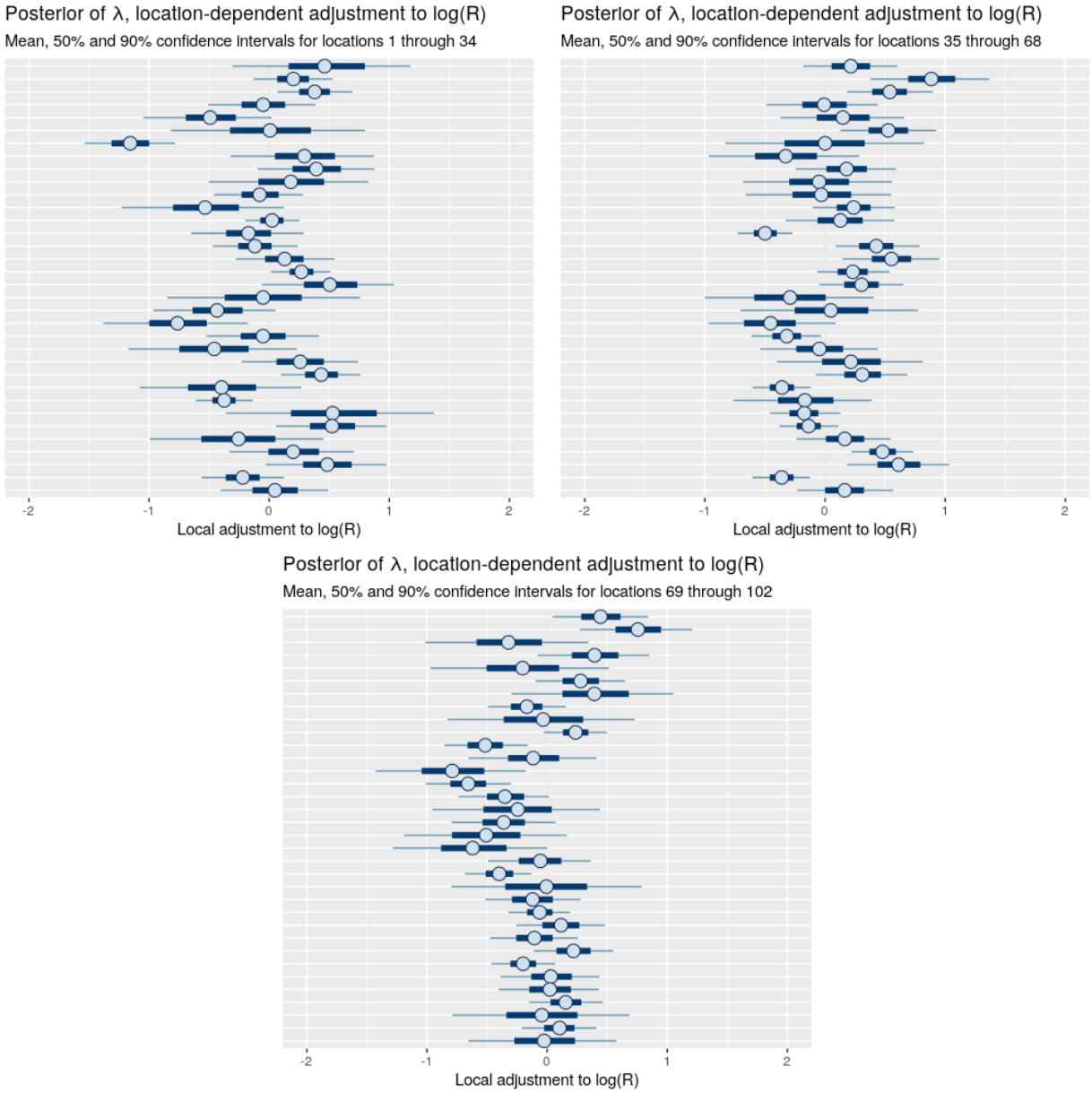
The posterior estimates of our model suggest that there is some weak evidence that with increasing temperature, the reproductive number R decreases. The effect of relative humidity on the reproductive number is much less conclusive. As we only used data from Italy in our model building, there is a limited range of weather information than we would like to have. Our model does a good job estimating the reproductive number R in various provinces with high confidence level and provides valuable information for policy makers. However, with no access to line-list data to estimate the serial interval number ourselves, our estimated reproductive number R is limited by the accuracy of the serial interval number we used. The estimated R values tend to have wide confidence interval in the beginning dates due to the imported cases in the beginning and the lack of testing.

For future research, we would like to look into the climate and infection data in different countries as the situation develops further. In particular, we would like to compare this with the results in Italy, and observe if the linear assumption between seasonal factors and R holds true across different climate regions. If not, we would like to try fitting a nonlinear relationship between R and seasonal factors.

We would also like to look into statistical simulation models which consider intercity travel and geographic proximity. As our estimated R is dependent on the serial interval number, it would also be worthwhile to see the sensitivity of our model with different values of serial interval number as did Flaxman et al. (2020)[6].

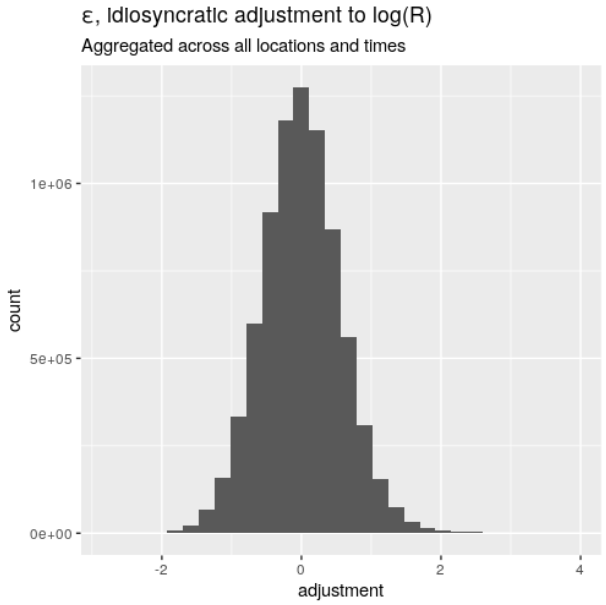
Appendix

Posterior of λ, location-dependent adjustment to $log(R)$



λ is an adjustment to the model based on the location. What we observe from the distribution of λ is that some locations have much more certainty in how they deviate from the mean than others. This can be explained by the unequal distribution of cases across provinces; provinces with more cases lend themselves to more accurate estimation of R.

Epsilon, idiosyncratic adjustment to log(R)



This histogram shows all samples of all entries in the ε matrix. Generally, we prefer a lower average spread of ε, because it means that more of the variation in reproduction number can be predicted from the seasonal factors, location, and time alone. Unfortunately, we found the observed spread of ε is quite high, indicating that there is substantial variance in reproductive numbers inferred from the data which cannot be explained by modeled features.

References

[1] Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (2019-nCoV) infections. medRxiv Published Online First: 2020. doi:10.1101/2020.02.03.20019497

[2] R.N. Thompson, J.E. Stockwin, R.D. van Gaalen, J.A. Polonsky, Z.N. Kamvar, P.A. Demarsh, E. Dahlqwist, S. Li, E. Miguel, T. Jombart, J. Lessler, S. Cauchemez, A. Cori. Improved inference of time-varying reproduction numbers during infectious disease outbreaks, Epidemics, Volume 29, 2019, 100356, ISSN 1755-4365, https://doi.org/10.1016/j.epidem.2019.100356.

[3] Wang, Jingyuan and Tang, Ke and Feng, Kai and Lv, Weifeng, High Temperature and High Humidity Reduce the Transmission of COVID-19 (March 9, 2020). http://dx.doi.org/10.2139/ssrn.3551767

[4] (PREPRINT) Abbott, S., Hellewell, J., Munday, J. D., Chun, J. Y., Thompson, R. N., Bosse, N. I., Chan, Y. D., Russell, T. W., Jarvis, C. I., ... & Funk, S. (2020). Temporal variation in transmission during the COVID-19 outbreak. CMMID. Available at CMMID: https://cmmid.github.io/topics/covid19/current-patterns-transmission/global-time-varying-transmission.html

[5] Presidenza del Consiglio dei Ministri - Dipartimento della Protezione Civile, Dati COVID-19 Italia, (2020), GitHub repository, https://github.com/pcm-dpc/COVID-19

[6] "Report 13 - Estimating the Number of Infections and the Impact of Non-Pharmaceutical Interventions on COVID-19 in 11 European Countries." Imperial College London, www.imperial.ac.uk/mrc-global-infectious-

disease-analysis/covid-19/report-13-europe-npi-impact/.

[7] Anne Cori, Neil M. Ferguson, Christophe Fraser, Simon Cauchemez, A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics, American Journal of Epidemiology, Volume 178, Issue 9, 1 November 2013, Pages 1505–1512, https://doi.org/10.1093/aje/kwt133

[8] Jacco Wallinga, Peter Teunis, Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures, American Journal of Epidemiology, Volume 160, Issue 6, 15 September 2004, Pages 509–516, https://doi.org/10.1093/aje/kwh255