Analysis Of COVID-19 Daily New Deaths Of Poland and Germany

E. Webb (34683518)*, A. Pham (34866191)†, A.Liang (37507494)‡

***Abstract*—In this report we analysed Polish and German daily new deaths, recorded from March, 2020 to January, 2021, by considering SARIMA models using an iterative Box-Jenkins approach. We conclude that for Polish data the best fitting model is $\mathbf{SARIMA}(3,1,4,2,1,1,\mathbf{s}=7)$, whilst for Germany $\mathbf{SARIMA}(3,1,2,1,0,1,\mathbf{s}=7)$. Our forecasting results suggest good performance across the models, validating their effectiveness. We discussed other important factors such as governments' response and also the limitations of our work. Finally, we explored ideas on how we could improve our analysis.**

I. Introduction

Coronavirus Disease 2019 (COVID-19) emerged in Wuhan, China in late 2019. By January 30th 2020, WHO declared a Public Health Emergency of International Concern [1] and by March 2020 a Global Pandemic had been established [2]. As the pandemic evolved, the epicentre shifted from the Hubei Province to Central Europe and, as of January 31st 2020, Europe's 5 largest countries are all in the World's top 10 for most COVID-19 cases - meanwhile China sits 83rd [3]. With that in mind, the European data is important to understanding the nature of the pandemic and how it might progress.

Throughout the course of history, humans have experienced the deadliest plagues, infectious diseases and epidemics, which have wiped out millions of people. Every outbreak has long-lasting effects on society and causes great suffering on humankind. However, through mathematical modelling we can learn from the past to advance our understanding of epidemiology, push medical innovation and reduce casualties.

Since the beginning of the outbreak, many mathematical and computational models have been published to analyse and predict the coronavirus case and death figures. These models are fundamental to government decision makers attempting to minimise the number of deaths and strain on health services, whilst ensuring the economy can function as best as possible. At the start of the pandemic, Zeynep Ceylan [4] modelled COVID-19 case figures in Spain, Italy and the UK using ARIMA models with second order differencing to help epidemiologists assess the worst affected areas. Meanwhile other models focused on forecasting the spread and death rate of coronavirus [5].

We analyse the Polish and German data of new daily deaths from March, 2020 to January 2021. Germany and Poland share a border and have similar levels of total deaths. This makes their comparison of interest because it could reveal how other factors such as governance approach, population density and country footfall affect the pandemic. Whilst we acknowledge countries have different methods of recording death figures, we choose to study new daily deaths as they are more reliable than case figures - you can't misdiagnose a death.

In our study, we build models to predict the number of new deaths, and compare the differences in these two series. Additionally, we look at how the restrictions and lockdowns impact these series as well as discussing the strength of our forecasts and the trends that lie within. The data has been obtained from Our World In Data (OWID) [6].

II. Methods

To begin the process of our analysis, we first layout our intentions. We wish to compare the new daily deaths of Germany and Poland. As a result, we first examine the raw data, to explore trends and stationarity, before fitting models in order to forecast how the new daily deaths might progress – all so we can assess the similarities and differences between the two countries.

A. Preliminary Analysis

To build our models we rely on useful results such as Wold's Decomposition, which allows us to write one time series as the sum of two time series - one deterministic and one stochastic. In order to use this result, we require our data to be independent and have constant stochastic properties, in other words we need stationarity. Upon our initial plots (Fig. 1), it is evident that the raw data does not meet this condition since the mean is not constant, whilst the variance and covariance differ across time. This observation is reaffirmed by an Augmented Dickey-Fuller Test (ADF) on the raw data, which yields p-values of 0.9286 and 0.99, for Poland and Germany respectively.

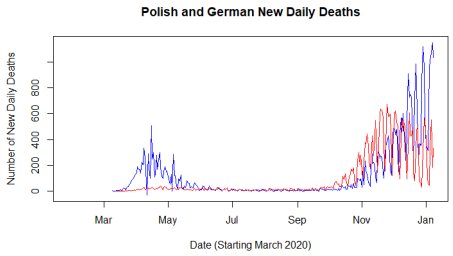

Fig. 1: Poland's (Red) and Germany's (Blue) new daily deaths.

We first difference the data, but it remains non-stationary. Likewise, further differencing doesn't yield stationarity. In transforming our data to become stationary, we take the logarithm of the data and then difference once. Since our raw data for both countries contains zero values, we clean our data before taking the log. To do this, we add a small non-negative

value, $\exp(-3)$, to each recording in the time series. From Fig. 2, we see that both data now look stationary as there are oscillating patterns around 0. The Augmented Dickey-Fuller Tests for both countries give the p-values less than 0.01, which affirm that the data now is stationary.

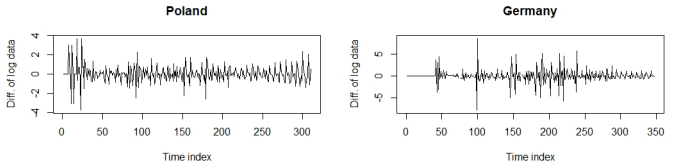

Fig. 2: Polish and German data after taking the logarithm and differencing.

With the stationary data, we then examine the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) for each time series. From this initial analysis, we gain insight into the potential Autoregressive and Moving Average components of each series as well as if there is any seasonality. Firstly, we notice that both data sets have significant peaks at lags of multiples of 7, suggesting a seasonality with period 7.

Using $\mathrm{ARIMA}(p, d, q)$ models, we see no clear cut-offs in the ACFs and PACFs. However, for Polish data (Fig. 3), we observe significant peaks on the ACF and PACF that suggest the p values of $2, 3$ or 4 and q values of $1, 2$ or 3. Whereas for Germany (Fig. 4), p values of the range 2 to 5 and q is 1 or 2.

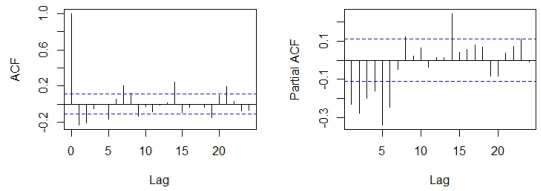

Fig. 3: ACF and PACF of differenced log Polish data.

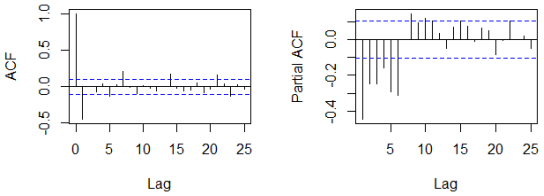

Fig. 4: ACF and PACF of differenced log German data.

B. Model Fitting

Once we have stationary data, we try to find the best fit SARIMA models by using a manual, iterative Box-Jenkins approach.

Before we begin building our models, we need to segment our data into training and test sets so we can assess the ability of our models to forecast the time series. Taking 10% of the data leaves a forecast horizon of 31 days for the Polish data and 35 days for the German data.

The modelling procedure is composed of five iterative steps: estimation of ARIMA parameters, diagnostic checking, estimation of SARIMA parameters, further diagnostic checking and finally prediction.

Since there is no clear cut off in either data's ACF and PACF, we start the process with an $\mathrm{ARIMA}(1, 1, 1)$ and iterate, changing the p and q values until the best model has been found. To check the models, we look at the p-value of the Ljung-Box statistic, which should be above the threshold of 0.05. This tests the randomness of our residuals. Then, by plotting the ACF and PACF of the residuals, we look for the resemblance of white noise. Lastly, we select the model based on information criterion, such as Akaike's Information Criterion (AIC), which measures the degrees of freedom/numbers of free parameters. As a result, we aim to choose a suitable model with the least AIC score in order to avoid over-fitting.

The seasonality to our data means that we can improve our ARIMA models by adding MA and AR components for seasonality - forming a SARIMA model. Hence we repeat the process above, but for the seasonal component, until settling on the best models for each data set.

Following this, we forecast our data (over the forecast horizon) using a fixed origin approach. We are then able to compare these forecasts to the real data to assess their ability. We finally use Error Matrix tests to measure the strength of, and error in, our predictions.

III. Results

A. Polish Model

1) ARIMA Selection: Using R, we choose different permutations of parameters, starting from $\mathrm{ARIMA}(1, 1, 1)$, and select the best fitting models as described in the previous section. The scores for each test of the most appropriate models are displayed in the Table I. Based on that, we consider $\mathrm{ARIMA}(2, 1, 2)$ and $\mathrm{ARIMA}(3, 1, 4)$ for the SARIMA process.

2) SARIMA Selection: Building on ARIMA, we repeat the process for the seasonal component and obtain two best performing models $\mathrm{SARIMA}(2, 1, 2, 0, 1, 2, s = 7)$ and $\mathrm{SARIMA}(3, 1, 4, 2, 1, 1, s = 7)$. The ACF and PACF plots of their residuals (Fig.5) are suitable for white noise, with a slightly better performance of the latter model.

3) Model Selection: For the final selection, we consider the complexity of the models – the first one is simpler than the second. For the Ljung-Box statistic - both pass the p-value threshold of 0.05 and residuals look like white noise. Finally, by comparing the AIC score, we choose $\mathrm{SARIMA}(3, 1, 4, 2, 1, 1, s = 7)$.

B. German Model

1) ARIMA Selection: Similarly, we use R to repeat this process for the German data. The best performing 3 models across the selection criteria were $\mathrm{ARIMA}(3, 1, 2)$, $\mathrm{ARIMA}(4, 1, 2)$ and $\mathrm{ARIMA}(5, 1, 2)$ (see Table I), which we iterate on for the seasonal component to find the best SARIMA model.

Data	Model	AIC	Log-Likelihood	Ljung-Box	MAE Av.	RMSE Av.	MAE Roll. O. (k=1)	RMSE Roll. O. (k=1)
Polish	ARIMA(2, 1, 2)	690.527	-339.26	0.8102	-	-	-	-
Polish	ARIMA(3, 1, 4)	696.4566	-339.23	0.9779	-	-	-	-
Polish	ARIMA(1, 1, 3)	714.0732	-351.04	0.7366	-	-	-	-
Polish	SARIMA(2, 1, 2, 0, 1, 2, s = 7)	642.42	-314.39	0.9978	1.177	1.346	0.381	0.504
Polish	SARIMA(3, 1, 4, 2, 1, 1, s = 7)	638.695	-308.35	0.943	1.283	1.419	0.371	0.498
German	ARIMA(3, 0, 2)	995.14	-490.57	0.7555	-	-	-	-
German	ARIMA(4, 0, 2)	992.63	-488.32	0.6898	-	-	-	-
German	ARIMA(5, 0, 2)	979.52	-480.76	0.8863	-	-	-	-
German	SARIMA(3, 1, 2, 1, 0, 1, s = 7)	957.27	-469.64	0.9999	1.045	1.092	0.272	0.342
German	SARIMA(5, 1, 2, 1, 0, 1, s = 7)	961.46	-469.73	0.9718	0.883	0.037	0.261	0.33

TABLE I: Score comparisons for the Polish and German models.

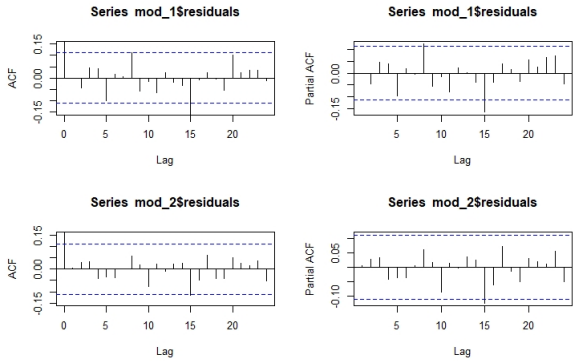

Fig. 5: ACF and PACF of residuals for both Polish models.

2) SARIMA Selection: Further iteration on the seasonal MA and AR parameters provides $\mathrm{SARIMA}(3,1,2,1,0,1,s=7)$ and $\mathrm{SARIMA}(5,1,2,1,0,1,s=7)$ as the most appropriate models.

3) Model Selection: In order to select the leading SARIMA model we require residuals that resemble white noise, so we plot the two model's respective residual's ACF and PACF. Both of these are suitable. Since both models have p-values close to one across many lags for the Ljung-Box statistic, we choose the simpler model $\mathrm{SARIMA}(3,1,2,1,0,1,s=7)$ which has a lower AIC score.

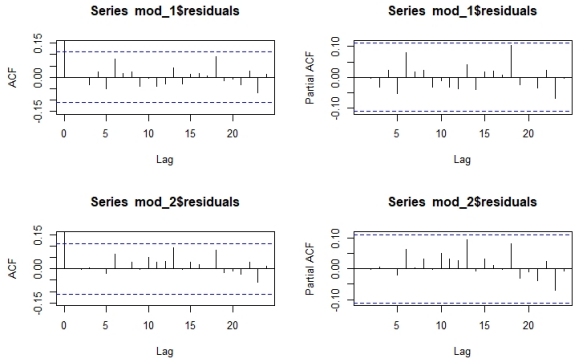

Fig. 6: ACF and PACF of residuals for both German models.

C. Forecasting Models For Both Countries

For the SARIMA models we built, we estimate future values x_{n+k} for $k = 1, 2, 3...$ leaving out 10% of the last observations for both Polish and German data, take exponents of these forecasts to transform to new deaths and compare the performance against their respective models. Then, we evaluate our forecasts by differencing the realised observations with the estimates and calculate the mean absolute error (MAE) and the root mean square error (RMSE). For good forecasts, these scores will be minimised. The plots, Fig. 7, Fig. 8, are from a fixed forecast origin n.

The forecast of future values x_{n+k} for $k = 1, 2, 3...$ for SARIMA(2, 1, 2, 0, 1, 2, s=7) is:

$$(1-\phi_1 B-\phi_2 B^2)(1-B)(1-B^7)(x_{n+k}-\mu) = (1+\theta_1 B+\theta_2 B^2)(1+\Theta_1 B^7+\Theta_2 B^{14})e_{n+k}$$

To derive an expression for the expected forecasts, we set $e_{n+1} = e_{n+2} = ... = 0$ and for simplicity $\mu = 0$. Expanding the left hand side, we obtain

$$\hat{x}_{n,k} = \hat{x}_{n,k-1}(\phi_1-1) + \hat{x}_{n,k-2}(\phi_2-\phi_1) - \phi_2\hat{x}_{n,k-3} + + \hat{x}_{n,k-7} - \hat{x}_{n,k-8}(1+\phi_1) + \hat{x}_{n,k-9}(\phi_1-\phi_2) + + \phi_2\hat{x}_{n,k-10}$$

Similar can be written for other models.

Starting from 9th Dec 2020 until 8th Jan 2021 for the Polish model prediction (red line), we can see that both SARIMA models perform well overall. However, we observe a much better accuracy at the start of predictions, which get worse over time. The same happens to German data with forecasts from 5th Dec 2020 to 8th Jan 2021.

The errors calculated in R, suggest that both Polish SARIMA models perform similarly with small difference. In contrary, for Germany, the complex model is more accurate but as the difference is insignificant, we maintain that the simpler model is better to avoid overfitting.

IV. Discussion

A. Impact of Lockdowns and Restrictions

Analysing the time series for both countries, we want to look at other factors that might be affecting daily new deaths such as governments' actions to tackle the pandemic.

In Poland, the nation was put into a national lockdown and closed their borders relatively early. As a result, they were saved from the first wave and did not record any significant peaks at the start. In October, as the daily new deaths increased during that time, Poland was put into 'red zone', equivalent to tier 4 in the UK, before going into the second national lockdown on 28th December. For the last month, there is a decreasing trend of reported death cases.

Unlike in Poland, we can see the first peak in COVID-19 daily deaths in Germany back in April. We can observe the lag of the death rate until the lockdown took effect and decreased

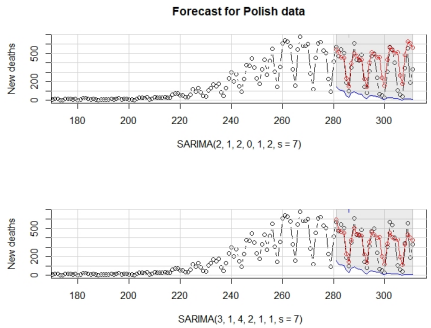

Fig. 7: Forecast for Polish data.

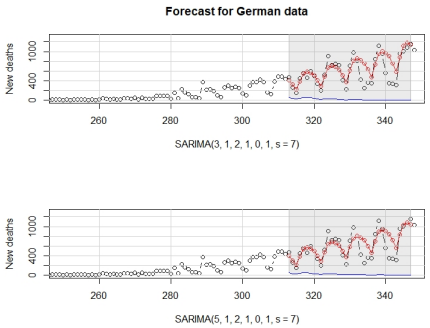

Fig. 8: Forecast for German data.

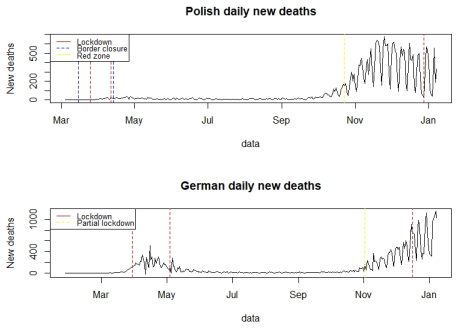

Fig. 9: Polish and German Data With Restirictions Added.

the number of deaths. Looking at the death rate for the last few months, the country is still tackling with the mortality rate indicated by the increasing trend.

Throughout summer and fall there was a growing number of anti-lockdown protests around both countries, with a greater severity in Poland. At the start of January, many businesses started reopening, rejecting their lockdown to survive. A map of 'free businesses' was created by the movement and some businesses have successfully overturned fines imposed by sanitary inspectors. Due to these circumstances, we might expect a higher death rate in the upcoming months in Poland.

B. Improvements

Reflecting on our work, we discuss our observations and how the analysis can be done differently.

Firstly, we added a small non-zero value to our observations at the start so that we can take the log of the data in the case of no recorded deaths. Noting that as the value tends to zero, the log of it tends to infinity, we choose a reasonably small value of $\exp(-3)$ that would not make a significant impact on stationarity.

The ever changing nature of the pandemic means that modelling over the full data set reduces our ability to accurately predict forthcoming death figures. To improve accuracy, we could use a shorter period for our model building. This could be further advanced by producing a rolling forecast, replacing with a new training period for each model prediction - updating the data in real-time.

The complexity of death cases during pandemic questions the effectiveness of our models. By using univariate time series, observations only depend on time, while there are multiple factors affecting mortality rate. Such examples include population: overall pre-existing medical conditions, demographic; social distancing measures or the number of medical staff and facilities per citizen. Furthermore, multivariate analysis would enable us to better relate the effect that neighboring countries, such as those in the EU Schengen Area, have on each other.

Therefore, we can consider multivariate time series models and involve more parameters. Based on a research article [7], we can use the Principal Component Analysis (PCA) to take into account other factors. In the paper, researchers combined the data from 56 countries on March 30, 82 countries on April 15, and 91 countries on April 25 2020 of total cases, total deaths, active cases, and critically ill cases into one single score, PC-1, and mortality recovery ratio, PC-2, at each time point. First, observations were converted to Z-scores, put into linear combinations, where each is a factor, and written as a correlation matrix. Then, using a clustering method grouped the countries accordingly to their score. A positive value of PC-1 indicates a high number of cases and deaths, whilst a positive PC-2 – a higher ratio of mortality to recovered cases.

One of the forecasting methods for multivariate time series is Vector Auto Regression [8], in which each variable is a linear function of the past values of itself and the past of other variables. For two time-dependent variables the equation is

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} x_1(t-1) \\ x_2(t-2) \end{bmatrix} + \begin{bmatrix} e_1(t) \\ e_2(t) \end{bmatrix}$$

a_1, a_2 - constant terms.

V. Conclusion

In conclusion, sadly, many lives have been lost and the consequences will be felt for many years. One way to help mitigate the impact of the pandemic is to understand the situation. Through time series analysis, we are able to find the best models to make predictions and to prepare for the repercussions. In our report, we analysed Polish and German daily death cases and built models to predict how these figures may evolve. These results can be used by decision makers to adjust restrictions and plan for the safe management of the rest of the pandemic. Whilst no models are perfect, some may be useful.

REFERENCES

[1] M. T. Meehan, D. P. Rojas, A. I. Adekunle, O. A. Adegboye, J. M. Caldwell, E. Turek, B. Williams, J. M. Trauer, and E. S. McBryde, "Modelling insights into the covid-19 pandemic," *Paediatric respiratory reviews*, 2020. 1

[2] "Coronavirus disease travel information (covid-19)." [Online]. Available: https://www.fitfortravel.nhs.uk/advice/disease-prevention-advice/coronavirus-disease-covid-19 1

[3] "Coronavirus cases woldometer data." [Online]. Available: https://www.worldometers.info/coronavirus/?utm_campaign=homeAdvegas1%3F#countries 1

[4] Z. Ceylan, "Estimation of covid-19 prevalence in italy, spain, and france," *Science of The Total Environment*, vol. 729, p. 138817, 2020. 1

[5] M. R. Mahmoudi, M. Maleki, and A. Pak, "Testing the difference between two independent time series models," *Iranian Journal of Science and Technology, Transactions A: Science*, vol. 41, no. 3, pp. 665–669, 2017. 1

[6] M. Roser, H. Ritchie, E. Ortiz-Ospina, and J. Hasell, "Coronavirus pandemic (covid-19) - statistics and research," Mar 2020. [Online]. Available: https://ourworldindata.org/coronavirus 1

[7] A. A.-S. N.-F. A.Ramadan, A. Kamel, "A multivariate data analysis approach for investigating daily statistics of countries affected with covid-19 pandemic," 2020, source accessed on 05-02-2021"https://www.sciencedirect.com/science/article/pii/S240584402032418X". 4

[8] A. Singh, "A multivariate time series guide to forecasting and modeling," 2020, source accessed on 05-02-2021 "https://www.analyticsvidhya.com/blog/2018/09/multivariate-time-series-guide-forecasting-modeling-python-codes/#:~:text=A%20Multivariate%20time%20series%20has,used%20for%20forecasting%20future%20values.&text=In%20this%20case%2C%20there%20are,considered%20to%20optimally%20predict%20temperature". 4