APPENDICES

A. Acknowledgements

This work builds upon internships on Covid-19 performed over the summer 2020, sponsored by ChaLearn and INRIA. We are grateful to Alian-Jacques Valleron, Sam Evans, Kristin Bennett, Paola Tubaro and John Erickson for help and advice.

B. References

(Yu et al, 2020) Yang Yu, Yu-Ren Liu, Fan-Ming Luo, Wei-Wei Tu, De-Chuan Zhan, Guo Yu, Zhi-Hua Zhou, COVID-19 Asymptomatic Infection Estimation. medRxiv, 2020 https://www.medrxiv.org/content/10.1101/2020.04.19.20068072v1

(Cepeda, 2020) Covid-19 risk mitigation. Master thesis. https://github.com/cepedus/COVID19-Risk-Mitigation. Submitted to JDSE2021.

(Pearl, 2009) Judea Pearl. Causality. Cambridge University Press, 2009.

(Sutton 2018) Sutton, R.S. and Barto, A.G. Reinforcement Learning. An Introduction. 2nd Edition, A Bradford Book, Cambridge, 2018. incompleteideas.net/book/RLbook2020.pdf

(Petersen, 2020) Eskild Petersen, Marion Koopmans, Unyeong Go, Davidson H Hamer, Nicola Petrosillo, Francesco Castelli, Merete Storgaard, Sulien Al Khalili, Lone Simonsen, Comparing SARS-CoV-2 with SARS-CoV and influenza pandemics, The Lancet Infectious Diseases, 2020.

(Goodman-Bacon, 2020) Andrew Goodman-Bacon and Jan Marcus. Using Difference-in-Differences to Identify Causal Effects of COVID-19 Policies. Techreport, 2020. https://papers.ssrn.com/abstract=3603970.

(Alkire, 2018) Blake C. Alkire, Alexander W. Peters, Mark G. Shrime, and John G. Meara. The Economic Consequences Of Mortality Amenable To High-Quality Health Care In Low- And Middle-Income Countries. Health Affairs, 2018.

(Javaid, 2015) Kiran Javaid et al. Morbidity and Economic Growth. Tech report Evans School Policy Analysis and Research Group, 2015. https://evans.uw.edu/policy-impact/epar/research/morbidity-and-economic-growth.

(Brodeur, 2020) Abel Brodeur and Andrew E. Clark and Sarah Fleche and Nattavudh Powdthavee. Assessing the impact of the coronavirus lockdown on unhappiness, loneliness, and boredom using Google Trends. ArXiv preprint, 2020. https://arxiv.org/abs/2004.12129

(Aron, 2020) Janine Aron and John Muellbauer. A pandemic primer on excess mortality statistics and their comparability across countries. Our World in Data, 2020. https://ourworldindata.org/covid-excess-mortality

(Woolf, 2020) Steven H. Woolf, Derek A. Chapman, Roy T. Sabo, Daniel M. Weinberger, and Latoya Hill. Excess Deaths From COVID-19 and Other Causes, JAMA, 2020. https://doi.org/10.1001/jama.2020.11787.

(Bai, 2018) Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. ArXiv preprint, 2018. https://arxiv.org/abs/1803.01271

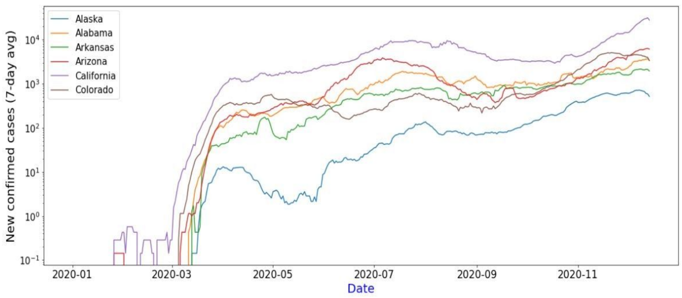
(Roser, 2020) Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell. Coronavirus Pandemic (COVID-19). https://ourworldindata.org/coronavirus. Our World in Data, 2020.

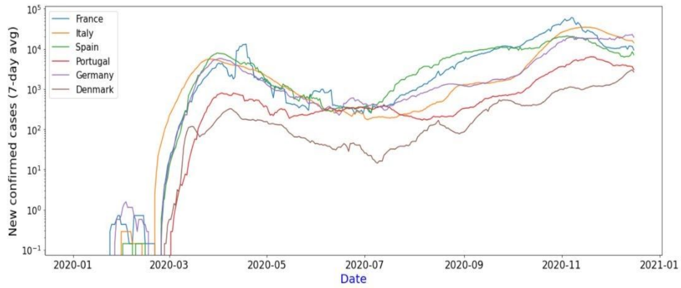
C. Supplementary material

I. Data Understanding

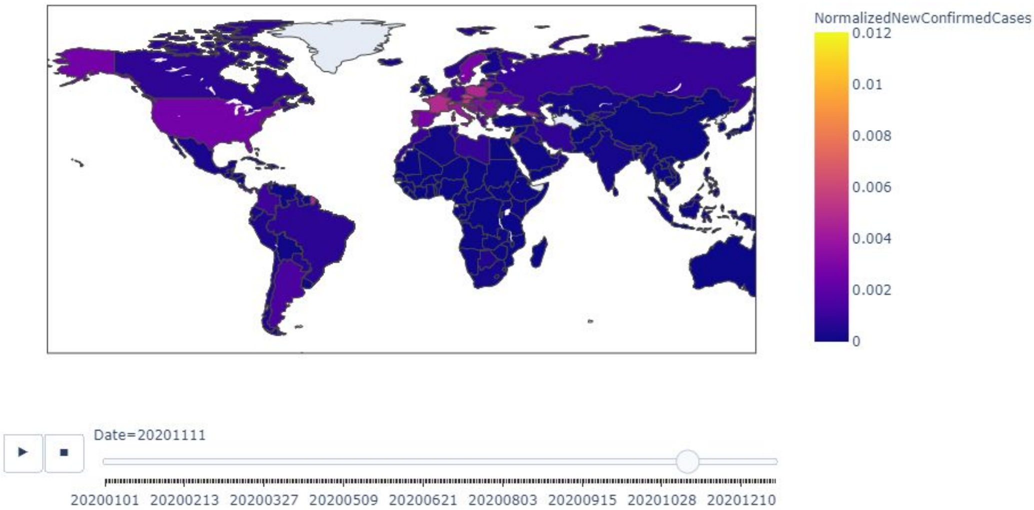
Similar evolution in geographically close countries/regions (log scale)

We observed two principal waves of outbreak with shifted starting periods depending on the countries. Countries geographically close or with a lot of commercial/tourism mobility will most likely show similar behaviors.





2nd wave visualization (ipynb available [here](#))



New cases normalized by country population.

Correlations when considering NPIs as ordinal data (using Kendall's Tau coefficient)

This correlation plot was to give some insight about the possible relations between the NPIs and the evolution of cases and deaths. Correlations between the NPIs themselves are understandable as their evolution is correlated but the correlation with the number of cases and/or deaths is not that obvious. Indeed, we would expect negative correlations at least for the PIs imposed even with a delayed impact.

	C1_School closing	C2_Workplace closing	C3_Cancel public events	C4_Restrictions on gatherings	C5_Close public transport	C6_Stay at home requirements	C7_Restrictions on internal movement	C8_International travel controls	H1_Public information campaigns	H2_Testing policy	H3_Contact tracing	H6_Facial Coverings	NewConfirmedCases_7davg	NewConfirmedDeaths_7davg
C1_School closing	1	0.62	0.64	0.55	0.48	0.57	0.56	0.47	0.52	0.3	0.32	0.34	0.43	0.43
C2_Workplace closing	0.62	1	0.67	0.63	0.52	0.61	0.59	0.42	0.48	0.29	0.3	0.34	0.43	0.44
C3_Cancel public events	0.64	0.67	1	0.66	0.47	0.57	0.59	0.47	0.54	0.34	0.36	0.39	0.45	0.43
C4_Restrictions on gatherings	0.55	0.63	0.66	1	0.46	0.56	0.52	0.37	0.5	0.35	0.34	0.4	0.45	0.43
C5_Close public transport	0.48	0.52	0.47	0.46	1	0.52	0.54	0.31	0.32	0.21	0.19	0.26	0.32	0.33
C6_Stay at home requirements	0.57	0.61	0.57	0.56	0.52	1	0.61	0.38	0.43	0.26	0.25	0.35	0.43	0.43
C7_Restrictions on internal movement	0.56	0.59	0.59	0.52	0.54	0.61	1	0.42	0.42	0.24	0.26	0.36	0.38	0.39
C8_International travel controls	0.47	0.42	0.47	0.37	0.31	0.38	0.42	1	0.46	0.29	0.31	0.19	0.23	0.2
H1_Public information campaigns	0.52	0.48	0.54	0.5	0.32	0.43	0.42	0.46	1	0.5	0.5	0.44	0.44	0.37
H2_Testing policy	0.3	0.29	0.34	0.35	0.21	0.26	0.24	0.29	0.5	1	0.53	0.44	0.45	0.37
H3_Contact tracing	0.32	0.3	0.36	0.34	0.19	0.25	0.26	0.31	0.5	0.53	1	0.42	0.32	0.25
H6_Facial Coverings	0.34	0.34	0.39	0.4	0.26	0.35	0.36	0.19	0.44	0.44	0.42	1	0.49	0.44
NewConfirmedCases_7davg	0.43	0.43	0.45	0.45	0.32	0.43	0.38	0.23	0.44	0.45	0.32	0.49	1	0.8
NewConfirmedDeaths_7davg	0.43	0.44	0.43	0.43	0.33	0.43	0.39	0.2	0.37	0.37	0.25	0.44	0.8	1

Features Importance
In order to understand better the relation between variables, we performed a Random Forest Regression model and then selected the feature importances of the variables. There was no important relevance on any of the variables in the OxCGRT dataset.

II. Model selection:

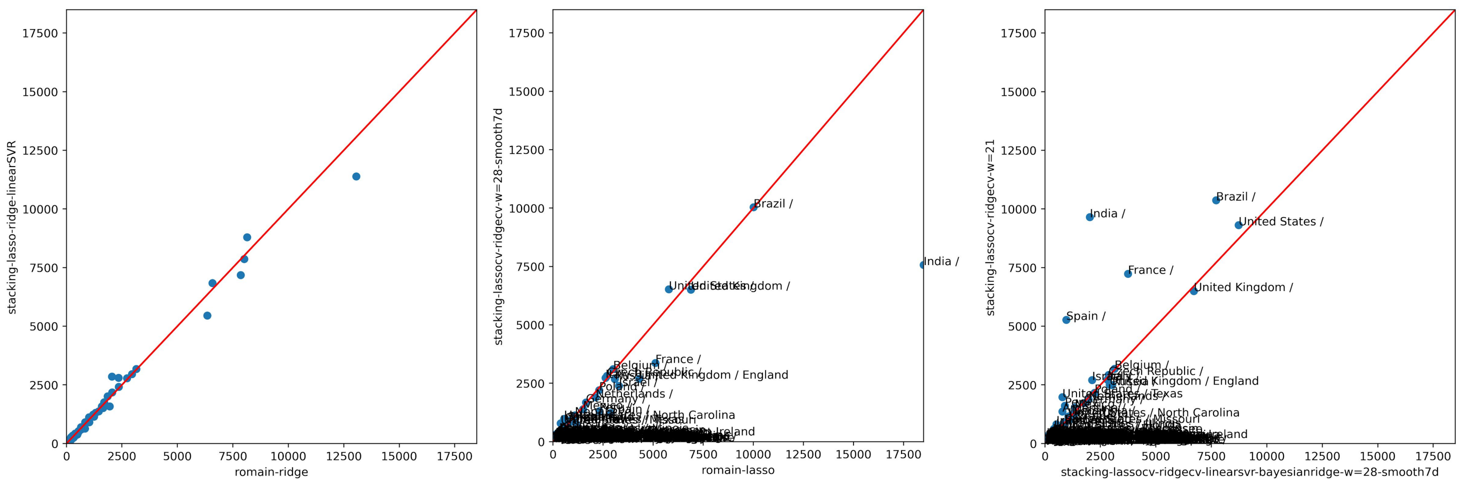
To select the final submission, we did the following:

1) TrainRidgeCV, LassoCV, BayesianRidge, LinearSVR and stack them with RidgeCV over data up to september.
2) Compute prediction over 1 month (October 2020)
3) Compute MAE over 20 days (length of testing phase) taking as reference the 7-day smoothed number of cases in reality (same as the testing phase). These MAE scores are computed per-country to see on which of them a particular model performs the best.
   Table: cumulative 20MAE on all countries, alongside with worst performing country/region

romain-lasso	113188.242857	'India / ‘
romain-ridge	114758.814016	'Brazil / '
stacking-lasso-ridge-linearSVR	111055.404530	'Brazil / '
stacking-lassocv-ridgecv-linearsvr-bayesianridge-w=28-smooth7d	84126.300332	'United States / '
stacking-lassocv-ridgecv-w=21	107380.492421	'Brazil / '
stacking-lassocv-ridgecv-w=28-smooth7d	86885.818699	'Brazil / '

4) Comparing models side by side: we scatter the countries over their 20MAE on 2 models (axes) to visualize which countries make the difference:

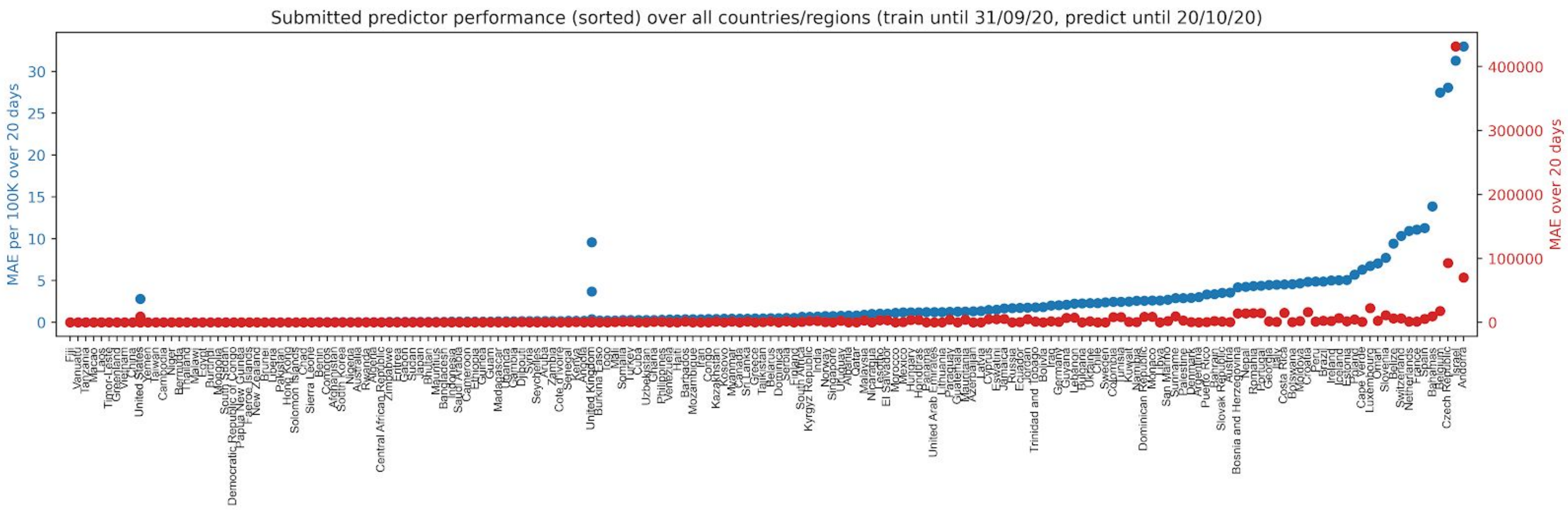
   Figures: model name on each axis, diagonal indicates equi-error. How to read: models in the upper triangle perform better on the x axis model and vice versa. The more aligned the points, the more similar both models perform. The axes are scaled to the max error across all models. Two very similar modelsoves error in countries in the upper triangle
   model on the y axes halves the error of India (worst error country) w.r.t model on x axis.
   As we measure in this case the “raw” MAE over 20 days, the bigger countries such as Brazil, India and the United States have the biggest error and their points are in the upper-right region of the plots.



At the left plot: 2 very similar models (points more or less aligned over the equi-error diagonal). At the center plot: Model in the y axis halves the error in India (worst performing country in the x axis model. At the right plot: Overall improvement of model on the x axis, specifically in India, Brazil, France, Spain and the US (farthest points from diagonal).

5) Final submission: as we work with general statistical models, it makes little sense to have specific countries for each one of them. We go for a stacking approach, where the final model has no specific regions. For more complex models such as SEIR/SEIRD we could give a super-well fitted to only one country, but it loses the interest of the challenge itself to help predictions and generate prescriptions on a global scale.

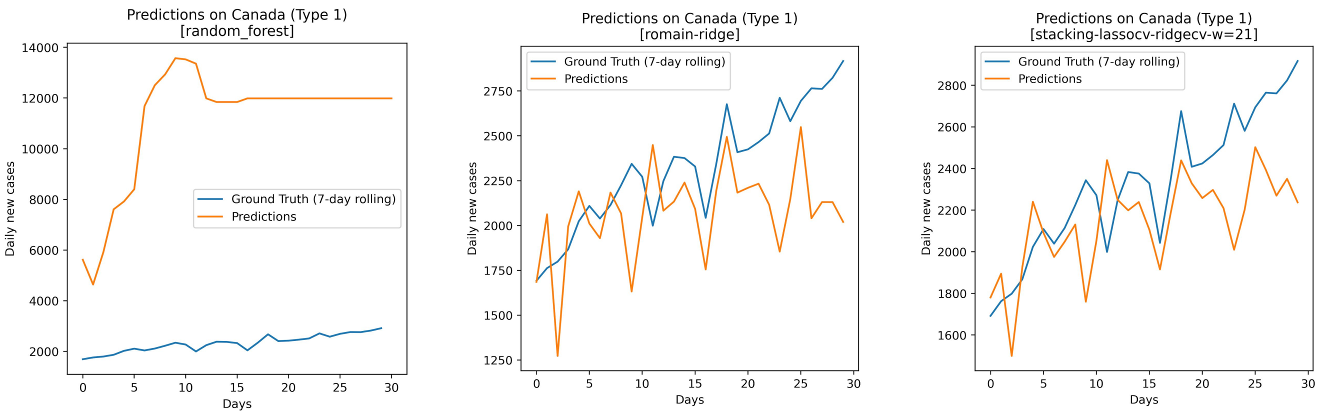
III. Best predictor performance over different countries:



Across all regions we reach a rather good performance (MAE per 100K less than 10). Particularly bad predictions are Bahamas, Belgium, Czech Republic, Israel and Andorra. Some of these countries have either a) very strong 2nd waves of cases and/or c) haven't reached the peak of a wave.

IV. Comparison test run in October

From left to right: Random Forest, Ridge Regression, Submitted model (Stacking ensemble). All models were trained until 30/09.



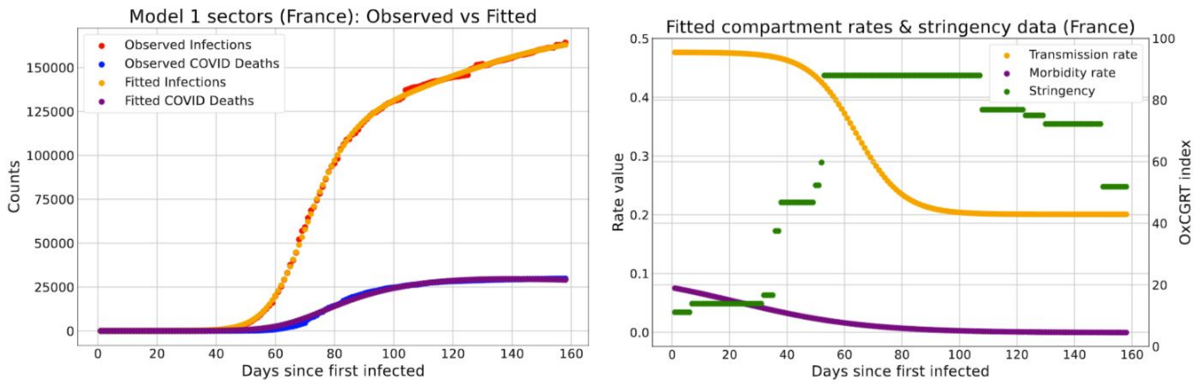
V. Other modelling choices:

On epidemiological models:

- Advantages:

1) Disease dynamics is accurate and has already been studied, in particular SEIR/SEIRD models are adapted to COVID-19 as the incubation period of the virus is important (4 days).
2) Comparatively few parameters to fit w.r.t. NN approaches and each parameter has a direct interpretation (contagion rate, death rate, etc.)
3) Once fitted, simulation is fairly fast depending on the ODE solver

- Disadvantages:
    1) It contains several latent variables (S, E, R) that are difficult to estimate from data, so to plug in a trained model on an arbitrary date implies estimaging all latent variables also. Best option then is to fit the model from ‘time 0’ and run it also from time 0, and start considering as a prediction when we surpass the end of the training data
    2) Highly country dependent: depends on context and time-shift (when the first infected person is detected). Country dependence is necessary as that’s what we’re trying to predict and in COVID-19 international transport is severely reduced.
    3) For COVID-19, the whole game is how to map NPIs to changes in transmission/death/incubation rates (as they’re clearly not constant, see case of the studies on R_0) . To do that, the rate modelling can be arbitrarily complex (piecewise linear, piecewise constant, arbitrary continuous function, etc.)
    4) To learn rates over time, the loss is not straightforward to compute (as the rate is the input of the actual model) and highly non linear, due to the coupled differential equations for the EPI model. Thus any training is bottlenecked by the ODE solver.
- Figures:



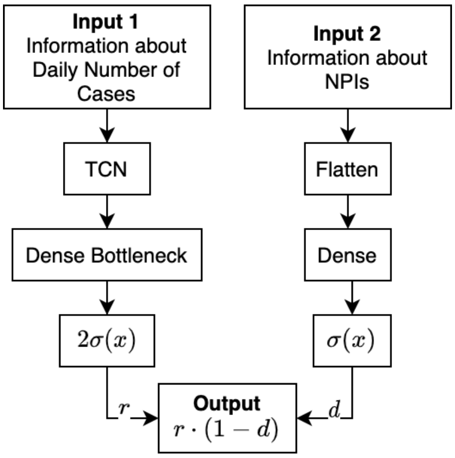
Data and fitted model in France from 25 January 2020 to 1 July 2020. At left, Infected and Dead compartments and at right, fitted transmission and morbidity rates compared with the value of the stringency index (sum of NPIs and non-NPIs). The accuracy on the model depends greatly on the assumption on transmission and morbidity rate. In this case, it was an affine sigmoid function.

On Neural Network-based models:

We tried NN models to predict the daily new number of cases (based on the linear baseline) or the normalized ratio of new daily cases (based on the LSTM baseline). In general these models performed worse than linear models. We noticed that TCN (Bai, 2018) models had a great potential to filter the input signal of the ratio of new daily cases. But, due to overfitting we did not use this model as our final submission.

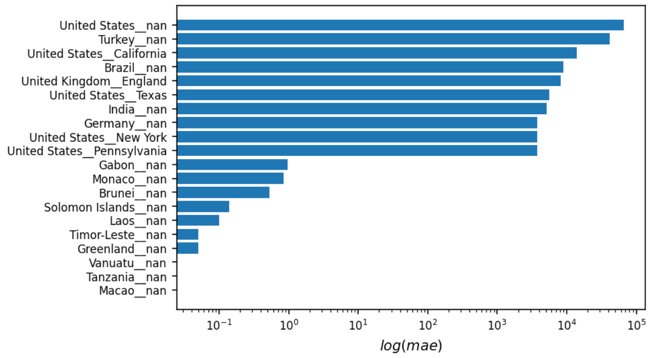
- Advantages:
    1) One of the main advantages is to be able to build a unique model which can learn from the data of all countries.
    2) A NN model can learn non-linear relationships between input and output variables.
- Disadvantages:
    1) It is hard to tune the hyperparameters and architecture of a NN and optimize its weights.
    2) Many overparameterized NN can have overfitting effects that can be hard to deal with.
    3) We noticed that adding a TCN on NPIS data was blocking all kinds of learning! However, adding it on the timeserie of new cases enables learning but massive overfitting.

- Figures:

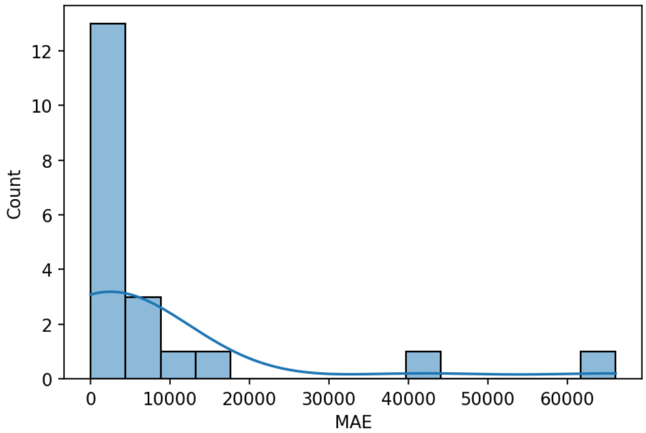


The TCN based model is inspired from the LSTM baseline. For the first input, instead of having an LSTM we filter the signal with a TCN, then a serie of dense layers (64, 32, 16, 8, 4, 2, and 1 units) to feed a rescale sigmoid function (to bound the output and avoid outliers).

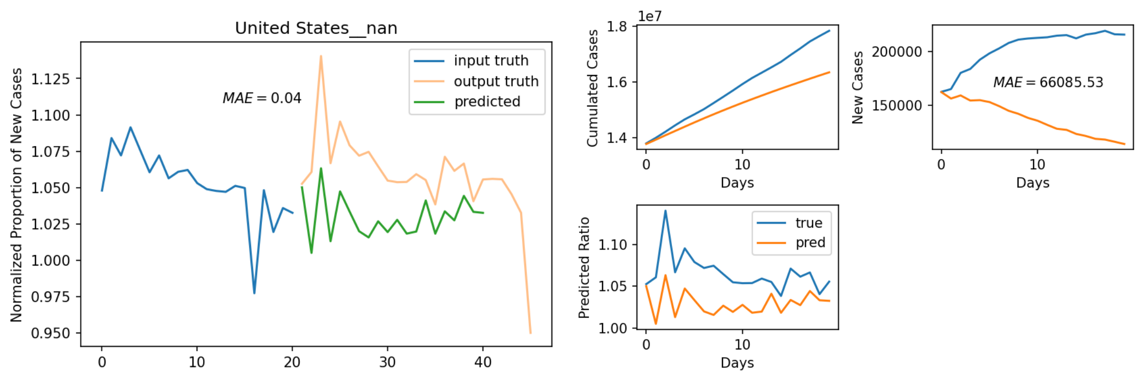
On the second input, we noticed simply use a single dense layer. In general, we were careful about doing a temporal split between our train/validation data (which is not done in the baselines), to avoid introducing a bias in our validation error.



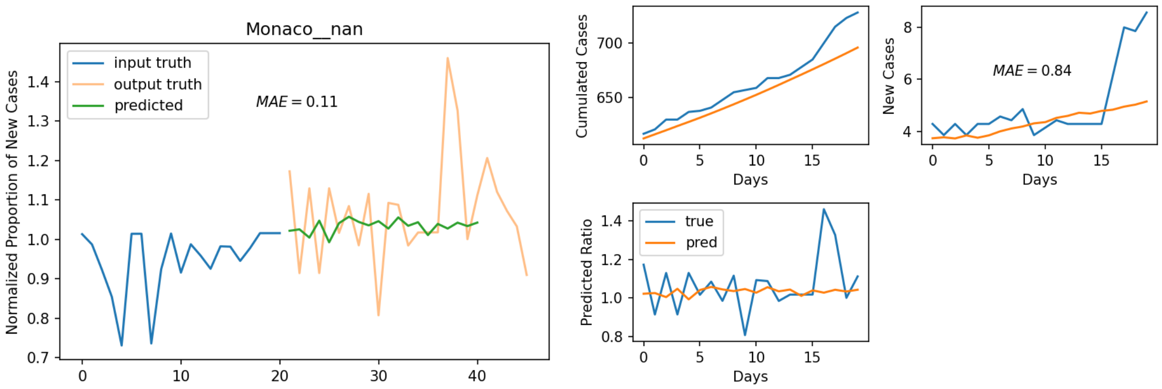
Prediction errors from 2020-12-01 to 2020-12-20, top-10 biggest errors (top) and top-10 smallest errors (bottom) are shown.



Histogram and estimated distribution of errors from 2020-12-01 to 2020-12-20, we can see that some countries are clearly outliers.



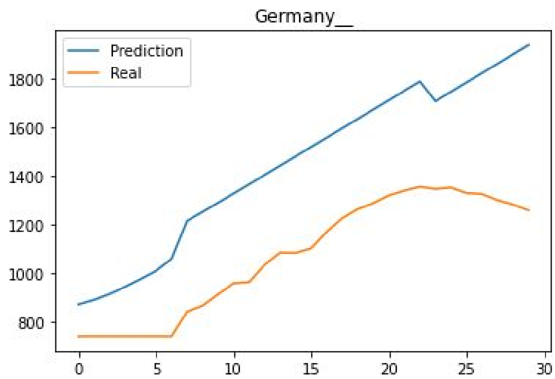
The United States is one of the worst performing countries with our TCN based NN. We can see that the ratio of new daily cases is clearly underestimated.

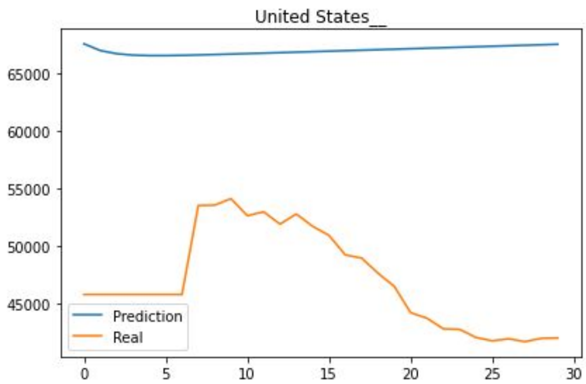


Monaco is one of the best performing countries, we can see that the general trend is followed without overfitting to the noise.

On ARIMA/X models:

- Advantages:
    1) Very fast to train and predict
    2) Adapted to auto-regressive time series such as the number of cumulated cases
    3) Idea: local context and look-back days will condition future evolution of the pandemic
- Disadvantages:
    1) On current data, it fails to capture trends and seasonality. (both in general and on specific countries)
    2) NPI context/NPI shifted context does not improve performance
- Figures: Prediction of daily new cases over 1 month (November 2020) in Germany and the US:





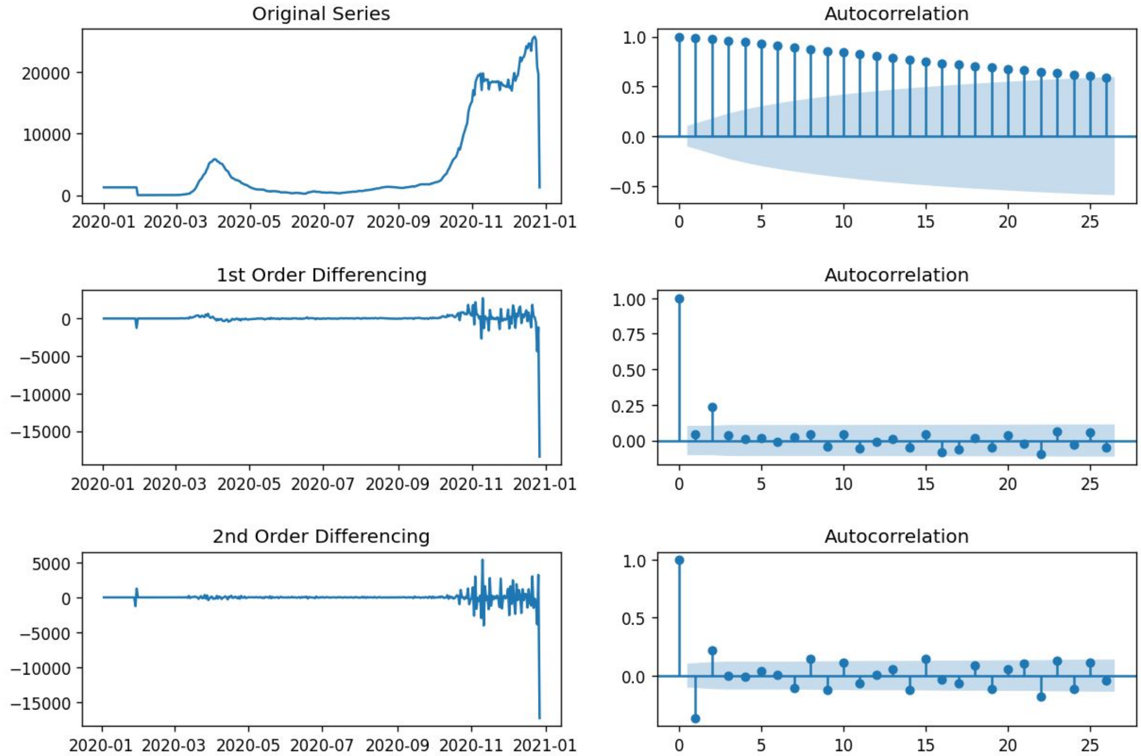
In the case of Germany, the ARIMAX model captures only the growth of cases, not the scale nor reaching the peak of a wave. In all, there's an MAE of almost 600. In the second figure, the model fails completely to predict new cases in the US, returning an almost constant prediction over the Month.

We also did an Augmented Dickey Fuller test (ADF) and Auto Correlation Function (ACF) plots analysis used in an attempt to fit Seasonal ARIMAX:

For instance for Germany:

ADF Statistic: -2.193192
p-value: 0.208726

Since p-value is greater than the significance level (p-value>0.05), we looked at the differencing and autocorrelation plots, to find the proper order of differencing.



From the plots above we see that a first order differencing makes the series of New Cases (smoothed over 7 days) in Germany stationary enough, no need for a second differencing.

On Random forest models:

We have tried one Random forest regressor model to try to predict daily new cases for the different countries. We have used the basic preprocessing. Hyperparameters selection includes the number of estimators, the maximum depth of the tree, and the maximum number of leaf nodes. Changing the criterion from “mse” to “mae” increased the training and testing time a lot (from minutes to several hours).

- Advantages:
    1) Can be very fast to train and predict.
    2) They are more understandable. We can use them to obtain feature importance information.
- Disadvantages:
    3) Hyperparameters have a big impact on the model; both in terms of results and time performance.
    4) The model quickly overfits the training data.
    5) They are not very good at capturing fluctuating trends in the medium/long term (as it can be observed in the figures).
- Figures: Results with our final model in 3 different countries. The model uses 100 estimators and “mse” as criterion. We train it with the data until 2020-09-30, and try to predict daily new cases from 2020-10-01 to 2020-10-31. The model takes 2’30’’ to train and 2’ to make the predictions. The training/test set split is 0.8/0.2, the MAE error in the training set is 48 and in the test set is 113.