Project 3: Understanding COVID

Pablo Freyria Dueñas

April 12, 2021

Abstract

On 2019, the first case of COVID-19 was identified, on March 2020 it was declared as a pandemic by the World Health Organization. This outbreak not only affected the health of millions of persons, but in an effort to control its spreading, it also affected the world economy and our every days activities. The globalized world not only facilitated the disease to spread rapidly, but also allowed for information to be gathered and shared on a global scale. The objective of this report is to build a model that describes and forecasts the outbreak development, and as a secondary objective to analyze the effect of policies, such as closures and travel bans, controlling by countries' characteristics such as age, income, diseases and infrastructure, on controlling disease spread.

1 COVID-19 cases and deaths forecast

1.1 Data description

The first database contained daily confirmed cases and deaths from 19 January 2020 to 10 June 2020 (140 days) and from 187 countries. Variables in database were: Location, country population, type of outcome (case or death) and target value. To pre-process the data, we removed the rows with negative target values and we aggregated by date and country all cases. This resulted in 26,180 rows (140 days for 187 countries). The total number of cases (and deaths) can be seen in Figure (1). We see a lot of heterogeneity

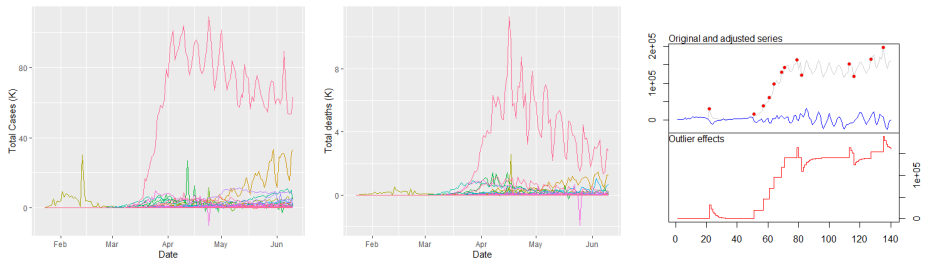

Figure 1: Total cases and deaths by country; Outliers analysis

in countries, particularly from the U.S.(pink), where the increase in cases is noticeably higher than other countries and we notice it accounts for the majority of cases. This suggests that information learned from one country cannot be extrapolated to another country: if U.S. is used as a training example, predictions for other countries will be off, and vice versa. Also, the third panel in Figure 1 shows a time series outliers analysis (tso function in R) that detected 13 outlier points: 7 of level shifts, 5 of temporary changes and 1 additive outliers.

1.2 Development of forecasting algorithm

We build the forecasting algorithm based on the Super Learner approach[1], where we train a library of candidate algorithms and evaluate a loss function on the validation set, repeating this process until all data has been evaluated in the validation set and thus getting a risk estimate for each algorithm. Finally, a linear combination of the predictions is built as the Super Learner prediction. This final regression is also assessed for overfitting with an additional layer of cross-validation. We use two schemes of cross validation for time dependent data: rolling origin and rolling window, and choose the Super Learner that achieves the least cross validated risk. In rolling origin, training set grows in time and validation set gets pushed forward, while in rolling window both sets have a fixed size and are pushed forward in time. In both cases, we use 15% of data for validation and a gap of 5% between sets. This resulted in cases being trained on rolling window and deaths on rolling origin.

The predictor variable used were: date, population and a basis expansion with a sine function of date with periodicity of 3, 5 and 7 days. The algorithms included in the learners library are: ARIMA models[2], splines, SVM (with sigmoid and a 3-degree polynomial), forward-feeding neural nets and gradient boosting algorithms. This last ones were trained over a parameter grid varying learning rate and max depth, letting the Super Learner choose. The "winning algorithms" for total cases was gradient boosting with 0.4 learning rate and 6 levels; which also had a big weight for predicting deaths, followed by a SVM with sigmoid function. The resulting fits and real points can be seen in Figure 2 and the proportion of variance explained ($1 - MSE/Var(Y)$) is 96.4% and 95.4% for total cases and deaths respectively.

[1] Targeted Learning in R: Causal Data Science with the tlverse Software Ecosystem. Mark van der Laan, Jeremy Cole, Nima Hejazi, Ivana Malenica, Rachael Phillips, Alan Hubbard

[2] Although we detected outliers and it is probably not a good model, it doesn't hurt to add more learners and this is a natural choice for the problem

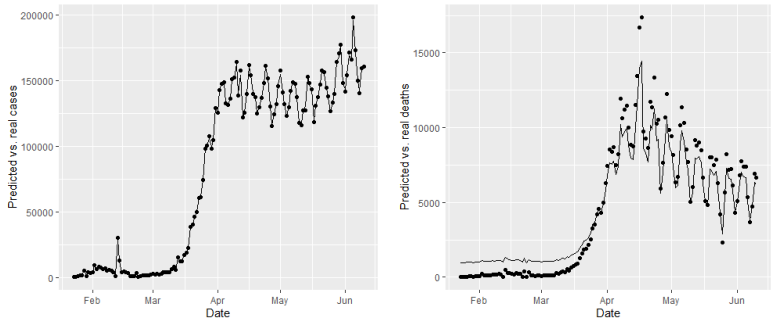

Figure 2: Super Learner predictions on cases and deaths

2 Impact of policies in controlling the disease

To assess the impact of policies, we used a database from "our world in data" which contained information on cases, deaths, number of tests, hand washing facilities, population, age (median, proportion over 65 and 70 years old), GDP and poverty, cardiovascular disease, diabetes, smoking, life expectancy and stringency index. NAs in intervention or outcome were removed, and NAs in variables were imputed with their median; if over 50% of a variable observations was NA, it was dropped. We propose to model new cases per million as a function of these covariates, stringency index and the average stringency index in the previous week and two weeks (based on the knowledge that disease can take up to 14 days to develop symptoms). We also verified with an ARMA model that series had no AR or MA components and periodicity is 1, meaning that we can ignore the time dependency to itself in the outcome. Then, a regression tree with minimum 20 observations per split and a minimum increase in fit of a 1.01 factor was run to identify main variables. Finally, a linear model with the splitting nodes and the interactions that could be read from the tree was run. The variables were standardized to interpret the effect of the variable regardless of scale.

The variables with the highest impact on new cases per millio were: new tests by thousand (0.57), date (0.17), stringency index (0.10) and cardiovascular death rate (0.09); the average stringency index two weeks before had the lowest coefficient. This results are meaningful because they remind us that the highest predictor of *confirmed* cases is the application of tests, highlighting the bias in how the data is collected; note also the coefficient with stringency index is positive, suggesting that countries with a more stringent response are also applying more tests and thus reporting more cases. Date being the second most meaningful variable suggests that all countries regardless of their response and baseline characteristics would experience an increase in cases. Finally, cardiovascular death rate being a predictor of cases is aligned with public guidelines that this is the population most at risk.