Ms in Data Science

EXPLORING AND ANALYZING DATA  
Final Project

**Instructor**

Dr. Milioris

A logo with text on it

Description automatically generated

**Submitted to**

Deree – The American College of Greece

**Submitted by**Kapsalis C. , Alimpertis D., Zouras I., Μargaritis K.

**Date:**

12/12/2023  
  
**Word Count:**

3,500

Table of Contents

[Introduction 5](#_Toc153214065)

[Dataset Features and Processing 6](#_Toc153214066)

[Methods and Techniques 9](#_Toc153214067)

[Results and Discussion 11](#_Toc153214068)

[Year 1: Mar 2020 – Mar 2021 11](#_Toc153214069)

[Year 2: Mar 2021 – Mar 2022 18](#_Toc153214070)

[Year 1 and Year 2 – Forecast Comparison 20](#_Toc153214071)

[Comparative Analysis 20](#_Toc153214072)

[Conclusions and Future Work 23](#_Toc153214073)

[REFERENCES 24](#_Toc153214074)

[APPENDICES 25](#_Toc153214075)

[Appendix 1. 25](#_Toc153214076)

[Appendix 2. 25](#_Toc153214077)

[Appendix 3. 26](#_Toc153214078)

**TABLES**  
[**Table 1: First 5 Observations of the Original Dataset** 6](#_Toc153176633)

[**Table 2: 12 Countries with Population Sizes,** **Table 3: NaN Values per Country Closest to the Median** 7](#_Toc153176634)

[**Table 4: Forecast Quality Metrics for Each Method** 12](#_Toc153176635)

[**Table 5: Regression Estimates** 13](#_Toc153176636)

[**Table 6: Accuracy Measures for Holt's and Holt-Winter's Method** 18](#_Toc153176637)

[**Table 7: Forecast Quality Metrics for 1st Year's Estimates** 20](#_Toc153176638)

[**Table 8: Forecast Quality Metrics for 2nd Year's Estimates** 20](#_Toc153176639)

[**Table 9: Monthly Averages for 11 Selected Countries, Mar 2020 - Mar 2021** 21](#_Toc153176640)

[**Table 10: MAD and MAPE of the Greek Forecasts, 12-point data, Mar 2020 - Mar 2021** 22](#_Toc153176641)

**FIGURES**

[**Figure 1: Forecasts vs Actual Cases** 11](#_Toc153176611)

[**Figure 2: Forecasts vs Actual Cases (Logarithmic Scale)** 12](#_Toc153176612)

[**Figure 3: Positive Cases in Greece, March 2020** 13](#_Toc153176613)

[**Figure 4: Fitted vs Actual Data for March 2020** 14](#_Toc153176614)

[**Figure 5: Actual vs Fitted Values, 2020** 14](#_Toc153176615)

[**Figure 6: Actual Cases, 2020** 15](#_Toc153176616)

[**Figure 7: Actual vs Decomposed Time-Series Values** 16](#_Toc153176617)

[**Figure 8: Holt and Holt-Winter Methods' Results** 16](#_Toc153176618)

[**Figure 9: Holt vs Holt-Winters, 2020** 17](#_Toc153176619)

[**Figure 10: Holt vs Holt-Winters in Log Scale, 2020** 17](#_Toc153176620)

[**Figure 11: Actual vs Estimated Cases (Log scale), Mar 2021 - Mar 2022** 18](file:////Users/chkapsalis/Desktop/Fall_23_projects/itc_6002_proj/report.docx#_Toc153176621)

[**Figure 12: Actual vs Estimated Cases, Mar 2021 - Mar 2022** 18](file:////Users/chkapsalis/Desktop/Fall_23_projects/itc_6002_proj/report.docx#_Toc153176622)

[**Figure 13: Actual Cases in March 2021** 19](file:////Users/chkapsalis/Desktop/Fall_23_projects/itc_6002_proj/report.docx#_Toc153176623)

[**Figure 14: Actual vs Fitted Cases, Mar 2021 - Mar 2022** 19](file:////Users/chkapsalis/Desktop/Fall_23_projects/itc_6002_proj/report.docx#_Toc153176624)

[**Figure 15: De-Trended Values, Mar 2021 - Mar 2022** 19](#_Toc153176625)

[**Figure 16: Actual vs Forecasted Values, Mar 2021 - Mar 2022** 19](file:////Users/chkapsalis/Desktop/Fall_23_projects/itc_6002_proj/report.docx#_Toc153176626)

[**Figure 17: Actual vs Forecasted Values, March 2021** 19](file:////Users/chkapsalis/Desktop/Fall_23_projects/itc_6002_proj/report.docx#_Toc153176627)

[**Figure 18: Actual vs Estimated Cases in Greece, 12-point dataset, Mar 2020 - Mar 2021** 21](#_Toc153176628)

[**Figure 19: Boxplot of Positive Cases Count in Greece** 25](#_Toc153176629)

[**Figure 20: Basic Residual Diagnostics** 26](#_Toc153176630)

[**Figure 21: Cases Boxplot, Greece, Mar 2021** 26](#_Toc153176631)

[**Figure 22: Cases Histogram, Greece, Mar 21** 27](#_Toc153176632)

# **Introduction**

Over the past three years, the COVID-19 pandemic drastically changed every aspect of the everyday lives of people all over the globe; from the way they worked and how they pursued their educational studies, up to how they interacted with others. One of the most interesting impacts of this shock to people’s behavior was that, for the first time, they started to highly esteem forecasts made by experts, as they directly influenced policy making and, subsequently, their lives.

Epidemiologists had a respected say on the decisions of national and supra-national policy makers. These decisions, in turn, often dictated how people would lead their lives, incorporating even lockdowns and horizontal bans on social events, aimed at containing the transmission of the virus.

The data on COVID-19 cases and related deaths that experts worked with were closely tracked by national public health organizations worldwide. As a result, we make use of these data that once played a significant role in our daily lives in order to implement some simple methods of time-series analysis and forecasting. We observe how these methods allow us to uncover the various developments in the pandemic’s progress, as well as make forecasts and compare their efficacy.

We focus on a relevant dataset shared by the European Centre for Disease Prevention and Control[[1]](#footnote-1). As Greek nationals, we found great interest in studying pandemic-related data on the EU, with an emphasis on Greece and its count of daily positive cases as our forecast variable. We chose cases over deaths because there has been a lot of controversy regarding the criteria that national agencies followed in order to track the number of deaths, at least during the first phases of the phenomenon, so the trustworthiness of the corresponding data could be questioned.

# **Dataset Features and Processing**

The dataset contains daily observations for the number of positive COVID-19 cases and deaths from 30 countries in the EU/EEA region. The time interval it spans lies from Jan 1, 2020 to Oct 26, 2022. The total number of observations is 28,715.

The variables tracked are daily counts of positive cases and deaths, the country that each observation refers to and the size of its population. There is either one or no record for each country on each day of the available data.

###### **Table 1: First 5 Observations of the Original Dataset**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **dateRep** | **day** | **month** | **year** | **cases** | **deaths** | **countriesAndTerritories** | **popData2020** |
| 23/10/2022 | 23 | 10 | 2022 | 3557 | 0 | Austria | 8901064 |
| 22/10/2022 | 22 | 10 | 2022 | 5494 | 4 | Austria | 8901064 |
| 21/10/2022 | 21 | 10 | 2022 | 7776 | 4 | Austria | 8901064 |
| 20/10/2022 | 20 | 10 | 2022 | 8221 | 6 | Austria | 8901064 |
| 19/10/2022 | 19 | 10 | 2022 | 10007 | 8 | Austria | 8901064 |

We begin by cleaning our dataset, and specifically by addressing a very significant sign of measurement error, which is negative counts of cases. Cases can never be negative, so we proceed to drop those records (NaN values appear in their position and we properly treat them later in our analysis,). That is because they will otherwise create significant problems in our forecasts. All in all, there are 14 negative values in our dataset.

An important inconsistency that we discovered in our data is that, even though the available observations for each country are of daily frequency (with the exception of some missing values), they do not necessarily lie in the same time interval. This creates imbalances in the number of missing or NaN values between the various countries, something that was addressed in the A/F ratio analysis we later performed on data from 11 countries.

To account for the previous, as well as for NaN and missing values in each country, we started by building a dataframe that contains daily records for every single country, containing NaN values where there is no observation for a specific (date, country) pair. Thus, it has one row for each possible pair. The resulting dataframe (called 'res') will be used to extract the data we'll use for our analysis.

In order to select a group of countries to incorporate A/F ratios in our analysis, we considered the size of each country’s population and the number of NaN values related to each. We deem as important to study countries relatively similar population sizes, because it could affect how the pandemic progresses. Significantly different populations could lead to our comparative metrics (A/F ratios in particular) relying on different fundamental conditions for each country, so they would not be trustworthy to revise our estimates based on them.

We observed that many countries in our dataframe have population sizes close to the (perceived) median. Thus, we initially found the 12 countries whose population sizes fall in the middle 50% of the population sizes' distribution. Greece was included, so we just needed to drop one country. By observing the count of NaN values for each country, we ultimately excluded Denmark from our dataset, because it had a much bigger count of NaN values than the rest and therefore would create inconsistencies in our data.

**Table 2: 12 Countries with Population Sizes** **Table 3: NaN Values per Country Closest to the Median**

|  |  |
| --- | --- |
| **countriesAndTerritories** | **popData2020** |
| Lithuania | 2794090 |
| Croatia | 4058165 |
| Ireland | 4964440 |
| Norway | 5367580 |
| Slovakia | 5457873 |
| Finland | 5525292 |
| Denmark | 5822763 |
| Bulgaria | 6951482 |
| Hungary | 9769526 |
| Portugal | 10295909 |
| Czechia | 10693939 |
| Greece | 10718565 |

|  |  |
| --- | --- |
| **countriesAndTerritories** | **NaN Values** |
| Lithuania | 34 |
| Croatia | 58 |
| Ireland | 72 |
| Norway | 51 |
| Slovakia | 76 |
| Finland | 6 |
| Denmark | 118 |
| Bulgaria | 68 |
| Hungary | 62 |
| Portugal | 96 |
| Czechia | 27 |
| Greece | 60 |

We also had to choose a specific 2-year period to study; we selected the interval from Mar 1, 2020 to Mar 1, 2022. We based that decision on the limited availability and willingness of people to perform diagnostic tests prior to that, which lead to very low case counts that did not reflect the pandemic’s reality.

Next, we detected all missing and NaN values from our dataframe containing all countries, and not just from data on Greece, so that it could later be immediately used for our comparative analysis. We filled in missing or NaN values with the next (in ascending chronological order) not-NaN value for each country.

Concentrating on the case count in Greece (for which we have 731 observations), we performed outliers classification using the Z-score method. We chose this method due to its popularity; however, we are negatively inclined toward it, because, at the beginning phase of the pandemic, people did not know how to react, so they tended to follow instructions given by experts, who, under those circumstances, greatly impacted their behavior. Thus, we expect that people’s behavior during that period was not random, and -as a consequence- the cases’ distribution departed from normality.

The Z-Score method works by applying a value to each data point, equal to its distance from the mean in terms of the standard deviation: , where X: the value of the random variable, μ, σ: the mean and standard deviation of its distribution. Using a threshold value, it we classify as outliers those data points that have Z-scores greater than that.

Here we set a threshold of 3, because, following the method’s normality assumption[[2]](#footnote-2) and the empirical rule, 99.7% of all values of a normally distributed random variable fall in ±3\*σ around its mean. This way, we truly only deem extreme values as outliers. We found 7 outliers, or 23% of the total observations.

We cannot consider them as cases of measurement error, so those outliers showcase the reality that held in that period, and we cannot take the freedom to drop them.

# **Methods and Techniques**

We decided to implement some simple time-series analysis methods on our data. These methods, following the order in which they are mentioned, were chosen based on the increasing level at which they fully consider all the characteristics of the time-series at hand. In applying them, we considered the data from March 2020 as the only ‘available’ data.

**Naïve Approach:** it accounts for no underlying factor of the time-series under study. Based on it, we extrapolate on March’s data by casting the same count of positive cases on the respective day of each month; for example, for February 10 we predict that there will be the same number of cases as in March 10.

**Simple Moving Average**: it produces estimates based on the assumption that the value of the forecast variable will be equal to the mean of its values over a specific number of past observations. The size of the ‘sliding window’ we decided to be 7 days. That is because experts at the beginning and maturity phases of the pandemic shared their latest updates on a weekly basis, so we suppose that it offers a good foothold for better tracking the development of the phenomenon under study. Thus, we expect it to give a better image on the progression of our forecast variable. We estimate it on the available data from March 2020, and then we extrapolate on future data points using the forecasts made on the previous 7 data points.

**Weighted Moving Average:** it works similarly to the previous, with the only difference that it does not weigh each observation equally; rather, it applies a higher weight the more recent an observation is. Using a 7-day sliding window, we apply a weight of 7 for the last day, 6 for the day prior to that, etc. As a result, it better represents the latest changes in the pandemic’s development, so we expect it to give an even better estimation than the simple moving average.

**Holt’s Method:** it works by fitting a line on our available data against a time index (which represents values corresponding to each day in our sample, i.e. 1 for Mar 1 and 10 for Mar 10). By doing so, we examine the existence of a statistically significant relationship between our forecast (‘dependent’) variable and time, which would signify the existence of a trend in the time-series. Accounting for that, we extrapolate by using the estimates we get from the model: , where a and b: our estimates for the intercept and slope coefficients of the simple linear regression model, and t: the time index acting as the regressor. The validity of this method is decided upon by studying the statistical significance of the slope coefficient by performing a corresponding t-test; rejecting the null hypothesis (i.e. having enough evidence showing that the real value of the slope coefficient is not zero) hints at the existence of a trend component in our data.

**Holt-Winter’s Method:** it expands on the previous method by decomposing the given time-series not based only on its trend component, but also its seasonality component; to approach the latter, we inspect the development of cases in the first year and see if there is any repetitive pattern related to time intervals in our data. After detecting such seasonal cycles, we compute and normalize the seasonality factors that our estimates need to be multiplied with to account for the seasonal effect in each period. We used the ‘mixed’ model of decomposition, which breaks down the observed values as follows:

, where . ‘Level’ signifies our de-trended data, which in this method are estimated by . After decomposing, we extrapolate using the formula: , where T: the time index corresponding to the last observation of our ‘available’ data. Since this method tries to completely break down the time-series into all its structural foundations, we expect–given that seasonal effects exist in our example– it to culminate in more accurate and less biased forecasts.

**A/F Ratios Method:** it fine-tunes the forecasts one makes based on a method using observations and estimates on other variables that are similar to the initial forecasting variable and have been studied using similar forecasting techniques. For each data point, the A/F ratio of a forecast is equal to . When applying this method, we calculate the mean A/F ratio for all complementary variables for every single data point considered, and we multiply that to our initial forecast. As complementary variables we chose the (cleaned up) daily cases (and corresponding forecasts) for 10 other EU/EEA countries, based on the reasoning we described in the previous chapter. Assuming that these complementary variables are actually similar to our forecast variable and the forecasts have been properly estimated, we expect that this method would positively refine our results, because it considers a fuller image of the variance of phenomena similar to our main forecast variable.

Regarding the measures we used to compare our forecasts.

**Mean Absolute Deviation (MAD):** it is the average of the absolute values of all the errors produced by a forecasting method. It is a measure of the accuracy of forecasts, meaning how much they differ (on average) from the respective actual values.

**Mean Absolute Percentage Error (MAPE):** it is the average of the absolute errors of our forecasts, expressed as their percentage distance from each actual value. It is also an accuracy measure.

**Tracking Signal:** the ratio of the sum of errors divided by MAD. It is a measure of the bias of a method, showing how much (on average) the estimates it produces are higher or lower than the actual values.

# **Results and Discussion**

## **Year 1: Mar 2020 – Mar 2021**

We begin showcasing our results by giving a graphical presentation of the estimates from each one of the 3 simplest methods discussed in the previous chapter:

A graph of different types of data

Description automatically generated

###### **Figure 1: Forecasts vs Actual Cases**

Due to their significant departure from the actual data, which has to do with the fact that the fundamentals changed significantly after March 2020, so that the transmission of the virus surged in a way not at all conveyed in the first month’s data, we also present the same plot in logarithmic scale:

A graph of different types of data

Description automatically generated

###### **Figure 2: Forecasts vs Actual Cases (Logarithmic Scale)**

We compare these techniques in 2 ways; the first one studies their relative accuracy based on the MAD measure, and the second one compares the bias of their estimates using the tracking signal.

###### **Table 4: Forecast Quality Metrics for Each Method**

|  |  |  |
| --- | --- | --- |
| **Forecasting Method** | **MAD** | **Tracking Signal** |
| Naive | 531.74 | 345.13 |
| Simple Moving Average | 518.39 | 335.63 |
| Weighted Moving Average | 519.99 | 337.75 |

We see that the naïve method provides the least accurate estimates, since it just reiterates March’s numbers; this would be accurate only if there were so strong seasonal cycles so that March’s outcomes would approximately repeat on each month. The S.M.A. and W.M.A. provided better estimates; in contrast to our expectations, the latter was slightly less accurate. This has to do with the fact that the W.M.A. is affected more by the most recent observations, and the forecast variable here showed a lot of unexpected behavior. This falsely affected the W.M.A. more than it should, through sudden movements that did not describe its development over time.

As for the forecasts’ bias, all of them performed poorly, with the most successful method (the S.M.A.) giving forecast values that where on average 335.63 times smaller than the actual values.

We then applied the Holt’s method to our data, because the clearly growing SMA values in the first month’s data made us assume that a (positive) trend might exist.

###### **A graph with a line Description automatically generated Figure 3: SMA of Positive Cases in Greece, March 2020**

To check this assumption, we created a ‘time index’ variable and fitted a simple linear regression model on our observation data against the time index[[3]](#footnote-3).

###### **Table 5: Regression Estimates**

|  |  |  |
| --- | --- | --- |
|  | **Estimated Value** | **p-value for t-test** |
| Intercept (a) | -6.548387 | 4.05E-01 |
| Slope Coefficient (b) | 3.048387 | 6.13E-08 |

A graph with purple dots and a blue line

Description automatically generated

###### **Figure 4: Fitted vs Actual Data for March 2020**

The slope coefficient is statistically significant, so our assumption seems to be valid; we accept the hypothesis that there is a trend in our data, and use the estimated linear model to extrapolate[[4]](#footnote-4):

A graph with green and purple dots

Description automatically generated

###### **Figure 5: Actual vs Fitted Values, 2020**

We then extended our analysis to check for seasonal effects, starting from plotting the observed values over time:

A graph with blue lines

Description automatically generated

###### **Figure 6: Actual Cases, 2020**

It is reasonable to expect the following behavior from a time-series of this nature; while winter nears (presumably from late October onwards), we would predict a surge in positive cases as people start to gather in closed spaces. We observe that in our data for the fourth quarter. As spring approaches, and the temperatures start rising, allowing for people to start mass gathering (even in open spaces), we would expect the cases to explode again. We see that in the first quarter's data (from late February to late March). Then, during the summer's hot season, we would expect to see an increase again as most people go on holidays - we observe an increase and subsequent decrease during the third quarter. In the second quarter there was no such clear movement due to the lockdown that had been imposed.

Taking the previous into consideration, we achieve the following decomposition of the time-series:

A graph with blue and orange lines

Description automatically generated

###### **Figure 7: Actual vs Decomposed Time-Series Values**

Ultimately, we use March’s 2020 data and get the following results for that month:

A graph of a graph showing the difference between a trend and actual

Description automatically generated

###### **Figure 8: Holt and Holt-Winter Methods' Results**

We then used the extrapolation formula discussed in Chapter 2 and got the following results:

A graph with numbers and lines

Description automatically generated

###### **Figure 9: Holt vs Holt-Winters, 2020**

A graph of different colored lines

Description automatically generated

###### **Figure 10: Holt vs Holt-Winters in Log Scale, 2020**

Regarding the accuracy of the 2 displayed methods:

###### **Table 6: Accuracy Measures for Holt's and Holt-Winter's Method**

|  |  |
| --- | --- |
| **Forecasting Method** | **MAD** |
| Holt's Method | -3 |
| Holt-Winter's Method | -17555 |

The latter one gave us too inaccurate forecasts. This is due to the fact that lockdowns broke the seasonality cycles that could arise otherwise, and our limited mathematical toolset doesn't allow for a better approximation. All in all, holt’s method is by far the best out of those we have applied in this study.

## **Year 2: Mar 2021 – Mar 2022**

We performed the exact same analysis on the second year’s data, based on the same (cleaned up from inconsistencies, NaN values and measurement errors) dataframe we used before. The same outlier classification method[[5]](#footnote-5) on the 2021 data returned 0 outlier values for cases in Greece, which could be attributable to the fact that no lockdown was in place then, so the pandemic’s development had characteristics closer to randomness.

Regarding the implementation of the three simple forecasting methods, we showcase the actual and fitted values for the second year in absolute and relative values:

A graph of different types of data

Description automatically generatedA graph of different types of data

Description automatically generated

###### **Figure 11: Actual vs Estimated Cases (Log scale), Mar 2021 - Mar 2022**

###### **Figure 12: Actual vs Estimated Cases, Mar 2021 - Mar 2022**

We also implemented the Holt’s method because there were hints of a trend existing due to the uptrend showed by the S.M.A. in the first month’s data, validated by the regression model we estimated. The actual data for March, together with the actual and fitted for the whole second year are shown below:

A graph with blue line

Description automatically generatedA graph with green dots and a purple line

Description automatically generated

###### **Figure 13: Actual Cases in March 2021**

###### **Figure 14: Actual vs Fitted Cases, Mar 2021 - Mar 2022**

By observing the de-trended data, we can see once again 4 potential seasonal cycles, matching the quarters the year, as in the first year’s data:

A graph of blue lines with months on it

Description automatically generated

###### **Figure 15: De-Trended Values, Mar 2021 - Mar 2022**

Below we showcase the actual and estimated cases in March 2021 and in the whole second year (due to the significant discrepancy between the two, we display the second plot in a logarithmic scale):

A graph of a graph showing the difference between a method and a method

Description automatically generated with medium confidenceA graph of different colored lines

Description automatically generated

###### **Figure 16: Actual vs Forecasted Values, Mar 2021 - Mar 2022**

###### **Figure 17: Actual vs Forecasted Values, March 2021**

## **Year 1 and Year 2 – Forecast Comparison**

It is evident that the Holt’s method was by far the most accurate in the 1st year, but also the most biased. That was due to the actual data of this dynamically developing phenomenon departing from the fitted line. At the same time, Holt-Winter’s method gave extremely inaccurate estimates compared to the rest, as a result of the inaccuracy of the trend estimate it builds on.

###### **Table 7: Forecast Quality Metrics for 1st Year's Estimates**

|  |  |  |
| --- | --- | --- |
| **Forecasting Method** | **MAD** | **Tracking Signal** |
| Naive | 531.74 | 345.13 |
| Simple Moving Average | 518.39 | 335.63 |
| Weighted Moving Average | 519.99 | 337.75 |
| Holt's Method | -3 | 886.62 |
| Holt-Winter's Method | -17555 | 357.36 |

Regarding the second year’s estimates, the actual data in that time period showed similar movements to the first year, and thus the ‘available’ data were again inadequate for efficiently applying the methods under study.

###### **Table 8: Forecast Quality Metrics for 2nd Year's Estimates**

|  |  |  |
| --- | --- | --- |
| **Forecasting Method** | **MAD** | **Tracking Signal** |
| Naive | 3807.46 | 363.39 |
| Simple Moving Average | 3694.45 | 363.34 |
| Weighted Moving Average | 3668.85 | 363.32 |
| Holt's Method | -1026 | 367.46 |
| Holt-Winter's Method | -471474 | 364.12 |

We can also detect two differences in the results; first, the MAD metric is significantly larger in the second year’s estimates. That is because the daily actual cases were many more, so, even with the same level of accuracy as in 2020, we would get errors of higher magnitude. Second, the W.M.A.’s bias in the second year is lower than the S.M.A.’s; this time, the extreme movements in the actual values were less severe, so the closer following of the current movements by the wma gave a better sense of the data.

## **Comparative Analysis**

Considering the first year’s data on all 11 selected countries as known, we consider a dataframe containing 12 points for each country, the mean of cases of each month:

###### **Table 9: Monthly Averages for 11 Selected Countries, Mar 2020 - Mar 2021**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **countriesAndTerritories\_x** | **(2020,10)** | **(2020,11)** | **(2020,12)** | **(2020,3)** | **(2020,4)** |
| **Bulgaria** | 1032.61 | 3081.87 | 1837.61 | 14.06 | 37.63 |
| **Croatia** | 1055.58 | 2637.53 | 2657.90 | 27.81 | 40.30 |
| **Czechia** | 8529.35 | 6274.33 | 6774.61 | 106.97 | 146.20 |
| **Finland** | 198.65 | 318.40 | 340.68 | 57.68 | 120.83 |
| **Greece** | 670.19 | 2200.67 | 1083.19 | 42.23 | 42.57 |
| **Hungary** | 1576.13 | 4726.70 | 3399.74 | 15.87 | 76.10 |
| **Ireland** | 876.26 | 369.60 | 620.48 | 104.32 | 579.23 |
| **Lithuania** | 386.39 | 1562.93 | 2712.74 | 14.81 | 30.67 |
| **Norway** | 225.81 | 516.37 | 436.90 | 164.03 | 89.33 |
| **Portugal** | 2006.16 | 4758.67 | 3815.87 | 233.90 | 562.67 |
| **Slovakia** | 1580.90 | 1574.57 | 2494.35 | 13.06 | 33.43 |

We began by fitting a line to each country’s data, and then we computed the A/F ratios for each month-country pair. Then, concentrating on the data on Greece once again, we computed the mean A/F ratio for all other countries for each month. Ultimately, we multiplied each forecast on Greece with the corresponding mean A/F ratio value.

A graph with green and orange lines

Description automatically generated

###### **Figure 18: Actual vs Estimated Cases in Greece, 12-point dataset, Mar 2020 - Mar 2021**

Below we reference a comparison of the basic forecasting accuracy metrics for the initial and revised forecasts:

###### **Table 10: MAD and MAPE of the Greek Forecasts, 12-point data, Mar 2020 - Mar 2021**

|  |  |  |
| --- | --- | --- |
| **Accuracy Metric** | **M.A.D.** | **M.A.P.E** |
| Initial | 0.001 | 315.014 |
| Final | 74.914 | 75.455 |

MAD gives a false image for the accuracy of the methods, because the revised estimate curve approaches the actual data much better, and it almost exclusively has positive deviations. On the other hand, the simple holt method gave both positive and negative, and they cancelled out. To address this, we also computed the M.A.P.E. for each method; the metric for the revised forecast is by far smaller than that for the initial, testifying to its higher accuracy. That is because it accounts for the mean error of all other European countries of similar characteristics population-wise, so we could say that it better takes into consideration the variance characteristics of somehow similar distributions, so it gives us a better sense of the random movements of our data.

# **Conclusions and Future Work**

Our decision to apply simple forecasting methods on a very intricate and unpredictable dataset lead us to estimating unsuccessful forecasts that heavily departed from the case counts observed in the years under consideration. The limited nature of the data we worked on (considered data just one month’s data as ‘available’) and the inability of the methods implemented to prove adequate in fully extracting the underlying information of the given time-series lead to invalid results. Our estimates using these methods would be accurate only under very strict conditions for the underlying data, such as showing very strong seasonal cycles or just following a steady trend over the years.

All in all, in future forecasting projects, we will need to apply more sophisticated methods able to cope with the inconsistencies and extreme behavior of our data.

# **REFERENCES**

Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and statistical modeling with Python. In 9th Python in Science Conference.* [*http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf*](http://conference.scipy.org/proceedings/scipy2010/pdfs/seabold.pdf)

# **APPENDICES**

## **Appendix 1.**

**A diagram of a positive cases

Description automatically generated**

###### **Figure 19: Boxplot of Positive Cases Count in Greece**

However, by observing a simple boxplot, it is clear that the distribution of values is not normal, since it is not even symmetric; the median is lower than the mean, due to the fact that most values are crowded in the lower end of the distribution, except for some extremely high ones that positively impact the mean.

## **Appendix 2.**

We performed some basic residual diagnostics to check for the validity of our forecasts for positive COVID-19 cases in Greece in the first year under consideration. In more detail, the relationship between the time index and # cases seems to be approximately linear in the initial period under study. The residuals don't seem to depart a lot from a straight line in the Q-Q plot, so it is not that we depart a lot from the normality assumption. The mean of residuals is (approximately) zero and, ignoring a few outliers, we can see than we have almost the same amount of residual values above and below the mean: sum(residuals>0)=15, sum(residuals<0)=16. The residuals seem to be independent, as there is no clear pattern in their values over time. Last, the variance (from inspecting the graph) seems to differ for the different regressor values, so the hypothesis of heteroscedasticity doesn't seem to hold. Thus, to get trustworthy forecasts, we would need to use heteroscedasticity-robust methods of estimating our model (e.g. GLS).

A group of graphs with numbers and lines

Description automatically generated

###### **Figure 20: Basic Residual Diagnostics**

## **Appendix 3.**

The distribution of cases in the ‘available’ data this time does is almost symmetric and does not depart much from randomness:

A diagram of a distribution of positive cases

Description automatically generated

###### **Figure 21: Cases Boxplot, Greece, Mar 2021**

A graph of cases

Description automatically generated

###### **Figure 22: Cases Histogram, Greece, Mar 21**

1. <https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-covid-19-eueea-country> [↑](#footnote-ref-1)
2. The validity of this hypothesis is tested in Appendix 1 [↑](#footnote-ref-2)
3. We made use of the ‘statsmodels’ python library for this purpose, due to the easy-to-use t-test functionality it offers for the coefficients under study. [↑](#footnote-ref-3)
4. For basic residual diagnostics, see Appendix 2. [↑](#footnote-ref-4)
5. For Z-Score method validation, see Appendix 3. [↑](#footnote-ref-5)