Understanding the Effect of Containment Measures on COVID-19 Rates

Project Report by Team 8

Klaudia Krawiecka, Vilda Markeviciute, Amaka Okafor,
Alina Petrova, and Sade Snowden-Akintunde

Team GitHub Repository: https://github.com/DS4All-2020-Team-8/Team8_DS4A/

1 Introduction

More than six months have passed since the World Health Organization categorized the spread of the COVID-19 virus as a global pandemic. Although everyone across the globe has been exposed to the same health threat, epidemiological implications vary greatly from country to country. In this report we aim to analyse these variations and to discover which factors influence them.

Several months of data on the worldwide spread of the virus as well as on how governments and healthcare systems counteracted it allow us to compare the situation in different countries and to identify factors and strategies that have proven effective in mitigating the pandemic. We start by looking at general socio-economic and geographic parameters of each country, such as GDP, population density or climate, and then move to analysing the specific anti-covid steps taken by each country. We **hypothesize** that although geographic and economic factors play a certain role in the spread of the virus, concrete measures and policies imposed by governments, such as school closure or travel bans, would be the most effective, as they would greatly reduce overall population mobility.

The **goals** of this report are:

- to understand what measures have been the most effective in decreasing the number of COVID-19 cases;
- to predict possible outcomes of enforcing specific mitigation measures; and
- to suggest which strategies are best for mitigating the number of new COVID-19 cases.

2 Datasets

The key step in our multi-factorial analysis is to integrate several heterogeneous data sources. In particular, we need country-level data on the number of confirmed COVID-19 cases, on the implemented policies and restrictions, on population mobility, as well as information about income level, average temperature, GDP etc.

For this analysis, we used the following **data sources**:

- The COVID-19 Data Repository provided by Johns Hopkins University,
- Coronavirus Government Response Tracker provided by the University of Oxford,
- Community Mobility Reports provided by Google,
- the World Bank, and
- Wikipedia.

From the above data source we used the following **data**: the total number of COVID-19 cases reported in each country, including infection and death rates, as well as socio-economic data for each country describing its overall population, population density, climate, GDP, GDP per capita, and wealth level. In addition, we aligned COVID-19 rates with the data on the governmental containment measures. Finally, we used the mobility dataset that for each country reports the percent change in visits to public places: retail and recreation places, groceries and pharmacies, parks, transit stations, workplaces, and residential areas.

The datasets have the following properties:

1. **COVID-19 infection and death rates** are reported in absolute numbers per country per day.
2. Countries may have different methodologies of reporting COVID-19 infected cases (this depends on how ubiquitous the testing system is).
3. Countries may also have different methodologies of reporting COVID-19 death cases (e.g., in Germany every patient that died while diagnosed with COVID-19 is counted as dead from COVID-19 vs. in Russia patients that died from comorbidities while also having COVID-19 are not part of the official statistics).
4. It is beyond the scope of this project to normalise COVID-19-related data with respect to the issues from (2) and (3). We rely on the official numbers provided by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University.

5. **Containment measures** are also reported per country per day. The data collected by the Oxford COVID-19 Government Response Tracker (OxCGRT) classifies all possible measures into 19 types, of which 13 types are of particular interest for us, as they deal with containment, closure policies, and the healthcare system, e.g., schools closure, contact tracing, or ban of international travels.
   Here is the full list of containment measures used in this report together with their codes that are used both in the original dataset:
   - C: containment and closure policies
     - C1: school closing
     - C2: workplace closing
     - C3: public events cancellation
     - C4: restrictions on gatherings
     - C5: public transport closure
     - C6: stay-at-home requirements

        - C7: restrictions on internal movement
        - C8: international travel controls
    - H: health system policies
        - H1: public information campaigns
        - H2: testing policy
        - H3: contact tracing
        - H4: emergency investment in healthcare
        - H5: investment in vaccines

6. For each containment measure the data is reported by country by case, both as a binary flag (a given measure was in place at a particular date) and a grade between 0 and 2 or 3 (0 meaning the given measure is not present and 3 being the strictest implementation of the measure).
7. We **do not** consider economic measures that helped sustain the economies; this is outside the scope of our project. We focus on containment measures only.

8. The percentage of missing values is very low, and we treat missing values as 0.
9. The **mobility data** is presented in the form of a time series between January and October 2020.

10. **Demographic data** is also reported per country. The population data is reported per country per year, up until 2019.
11. **Income** is reported using one categorical value per country: each country is classified either as *High income*, *Upper middle income*, *Lower middle income*, or *Low income*. Additionally, for each country we used the 2019 figures of GDP and GDP per capita, as reported by the World Bank.
12. Finally, we factored in the **climate** of each country, as some scientists claim that the virus is more likely to spread in colder temperatures. We used historic climate data from the Climatic Research Unit that reports the average yearly temperature per country as a single value.

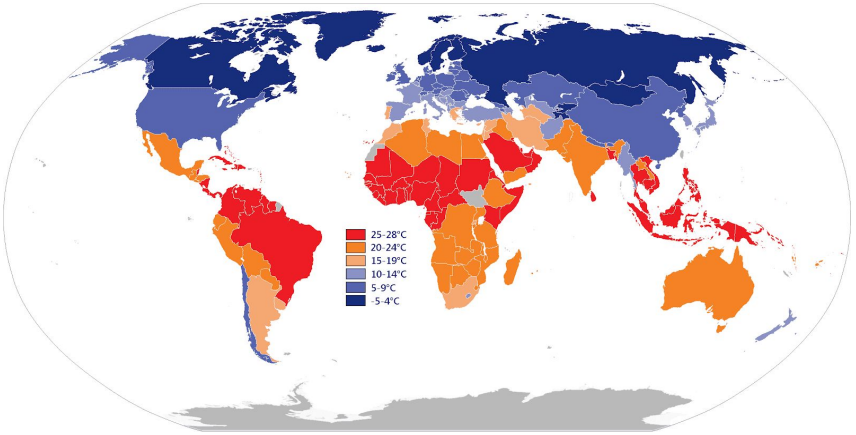


Figure 1. Yearly temperature average [8]

3 Exploratory Data Analysis

First we look at the global data and analyse how fundamental geopolitical parameters, such as GDP, location, climate or population density, correlate with the spread of the virus in each country.

3.1 Random Variables Used

We use the following **variables** taken directly from the original dataset:

- *population size,*
- *population density,*
- *absolute GDP,*
- *GDP per capita,*
- *income group,*
- *average yearly temperature,*
- *number of COVID-19 cases,*
- *number of COVID-19-related deaths.*

Additionally, we derived the following three variables:

- *mortality rate* (calculated as a the ratio of death cases to all COVID-19 cases),
- *normalized number of cases* (number of cases divided by population),
- *normalized number of deaths* (number of deaths divided by population).

These three variables are our target variables: we are interested in whether there is any correlation between factors like income, temperature, or population density and the normalized infection rates. Figure 3 depicts a sample of the integrated dataset with all variables, both original and derived.

	Country name	Confirmed cases	Deaths	Mortality rate	Normalized cases	Normalized deaths	Population	Density	Income group	GDP	GDP per cap	Temperature
0	Afghanistan	39192.0	1453.0	3.707389	0.001030	0.000038	38041754.0	56.937760	Low income	1.910135e+10	2293.551684	12.60
1	Angola	4672.0	171.0	3.660103	0.000147	0.000005	31825295.0	24.713052	Lower middle income	9.463542e+10	6929.678158	21.55
2	Albania	13153.0	375.0	2.851061	0.004608	0.000131	2854191.0	104.612263	Upper middle income	1.527808e+10	14495.078514	11.40
3	Andorra	1836.0	53.0	2.886710	0.023800	0.000687	77142.0	163.842553	High income	3.154058e+09	NaN	7.60
4	United Arab Emirates	90618.0	411.0	0.453552	0.009275	0.000042	9770529.0	135.609110	High income	4.211423e+11	69900.877848	27.00

Figure 3. A sample of the dataset used for EDA

3.2 Data Distribution

In total, we collected data for 182 countries. Table 1 summarizes the descriptive statistics of the original numeric variables. The data points are very diverse with respect to all variables, and with all variables but *temperature* having high variance and standard deviation. We tried removing the outliers for prior to Correlation Analysis (Section 3.3), but the changes in correlations were insignificant, therefore we kept the full dataset for the subsequent analysis.
The overall amount of missing data in the dataset is 1.2%, with *population density*, *GDP*, *GDP per capita*, and *average yearly temperature* having 2.2%, 7.1%, 7.7%, and 1.6% missing values, respectively.

Variable	# values	min	max	median
Confirmed cases	182	19	7078039	13379.5
Deaths	182	0	228	204486
Population	182	33860	9758033	1397.7 mio
Population density	178	2.041	83.949	7952.998
GDP	169	429 mio	47319 mio	21374419 mio
GDP per cap	168	782.82	14747.95	121292.74
Temperature	179	-5.1	21.85	28.29

Table 1. Descriptive statistics of the original numeric variables

3.3 Correlation Analysis

Figures 4 and 5 depict the **correlation matrices** of all numerical variables that use Pearson's and Spearman's coefficients, respectively. The results between the two matrices are consistent, with Spearman's coefficient producing slightly stronger correlation scores. We are primarily interested in the areas of the matrices that lie at the intersection of *mortality rate*, *normalized cases* and *normalized deaths*, and the rest of the variables, although the remaining parts of the matrices also contain interesting insights (e.g, there is a tangible negative correlation between the average temperature and the country's GDP or GDP per capita).

The results in Figures 4 and 5 suggest that there is a weak negative correlation between COVID-19 rates, in particular death rates, and the country's **climate**. This backs the observation of many scientists and medical practitioners that coronaviruses have seasonality and typically spread faster in colder temperatures (i.e., in colder

months and colder climates) [10]. However, temperature is not the factor that alone is able to counteract the virus; it can only alleviate the situation.

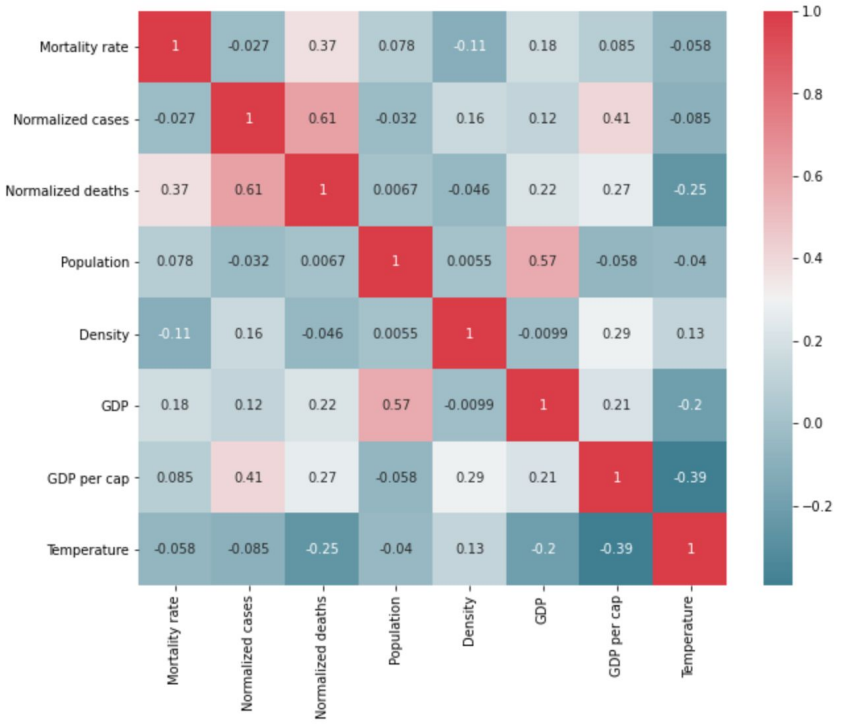


Figure 4. A heatmap of Pearson's correlation between variables

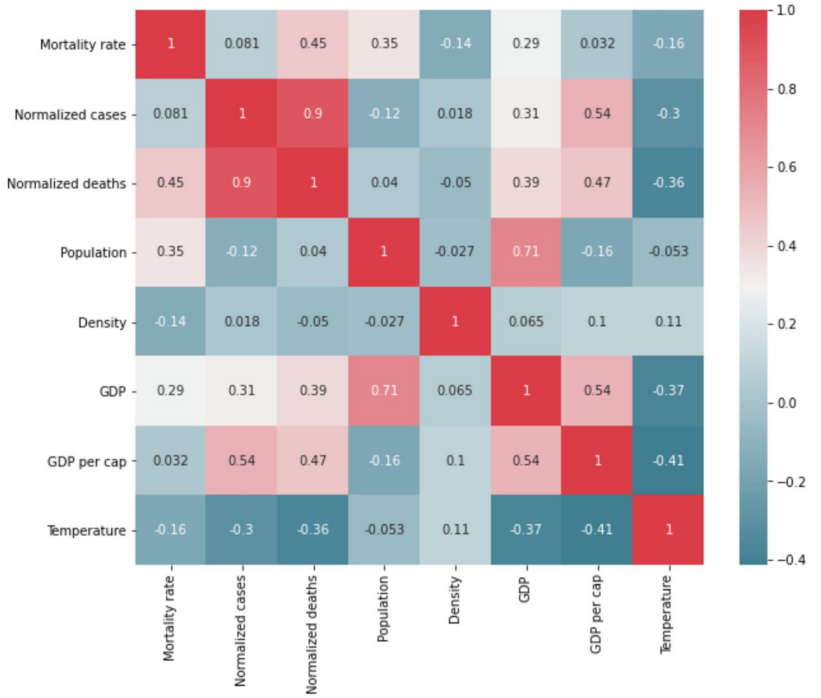


Figure 5. A heatmap of Spearman's correlation between variables

Economic measures, such as the gross domestic product (GDP) and the per capita GDP, have moderate positive correlation with COVID-19 rates, in particular with the number of confirmed cases. As we have already mentioned in Section 2, the testing capacities as well as the guidelines for keeping the statistics of COVID-19-related deaths vary from country to country. In fact there is a linear dependency between the GDP per capita and the number of COVID-19 tests performed per 1000 people (see Figure 6), which explains the positive correlation.

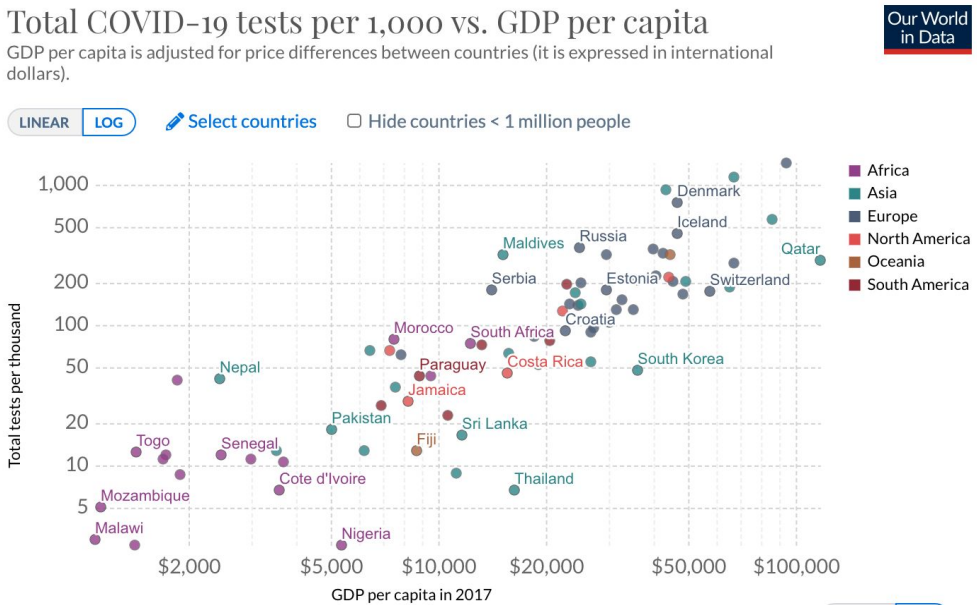


Figure 6. Total COVID-19 tests per 1,000 vs. GDP per capita [9]

Most interestingly, population **density** has little to no correlation with infection rates. This may seem counterintuitive as the virus is expected to spread faster when there are more face to face interactions. In fact, in cases when a group of people leave in a tight, restricted environment, e.g., a prison or a nursing home, the COVID-19 virus can cause devastating outbreaks [11].

Then why is it that the data tells us differently? Our **hypothesis** is that the virus has not been spreading 'at full capacity', as its circulation has been mitigated with various restriction measures that play the key role in how the pandemic is handled country by country. We are going to test this hypothesis in Section 4.

3.4 EDA Takeaways

The epidemiological situation in each country correlates poorly with fundamental geopolitical parameters such as GDP, location, population density etc. Countries of

the same size and economic cohort, in the same region and with the climate have handled the pandemic in dramatically different ways. This implies that on the one hand, the virus does not choose, it attacks people and societies of any socio-economic background. And on the other hand, it also means that the spread of the virus depends primarily on the immediate actions and strategies of the governments rather than on geopolitical factors and predefined settings. **Everything is in our hands**.

We assume that the spread depends predominantly on the measures taken by each government. In the next section we will analyse how various containment measures affect the spread of the disease. Due to the exponential nature of the problem, we are unable to analyse all measures implemented in all countries. Therefore, in the next section we will focus on the European region, specifically on the **following 9 countries**: United Kingdom, Russia, Poland, Italy, Lithuania, France, Sweden, Spain, and Czech Republic. The choice of the countries is motivated by their diversity: not only do these countries have different climates, economies and geographies, but they also reacted very differently to the spread of COVID-19, with complete lockdows in the UK and Italy, to moderate temporary measures in Poland and Russia, to minimal government response in Sweden.

4 Statistical Analysis & Machine Learning

After the initial data exploration phase, we decided to construct prediction models based on three baseline datasets that helped us to capture the relationships between the enforcement of specific COVID-19 prevention measures, mobility, and the number of daily confirmed COVID-19 cases.

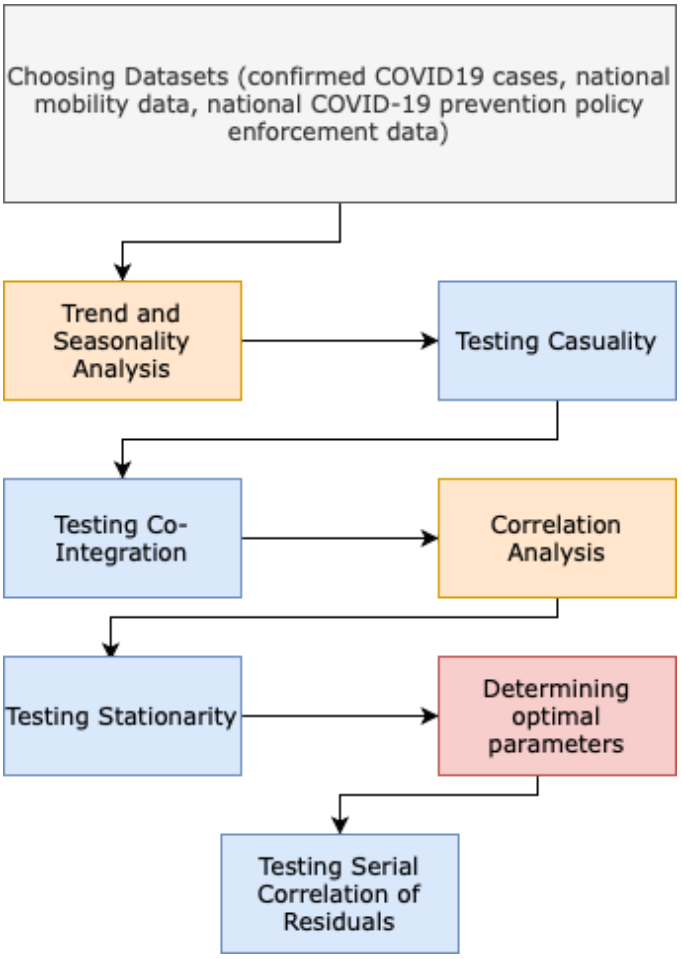


Figure 7. The Data Analysis and Forecasting Pipeline

Figure 7 shows the data analysis and forecasting process flow of this project. In the following subsections, we will discuss in detail every step of the process and present the results.

Disclaimer: Due to the space limitation, we present and discuss only a subset of the countries we analysed. The remaining data, analysis, and forecast results are uploaded on our Github.

4.1 Seasonal and Trend Decomposition

Our main goal is to advise on the enforcement of specific COVID-19 prevention policies to decrease the number of COVID-19 cases in various countries. In order to achieve this goal, we need to predict the trend of the data that describe confirmed COVID-19 cases after the potential enforcement of a proposed measure. We begin our analysis by looking at the seasonal decomposition of daily confirmed cases in a subset of European countries.

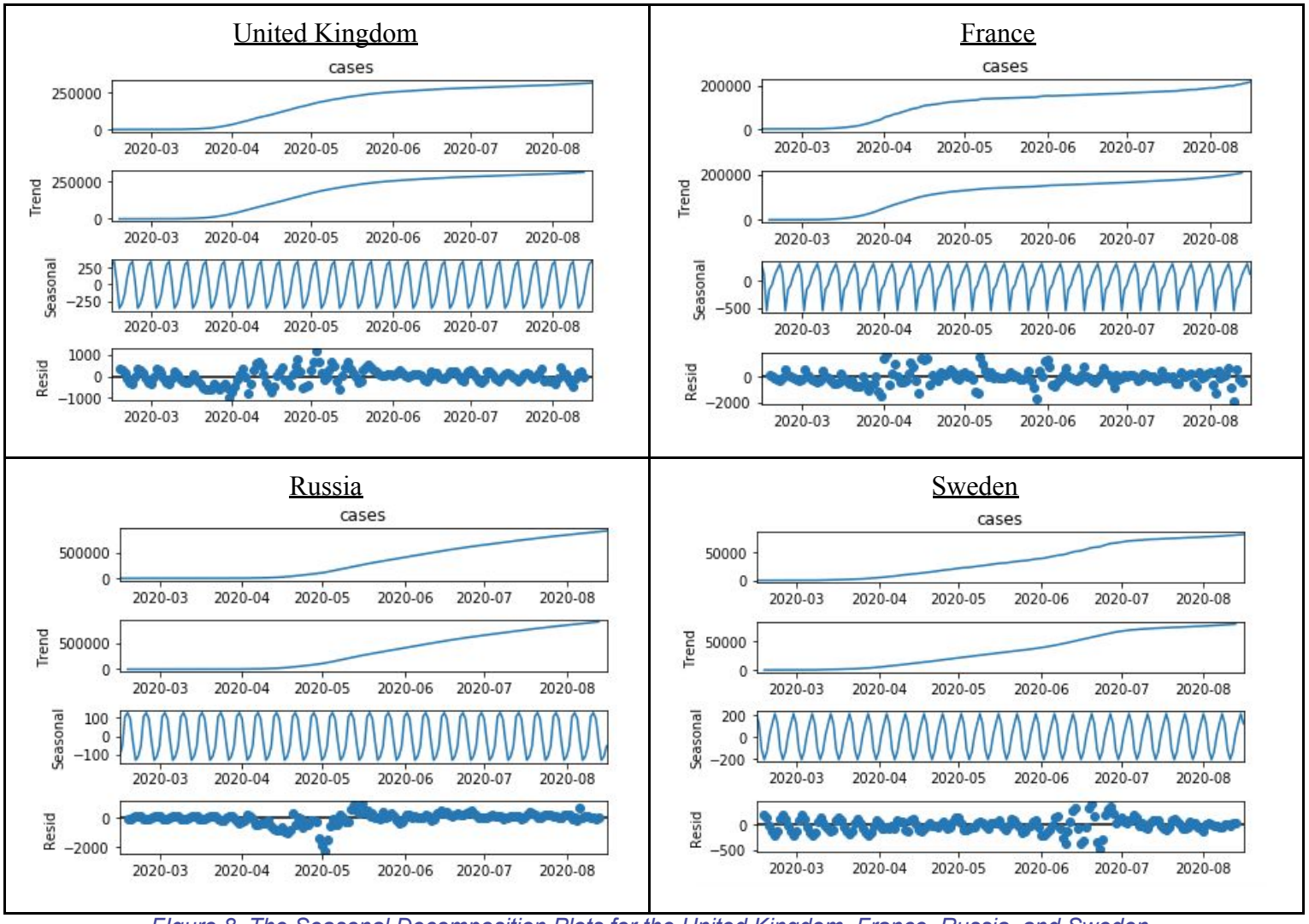


FIgure 8. The Seasonal Decomposition Plots for the United Kingdom, France, Russia, and Sweden

FIgure 8 shows additive seasonal decomposition plots for selected countries. We chose the additive model because the seasonal variation seemed constant, regardless of the increase in the number of reported cases. These plots confirm the increasing trend and expected weekly seasonality. It also shows that there is some irregular fluctuation of residuals in certain periods (different for each analysed country).

4.2 Granger's Causality Test

The next step was to verify if three types of time series data influence each other. We hypothesized that the enforcement of the COVID-19 prevention policies will have a direct impact on national mobility (i.e. movement in parks, shops, and public transportation), and mobility will influence the number of reported COVID-19 cases. In order to verify this, we tested causation [2]. We could just build a model and evaluate this to disregard or accept this hypothesis, however, it is a good practice to do that before building the prediction model.

The null hypothesis (H_0) stated that the values of the first time series do not cause the other one. We chose the standard threshold (i.e. significance level) of 0.05 for p-value. We computed the results for up to 12 lags.

During the first phase, we tested causation for the number of daily cases and different types of mobility data. As an example, we present the results for retail and recreation mobility data in the following table.

Country Name	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
United Kingdom	.0	.3058	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
France	.0	.0047	.0003	.0052	.0516	.0744	.0863	.1137	.0072	.0202	.0196	.0394
Russia	.0	.004	.0011	.0043	.0006	.0	.0001	.0002	.0006	.0026	.0006	.0021
Sweden	.0767	.0001	.0	.0	.0	.0	.0012	.002	.0033	.0078	.0126	.0003

Table 2. The p-values obtained as a result of applying Granger's Causality test to the COVID-19 cases and retail_recreation mobility datasets

As expected, we observed that causation is present. For instance, Table 2 shows that we can reject our H_0 given lags of 1-4 and 9-12 for France. This is important because the performance of prediction models depends on selecting appropriate parameters. Thus, we noticed that the number of lags for every country have to be individually adjusted (i.e. different numbers of lags will be selected for different countries). Also, this helped us to determine which combinations of time series are best (i.e. we observed that each type of mobility impacts the number of cases, however, not every type of COVID-19 prevention policy will affect mobility).

Policy Type	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12
C4	.1929	.4787	.619	.7474	.7734	.8562	.8618	.883	.9293	.9444	.9242	.8355
C6	.0006	.0095	.0122	.0158	.0292	.0114	.0001	.034	.0347	.0339	.1187	.1156
C7	.4496	.0306	.0103	.0007	.0004	.0004	.0	.0001	.0002	.0003	.0017	.0033
H3	.1423	.1136	.1705	.2165	.3794	.3693	.2472	.6951	.6109	.5417	.3095	.2208

Table 3. The p-values obtained as a result of applying Granger's Causality test to retail_recreation mobility and policy datasets for France

Table 3 shows the results of the causality tests for retail/recreation mobility time series and different policies in France. It is crucial to identify and understand how the different types of prevention measures affect certain types of mobility. We identified that, for instance, the policies C4 and H3 do not cause changes in retail and recreation mobility time series. However, the policies C6 and C7 do affect the retail/recreation mobility time series. We ran such tests for every possible combination, excluding the policies for which the time series values were constant.

Disclaimer: We would like to highlight that we obtained p-values of 0.0 are not necessarily equal to 0. The rounding operation is applying to the results of this test (i.e. it means that the p-value was extremely small).

4.3 Johanson's Cointegration Test

The next step in our journey was to determine if a set of non-stationary time series can result in a stationary linear combination. Thus, we implemented Johanson's cointegration test [3] and ran it for different numbers of lags, using various combinations of <cases, mobility_{xi}, policy_{xi}>. We used this metric to, for instance, determine the cointegration rank for a VECM (this model is discussed in one of the subsections of this report).

The null hypothesis (H_0) indicated that time series were not cointegrated. We chose, once again, the standard significance level of 0.05. We computed the results for up to 12 lags, however, we report up to 6 for the demonstration and explanation purposes. Moreover, we use a trace statistic to report these results.

Mobility Type	**L1** (trace > C(95%))	**L2** (trace > C(95%))	**L3** (trace > C(95%))	**L4** (trace > C(95%))	**L5** (trace > C(95%))	**L6** (trace > C(95%))
Retail/Recreation	64.48 > 24.2761 17.37 > 12.3212 0.08 < 4.1296	55.41 > 24.2761 12.68 > 12.3212 0.13 < 4.1296	37.53 > 24.2761 10.4 < 12.3212 0.03 < 4.1296	25.91 > 24.2761 6.92 < 12.3212 0.0 < 4.1296	20.44 < 24.2761 4.87 < 12.3212 0.06 < 4.1296	24.81 > 24.2761 4.05 < 12.3212 0.15 < 4.1296
Parks	50.29 > 24.2761 12.92 > 12.3212 0..13 < 4.1296	35.24 > 24.2761 9.64 < 12.3212 0.15 < 4.1296	32.87 > 24.2761 8.07 < 12.3212 0.16 < 4.1296	26.33 > 24.2761 6.73 < 12.3212 0.2 < 4.1296	20.55 < 24.2761 4.38 < 12.3212 0.26 < 4.1296	19.23 < 24.2761 2.56 < 12.3212 0.4 < 4.1296
Residential	110.34 > 24.2761 17.21 > 12.3212 0.03 < 4.1296	61.71 > 24.2761 12.12 < 12.3212 0.03 < 4.1296	57.87 > 24.2761 9.72 < 12.3212 0.04 < 4.1296	68.82 > 24.2761 6.7 < 12.3212 0.07 < 4.1296	52.65 > 24.2761 4.4 < 12.3212 0.09 < 4.1296	33.2 > 24.2761 2.77 < 12.3212 0.01 < 4.1296
Workplaces	105.07 > 24.2761 16.22 > 12.3212 0.04 > 4.1296	55.36 > 24.2761 11.56 > 12.3212 0.03 > 4.1296	51.37 > 24.2761 10.56 > 12.3212 0.04 > 4.1296	54.17 > 24.2761 7.28 > 12.3212 0.06 > 4.1296	37.02 > 24.2761 4.9 < 12.3212 0.1 < 4.1296	26.97 > 24.2761 3.31 < 12.3212 0.07 < 4.1296

Table 4. The values obtained as a result of applying Johanson's Cointegration test to the number of cases, the C3 policy, and a specified mobility for Czech Republic

The first value in each column and row (highlighted in red) indicates that the time series of confirmed COVID-19 cases is or isn't a linear combination of at least one of the remaining time series (i.e. mobility_{xi} and policy_{xi}). The trace values smaller than the critical value at 95% confidence level show that the cointegration doesn't exist; thus, we can confirm H_0. Other values indicate the same property for, successively, the mobility and policy time series.

Table 4 shows the trace test results for Czech Republic with selected mobility types (i.e. retail/recreation, parks, residential, and workplaces) and the C3 policy. We observed that the COVID-19 cases time series are indeed linear combinations of the workplace and residential mobility and the C3 policy. We also rejected H_0 for the park mobility up to the 4th lag and retail/recreation, excluding the 5th lag.

4.4 Autocorrelation and Partial Autocorrelation Analysis

We examined the correlations between the current observation and the different number of adjacent observations, lags, denoted on the *x*-axis. All values on the *y*-axis close to 0 indicate a lack of correlation and the values approaching 1.0 indicate a strong correlation.

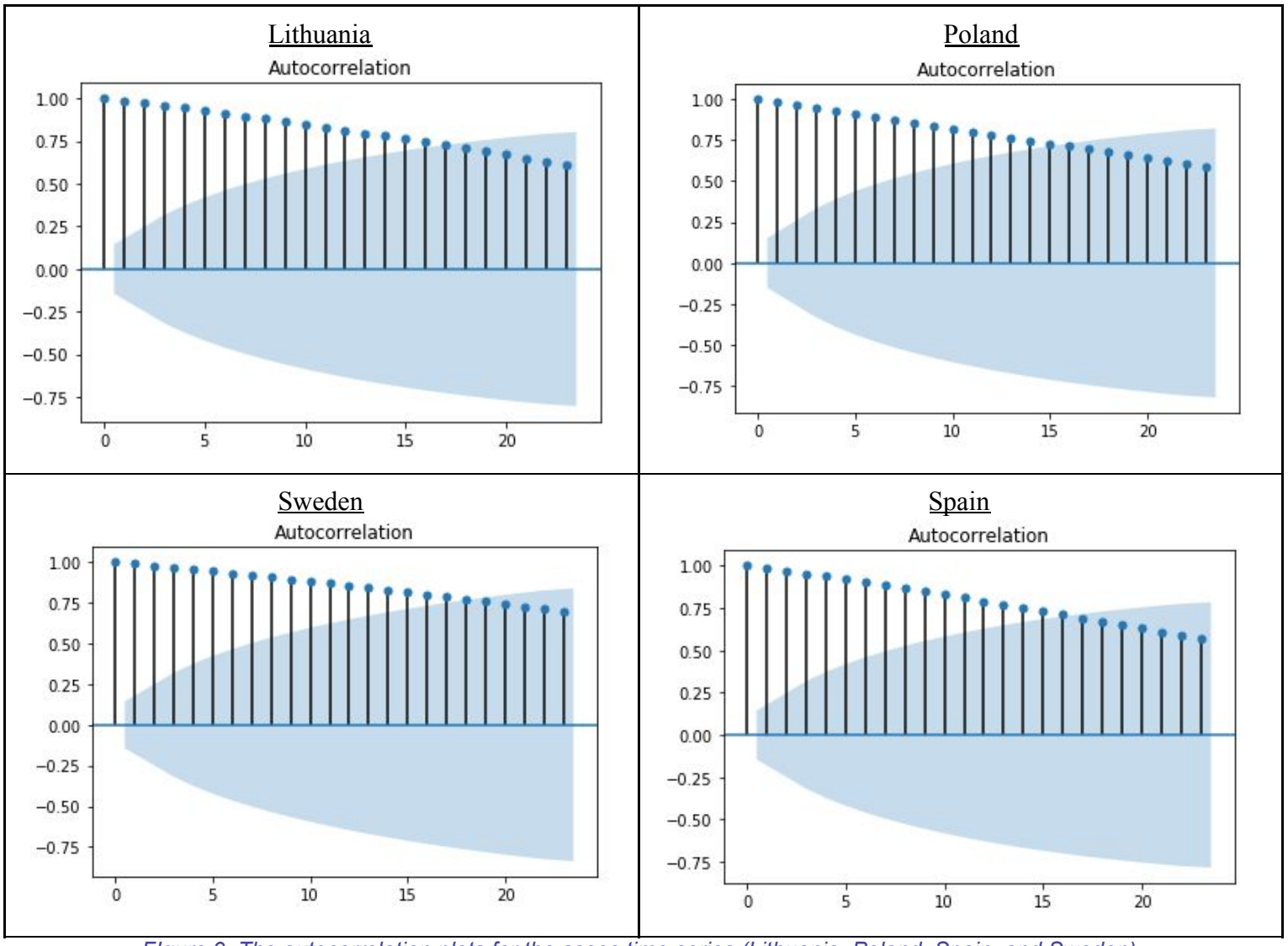


FIgure 9. The autocorrelation plots for the cases time series (Lithuania, Poland, Spain, and Sweden)

Figure 9 shows autocorrelation plots for Lithuania, Poland, Spain, and Sweden. The serial correlation analysis yielded similar results for other European countries we analyzed, including the United Kingdom, Czech Republic, Russia, Italy, and France. The blue areas marked on the plots denote 95% confidence intervals. This means that all the values outside these areas can be treated, indubitably, as a correlation.

After investigating the autocorrelation plots, we can determine that there are strong correlations for the lag lengths of 1-15. This was later confirmed by the greedy search of optimal hyperparameters for some prediction models we implemented.

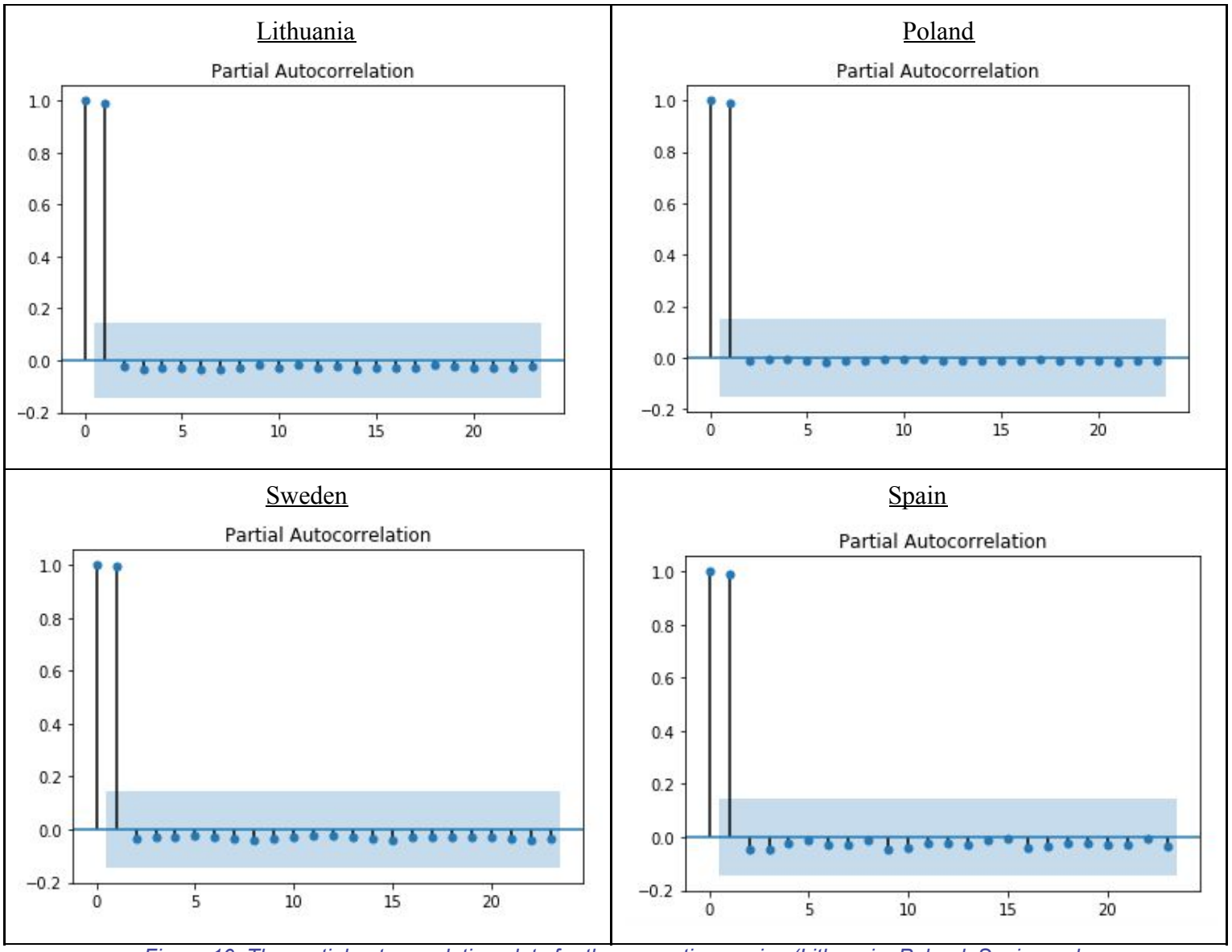


Figure 10. The partial autocorrelation plots for the cases time series (Lithuania, Poland, Spain, and Sweden)

The partial autocorrelation function removes the effects of the correlations from the previous lags of shorter lengths. In this case, we observed that there is only one significant correlation at lag 1. We obtained similar results for other analyzed countries. This helped us determine the order of autoregressive models.

4.5 Augmented Dickey-Fuller Test

The final step before implementing the prediction models was to test if the time series is stationary. Various forecasting models (e.g. VAR and SARIMAX) require the time series to have no change in variance and mean over time. This simply means that we should remove certain properties of the time series, including a trend.

The time series variables we used typically required 0 or 1 differencing in order to be applied to the prediction models. We verified that by running an Augmented Dickey-Fuller (ADF) [4] test on our data. Every time we differenced the data, we checked again for stationarity.

Country	Cases	Residential	Workplaces	Transit	C2	C6
United Kingdom	0	1	1	1	1	1
Russia	0	1	1	1	1	1
Italy	0	1	1	0	0	0
France	2	1	1	1	1	1

Table 5. The number of times the differencing is required for time series variables. The green columns indicate selected measures, the blue columns indicate mobility variables, and the pink column indicates the time series of cases

Table 5 shows that differencing had to be applied twice to the French time series of COVID-19 cases. The values of 1 indicate that differencing had to be applied only once to make the time series stationary (e.g. the residential time series of selected countries had to be differenced once in each case). The value of 0 suggests that the differencing was not needed; thus, the time series was already stationary (e.g. the Italian C2 policy time series).

4.6 Forecasting and Determining Optimal Parameters for Various Models

4.6.1 SARIMAX

The first model we implemented and trained was Seasonal Autoregressive Integrated Moving Average with eXogenous variable (SARIMAX) [5]. This model uses 7 different parameters to construct its predictions:

1) **(p, d, q)** -> (the order of the autoregressive (AR) model, the degree of differencing, and the order of the moving average (MA) model)
2) **(P, D, Q, m)** -> (the order of the seasonal component for the AR model, the integration order of the seasonal process, the order of the seasonal component for the MA model, and the number of observation per seasonal cycle)

Since this model supports only one additional variable as an external predictor, we used pairs of <cases, mobilityxi> and <mobilityxi, policyxi>. First, we trained a model to predict a certain type of mobility given a policy time series and used a grid search to find optimal hyperparameters. The grid search confirmed the outcomes of the analysis discussed in previous sections.

Figure 11 shows plots of true values versus predictions for grocery/pharmacy mobility based on different prevention measures in the last 30 days. Our aim was to determine which measures could potentially help decreasing COVID-19 cases in various regions, so predicting accurately how certain policies influence mobility in these regions is crucial. We implemented a separate model for each country, which is available on our Github page. In some cases like Poland, brute force search had to be performed to find better parameters.

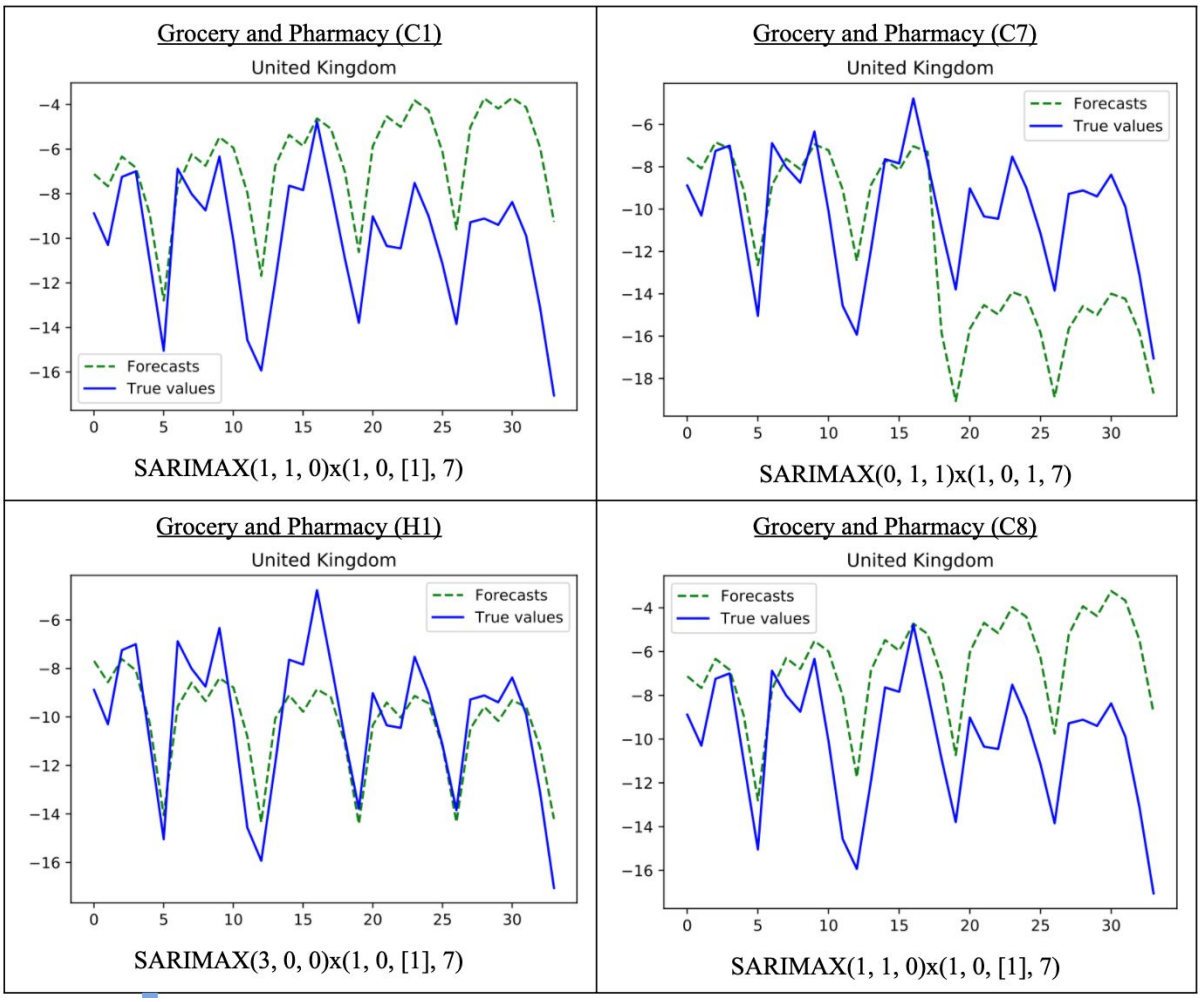


Figure 11. The forecasts of grocery/pharmacy mobility based on enforcing different COVID-19 prevention measures (C1, C7, H1, C8)

Figure 12 shows the residual plots to evaluate the model's predictions. Generally, we can see that the residuals are normally distributed and they have a mean around 0. This indicates that these values don't store additional information that could be used by the model to predict mobility values. Also, it seems that there is no correlation pattern left. Overall, it indicates pretty good performance of the model.

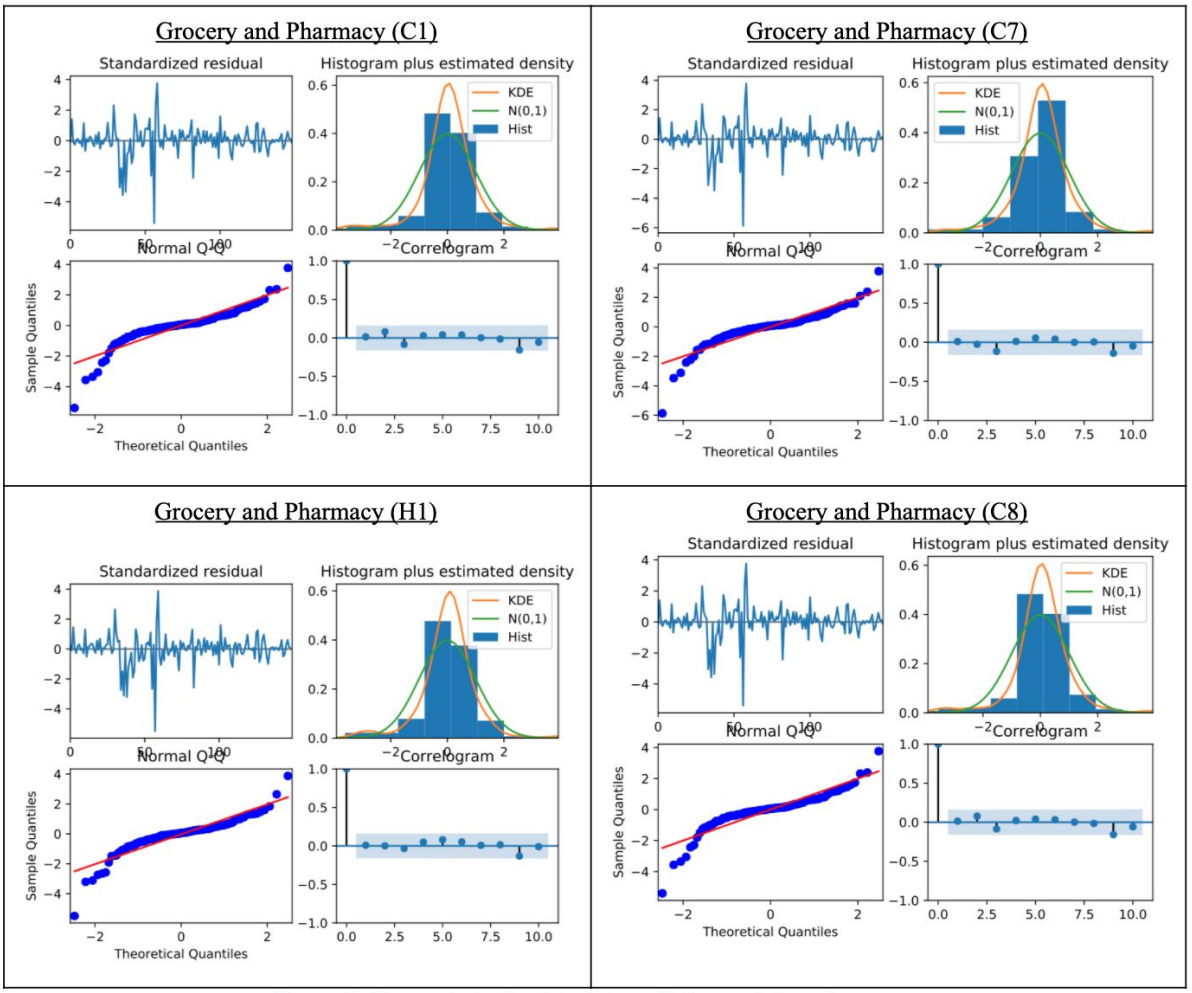


Figure 12. The analysis of residuals to verify the model's performance

Finally, we can predict cases based on the grocery and pharmacy mobility data.

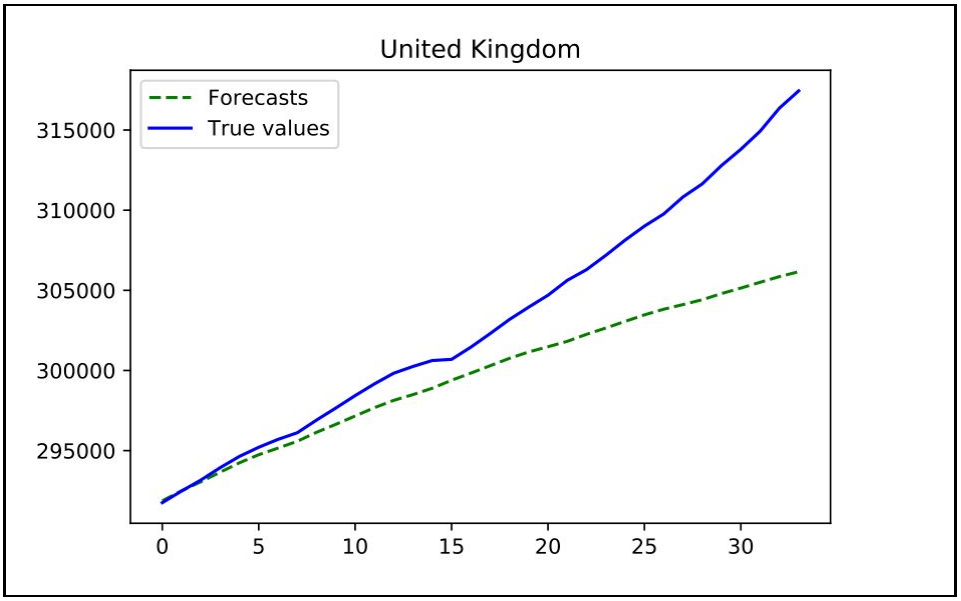


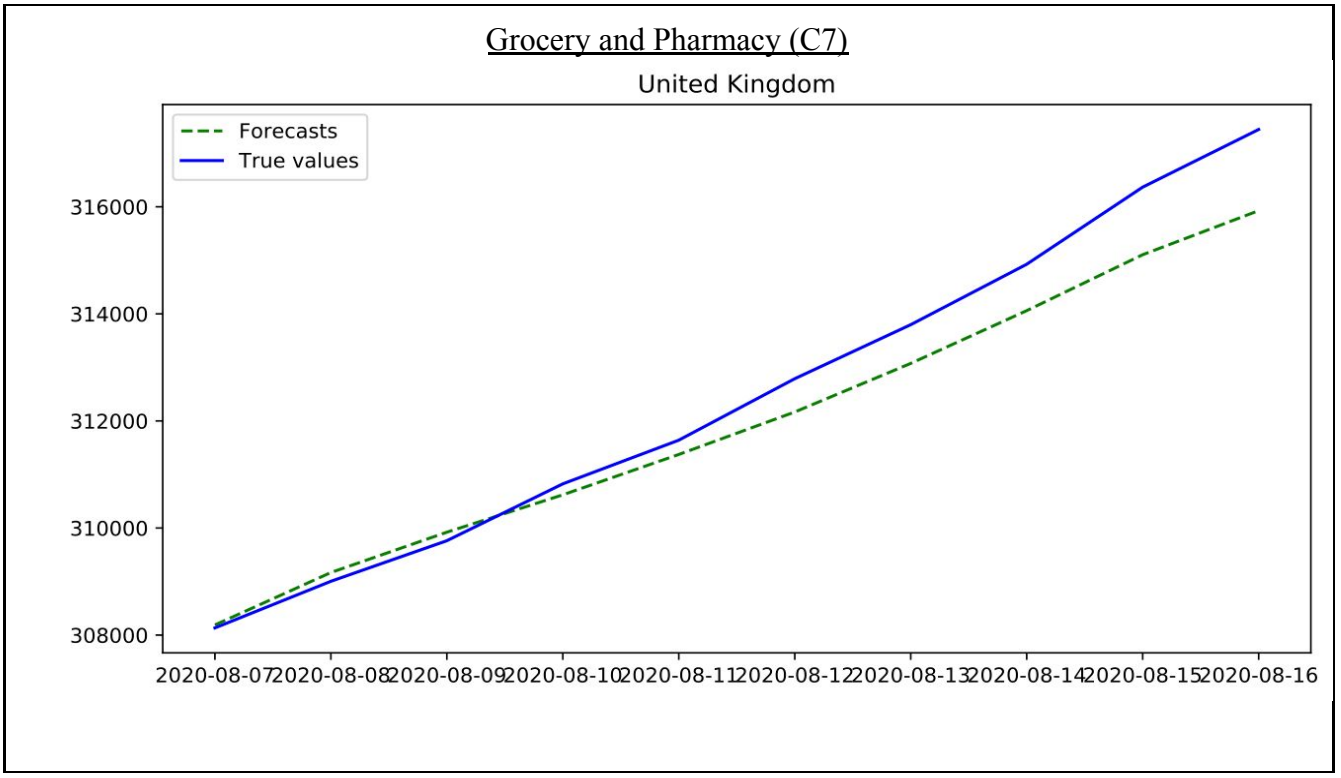
Figure 13. The prediction of the COVID-19 cases given the grocery and pharmacy mobility data

Figure 13 presents the final prediction for the COVID-19 cases given the grocery and pharmacy mobility data. Both Johansen's cointegration and Granger's causality tests reject the null hypothesis for this combination, which means that the pair of <cases, grocery_pharmacy_mobility> is cointegrated and has a causation relationship.

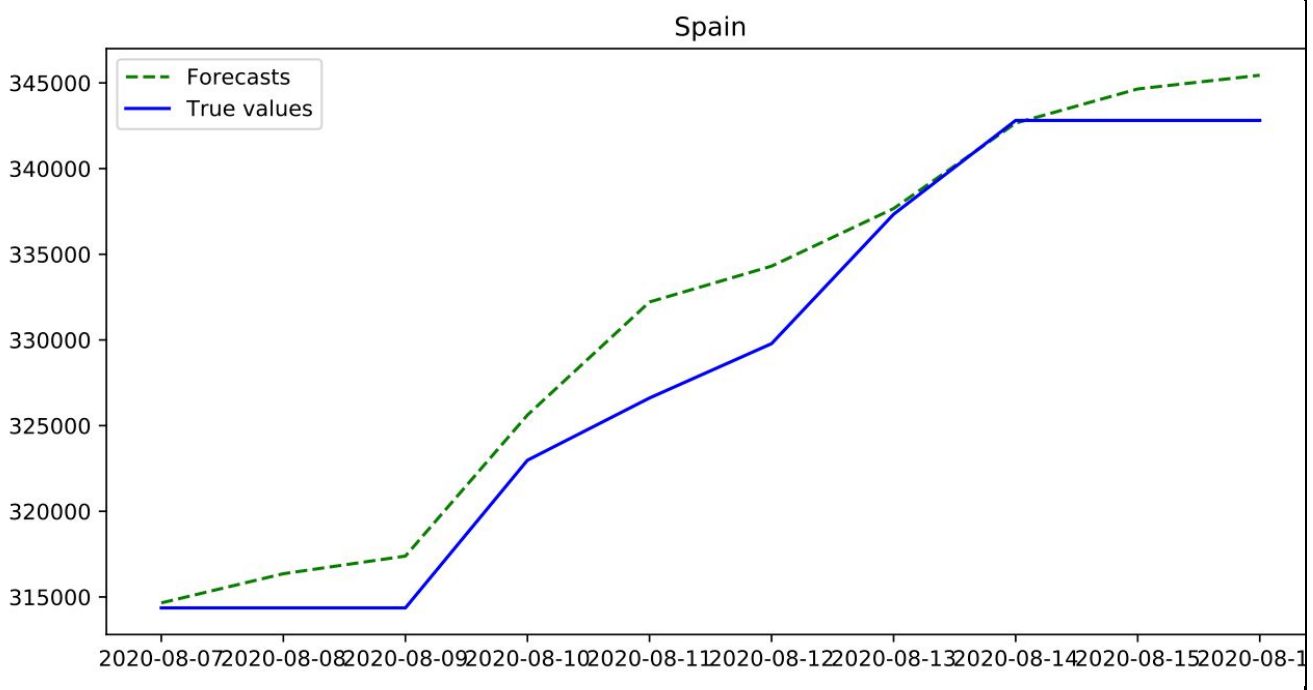
4.6.2 VECM

The next model we implemented was the Vector Error Correction Model (VECM) [6] that allows using multiple non-stationary time series for forecasting. Thus, the differencing procedure is not required. As input parameters, this model takes the number of lagged differences in the model as well as the cointegration order (i.e. the number of cointegrating relationships). We ran a select order test to choose the best fit for the lag parameter based on the AIC metric as we intended to develop a forecasting model.

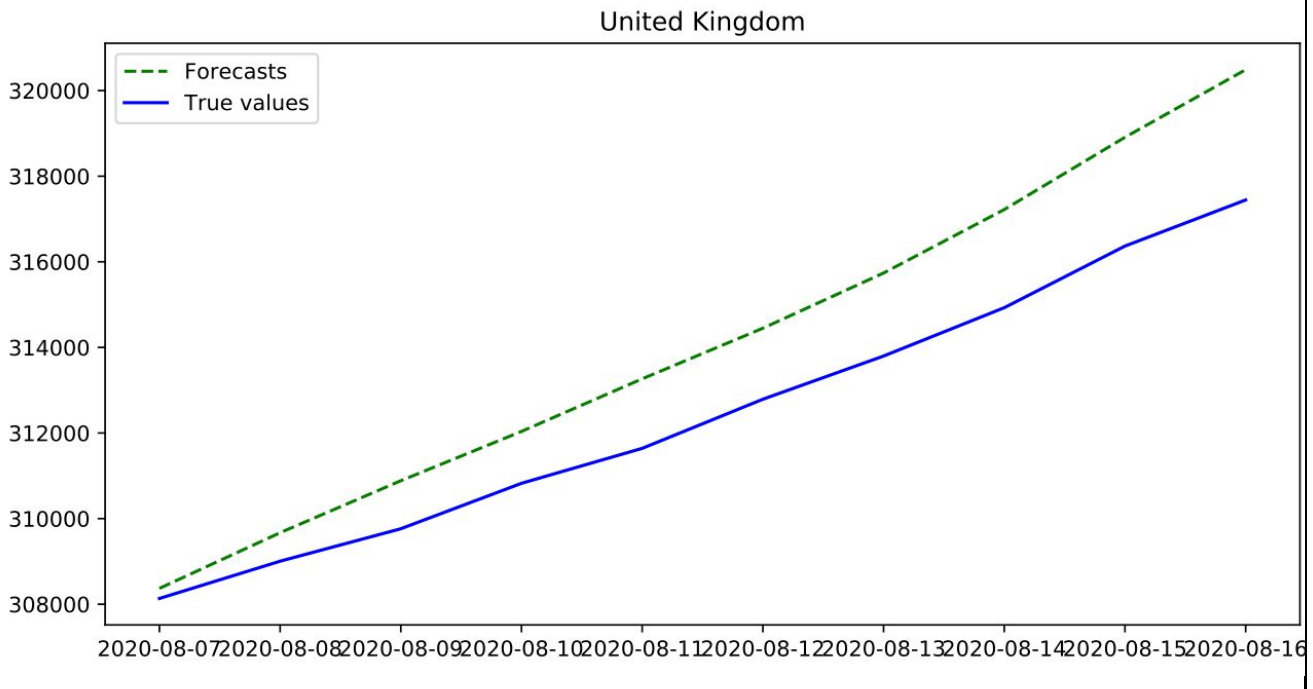
We implemented and trained our model to use triples of <cases, mobilityxi, policyxi> for forecasts of the number of confirmed COVID-19 cases in selected countries, given different combinations of mobility and policy data.



Grocery and Pharmacy (C7)



Parks (C7)



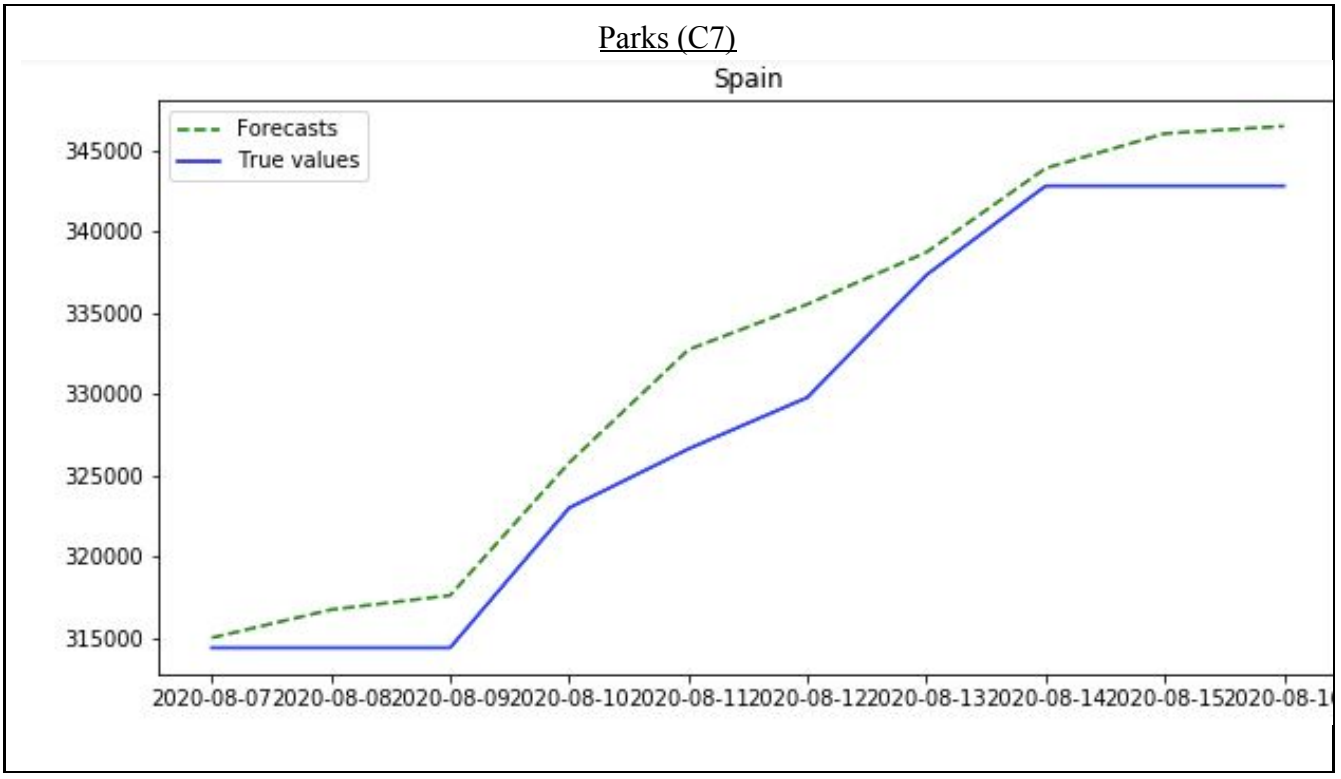


Figure 14. The VECM predictions of the number of COVID-19 cases for 10 days given the parks and grocery/pharmacy mobility data as well as the C7 policy in Spain and the United Kingdom

Figure 14 shows the predictions of the COVID-19 cases made by the VECM model for 10 consecutive days in Spain and the UK. The plots present the actual values versus the prediction results given the enforcement of C7 policy and taking into account the grocery/pharmacy and park mobility data. The cointegration tests confirmed that the cointegration exists with the time series of COVID-19 cases for the specified number of lags. We also executed a Durbin-Watson test to verify whether residuals contain unwanted patterns. In each case, we obtained the value of ~2.0, which means that there are no patterns in the residuals.

4.7 Building a Forecasting Model using Random Forests

In order to be able to advise the governments on the best policies to impose today, we tackle the following question: *"Given the level of mobility in the country 1-4 weeks ago as well as the combination of policies imposed during the same time frame, can we predict the number of COVID-19 cases in the next month?"* In this section, we adopt the Random Forest approach to achieve this task [7]. Random forests were shown to have stronger predictive ability than the usual ARIMA model when applied to data of avian influenza outbreak in Egypt [1]. Moreover, the random forest model is highly intuitive and offers fast performance while avoiding overfitting. Due to these reasons, random forest was a natural choice of predictive models in our situation. The

data used for this model is the mobility time series (features) and the change of the number of confirmed COVID-19 cases time series (output) expressed as the percentage change every day. The percentage change form was chosen to account for different population sizes between different countries, however, some issues of non-stationarity of data were encountered which did not prevent the model from working and will be discussed in this section.

4.7.1 Naive Model

We started our analysis by building a simple random forest model for the United Kingdom which used 1000 decision trees and included only the mobility variables without any time lag. The model was trained on 120 days of data and tested on another 64 days. Figure 15 shows excellent predictive ability of the model confirming the correlation between the mobility data and the number of COVID-19 cases. However, this framework is not suitable for *predictive* model as it is only comparing the *current* mobility values with the *current* number of cases. Next, we build a random forest model taking into account only the *previous* mobility data and we also extend the model to include the policies imposed by the government.

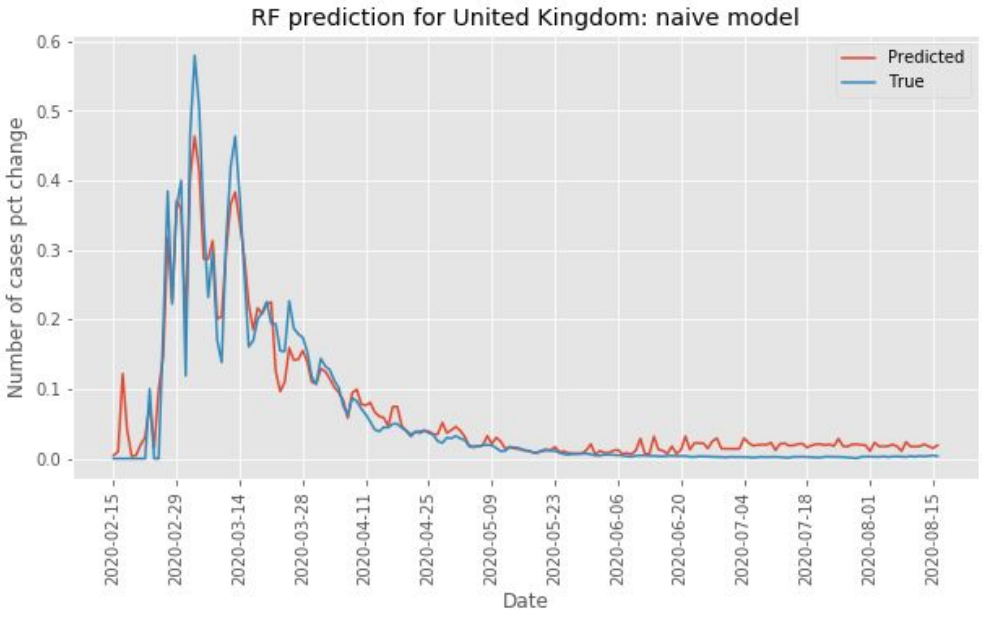


Figure 15. Results of the Naive Random forest applied to the data for the United Kingdom

4.7.2 Improved model: adding lagging data

To build a predictive model, lagging variables were chosen as the predictive features. All variables including the mobility data and the policy data were evaluated with 14, 21 and 28 days lag. In addition to this, rolling averages of each variable were considered as a potential feature but were later rejected to avoid including features which were directly derived from other features used in the model. The lagging data was chosen over rolling averages as it showed better performance. This added up to 51 features in total as an input to the random forest. To avoid overfitting the model, the minimum sample leaf size was restricted to 7 and the number of features allowed in each tree was restricted to 25 features (a half of total number of features). Again, random forest was made from 1000 estimators and 150 days of data were used for training the model. The last month of the data was used to test the prediction power of the model. In what follows, we discuss the performance of this model trained for the United Kingdom, Poland and Spain.

4.7.3 United Kingdom

While the model captures the main trends of the change of the number of cases (R squared error 0.92), during the prediction window it significantly overestimates the change of the number of cases (see Figure 16). This can also be seen in the plot of the percentage error of the prediction: the error becomes very large in the prediction window. One reason for this behaviour could be the clear non-stationarity of the time series, even after transforming it to the percentage change of the number of cases. Another reason would be overfitting the model while training it at high values of the change in the number of cases while testing the model in the prediction window of low values of the percentage change in the number of cases.

Nevertheless, the feature importance analysis reveals interesting results. The most significant features are shown in Table 6 together with the impurity-based feature importance coefficient, also known as the Gini index. The mobility data in workplaces, residential areas, transit stations, grocery stores, retail and recreational spaces 14 or 21 days ago are all significantly important predictors of the change in the number of cases with only three policies (workplace closing, restrictions on gatherings and domestic travel restrictions) contributing to the top 10 predictors. This confirms the expectation that accurate mobility data can be used to model an outbreak.

4.7.4 Poland

Application of random forest technique to Poland's data shows different results - see Figure 17. First of all, the time series of change in the number of cases is more stationary for this country, fluctuating only up to 0.1% (compared to 0.2% for the UK). As a result, a much more robust and well-behaved prediction model is obtained scoring a more realistic 0.87 R squared value. Moreover, the percentage error of the model shows homogeneous behaviour throughout both training and testing sets and stays very low in comparison to the previous case.

Surprisingly, the set of the most significant features in Poland's predictive model contains mainly various policies which restrict the mobility of the citizens (restrictions on gatherings, international travel, workplace and transport closure) in addition to the testing policy (see Table 7). The mobility data itself accounts only for 2 out of 10 most significant features and scores about 5 times smaller Gini index compared to the most significant feature. This raises the question whether the mobility data in this case was flawed and gives an insight into the policies that have the most significant impact onto the number of cases.

4.7.5 Spain

As the last example, consider the random forest model trained on the data from Spain - see Figure 18 for the results. The time series for the UK and Spain hold similarities in terms of high non-stationarity and result in a similarly performing model. Again, the R squared value is 0.92 and the model is over-predicting the number of cases during the prediction window. The most significant features together with their Gini index are shown in Table 8. There is a lot of overlap between the features included in the UK model and the Spain model such as the mobility data from retail and recreation spaces, grocery and pharmacy stores, transit stations. However, the set of features in the Spain model is more balanced between the mobility features and the policy features: 4 out of 10 of these features were policies: school closing, restrictions on gatherings and cancelation of public events. Thus, this model becomes a good candidate for a universal predictor model that could be used on the data from a variety of countries.

.

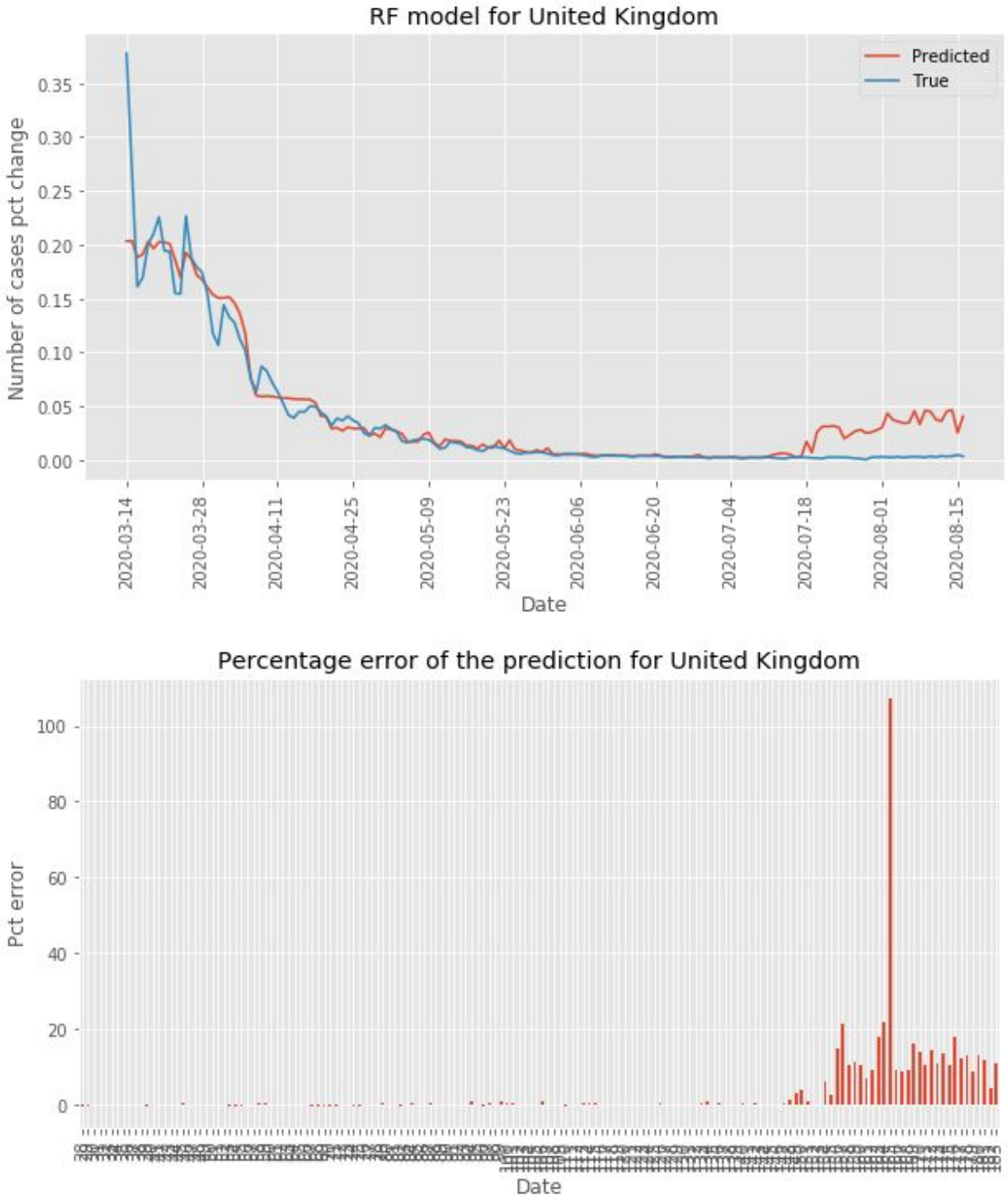


Figure 16. Predictive model for the United Kingdom and its percentage error

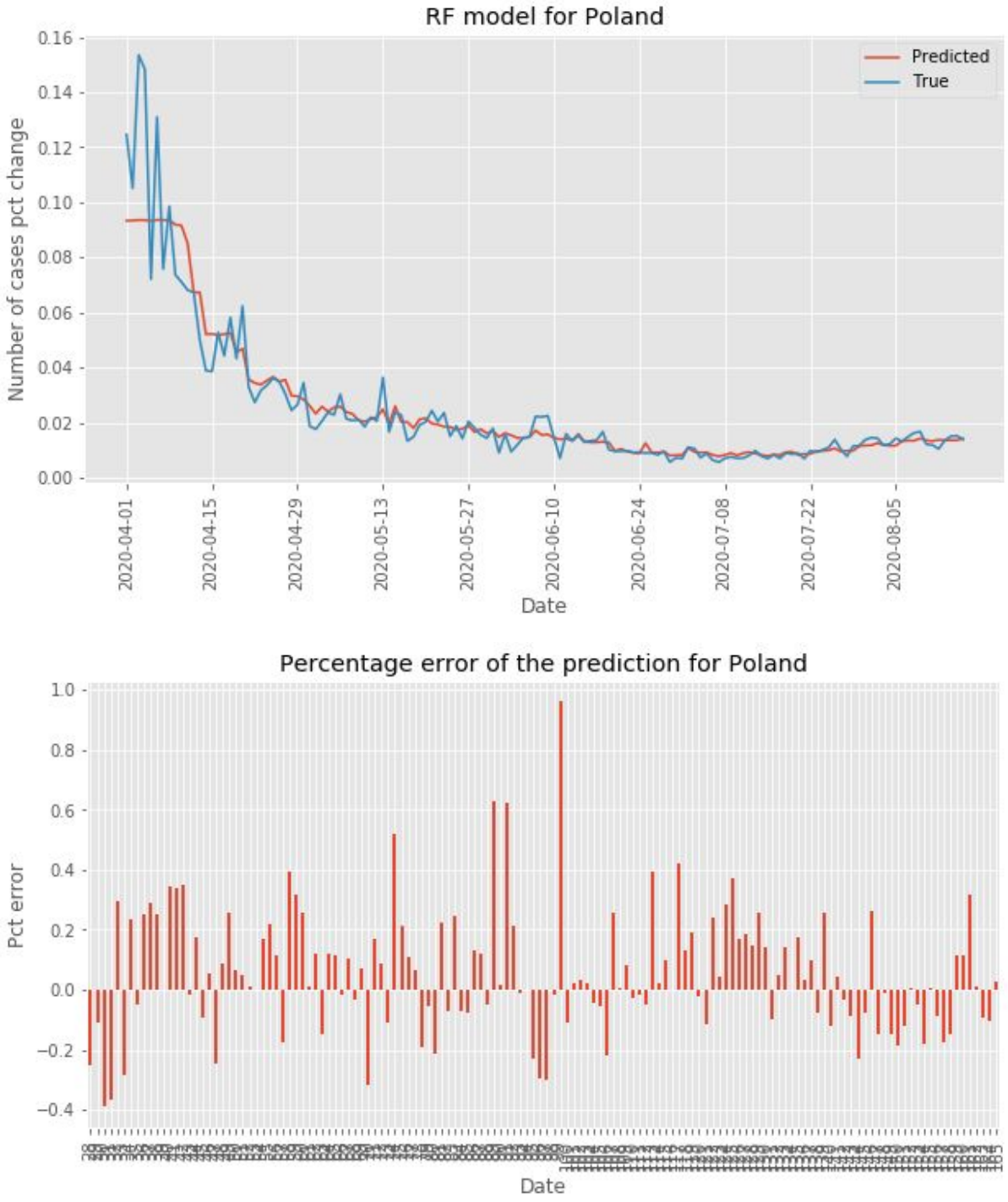


Figure 17. Predictive model for Poland and its percentage error

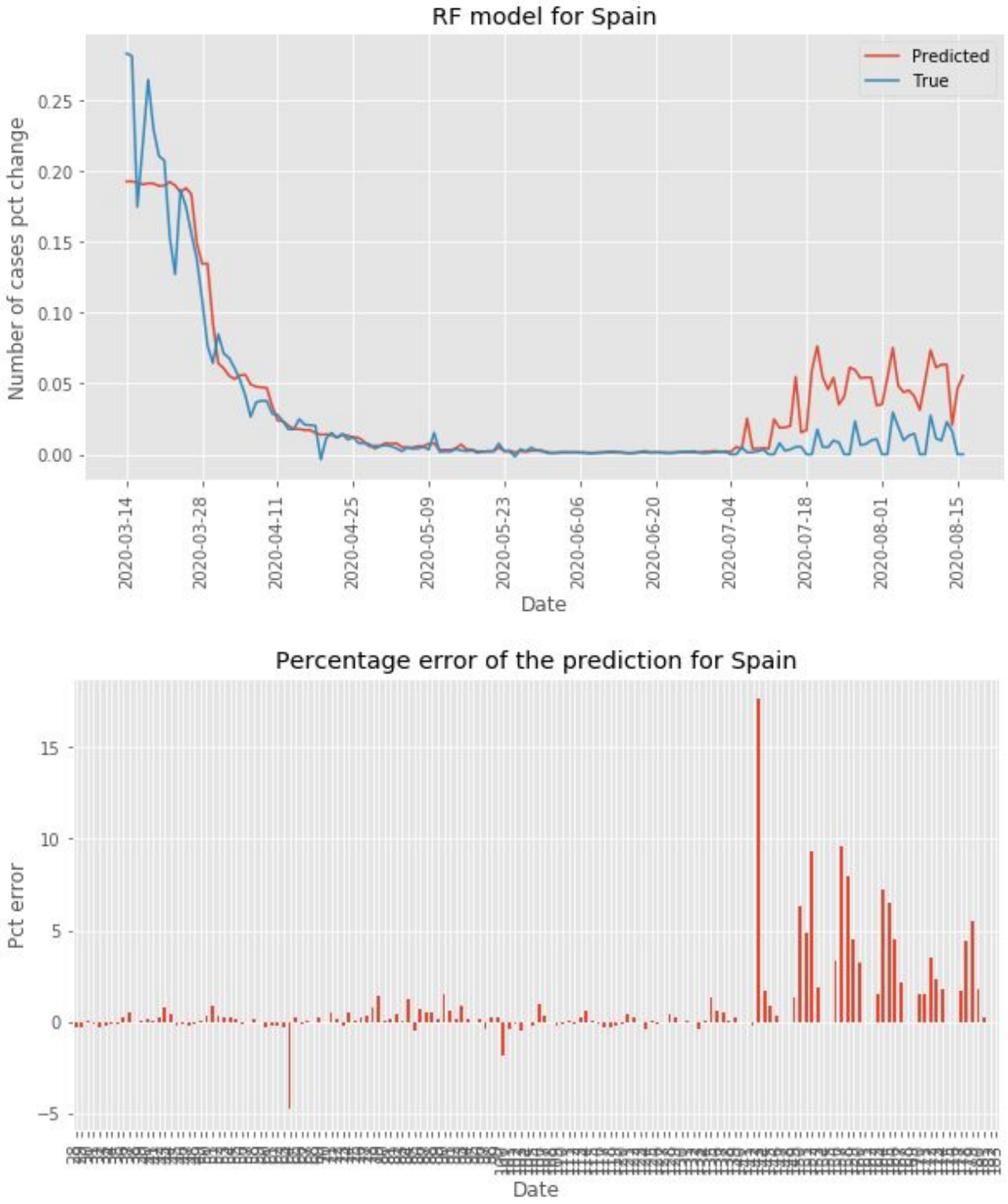
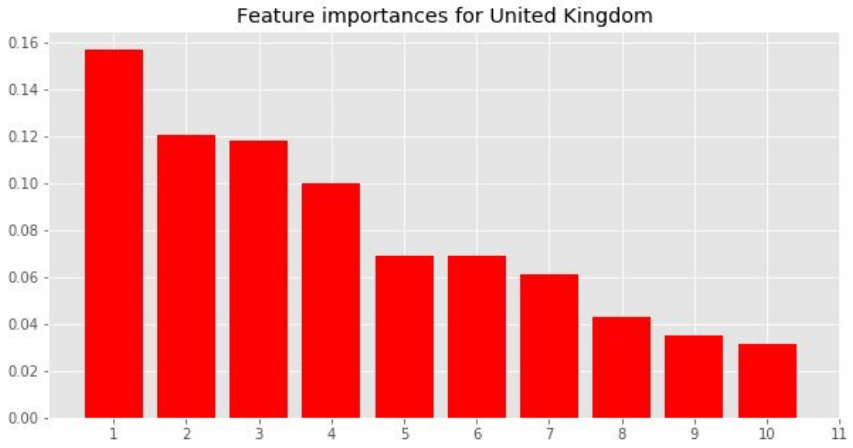


Figure 18. Predictive model for Spain and its percentage error

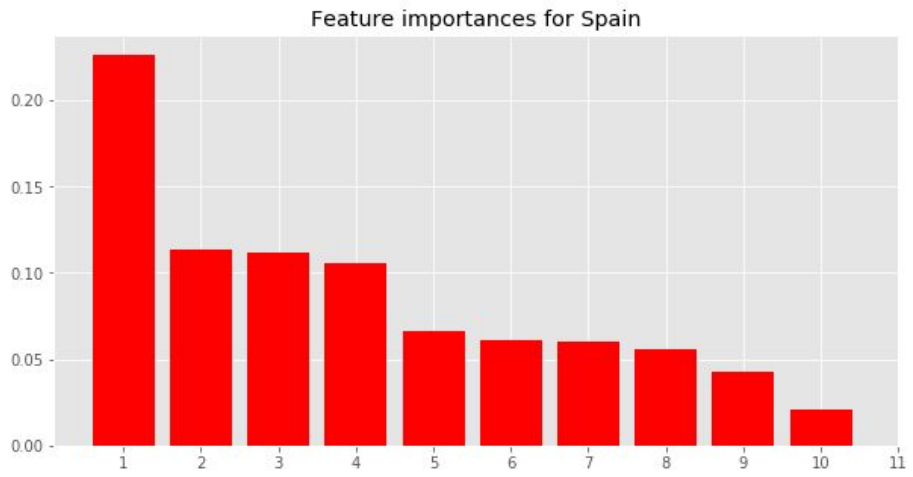


1. Workplaces percent change from baseline t-21 (0.157)
2. C2: workplace closing t-21 (0.121)
3. Residential percent change from baseline t-21 (0.118)
4. C4: restrictions on gatherings t-14 (0.100)
5. Transit stations percent change from baseline t-21 (0.069)
6. C7: domestic travel t-14 (0.069)
7. Grocery and pharmacy percent change from baseline t-14 (0.061)
8. Retail and recreation percent change from baseline t-21 (0.042)
9. Retail and recreation percent change from baseline t-14 (0.034)
10. Residential percent change from baseline t-21 (0.035)

Table 6. Most significant features in the United Kingdom Random Forest Model

1. C4: restrictions on gatherings t-14 (0.210)
2. C4: restrictions on gatherings t-21 (0.190)
3. C8: international travel t-28 (0.118)
4. H2: testing policy t-21 (0.117)
5. H2: testing policy t-28 (0.098)
6. C4: restrictions on gatherings t-28 (0.057)
7. C2: workplace closing t-28 (0.056)
8. Transit stations percent change from baseline t-14 (0.021)
9. Parks percent change from baseline t-14 (0.015)
10. C5: close public transport rolling t-21 (0.015)

Table 7. Most significant features in Poland Random Forest Model



1. C1: school closing t-21 (0.226)
2. Retail and recreation percent change from baseline t-14 (0.114)
3. Grocery and pharmacy percent change from baseline t-14 (0.112)
4. Transit stations percent change from baseline t-14 (0.106)
5. C4: restrictions on gatherings t-21 (0.067)
6. C3: cancel public events t-21 (0.061)
7. Transit stations percent change from baseline t-21 (0.059)
8. C8: international travel t-21 (0.056)
9. Retail and recreation percent change from baseline t-21 (0.042)
10. Transit stations percent change from t-28 (0.021)

Table 8. Most significant features in Spain Random Forest Model

4.7.6 Application to multiple countries

So far, we have learnt that random forests can predict the trend of the growth of COVID-19 cases reasonably well despite the non-stationarity of the time-series. Next, we want to answer the following question: *"Can we build a universal Random Forest Model which would work for multiple countries?"* As suggested in the previous section, we choose the model trained on Spain's data to test this possibility.

The model trained explicitly on Spain's dataset was used to predict the evolution of the change in the number of COVID-19 cases throughout the duration of the pandemic starting in March in the United Kingdom, Poland and France. The results are shown in Figure 19. The model shows remarkable predictive power when applied to the data from the United Kingdom. While it's incapable to predict sudden spikes, it follows the overall trend and maintains level predictions close to the true value throughout the available window of time. Meanwhile, the model shows extreme success in the beginning of the pandemic in the case of Poland and France, but after about three months, the predicted and true values diverge significantly. While this is still a very

productive result, there is some room for improvement. The most convincing reason for the divergence of the predicted values from the true values is the difference in the most significant features between the corresponding models.

4.7.7 Limitations and conclusions

While random forests proved to be a good model for predicting the increase in the COVID-19 cases on a timescale of a month based only on the past mobility and policy data, there is a lot of room for improving the model which would enable scaling up the model for development of the policy recommendation tool:

- one of the limitations of the data was the high level of non-stationarity of the time series used. While using the first time derivative (expressed as percentage change) of the time series accounted for the differences in the countries' population and proved to work in practice, it is clear that the models would be improved if more sophisticated techniques were used. For example, calculating the number of cases per capita would allow us to calculate the differences as opposed to the percentage change of the number of data which would result in more stationary series. If stationarity persists, appropriate non-linear trends could be fitted and corresponding residuals used as the output time series.
- A more rigorous training and testing procedure should be employed when longer periods of data become available. For example, training the model on a sliding window would help the model to adjust to changing trends but was not possible in this case as only 6 months of data were available.

Despite these limitations, random forest analysis revealed a broad range of possibilities to model the change of the number of cases for various countries. For some countries such as the United Kingdom, the mobility data accounted for the most important predictors, while for countries such as Poland, the policy history accounted for the success of high-level prediction. Some countries with more balanced models such as Spain emerged offering more universal models which work well when applied to other countries.

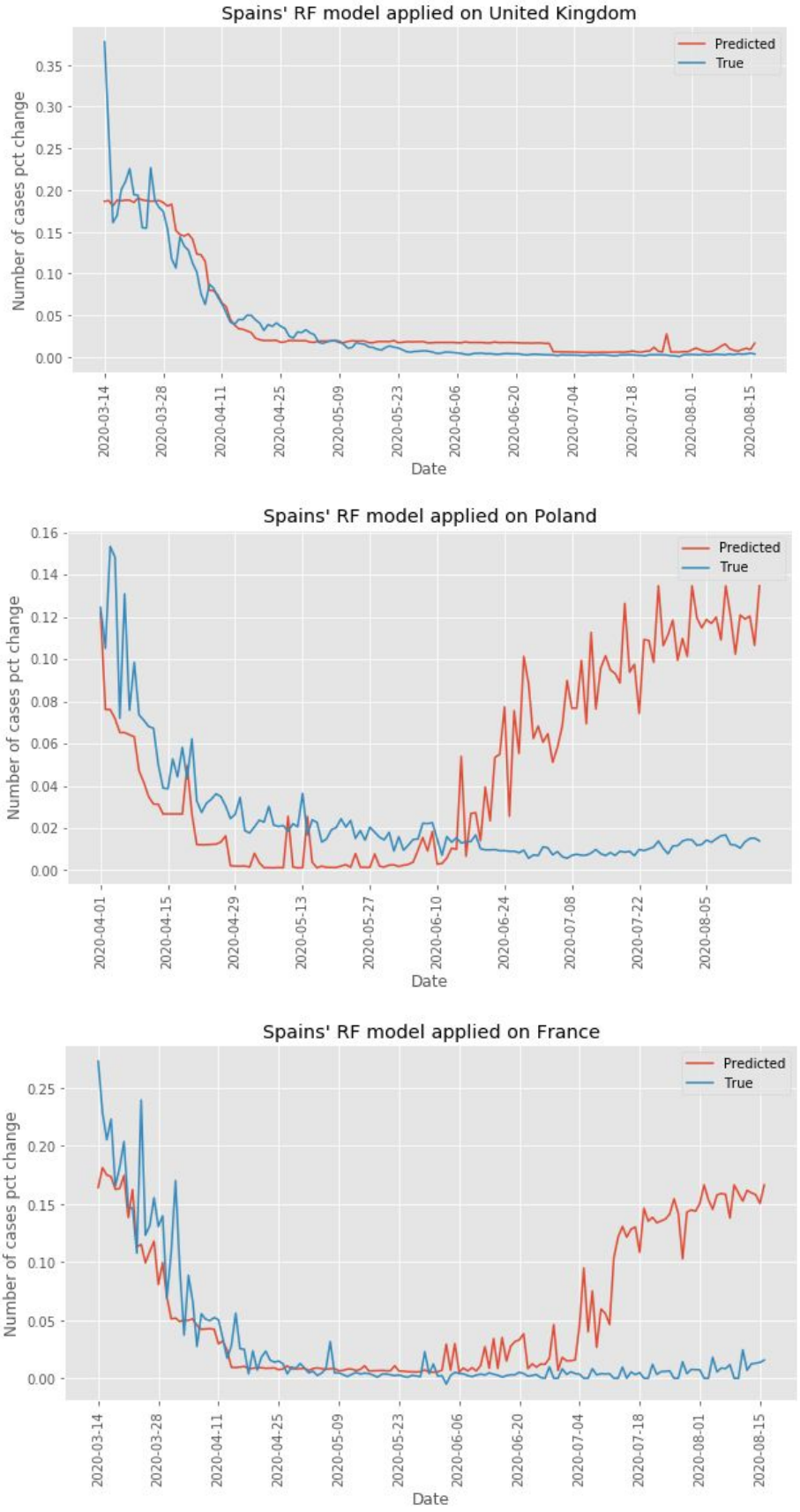


Figure 19. Random forest model trained on Spain's data applied used to predict the evolution of the change in the number of COVID-19 cases throughout the whole duration of the pandemic in the United Kingdom, Poland and France.

5 Conclusions

Containment measures implemented by the governments of various countries proved to be crucial for the control over the coronavirus spread. While the time series analysis of measures, COVID-19 infection rates, and population mobility yielded many interesting insights for individual countries, we are still able to make some general conclusions, which we list below.

- Major **socio-economic factors**, such as country population, population density, income or climate, play a secondary role when it comes to the spread of the virus.
- The main influential factor is the **policies and restrictions** imposed by the government of every country.
- The policies affect **population mobility**, which in turn affects the rates of infection spread.
- **Country-specific models** cannot always be directly applied to other countries without re-training; in some countries certain restriction measures are more effective than the others, and vice versa. For example, certain policies (usually public gathering restrictions, school closing) are highly important features for the model, while for other countries, mobility data is more important.
- However, the main principle persists: **restriction measures mitigate the virus spread the most**.
- The measures that positively influence the COVID-19 rates the most are **travel bans**, **school closure**, and **cancellation of public events**.
- It is more difficult to predict the number of cases for countries with larger stationarity of cases, as they lack seasonal and trend components. The majority of models presented in this report are based on **seasonality**.

6 References

[1] Kane, M.J., Price, N., Scotch, M. et al. *Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks*. BMC Bioinformatics. 2014 Dec 1;15(1):276.

[2] Granger CW. *Investigating causal relations by econometric models and cross-spectral methods*. Econometrica: Journal of the Econometric Society. 1969 Aug 1:424-38.

[3] Johansen S. *Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models*. Econometrica: journal of the Econometric Society. 1991 Nov 1:1551-80.

[4] Cheung YW, Lai KS. *Lag order and critical values of the augmented Dickey–Fuller test*. Journal of Business & Economic Statistics. 1995 Jul 1;13(3):277-80.

[5] Vagropoulos SI, Chouliaras GI, et al. *Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting*. IENERGYCON. 2016 Apr 4:1-6.

[6] Mukherjee TK, Naka A. *Dynamic relations between macroeconomic variables and the Japanese stock market: an application of a vector error correction model.* Journal of financial Research. 1995 Jun;18(2):223-37.

[7] Liaw A, Wiener M. *Classification and regression by randomForest*. R news. 2002 Dec 3;2(3):18-22.

[8] *List of countries by average yearly temperature*. Wikipedia, 2020. https://en.wikipedia.org/wiki/List_of_countries_by_average_yearly_temperature

[9] *Coronavirus (COVID-19) Testing*. Our World In Data , 2020. https://ourworldindata.org/grapher/tests-of-covid-19-per-thousand-people-vs-gdp-per-capita

[10] *COVID-19 worse in colder weather*. King's College London 2020. https://www.kcl.ac.uk/news/covid-19-worse-in-colder-weather

[11] *Invisible deaths: from nursing homes to prisons, the coronavirus toll is out of sight – and out of mind?* The Guardian 2020. https://www.theguardian.com/us-news/2020/may/16/coronavirus-pandemic-deaths-grieving-invisible