

Global Short-Term Forecasting of Covid-19 Cases



Thiago de Paula Oliveira

thiago.paula.oliveira@insight-centre.org

<https://prof-thiagooliveira.netlify.com>

Rafael de Andrade Moral

rafael.deandrademoral@mu.ie

30th July 2020



NUI Galway
OÉ Gaillimh



**Maynooth
University**
National University
of Ireland Maynooth

Contents

1 Introduction and Motivation

- Social Impact
- Economic Impact
- Objectives

2 Data Acquisition

3 Methods

- Modelling framework
- Model Validation
- Clustering

4 Results and Discussion

5 Final Remarks

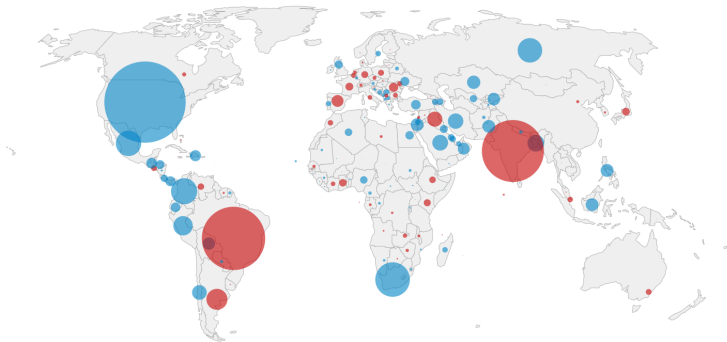
6 References

Introduction and Motivation

- The virus, which causes the respiratory infection Covid-19, was first detected in the city of Wuhan, China, in late 2019.
- Social and Economic Impact

Number of new cases, last 14 days

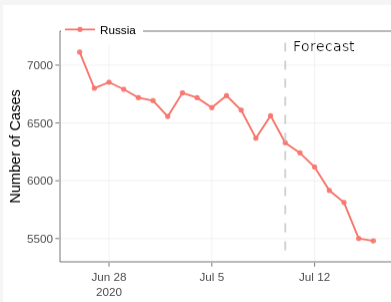
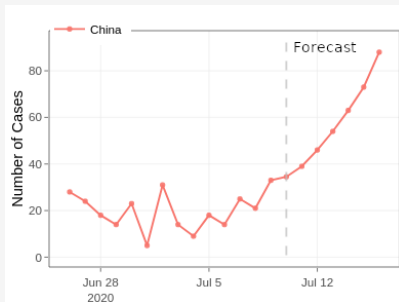
Countries where cases **rose** or **fell** last week, compared with the previous week



Source: Johns Hopkins University

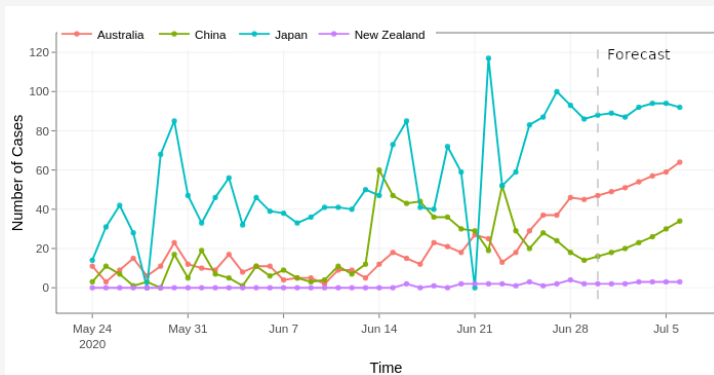
Introduction and Motivation - Forecast Models

- New disease with many factors acting in concert:
 - Spreading with great speed
 - Human behaviour
 - Government intervention/policies
 - Data quality - number of cases did not reflect correct numbers
- Forecasting with great accuracy under these circumstances is very difficult, and, consequently, would prove itself invaluable



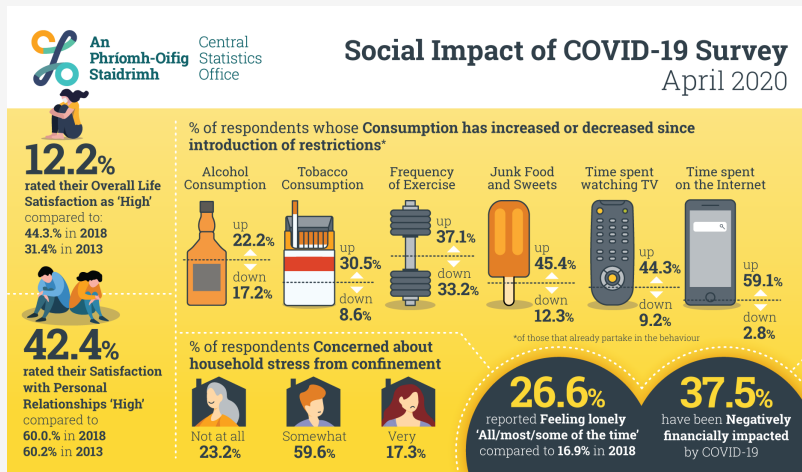
Introduction and Motivation - Forecast Models

- Short-term forecasts can give a good idea about the **trend** of the outbreak, and **can be crucial to assist planning**
- Could help low-and-middle-income countries/cities
- Develop **strategic planning** in the public health system to **avoid deaths**



Introduction and Motivation - Social Impact

- Impact COVID-19 has had on Irish society in April 2020
- Sample of 4,033 persons 18 years and over

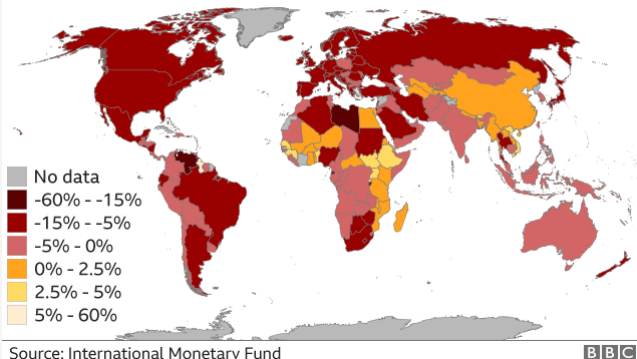


Introduction and Motivation - Economic Impact

- Outbreaks of the COVID-19 pandemic have been causing worldwide socioeconomic and health concerns

Majority of countries on the brink of recession

Real GDP growth, Q1 2020



- IMF says that the global economy will shrink by 3% this year

Objectives

- Propose a new modelling framework to handle with the behaviour of reported number of cases by country
 - State-space hierarchical model
 - Generate forecasts with very good accuracy for up to seven days ahead
- Propose clustering the countries based on the behaviour of their estimated autoregressive parameter over the last 60 days
- Provide all results as an R Shiny Dashboard
 - Including week-long forecasts for every country in the world
 - Point forecasts
 - Prediction intervals to express the uncertainty in the forecasts

Data and model access:

- https://github.com/Prof-ThiagoOliveira/covid_forecast

Data Acquisition

Data Acquisition

- European Centre for Disease Prevention and Control (ECDC)
- Data from the current day is removed, since it can be updated.

Other sites:
ECDC
European Antibiotic Awareness Day
ESCAIDE - Scientific conference
Eurosurveillance Journal
EVIP - Vaccination portal



European Centre for Disease Prevention and Control

An agency of the European Union

Home
All topics: A to Z
News & events
Publications & data
Tools
About us

Home > All topics: A to Z > Coronavirus > Threats and outbreaks > COVID-19 > Situation updates on COVID-19 > Situation update worldwide

< Situation updates on COVID-19

Situation update for the EU/EEA and the UK

Situation update worldwide

COVID-19 country overviews

Weekly surveillance report on COVID-19

COVID-19 situation update worldwide, as of 11 July 2020

Epidemiological update






The data presented on this page has been collected between 6:00 and 10:00 CET

Disclaimer: National updates are published at different times and in different time zones. This, and the time ECDC needs to process these data, may lead to discrepancies between the national numbers and the numbers published by ECDC. Users are advised to use all data with caution and awareness of their limitations. Data are subject to retrospective corrections; corrected datasets are released as soon as processing of updated national data has been completed.

Download today's data
How is the data collected?

Data Acquisition

Table: Data structure

	dateRep	cases	deaths	countries	popData2019	continent
1	29/07/2020	103	1	Afghanistan	38041757	Asia
2	28/07/2020	105	1	Afghanistan	38041757	Asia
3	27/07/2020	106	10	Afghanistan	38041757	Asia
4	26/07/2020	121	13	Afghanistan	38041757	Asia
5	25/07/2020	108	35	Afghanistan	38041757	Asia
6	24/07/2020	13	0	Afghanistan	38041757	Asia
:	:	:	:	:	:	:

- 1 Number of Rows: 32,850
- 2 Number of Columns: 6
- 3 countries is a factor with **210** levels

Methods

Modelling framework

- We introduce a class of state-space hierarchical models for overdispersed count time series

$$Y_{it} | Y_{i,t-1} \sim \text{NB}(\mu_{it}, \psi)$$

$$\log \mu_{it} = \gamma_{it} + \Omega_{it}$$

$$\gamma_{it} = \phi_{it} \gamma_{it-1} + \eta_{it}, \text{ with } \eta_{it} \sim \text{N}(0, \sigma_{\eta}^2)$$

$$\phi_{it} = \sum_{q=0}^Q (\beta_q + b_{iq}) P_q(t), \text{ with } \mathbf{b}_i \sim \text{N}_Q(\mathbf{0}, \Sigma_b)$$

$$\Omega_{it} = \lambda_{it} \omega_{it}$$

where $\lambda_{it} \sim \text{Bernoulli}(\pi)$ and $\omega_{it} \sim \text{N}(0, \sigma_{\omega}^2)$.

- When $\lambda_{it} = 1$, then observation y_{it} is considered to be an outlier, and the extra variability is modelled by σ_{ω}^2

- ϕ_{it} varying by country: more flexible autocorrelation function
- Iterating γ_{it} we obtain

$$\gamma_{it} = \left(\prod_{k=2}^t \phi_{ik} \right) \gamma_{i1} + \sum_{j=2}^{t-1} \left[\left(\prod_{k=j+1}^t \phi_{ik} \right) \eta_{ij} \right] + \eta_{it} \quad (1)$$

for $t = 3, \dots, T$.

- When $\phi_{it} = \phi_i = \beta_0 + b_{0i}$ (country-specific AR(1) process):

$$\gamma_{it} = \phi_i^{t-1} \gamma_{i1} + \phi_i^{t-2} \eta_{i2} + \phi_i^{t-3} \eta_{i3} + \dots + \phi_i \eta_{it-1} + \eta_{it} \quad (2)$$

- When $\phi_{it} = \phi_i = \beta_0$ (same autocorrelation parameter):

$$\gamma_{it} = \phi^{t-1} \gamma_{i1} + \phi^{t-2} \eta_{i2} + \phi^{t-3} \eta_{i3} + \dots + \phi \eta_{it-1} + \eta_{it} \quad (3)$$

Forecast future observations $y_{i,t+1}^*$

- Median of the posterior distribution of $Y_{i,t+1}|Y_{it}$.
- Reasonable for short-term forecasting
 - error accumulates from one time step to the other
- Seven days ahead



<https://revenue-hub.com>

Model implementation - Bayesian framework

The model is estimated using a Bayesian framework, and the prior distributions used are

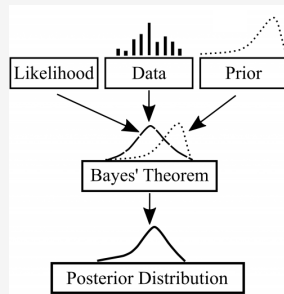
$$\beta_i \sim N_Q(\mathbf{0}, \mathbf{I}_Q \times 1000)$$

$$\sigma_{b_q}^{-2} \sim \text{Gamma}(0.001, 0.001)$$

$$\sigma_{\eta}^{-2} \sim \text{Gamma}(0.001, 0.001)$$

$$\sigma_{\omega}^{-2} \sim \text{Gamma}(0.001, 0.001)$$

$$\pi \sim \text{Uniform}(0, 1)$$



<https://medium.com>

- 3 MCMC chains
- 2,000 adaptation iterations
- 50,000 as burn-in
- 50,000 iterations per chain with a thinning of 25

Model Validation

- Last seven observation by country as test set




Training set



Test set

	Observed values						
Country A	y_{A1}	y_{A2}	\dots	y_{AT}	$y_{A,T+1}$	$y_{A,T+6}$	$\dots y_{A,T+7}$
Country B	y_{B1}	y_{B2}	\dots	y_{BT}	$y_{B,T+1}$	$y_{B,T+6}$	$\dots y_{B,T+7}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
Country N	y_{N1}	y_{N2}	\dots	y_{NT}	$y_{N,T+1}$	$y_{N,T+6}$	$\dots y_{N,T+7}$

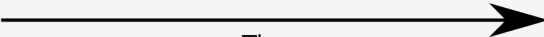

 Time

- Compared the forecasts with the true observations y_{it} for each day ahead;

- Concordance correlation coefficient (Lin, 1989)
- ρ - Pearson Correlation Coefficient (precision)
- C_b - bias corrector factor (accuracy)

$$\rho_t^{(CCC)} = 1 - \frac{E[(Y_t^* - Y_t)^2]}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \frac{2\sigma_{12}}{\sigma_1^2 + \sigma_2^2 + (\mu_1 - \mu_2)^2} = \rho C_b$$

	Observed		Forecast			Observed		Forecast		
	$\rho_{T+1}^{(CCC)}$		$\rho_{T+2}^{(CCC)}$...	$\rho_{T+7}^{(CCC)}$				
Country A	$y_{A,T+1}$	$y_{A,T+1}^*$	$y_{A,T+2}$	$y_{A,T+2}^*$		$y_{A,T+7}$	$y_{A,T+7}^*$			
Country B	$y_{B,T+1}$	$y_{B,T+1}^*$	$y_{B,T+2}$	$y_{B,T+2}^*$...	$y_{B,T+7}$	$y_{B,T+7}^*$			
⋮	⋮	⋮	⋮	⋮		⋮	⋮			
Country N	$y_{N,T+1}$	$y_{N,T+1}^*$	$y_{N,T+2}$	$y_{N,T+2}^*$		$y_{N,T+7}$	$y_{N,T+7}^*$			



Time

Clustering

- Aim: obtain sets of countries that presented a similar recent behaviour
- Last 60 values of the estimated autoregressive component

Clustering

- Aim: obtain sets of countries that presented a similar recent behaviour
- Last 60 values of the estimated autoregressive component

Dynamic time warp (DTW) - Muller (2007)

Let M be the set of all possible sequences of m pairs preserving the order of observations in the form $r = ((\hat{\gamma}_{i1}, \hat{\gamma}_{i'1}), \dots, (\hat{\gamma}_{im}, \hat{\gamma}_{i'm}))$.

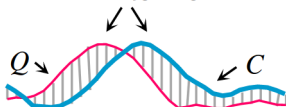
- minimise the distance between the coupled observations $(\hat{\gamma}_{it}, \hat{\gamma}_{i't})$

$$d(\hat{\gamma}_i, \hat{\gamma}_{i'}) = \min_{r \in M} \left(\sum_{t=1}^m |\hat{\gamma}_{it} - \hat{\gamma}_{i't}| \right).$$

- Recognise similar shapes in time series, even in the presence of shifting and/or scaling (Montero & Vilar, 2014).
- M is a set of all possible sequences of m pairs preserving the observation order in r

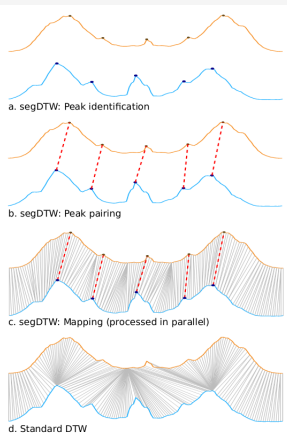
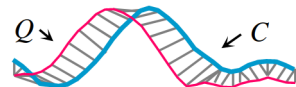
Clustering

Similar, but out of phase peaks ...



... produce a large Euclidean distance.

However this can be corrected by DTWs nonlinear alignment.



Rakthanmanon et al. 2012

- Picks the deformation of the time axes of $\hat{\gamma}_i$ and $\hat{\gamma}_{i'}$ which brings the two time series as close as possible to each other

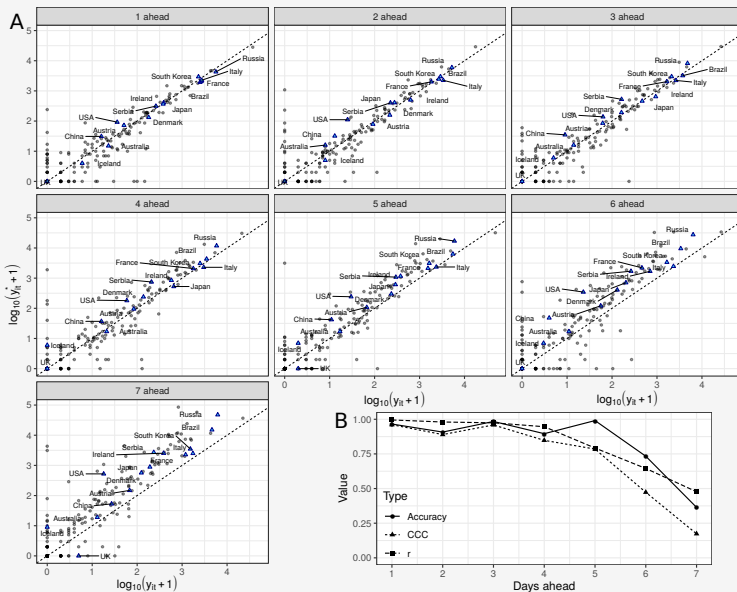
Clustering

Minimising the variability within clusters

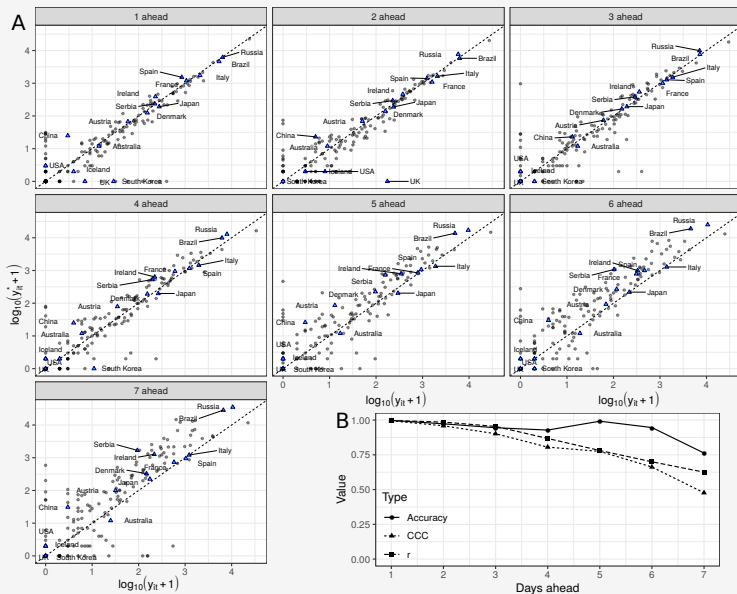
- Hierarchical clustering using the matrix of DTW distances using Ward's method (Murtagh & Legendre, 2014)
- Produce a dendrogram using hierarchical clustering analysis

Results

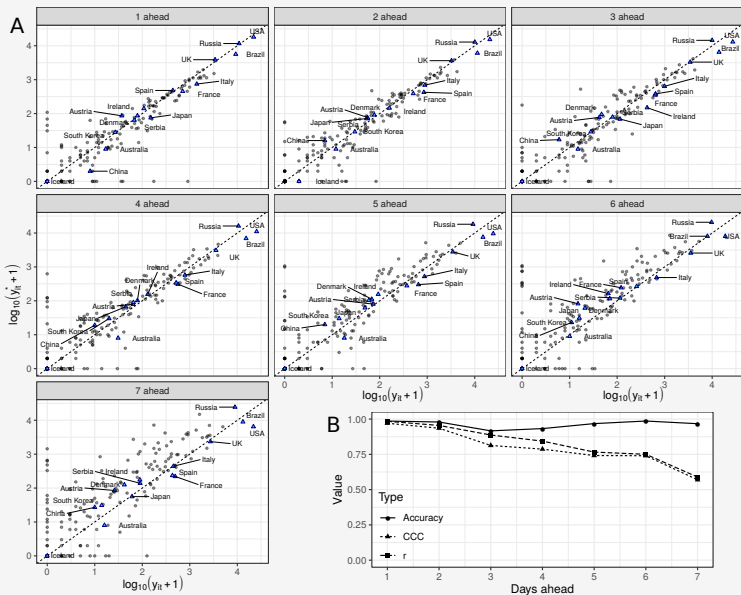
Predictive Performance - [30-April-2020, 06-May-2020]



Predictive Performance - [07-May-2020, 13-May-2020]



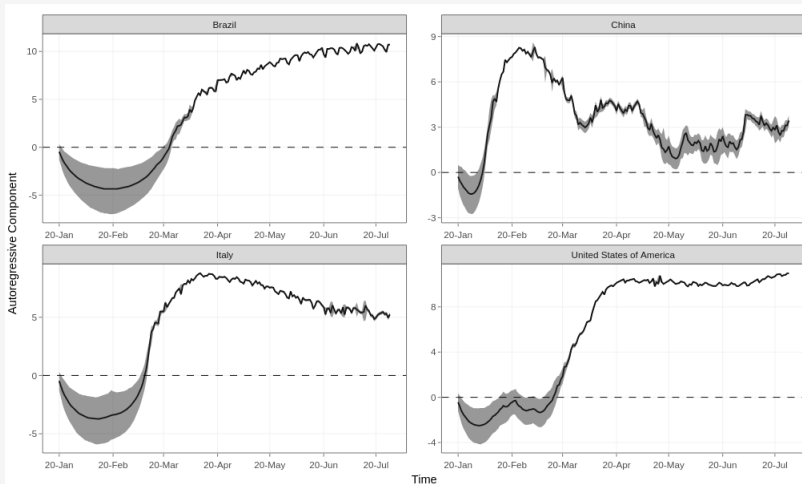
Predictive Performance - [14-May-2020, 20-May-2020]



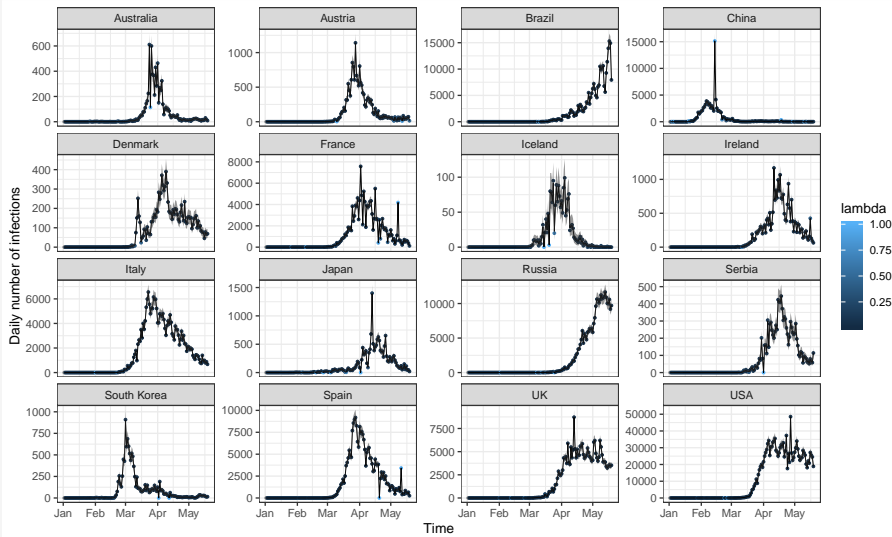
Autoregressive component - Pandemic behaviour

- Proportional to the natural logarithm of the daily number of cases

$$\gamma_{it} | \gamma_{i,t-1}$$



Overdispersion Parameter $\Omega_{it} = \lambda_{it}\omega_{it}$



- large number of accumulated suspected cases that were then confirmed

Estimates

Table: Parameter estimates and associated 95% credible intervals (CI) for the fitted autoregressive hierarchical state-space negative binomial model.

Parameter	Estimate	95% CI [lower; upper]
β_0	0.9993	[0.9974, 1.0012]
β_1	-0.1658	[-0.1871, -0.1447]
β_2	0.4090	[0.3630, 0.4080]
σ_{b_0}	0.0072	[0.0061, 0.0085]
σ_{b_1}	0.0325	[0.0168, 0.0565]
σ_{b_2}	0.2392	[0.1979, 0.2821]
σ_η	0.5206	[0.5071, 0.5341]
π	0.1080	[0.0983, 0.1180]
σ_ω	3.3797	[3.1391, 3.6544]
ψ	0.0009	[0.0002, 0.0025]

- about 11% of the number of reported cases can be viewed as contributing to extra variability

Final Remarks

Final Remarks

- We introduces statistical novelty in terms of modelling the autoregressive parameter as a function of time

Final Remarks

- We introduces statistical novelty in terms of modelling the autoregressive parameter as a function of time
- Translates directly into improved predictive power in terms of forecasting future numbers of daily cases

Final Remarks

- We introduces statistical novelty in terms of modelling the autoregressive parameter as a function of time
- Translates directly into improved predictive power in terms of forecasting future numbers of daily cases
- Model can be adapted to other types of data:
 - number of deaths
 - inclusion more variable in the linear predictor
 - obtain forecasts for smaller regions within a country

Final Remarks

- We introduces statistical novelty in terms of modelling the autoregressive parameter as a function of time
- Translates directly into improved predictive power in terms of forecasting future numbers of daily cases
- Model can be adapted to other types of data:
 - number of deaths
 - inclusion more variable in the linear predictor
 - obtain forecasts for smaller regions within a country
- Model will obtain forecasts based on what is being reported
 - direct reflection of the data collection process
 - be it appropriate or not

Final Remarks

- We introduces statistical novelty in terms of modelling the autoregressive parameter as a function of time
- Translates directly into improved predictive power in terms of forecasting future numbers of daily cases
- Model can be adapted to other types of data:
 - number of deaths
 - inclusion more variable in the linear predictor
 - obtain forecasts for smaller regions within a country
- Model will obtain forecasts based on what is being reported
 - direct reflection of the data collection process
 - be it appropriate or not
- Predictive models for large countries, such as the US, are even more problematic because they aggregate heterogeneous subepidemics in local areas

References

- Holmes, E. A. et al. Multidisciplinary research priorities for the COVID-19 pandemic: a call for action for mental health science. **The Lancet Psychiatry**, v.7, 547-560 (2020)
- Lin, L. I. A Concordance Correlation Coefficient to Evaluate Reproducibility. **Biometrics**, v. 45, 255-268 (1989)
- Muller, M. Dynamic Time Warping, v. 2, 69-84 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2007).
- Montero, P. & Vilar, J. A. TSclust: An R package for time series clustering. **J. Stat. Softw.**, v. 62, 1-43 (2014)
- Murtagh, F. & Legendre, P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? **J Classif**, v.31, 274-295 (2014)
- Rakthanmanon, T et al. Searching and Mining Trillions of Time Series Subsequences under Dynamic Time Warping. **ACM**, 2012, DOI: 10.1145/2339530.2339576