Data pooling mechanism for forecasting pandemic time series
+
Briefing on Spanish and Australian ‘Forecast Hubs’

Pablo Montero-Manso, University of Sydney

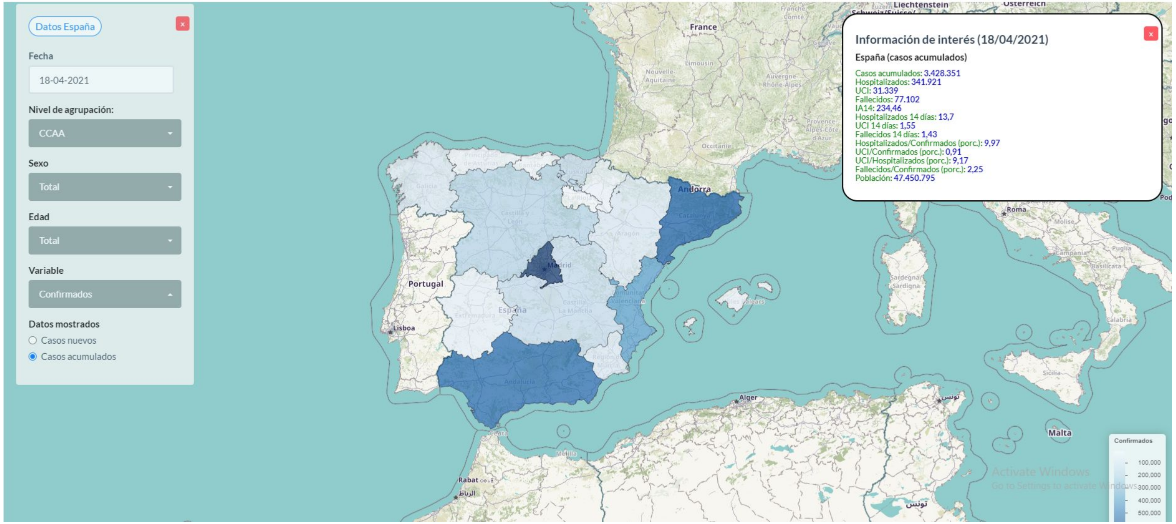
13 July 2021

https://covid19.citic.udc.es/

Mathematics against coronavirus

The Spanish Committee for Mathematics, CEMat, is promoting the initiative *Mathematics against coronavirus*. In this initiative, our goal is to use the analysis and modelling skills of our community in order to create a better understanding of the COVID-19 health crisis. Currently, the activities of this initiative include:

1. To collect links and contributions of the Spanish mathematical community about the virus spread on the website.
2. To promote discussion in the community using the contributions from researchers and groups, and involve a variety of models and techniques.
3. To establish a [Committee of Experts](#) to evaluate the collaborations and, eventually, will report conclusions and suggestions to the authorities.



Spanish Forecast Hub

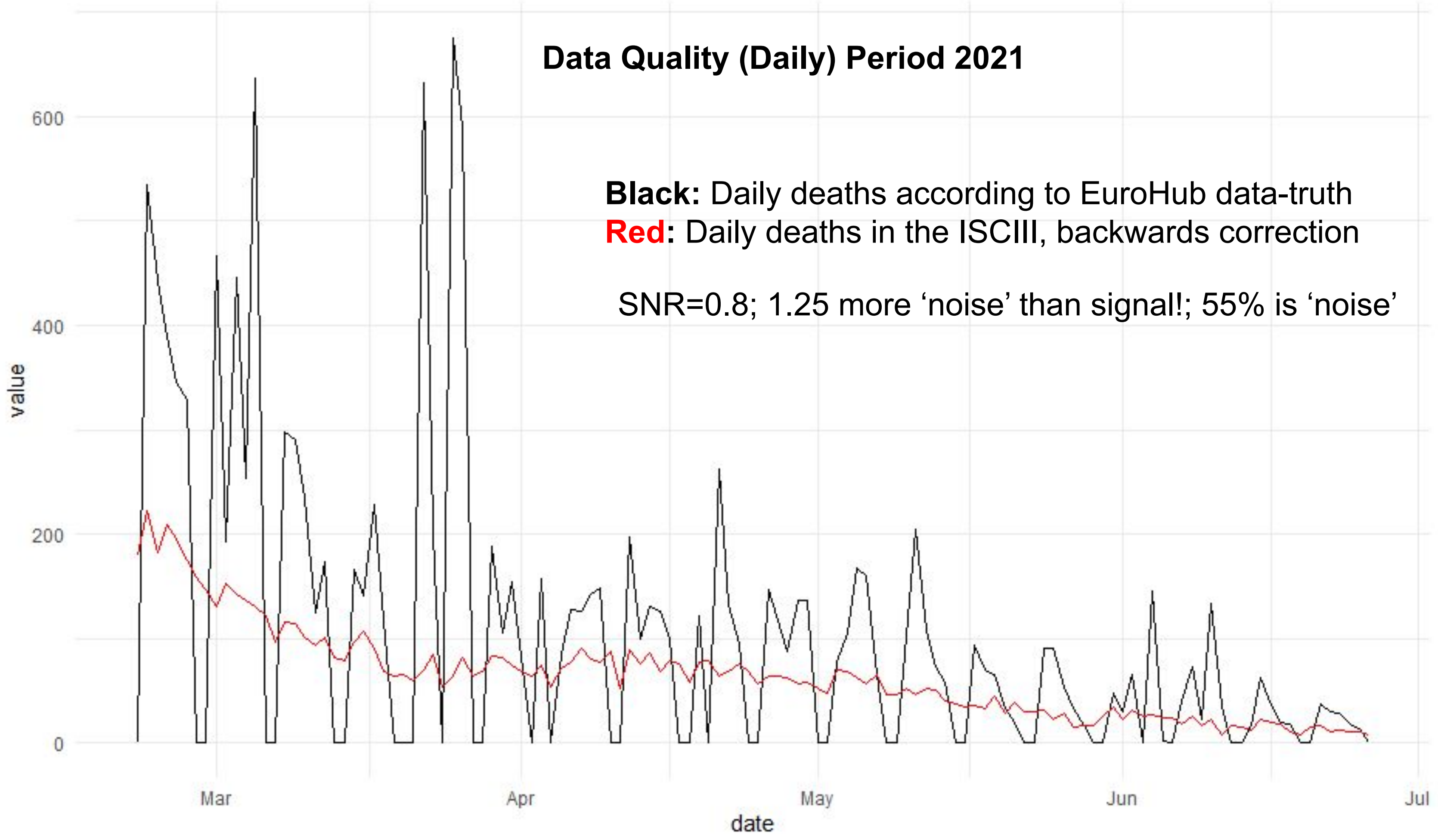
- Forecasts since April 1st 2020
- Cases, hospitalizations, ICU, deaths
- Every day, 7 days ahead
- State level and regions (19 regions x 4 variables = 76 time series)

- 46 teams submit regularly for *at least one* time series
- SEIR and variants, generalized regression, functional data, kernel smoothing, hidden markov, expert systems, bayesian, ML(random forests), time series, agent based

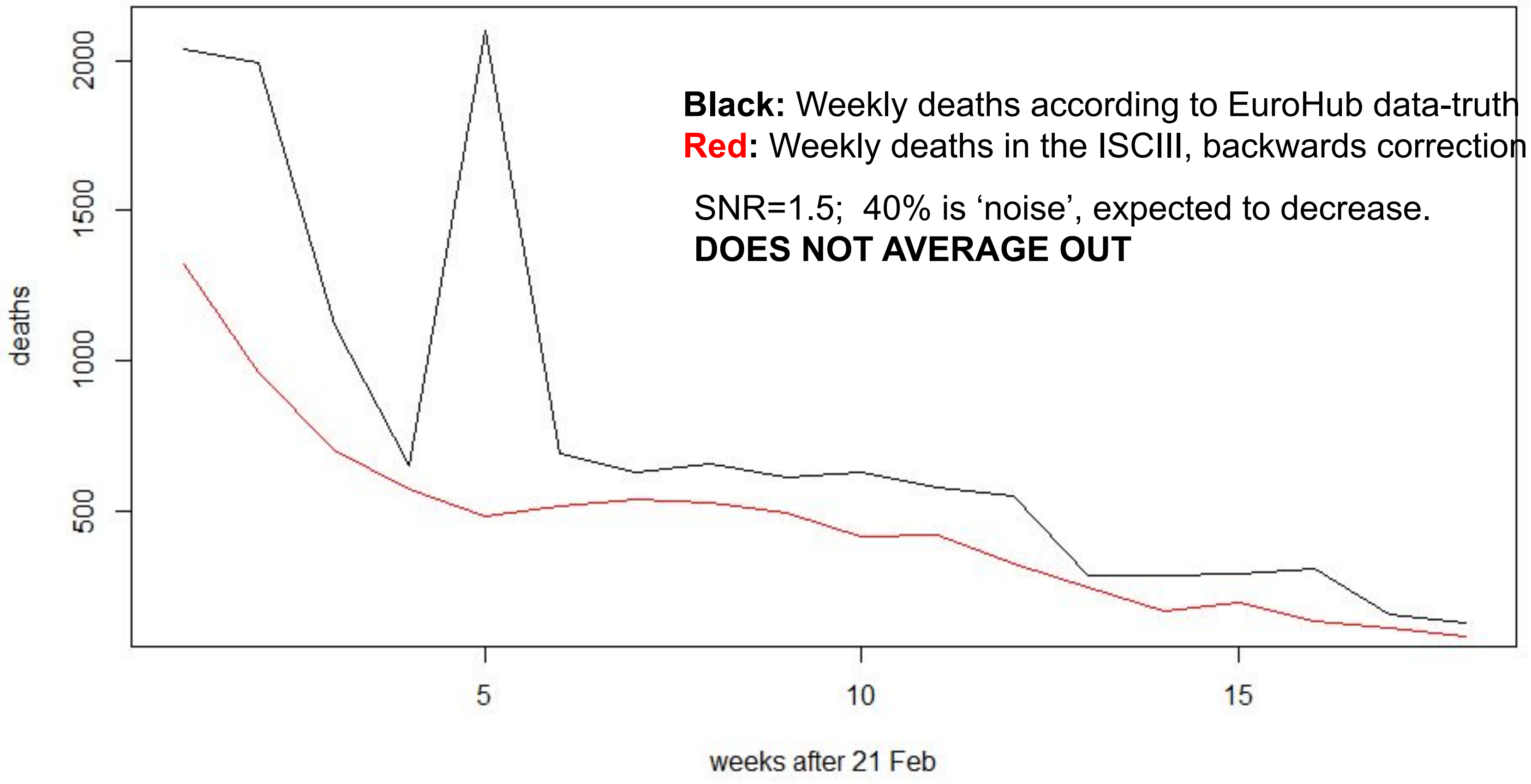
Results of the Ensemble

- **Basic Model combinations:** Simple mean, trimmed mean, winsorized mean, median
- **Weights based on past performance:** Bates-Granger, local smoothing.

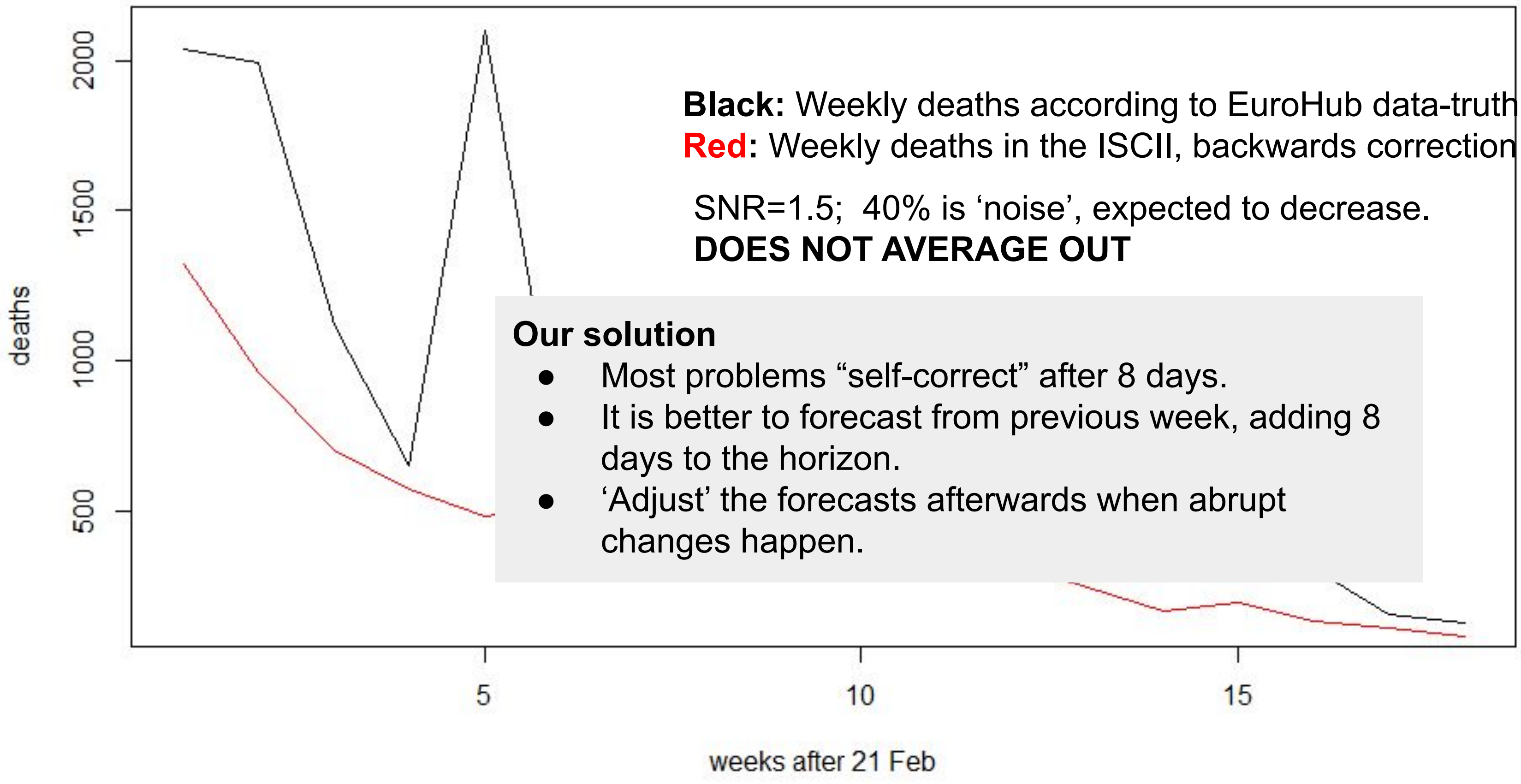
- The best were median, trimmed mean, winsorized mean
- Weight-based perform relatively well with a lot of fine tuning
- Simple Mean the worst performer because of ‘outlier’ forecasts


Data Quality (Daily) Period 2021
Black: Daily deaths according to EuroHub data-truth
Red: Daily deaths in the ISCIII, backwards correction
SNR=0.8; 1.25 more 'noise' than signal!; 55% is 'noise'
600
400
200
0
value
Mar
Apr
May
Jun
Jul
date

Data Quality (Weekly) Period 2021



Data Quality (Weekly) Period 2021



Black: Weekly deaths according to EuroHub data-truth
Red: Weekly deaths in the ISCII, backwards correction

SNR=1.5; 40% is ‘noise’, expected to decrease.
DOES NOT AVERAGE OUT

Our solution

- Most problems “self-correct” after 8 days.
- It is better to forecast from previous week, adding 8 days to the horizon.
- ‘Adjust’ the forecasts afterwards when abrupt changes happen.

Acción matemática contra el coronavirus

Comité Español de Matemáticas

Selecciona variable:

Confirmados

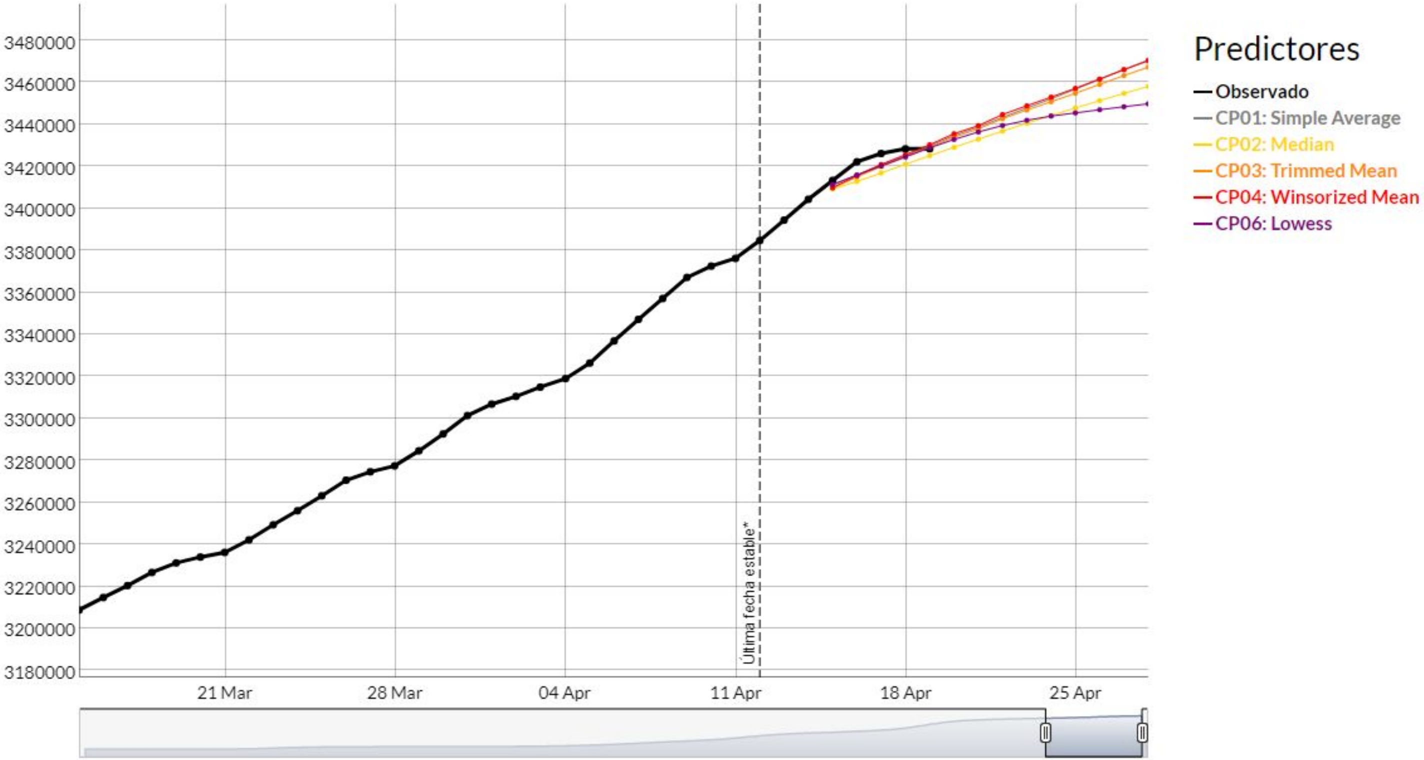
Comunidad autónoma:

España

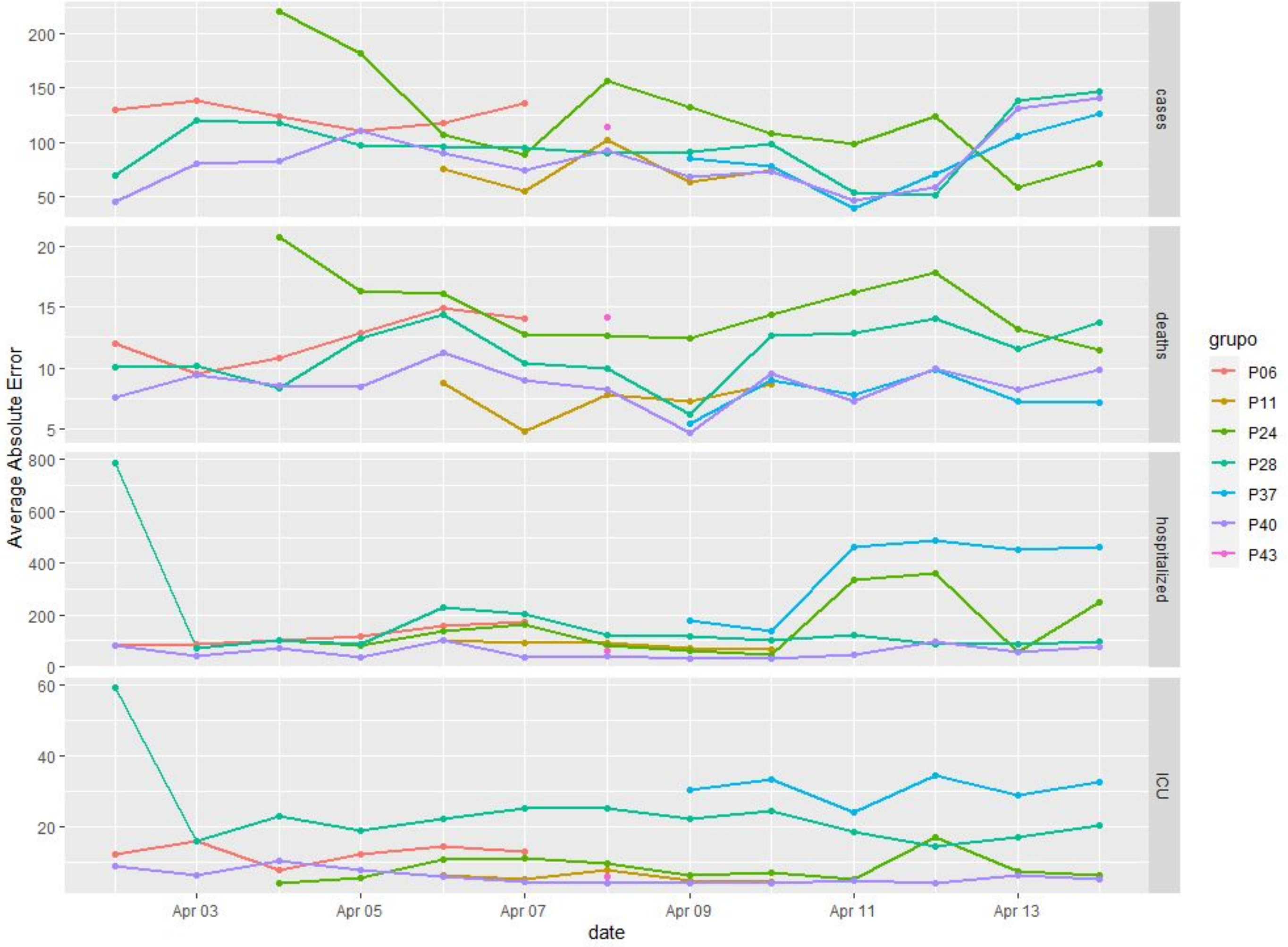
Fecha de predicción

15-04-2021

Gráfica Predicciones



errors during first wave 2020, (this presentation talks about P40)



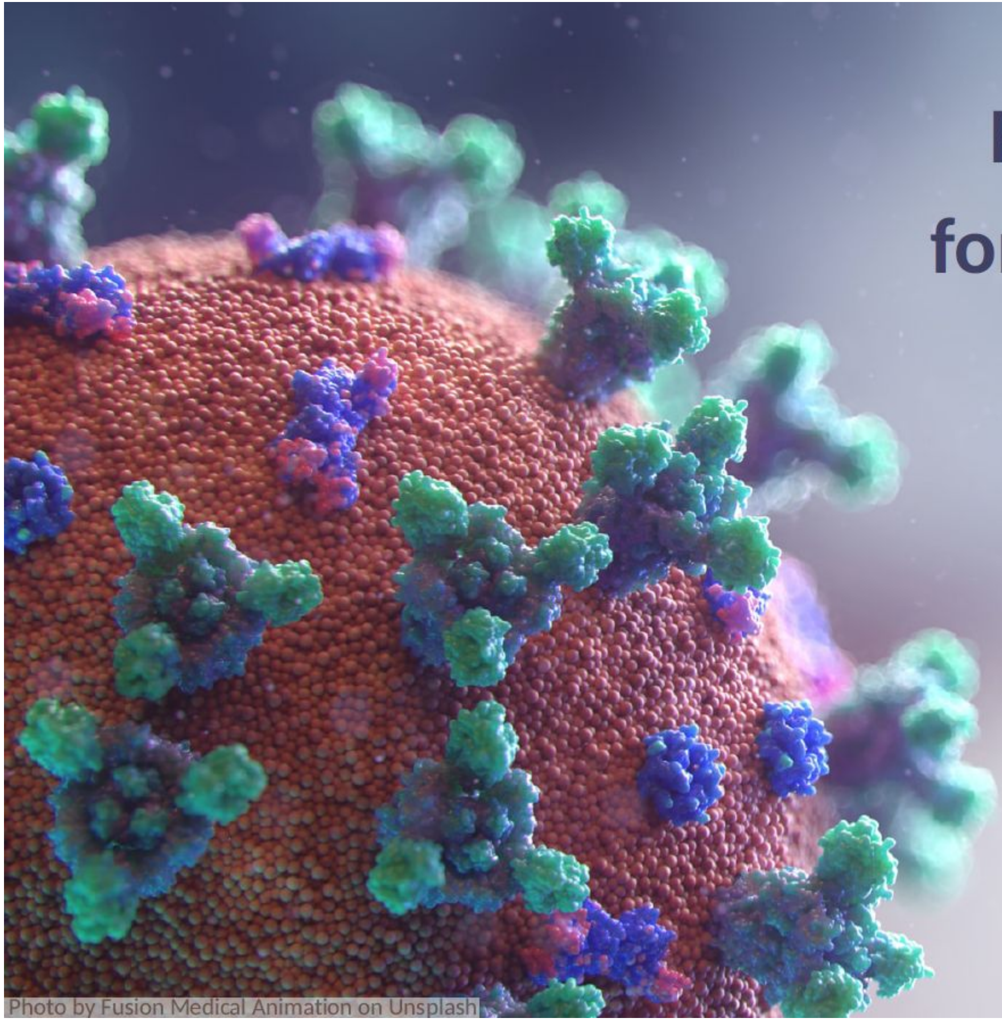


Photo by Fusion Medical Animation on Unsplash

Probabilistic ensemble forecasting of Australian COVID-19 cases

Rob J Hyndman

robjhyndman.com/covidtalk

MONASH University

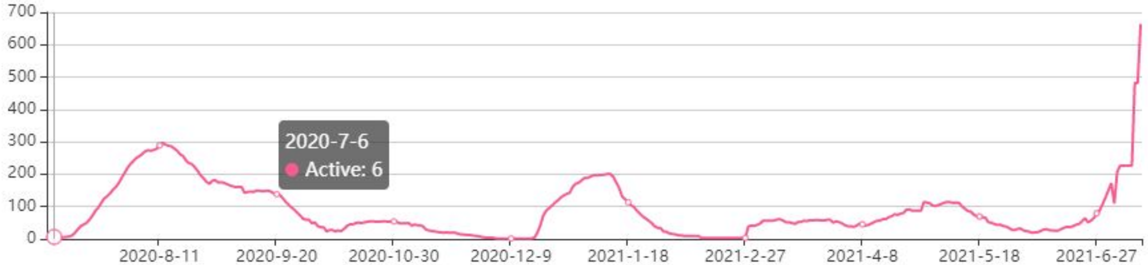
Australian Health Protection Principal Committee

The **Australian Health Protection Principal Committee** is the key decision-making committee for national health emergencies. It comprises all state and territory Chief Health Officers and is chaired by the Australian Chief Medical Officer.

COVID-19 forecasting group

- Peter Dawson
- Nick Golding
- Rob J Hyndman
- Dennis Liu
- James M McCaw
- Jodie McVernon
- Pablo Montero-Manso
- Robert Moss
- Mitchell O'Hara-Wild
- David J Price
- Joshua V Ross
- Gerry Ryan
- Freya M Shearer
- Tobin South
- Ruarai Tobin

Slide by Rob J Hyndman robjhyndman.com/covidtalk



ardian
by readers
Search jobs Sign in Search Australian editio

The Guardian
For 200 years

n Sport Culture Lifestyle More

politics Environment Football Indigenous Australia Immigration Media Business Science Tech

NSW Covid outbreaks: Gladys Berejiklian locks down Sydney, Central Coast, Blue Mountains and Wollongong

FINANCIAL REVIEW
PLATINUM 70 YEAR

Home Companies Markets Street Talk Politics Policy World Property Technology Opinion Wealth Work

Policy Economy Coronavirus pandemic

NSW lockdown will deliver $2b economic hit

NSW GOVERNMENT

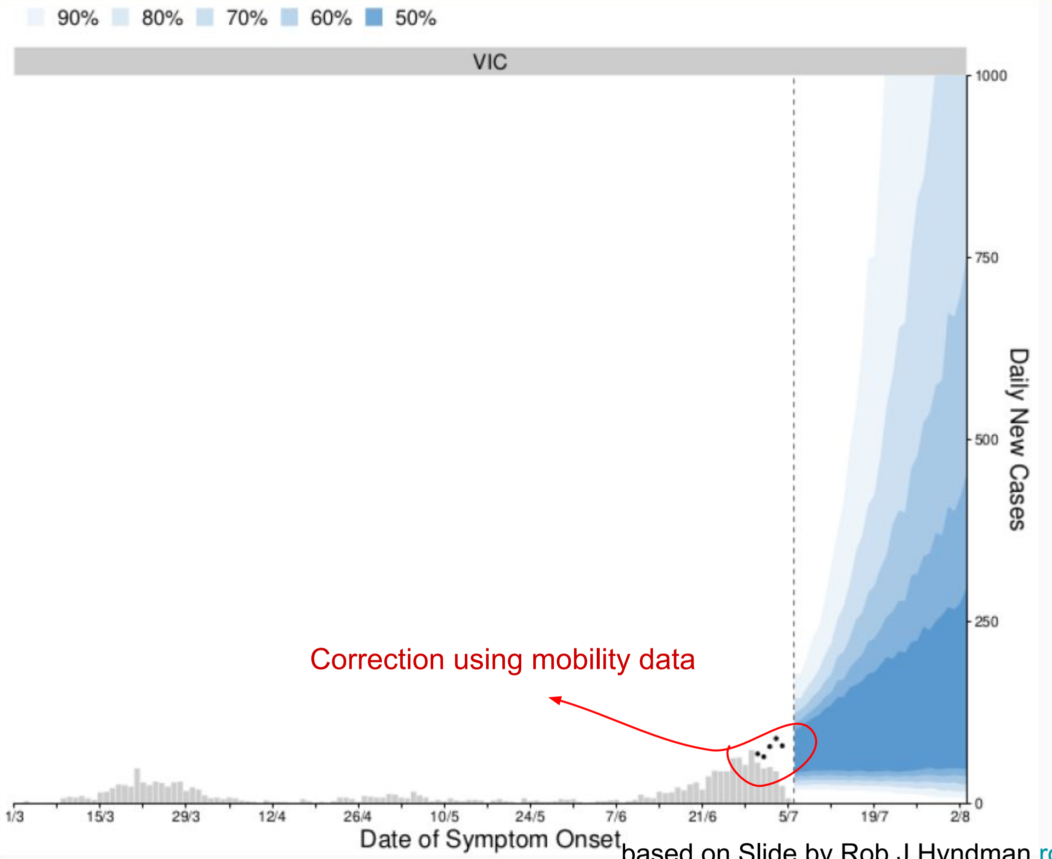
Living in NSW Working and business What's happening Have your say COVID-19

Home › Media releases › NSW lockdown extended until Friday, 16 July

NSW lockdown extended until Friday, 16 July

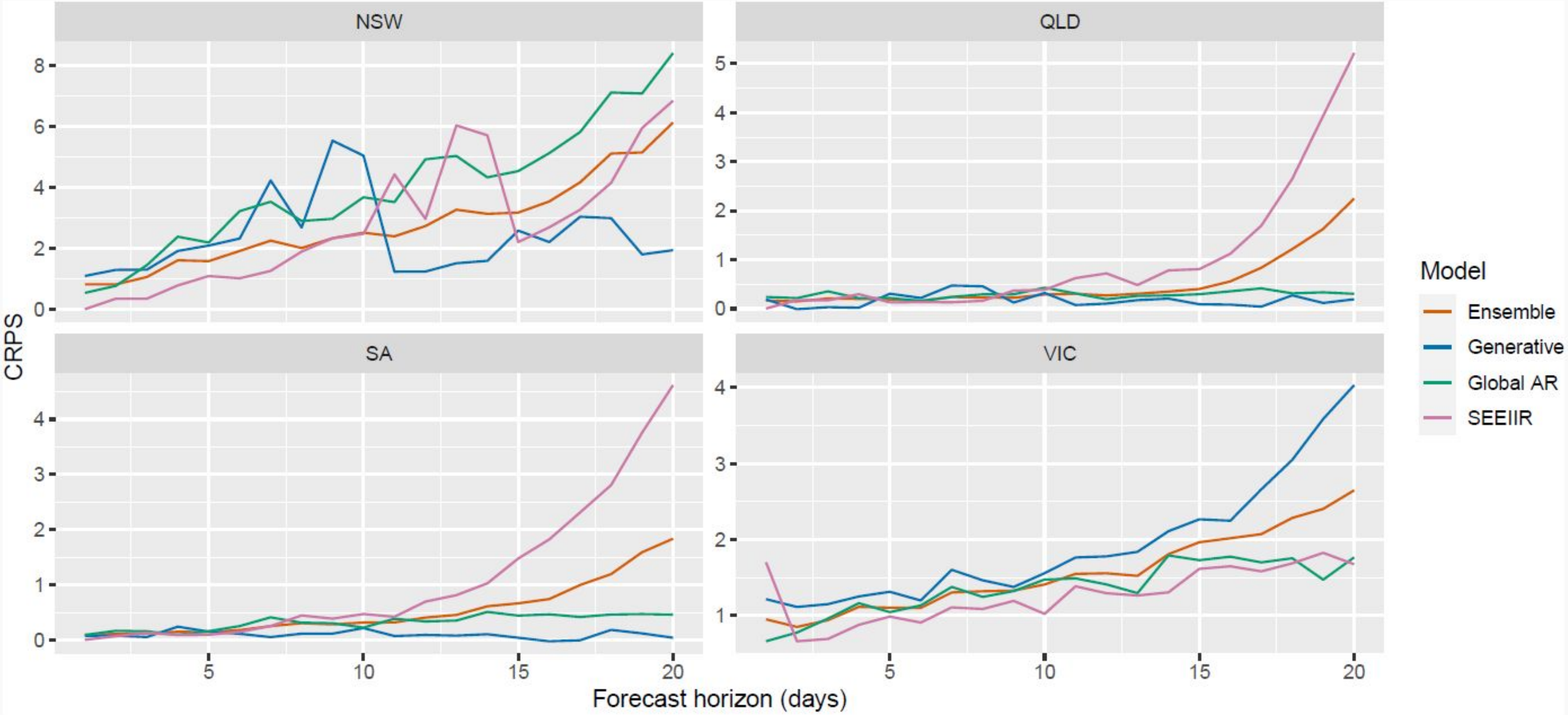
Published: 7 Jul 2021 · Released by: The Premier, Minister for Health and Medical Research

Ensemble forecasts: Victoria



based on Slide by Rob J Hyndman robjhyndman.com/covidtalk

CRPS: Continuous Ranked Probability Score



For weekly forecasts created from 17 September 2020 to 15 June 2021

Slide by Rob J Hyndman robjhyndman.com/covidtalk

What have we learned?

- Diverse models in an ensemble are better than one model, especially when they use different information.
- Understand the data, learn from the data custodians.
- Have a well-organized workflow for data processing, modelling and generation of forecasts, including version control and reproducible scripts.
- Communicating probabilistic forecasts is difficult, but consistent visual design is helpful.

Slide by Rob J Hyndman robjhyndman.com/covidtalk

Methodology

Methodology

- Statistical / Machine Learning Time Series model
- Autoregressive / Convolutions / Discrete time dynamical systems
- Time-delay Embedding: Dynamics are implicit in each time series

- Fundamental contribution is the **data pooling mechanism**

- Motivated by the similarity to growth curves

Data pooling

- Combine data from different sources to improve estimation
- Data ‘external’ to our specific problem

- Example: In the COVID pandemic, using data from another country/region to estimate parameters of the model for our region of interest.
- In early 2020, data from China to estimate transmission rate. Italy to estimate the effect of lockdowns, data from the UK to estimate the new variant, effect of vaccination, mortality...

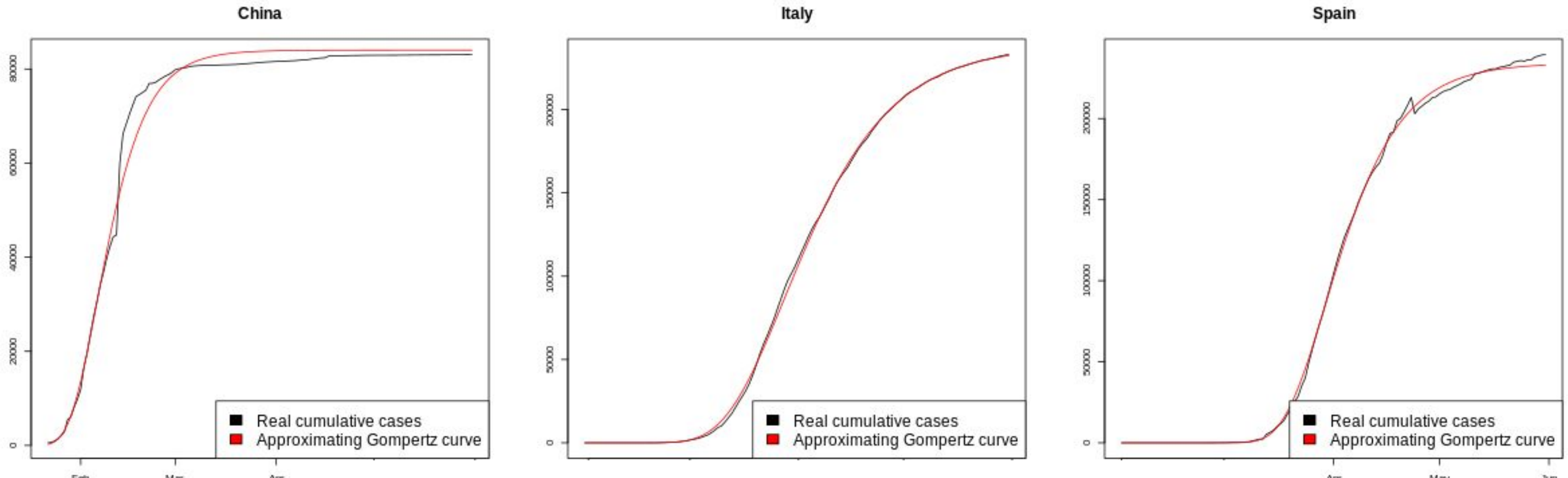
- Based on expert knowledge
- Underlying assumption of similarity, not always right but a positive tradeoff

Contribution

"A time series model that enables <u>perfect</u> data pooling from multiple growth curves."

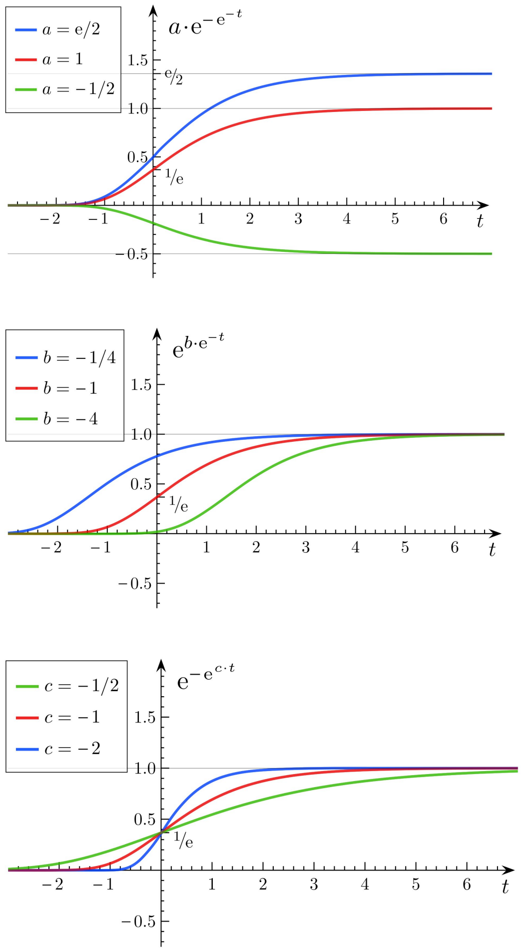
Growth Curves describe the evolution of a epidemic

- We use **Gompertz** for their **simplicity (explicit solution)** and **popularity**, results in this talk can be extended to SEIR Compartmental models (only numerical simulations)



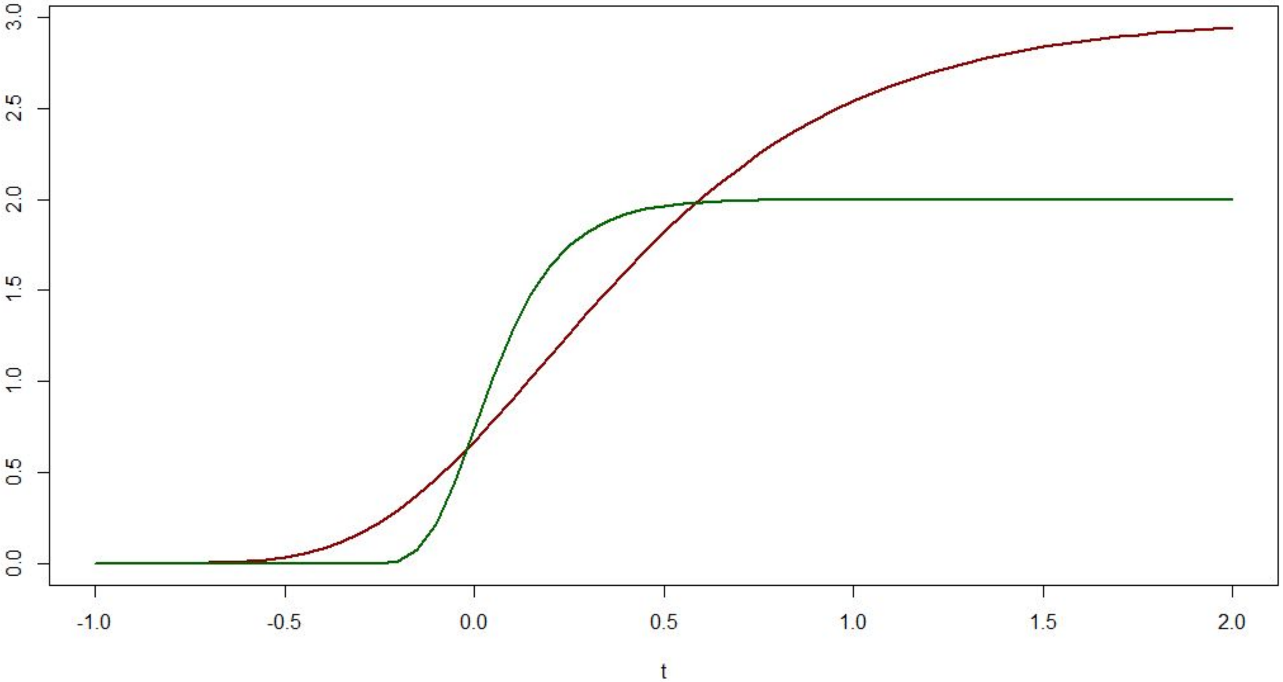
See fits for 73 countries: https://www.mdpi.com/819322

$$f(t) = Ae^{-Be^{-Ct}}$$



figures from wikipedia

Are these two curves equal?



- As Gompert curves, they differ on their values of parameters A, B, C
- We will show that they follow **the same process, with the same parameters**
- This **process** can be estimated from data **using both curves (2x more data!)**

STEP 1: Autoregressive parameterization of Gompertz

Unroll the process to express the value of **next time step as a function of current time-step**, **instead of as a function of time**

$$f(t) = Ae^{-Be^{-Ct}}$$

$$f(t+1) = Ae^{-Be^{-C(t+1)}} = Ae^{-Be^{-Ct}e^{-C}}$$

$$f(t+1) = A\left(\frac{f(t)}{A}\right)^{e^{-C}}$$

STEP 1.b: Linear Autoregressive of Gompertz

$$f(t+1) = A\left(\frac{f(t)}{A}\right)^{e^{-C}}$$

$$log(f(t+1)) = e^{-C} log(\frac{f(t)}{A}) + log(A)$$

A linear autoregressive model / ‘Convolution’ expresses the evolution of the logarithm of a Gompertz curve

$$l(t+1) = \alpha l(t) + \beta$$

Step 2: Autoregressive parameterization of **ALL** Gompertz

Unroll the process one more time, **solve the system for alpha and beta**

$$l(t+1) = \alpha l(t) + \beta$$
$$l(t+2) = \alpha l(t+1) + \beta$$

$$\alpha = \frac{l(t+1) - l(t+2)}{l(t) - l(t+1)} \qquad \beta = \frac{l(t)l(t+2) - l(t+1)^2}{l(t) - l(t+1)}$$

Step 2: Autoregressive parameterization of **ALL** Gompertz

Unroll the process one more time, substitute in alpha and beta

$$l(t+3) = \alpha l(t+2) + \beta$$

$$\alpha = \frac{l(t+1) - l(t+2)}{l(t) - l(t+1)} \quad \beta = \frac{l(t)l(t+2) - l(t+1)^2}{l(t) - l(t+1)}$$

$$\Downarrow$$

$$\boxed{l(t+3) = \frac{l(t+1)l(t+2) - l(t+2)^2 + l(t)l(t+2) - l(t+1)^2}{l(t) - l(t+1)}}$$

Parameters have disappeared: **ALL** Gompertz curves can be expressed in this form!

Keeping it purely linear

- A similar result with only linear autoregressive models (‘convolutions’)
- Less powerful but linear models are ‘well behaved’ and ‘interpretable’
- The idea is to ‘unroll’ even more (add lags, increase the order of autoregression). By overparameterization, we get many solutions for the system. Some of these solutions will be the same for different curves.

$$l(t+2) = \alpha l(t+1) + \beta$$
$$l(t+2) = \mathbf{a}l(t+1) + \mathbf{b}l(t) + \mathbf{c}$$

- There is a linear model of lag 1 for each Gompertz. Then there is a linear model of lag 2 for *any pair* of Gompertz.
- In general, we will need as much lags as curves we want to pool, which is bad
- **In practice, we will need only a few lags to approximate many curves**

This is the parameterization we have shown:

$$l(t+2) = \alpha l(t+1) + \beta$$

We can express it this way, in fact many such **a**, **b**, **c** will satisfy the equation

$$l(t+2) = \mathbf{a} l(t+1) + \mathbf{b} l(t) + \mathbf{c}$$

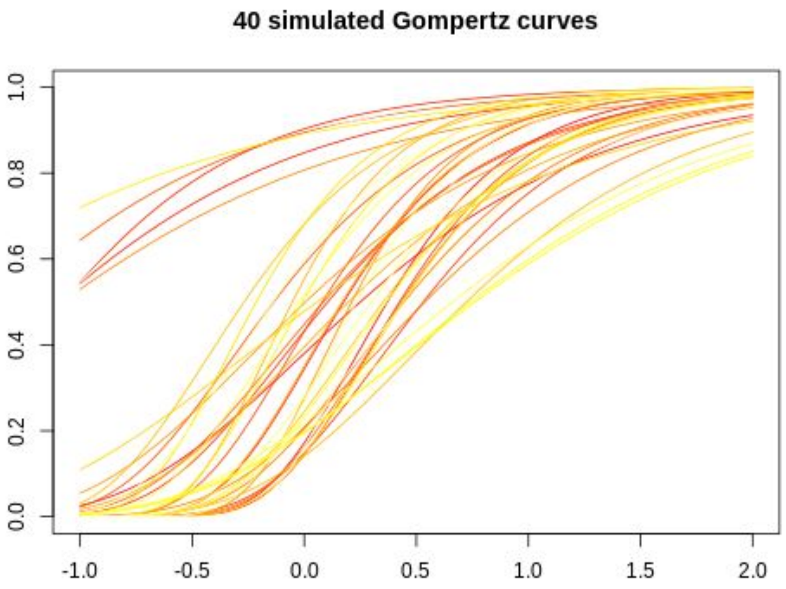
We can do the same for a different process, g

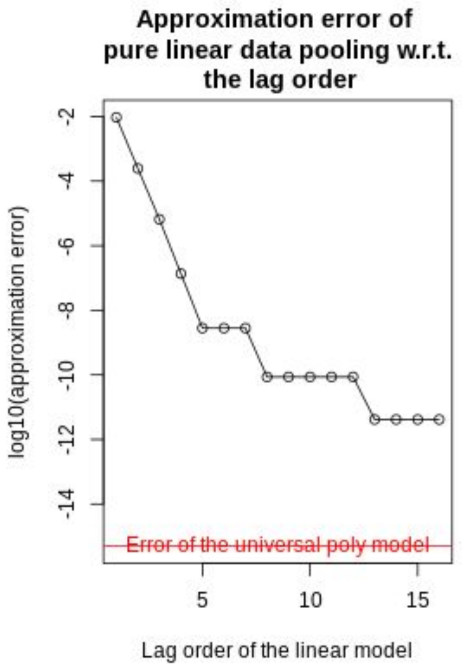
$$g(t+2) = \gamma g(t+1) + \delta$$
$$g(t+2) = \mathbf{x} g(t+1) + \mathbf{y} g(t) + \mathbf{z}$$

We can choose the solution that solves both systems

$$\mathbf{a} = \mathbf{x}, \mathbf{b} = \mathbf{y}, \mathbf{c} = \mathbf{z}$$

Simulation (please see interactive notebook Section 3.2)

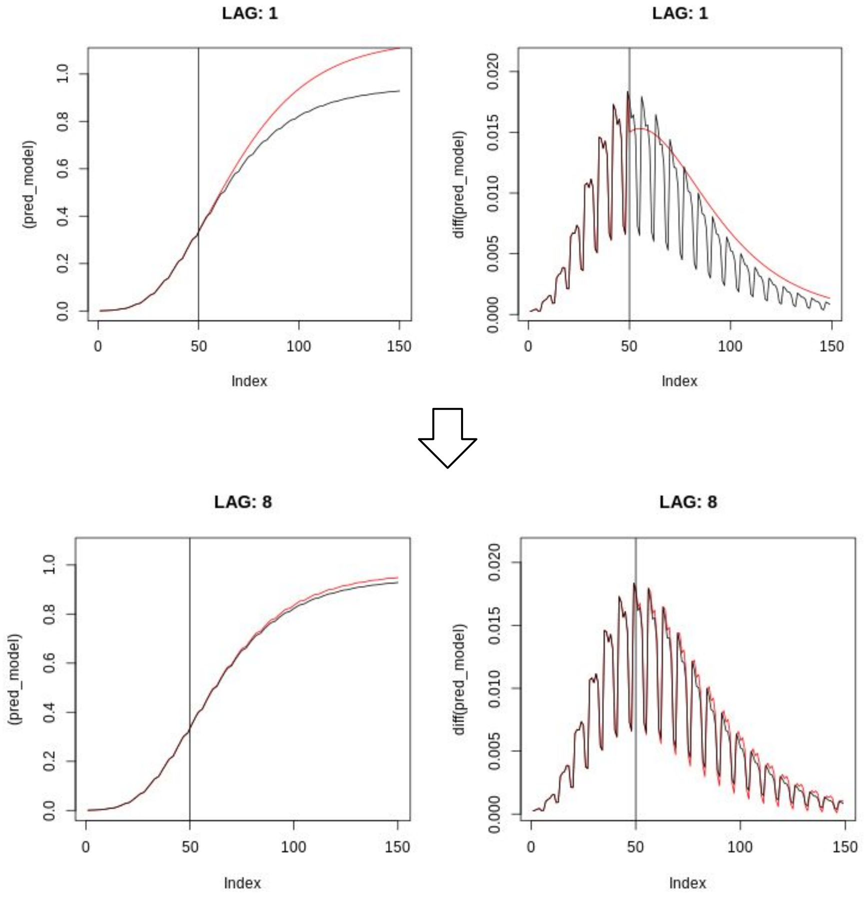




A purely linear autoregressive data pooled model is able to capture many curves with a few lags.
Consequence: The data pooled model has less parameters than individual models for the same level of approximation.

Deviations from ideal Gompertz: Periodic ‘perturbations’

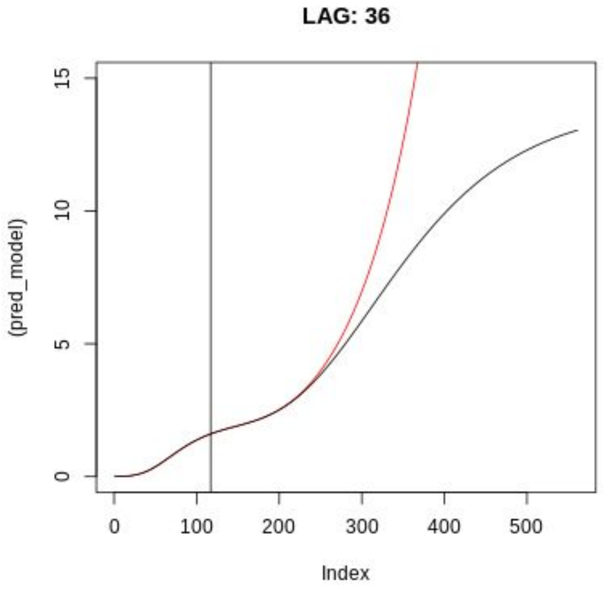
Interactive notebook
Section 4.1

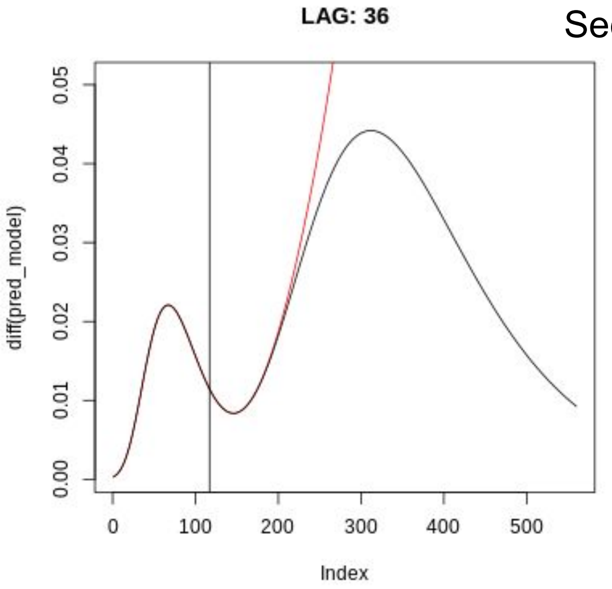


Autoregressive models are more general than growth curves, and can adapt to well known perturbations such as weekly effects.

Deviations from ideal Gompertz: Overlapping waves

[Interactive notebook](#)
Section 4.2





The flexibility of autoregressions can work to capture other important deviations such as multiple waves. No need to do ‘piecewise’ approximations. Very difficult in practice due to noise.

Summary

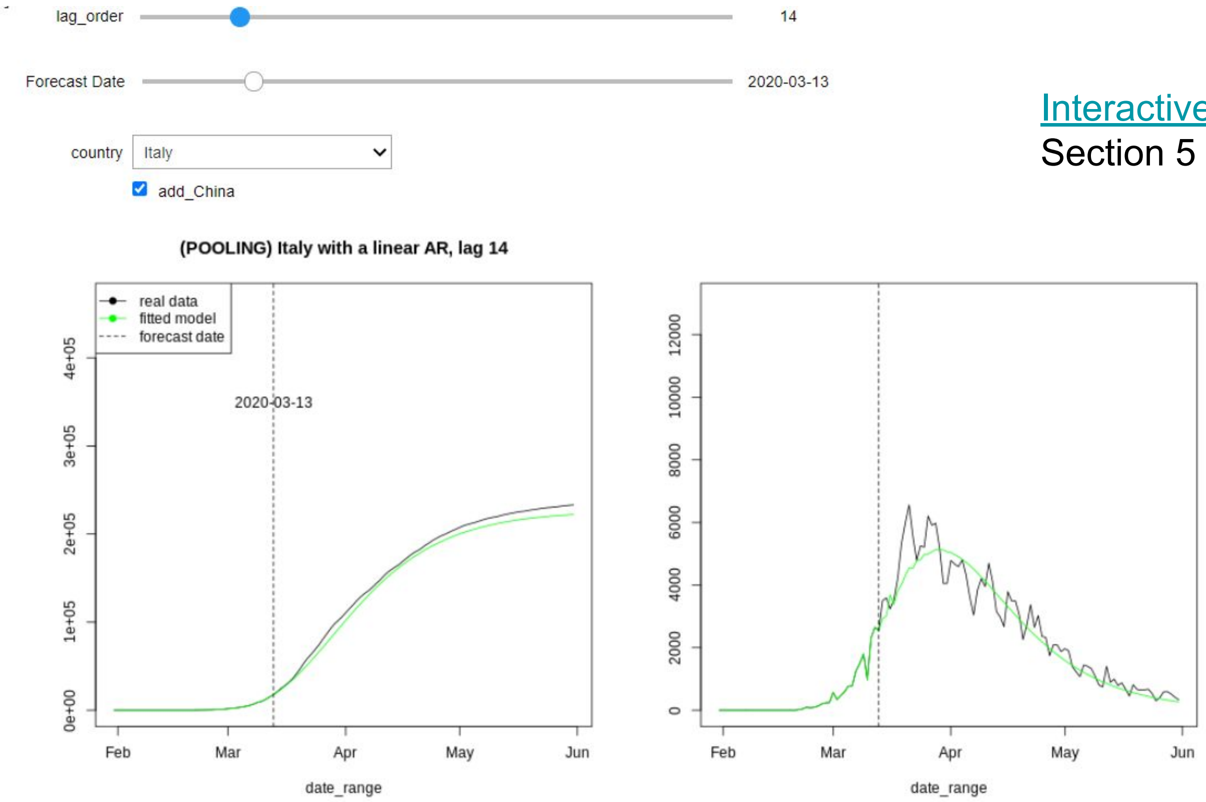
1. We use Gompert curves as motivation that pandemics roughly follow a class of parametric curves. Their approximation is good enough.
2. We show that each Gompertz curve can be expressed as an autoregressive process
3. We show that there is a 'special' class of autoregressive that expresses all possible Gompertz curves.
4. In practice, Gompertz curves **have limitations for prediction,** it is difficult to find the right curve 'beforehand'. Noise or more fundamental perturbations.

We can try to find the 'special' autoregressive model in the data. We can use all available time series to fit this model, because it captures all curves.
This model has better statistical properties under noise and can adapt to perturbations.

Results in Europe:

Hand-picked Example (2020)
Euro Forecast Hub (2021)

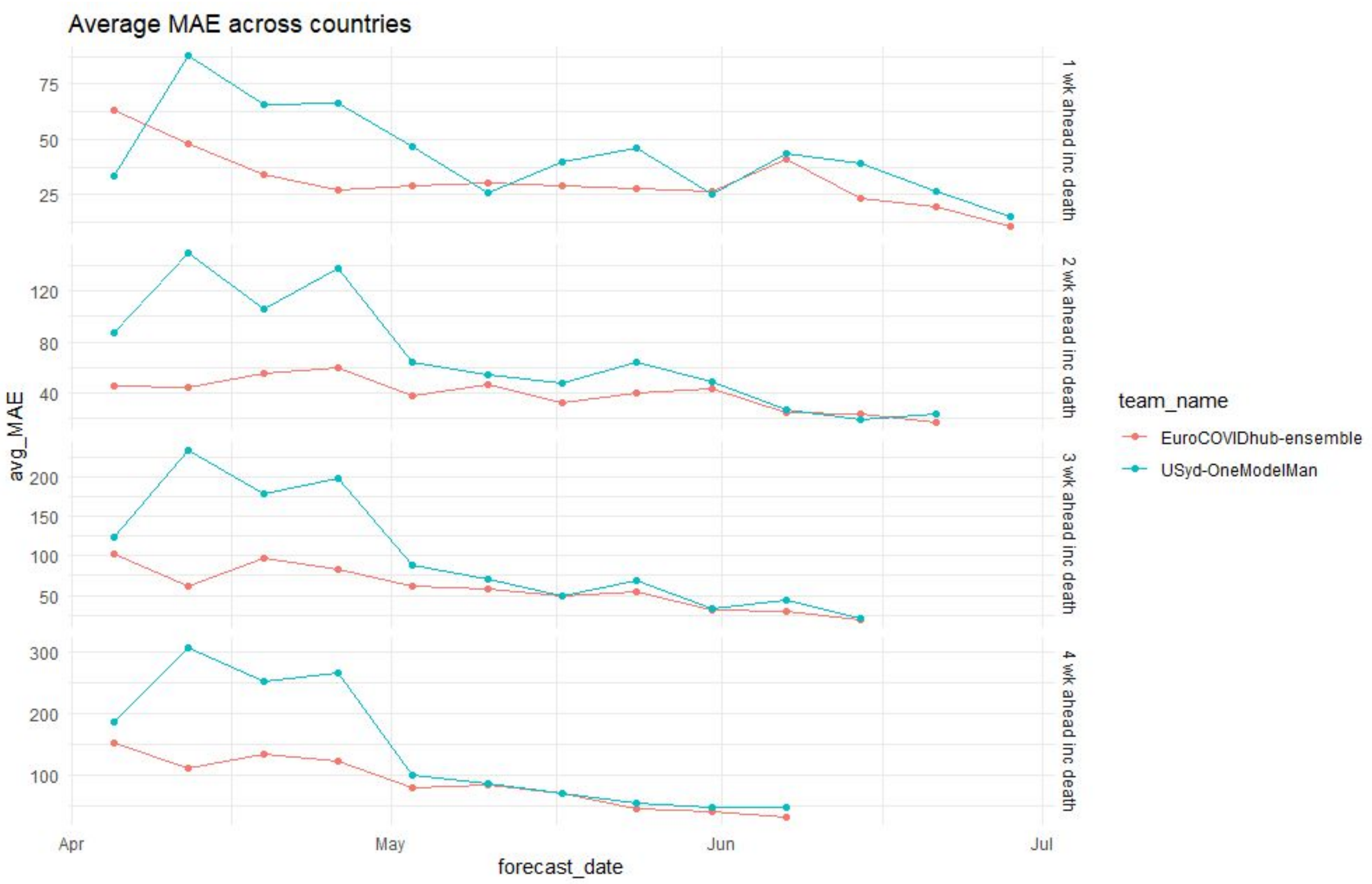
Demo of forecasting the first wave (2020)



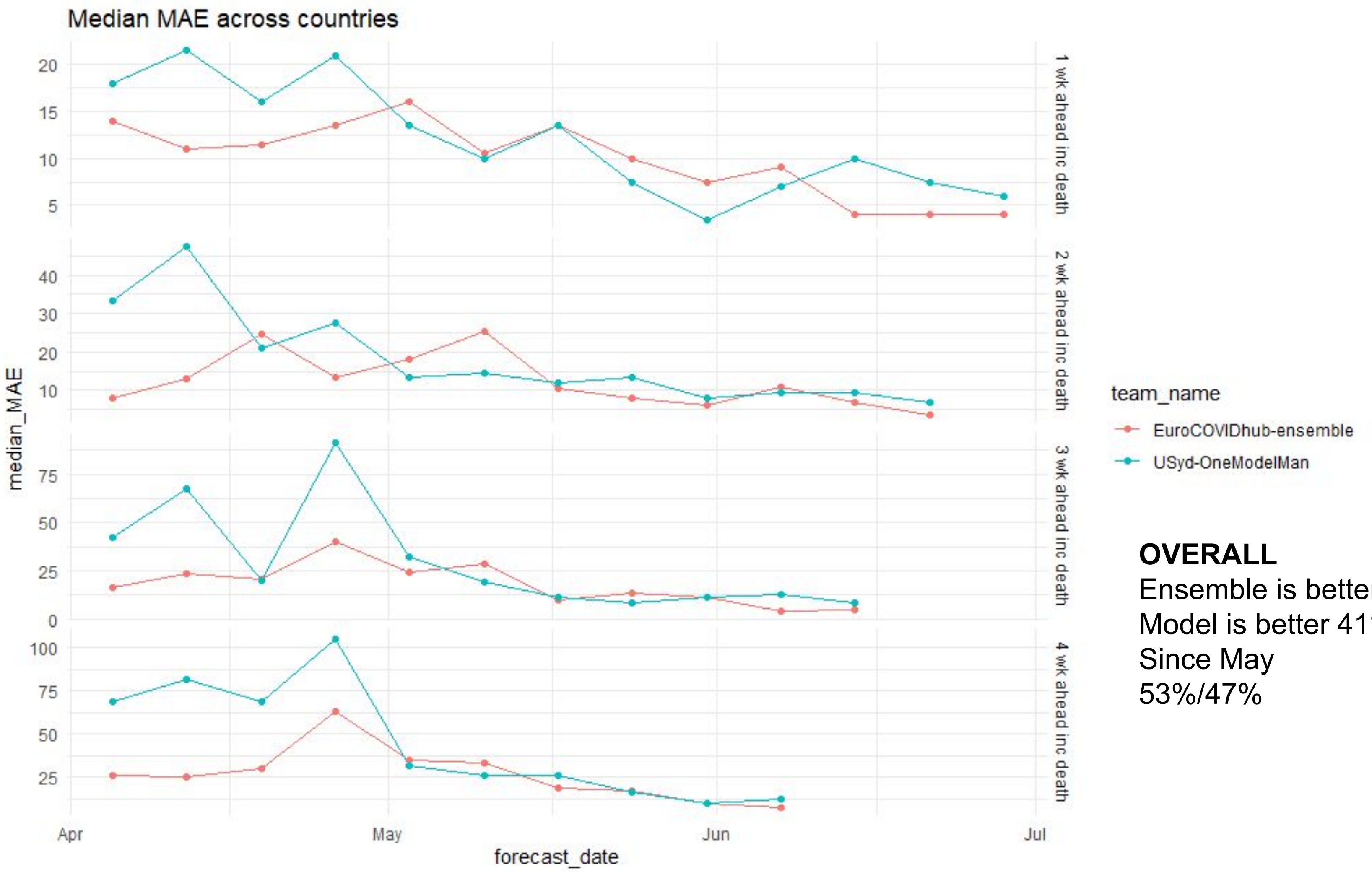
[Interactive notebook]
Section 5

Results in the Euro Hub: Implementation of the model

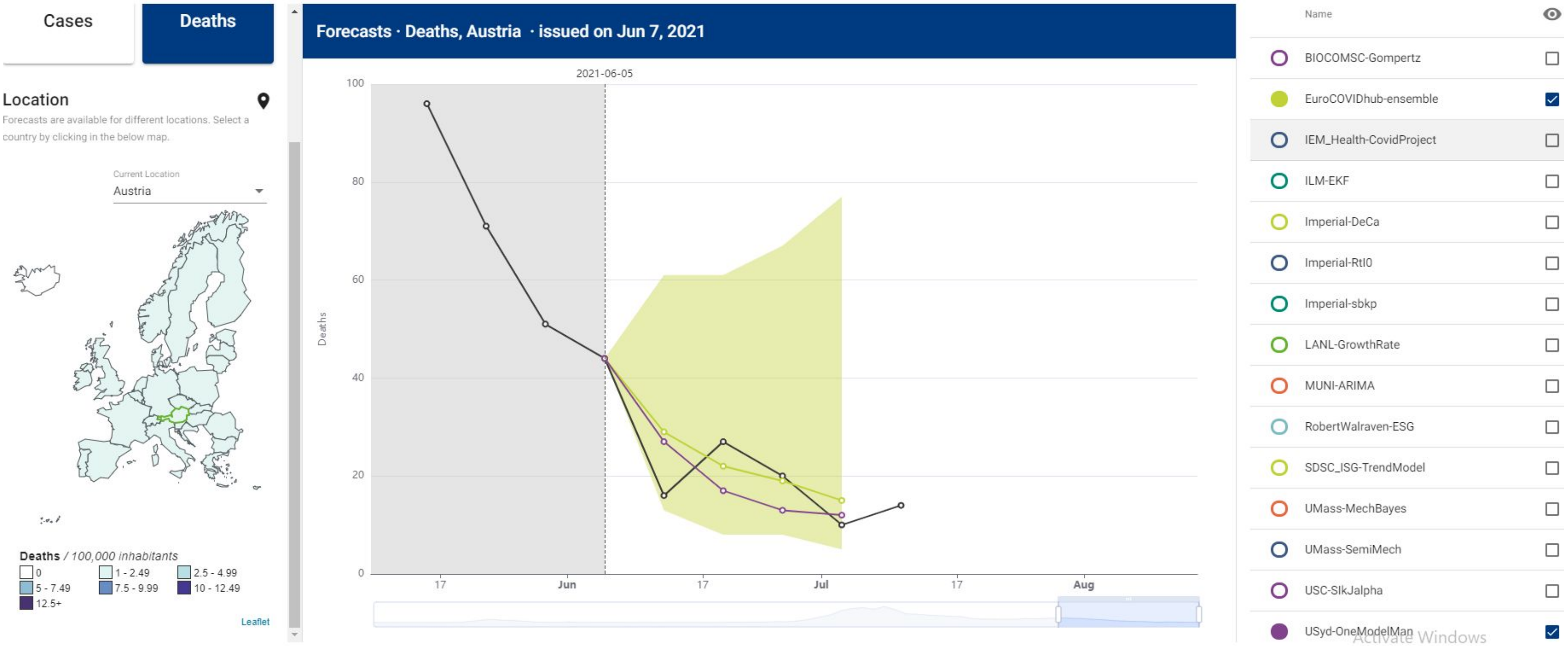
- We take the data of deaths from the European Forecast Hub
- We add time series of the top countries and regions
- We normalize the scale of each time series
- Fit a linear autoregressive model to the pool of all series
- Chose the number of lags by holdout validation (best model last month)
- Predict each time series with the fitted function



Average MAE across countries
avg_MAE
forecast_date
Apr
May
Jun
Jul
75
50
25
120
80
40
200
150
100
50
300
200
100
1 wk ahead inc death
2 wk ahead inc death
3 wk ahead inc death
4 wk ahead inc death
team_name
EuroCOVIDhub-ensemble
USyd-OneModelMan



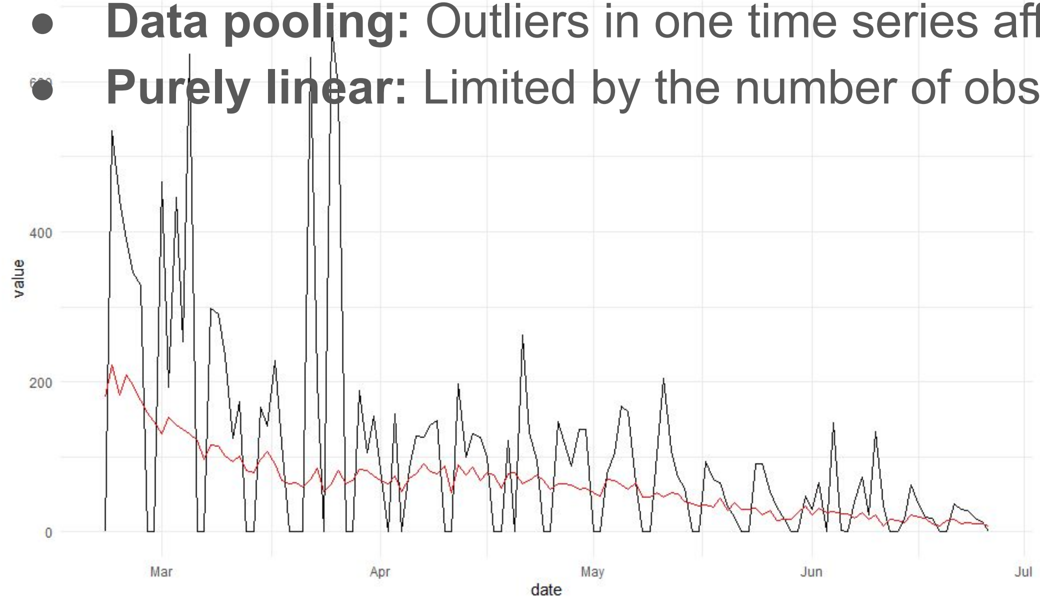
OVERALL
Ensemble is better 59%
Model is better 41%
Since May
53%/47%



Highlight problems in the UK, France, good forecasts for the increasing phases in late March/early April

Limitations

- **Autoregression:** Noisy input to the predictive function, as opposed to growth curve model that parameterize as a function of time (no noise).
- **Autoregression:** Accumulation of error for long horizons.
- **Data pooling:** Outliers in one time series affect the predictions for all series.
- **Purely linear:** Limited by the number of observations for large datasets



Solutions:
Explore non-linear by robust autoregressive models
Automatic ‘outlier’ detection
Automatic grouping
Tweaks:
Scale normalization, model selection, model combination, regularization, data cleaning

Take aways

- There is a special parameterization that enables data-pooling of pandemic time series while maintaining accurate approximations of individual time series, unlike traditional forms of data pooling.
- This parameterization might:
  - Have better statistical properties, much more data to fit the model.
  - Capture information ‘ahead’ of time from the more advanced countries w.r.t to the others.
- Useful for predictions. Data-driven, not mechanistic.
- Epidemiology experts can manually explore ‘what if’ similarities using this model, e.g. how ‘pooling’ with Israel (more advanced into vaccinations) affects predictions, pooling countries that have similar characteristics. Then **recover a mechanistic interpretation by estimating a mechanistic model to the forecasts of the data-pooled method in a individual time series.**

> *“There is a single universal function that predicts everything and we can find it in the data.”*

Resources

For further discussion, questions or **collaboration** please contact me!

- Links to notebook for interactions
- Applied 2019 to large sets of heterogeneous time series
- ‘Theorems’ about Equivalence and Statistical tradeoffs in paper (https://arxiv.org/abs/2008.00444)
- Use of Gompertz curves by colleagues in the hub, additionally (https://www.medrxiv.org/content/10.1101/2020.08.12.20173328v1), for the ‘universality of Gompertz in COVID: https://academic.oup.com/ptep/article/2020/12/123J01/5917637
- Related to step methods in differential equations, Taylor expansions, the relationship between lags and higher order derivatives etc.