engineering proceedings

MDPI

Proceeding Paper

Epidemiology SIR with Regression, Arima, and Prophet in Forecasting Covid-19 †

Pedro Furtado

Departamento de Engenharia Informatica/CISUC, Universidade de Coimbra/Polo II, 3030-790 Coimbra, Portugal; pnf@dei.uc.pt

† Presented at the 7th International Conference on Time Series and Forecasting, Gran Canaria, Spain, 19–21 July 2021.

Abstract: Epidemiology maths resorts to Susceptible-Infected-Recovered (SIR)-like models to describe contagion evolution curves for diseases such as Covid-19. Other time series estimation approaches can be used to fit and forecast curves. We use data from the Covid-19 pandemic infection curves of 20 countries to compare forecasting using SEIR (a variant of SIR), polynomial regression, ARIMA and Prophet. Polynomial regression deg2 (POLY d(2)) on differentiated curves had lowest 15 day forecast errors (6% average error over 20 countries), SEIR (errors 25–68%) and ARIMA (errors 15–85%) were better for spans larger than 30 days. We highlight the importance of SEIR for longer terms, and POLY d(2) in 15-days forecasting.

Keywords: time series forecasting; epidemiology; SIR

Citation: Furtado, P. Epidemiology SIR with Regression, Arima, and Prophet in Forecasting Covid-19 . *Eng. Proc.* **2021**, *5*, 52. https://doi.org/10.3390/engproc2021005052

Academic Editors: Ignacio Rojas, Fernando Rojas, Luis Javier Herrera and Hector Pomare

Published: 13 July 2021

1. Introduction

The typical initial evolution of an epidemic when the population has no immunity and the pathogen has high virulence and high death rate is a frightening exponential curve. Scientists still lack a lot of knowledge about the SARS-CoV-2 virus and variants, and recently new virus variants have increased its contagiousness. Nevertheless, we can say that its reproduction number (rate of growth) should be around 3 and its death rate could be around 0.3%. Since a reproduction number higher than 1 already means exponential growth, a value of 3 indicates a significant virulence and containment is necessary if no vaccination is available or significant acquired immunity in the population. In order to study and predict the evolution of the curve and to decide how to act at each moment, epidemiologists and mathematicians frequently use variants of the Susceptibility Infected Removed (SIR). SIR or the alternative SIR with an additional state called Exposed (SEIR) we use here, is actually a simple model. Given four possible states (S, E, I and R), at any given moment each individual from a population can be in any of those states. At the start, with zero immunity, all population is in state S (Susceptible). In the model individuals transition from S to I (infected) and from there to recovered or dead (R). The model also uses some other parameters to describe rates of transition in three differential equations.

There are many other ways to fit and forecast curves in general. Approaches such as linear regression, polynomial regression, ARIMA [1] or other time series forecasting approaches could in principle be used in this context as in any other context, but the question is whether they would stand any chance when compared with SEIR that integrates epidemic-specific parameters. Lacking all the specific model parameters that SEIR can include, we expect those more generic time series analysis models to miss important information that leads to future changes in the curve, but on the other hand it is a possibility that they could be useful in short-term (few days) forecasting, with more constant conditions.

We review the approaches, setup variants of polynomial regression, ARIMA, Prophet and SEIR, collect the Covid-19 evolution curves of the 20 hardest hit countries at the end of March and compare their performance forecasting the last 15 days of the curves and

different lengths as well. This setup allowed us to reach conclusions regarding their relative merits. In this work we were slightly constrained by the fact that, except for China where the outbreak started much earlier, we had only around 45 days of data for most countries (the time from the start of the outbreak in most countries to this study), but on the other hand this study is especially interesting because it deals with an ongoing outbreak of hard consequences. Future work would be interesting in generalizing these results with more diseases and more forecasting evaluation alternatives.

2. Related Work

Epidemiological modelling has been discussed extensively in works such as [2–6]. The Susceptible-Infected-Removed model (SIR) is reviewed in [7] and an analytical solution to it is discussed in [8]. According to the definition, infection transitions between the three states given in the name itself, and a set of three equations describes the transitions between those states. Parameters include the average contacts of individuals (Beta) times transmission probability per contact (I), rate of deaths and recoveries, time a person is infected D and its inverse γ. There are many other models that evolved from SIR, including those in [9].

Polynomial regression is a fitting procedure that tries to approximate a given curve using a polynomial of degree n. Reference [10] describes numerical methods for curve fitting. Regression analysis [11,12] focuses especially on statistical inference related to curve fitting and associated uncertainty. These kinds of approaches can be helpful for abstracting trends and forecasting into the near future in different contexts.

The model Autoregressive Integrated Moving Average is reviewed in [1,13]. It uses differentiation iterations to solve non-stationarity, plus regression, moving averages and integration to successively improve data fitting. A simpler but also effective model [14] was proposed by Facebook^{TM}. In [15] we used both Prophet and a modified ARIMA to predict evolution of business performance indicators in Telecom. In that specific application, we concluded that ARIMA outperformed Prophet.

3. Curve Fitting and Forecasting Models

3.1. The SEIR Model Plus a Social Distance Factor

The SEIR model [5,16] has susceptible (S), exposed (E), infected (I) and recovered (R) states and describes the dynamics of the population successively moving from one of the states to the next. As soon as individuals reach state R they are no longer able to become infected. Initially, the whole population is in state S. The following differential equations model how the individuals of a population evolve between these states in SEIR. For instance, $S'(t)$ is the change in the number of people in state S from moment t to $t + 1$. A social distancing factor is also added to model the degree of distance between people, which has the potential to decrease the rate of contagions:

$$S_{t+1} = \rho \times \beta \times S_t \times I_t \quad (1)$$

$$E_{t+1} = \rho \times \beta \times S_t \times I_t - \propto \times E_t \quad (2)$$

$$I_{t+1} = \propto \times E_t - \gamma \times I_t \quad (3)$$

$$R_{t+1} = \gamma \times I_t \quad (4)$$

In these four equations, α is the inverse of the incubation time ($1/\mathrm{d}\alpha$), estimated to be 5 days in average for Covid-19 (varying between 1 and 14 days); $\beta = \tau \times c$ is transmissibility (τ = infection probability with contact with infected) and the average rate of contact between susceptible and infected individuals c, obtained by curve fitting; γ is the inverse of the mean infectious time ($1/\mathrm{d}\gamma$), estimated to be 10 days; ρ is the social distancing factor, varying between 0 and 1, observable by curve fitting. We have coded this model together with least squares fitting to find the term "social distancing $\times$ Beta ($\rho \times \beta$) that minimizes the average root mean squared error (RMSE) between the SEIR curve and the official

country curve. Forecasting was done by assuming that ($\rho \times \beta$) remains the same for the next days.

3.2. Polynomial Regression

Polynomial regression (POLY) uses least-squares fitting to find the coefficients of a polynomial of degree n that best fit a curve. Equation (1) shows the polynomial, where some curve Y is to be approximated by the polynomial function with coefficients C_0 to C_n,

$$Y_a = C_0 \times x^n + C_1 \times x^{n-1} + \ldots + C_{n-1} \times x + C_n \quad (5)$$

In our case the x is the time unit (the forecast is Y_a value for time unit ti. Different polynomial degrees were tested in our experiments. We will show in our experiments that forecasting with POLY over differentiated curves of "number of active patients" instead of the original followed by an inverse transformation to reconstruct the curve yielded best results (ARIMA also uses differentiation to work on stationary curves).

3.3. Time Series Forecasting with ARIMA

For our own review of time series forecasting using ARIMA, please refer to [15], another reference is [9]. In [15] we explain how ARIMA uses Auto-Regressive (AR) and Moving Average (MA) models and how a set of parameters are applied ineach of those component models. Please, refer to [15] for more details on ARIMA.

3.4. Automated Parameterization of ARIMA

In automated parameterization of ARIMA, a set of three parameters are tuned (p, d, q) and seasonality as well are obtained by automatically testing the fit of the curve for each combination of those values. The Akaike criteria (AIC), estimating the prediction error, provides a relative metric for the model quality. Thus, AIC provides a means for model selection. In the following example, the combination with lowest AIC is chosen.

pdq = 0, 0, 0 resulting in AIC = 705.7393610322358
...
pdq = 0, 1, 1 resulting in AIC = 456.58099826482464
...
pdq = 1, 1, 1 resulting in AIC = 304.90034871700635

3.5. Time Series Forecasting with Prophet

Forecasting using Prophet is described in [17] and in detail in [18], or in our own prior work [15]. It uses three sub-models, one analyzing the trend, another one analyzing the seasonality and the third one taking into account festivity periods. Each of those sub-models is modelled by a function (logistic for trend, Fourier for seasonality and an adjustment for festivity periods.

4. Experimental Work

Our experimental setup was created using the pandemics curves (number of active cases) up to a specific day (27 March 2020). This data can be obtained for instance in [19]. Figure 1 illustrates the curves, showing the per-million aligned active cases in 5 European countries (we show only 5 countries to avoid cluttering). The number of active cases of the 20 most hit countries except China (27 March) were used for curve fitting, and Chinas's curve was used for testing longer forecasting spans.

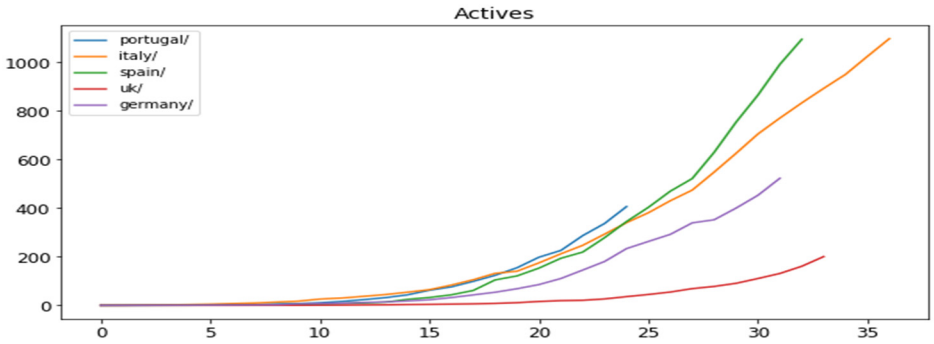


Figure 1. Active cases up to 27 March, 5 Europe countries align start of outbreak (number of cases > 0).

For our experimental setup we implemented each of the forecasting approaches (polynomial regression, ARIMA, Prophet and SEIR) and all necessary data loading and data transformations in python. The full list of countries in our setup included 21 countries. China was used in forecasting larger lengths. We tested polynomial degrees 1 to 4 and differentiation as well. The experimental procedure is simple: extract x days from the curve, fit the model to the truncated curve and finally forecasting the last x days using the model. All results report the average error using *MAE* relative to the values (*MAEr*) for the new forecasted segment, shown in Equation (5):

$$MAEr = \frac{\sum_1^n \left| \frac{Y_i - YY_{ei}}{Y_i} \right|}{n} \quad (6)$$

4.1. Fifteen-Day Forecast over 20 Countries

The 15-day forecast is done on each country by extracting 15 days from the curve, then fitting the model to the truncated curve and finally forecasting the last 15 days using the model. Figure 2 and Table 1 show the results for 20 countries (POLY d(n) = polynomial regression with degree n on differentiated curve, Prophet d = Prophet on differentiated curve).

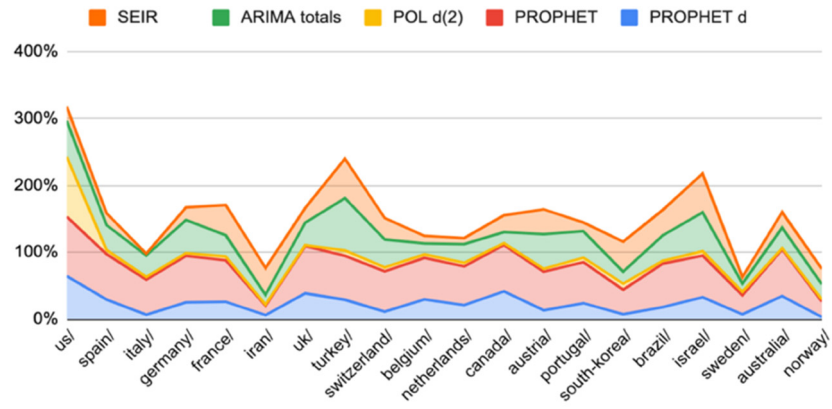


Figure 2. Comparison of methods over 20 countries (MAER, stacked chart).

Note that in the previous results we have chosen POLY d(2) because it had the best forecasting results. Table 2 shows the comparison of forecasting results of different polynomial regression options for the same experimental setup. In that table the number in parenthesis is the polynomial degree and d stands for differentiation.

Table 1. Comparison of methods over 20 countries (MAER).

	PROPHET d	PROPHET	POLY d(2)	ARIMA	SEIR
Countries avg	23%	57%	9%	35%	28%
stdev	15%	18%	19%	16%	16%
US	64%	89%	89%	54%	21%
Spain	29%	68%	6%	37%	18%
Italy	7%	52%	4%	32%	3%
Germany	25%	69%	4%	50%	19%
France	26%	62%	6%	32%	45%
Iran	6%	14%	2%	15%	40%
UK	39%	71%	1%	34%	22%
Turkey	29%	66%	8%	78%	59%
Switzerland	11%	60%	6%	42%	32%
Belgium	29%	62%	5%	16%	11%
Netherlands	21%	58%	5%	28%	9%
Canada	41%	69%	3%	17%	25%
Austria	13%	58%	4%	52%	37%
Portugal	24%	61%	7%	40%	13%
South-Korea	7%	37%	9%	18%	45%
Brazil	18%	65%	4%	38%	38%
Israel	33%	62%	7%	58%	58%
Sweden	7%	28%	5%	12%	10%
Australia	35%	70%	2%	31%	23%
Norway	3%	23%	3%	23%	23%

Table 2. Summary of MAER for polynomial regression, 20 countries 15-day forecast.

POLY 4	POLY 3	POLY(2)	POLY(1)	POLY d(4)	POLY d(3)	POLY d(2)
44%	28%	21%	49%	14%	11%	9%

4.2. Discussion of Results

The results in Figure 2 and Table 1 show that POLY d(2) had the best 15-day forecasts (MAEr 9 $\pm$ 19%), less than half the next competitor, then Prophet d (22 $\pm$ 15%) and SEIR (28 $\pm$ 16%). ARIMA had (35 $\pm$ 16%) and Prophet without differentiation was the worst (55 $\pm$ 18%). All techniques had a similar degree of variation between countries (stdevs 15% to 19%). The experiment with multiple alternatives of polynomial regression show that differentiating was useful and the error was smaller with smaller polynomial degree in the tested interval 2 to 4. In essence, the polynomial regression of degree 2 is fitting the curve by a parabola that mimics the initial steep increase of the number of daily cases, then as confinement and social distancing kicks in, the change first to a stabilization and then a decrease of the number of daily cases. But degree 3 or 4 on the differentiated curve also had relatively small errors.

The fact that SEIR did not achieve the best 15-days forecast may seem surprising, since SEIR is the preferred epidemic modeling approach. However, although its results were still reasonable and we still expect SEIR to be the best for forecasting the whole epidemic curve, some of its parameters are abstractions that mean it may not fit official epidemics curves perfectly. One such parameter is the initial population of Susceptibles (S), and the variability is due to the degree of susceptibility of individuals to the contagion and to transmission varying in reality. There is also a problem with the official account of quantity of infected actually infected, since there are many asymptomatic patients and knowing the actual quantity of infected would require extensive continuous testing. The dynamics of transmission and social distancing at a country level is also a coarse approximation, since there exist high-density, highly populated cities and lower-density zones in every country.

Prophet on the differentiated curve scored 22% error, and both Prophet and POLY were much better if used on the differentiated curve. The best ARIMA results were obtained

with the input curve not differentiated (35%), but note that ARIMA itself differentiates through the parameter d for stationarity.

4.3. Testing the Approaches on Larger Spans (China)

China was the only Covid-19 worst-hit countries curve for which there were considerably more days (around 90). The next experiment consisted in removing a variable number of days from that series, building the model and forecasting those removed days using the model. The results are shown in Figure 3 (stacked chart) and Table 3.

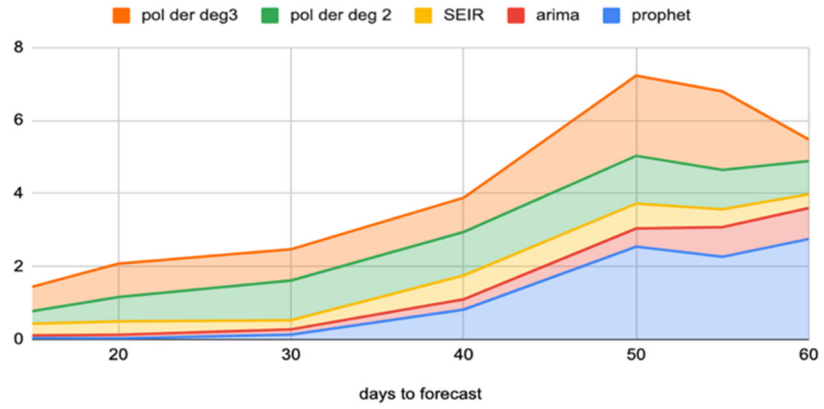


Figure 3. China with different spans (MAER, stacked chart) (der = differentiated, deg = degree).

Table 3. MAER data for China spans.

Days to Forecast	Prophet	ARIMA	SEIR	Poly der Deg 2	Poly der Deg 3
15	4%	7%	32%	34%	67%
20	3%	10%	37%	67%	92%
30	13%	15%	25%	109%	85%
40	82%	28%	65%	119%	94%
50	254%	50%	68%	131%	220%
55	226%	81%	49%	108%	215%
60	275%	85%	37%	91%	59%
AVG	123%	39%	45%	94%	119%
STD	125%	33%	17%	33%	69%

4.4. Discussion on Larger Forecasts

SEIR was clearly superior for this case of longer forecast periods, it had the best scores overall and least variance (error between 32% and 45%). The specific modeling that considers important epidemiology concepts overcame the results of generic models when modeling on longer spans. ARIMA was next (7% to 85% errors), then POLY der deg 2 (34% to 131%), while the polynomial of degree 3 was much worse (59% to 220%). Prophet (4% to 275%) had the smallest errors up to forecast length 30 and then the largest errors for the remaining lengths. The advantage of ARIMA over Prophet on longer spans could be related to ARIMA adjusting its p, d, q parameters.

4.5. Visualization of Some Results

We do not have space to show most visualizations, however, we show a few illustrative examples next. Figure 4 is the daily cases in Spain together with the POLY d(2) forecast, and Figure 5 is the daily cases in China together with the SEIR forecast, both showing the actual values (blue curve) together with the estimations. Figure 6a is Spain's 15 day forecast using ARIMA and Figure 6b is the forecast for the same case using Prophet. We can see that POLY d(2) was near the actual curve in Figure 4, although at the end of the interval the divergence increased, SEIR forecast was also quite good in Figure 5, while in

ARIMA the forecast for the last days was diverging significantly upwards and in Prophet it was diverging downwards.

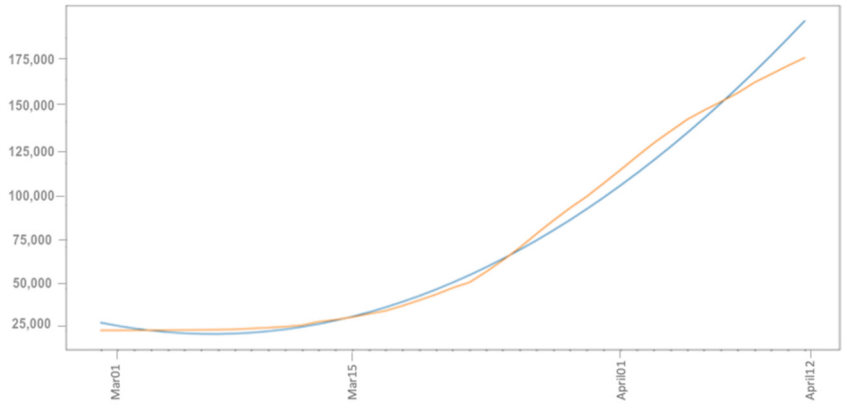


Figure 4. Example of POLY d(2) Spain 15-days forecasting (blue) over original line (orange), daily cases.

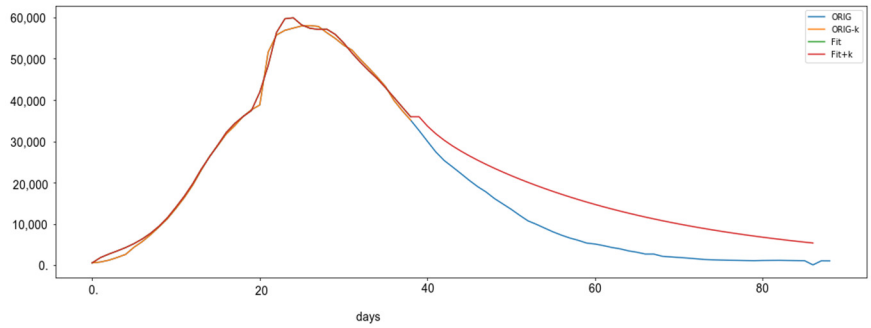
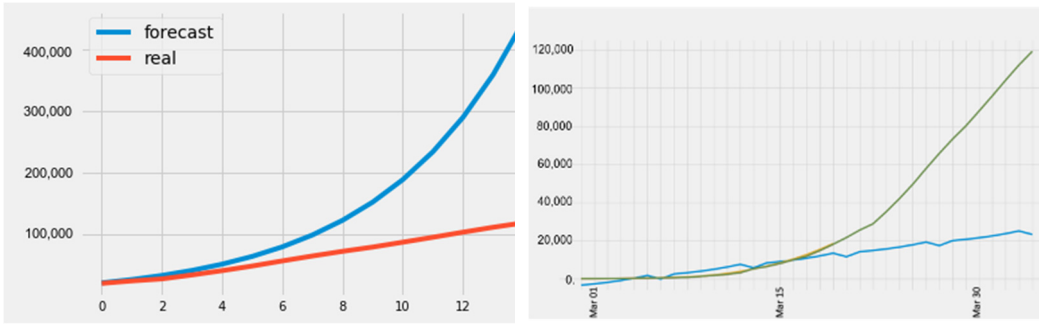


Figure 5. SEIR forecasting (red) and original curve (orange) plus cut (blue) for China, daily cases.



(**a**) (**b**)

Figure 6. ARIMA and Prophet 15-days forecasting of Spain (days versus number of infected). (**a**) ARIMA (forecast in blue); (**b**) Prophet (forecast in blue).

5. Conclusions

In this article we investigated the issue of short and longer-term forecasting over epidemiological curves of Covid-19 using both generic forecasting approaches and the more specific epidemiological SEIR model, with the objective of confronting the alternatives. After describing the approaches used we created an experimental setup with the alternatives and tested over 20 countries, plus longer-term forecasts on the longest curve (China). We concluded that, in average, polynomial regression of degree 2 was the best for short term (15 days or less), but on longer term SEIR was clearly superior to the competition, which is explained by its use of more specific epidemiological parameters. The use of Covid-19

curves makes the work very up-to-date, but on the other hand we would like to experiment with other epidemics curves and to test different segments on multiple spans. Our own current and future work deals also with automatic fitting, parameter optimization and what-if analysis with the SEIR model.

Funding: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All the data used in this study is publicly available in https://www.worldometers.info/coronavirus/, accessed on 1 June 2020.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Brockwell, P.J.; Davis, R.A. *Time Series: Theory and Methods*, 2nd ed.; Springer: New York, NY, USA, 2009; p. 273. ISBN 9781441903198.
2. Hethcote, H. The Mathematics of Infectious Diseases. *SIAM Rev.* **2000**, *42*, 599–653. [CrossRef]
3. Bailey, N.T. *The Mathematical Theory of Infectious Diseases and its Applications*, 2nd ed.; Griffin: London, UK, 1975; ISBN 0-85264-231-8.
4. Altizer, S.; Nunn, C. Infectious diseases in primates: Behavior, ecology and evolution. In *Oxford Series in Ecology and Evolution*; Oxford Univers Press: Oxford, UK, 2006; ISBN 0-19-856585-2.
5. Brauer, F.; Castillo-Chávez, C. *Mathematical Models in Population Biology and Epidemiology*; Springer: New York, NY, USA, 2001; ISBN 0-387-98902-1.
6. Anderson, R.M. *Population Dynamics of Infectious Diseases: Theory and Applications*; Chapman and Hall: London, UK, 1982; ISBN 0-412-21610-8.
7. Kermack, W.O.; McKendrick, A.G. A Contribution to the Mathematical Theory of Epidemics. *Proc. R. Soc.* **1927**, *115*, 772.
8. Is Prophet Really Better than ARIMA for Forecasting Time Series Data. Available online: https://blog.exploratory.io/is-prophet-better-than-arima-for-forecasting-time-series-fa9ae08a5851 (accessed on 16 August 2018).
9. A Comprehensive Beginner's Guide to Create a Time Series Forecast. Available online: https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python (accessed on 15 March 2018).
10. Guest, P.G.; Guest, P.G. *Numerical Methods of Curve Fitting*; Cambridge University Press: Cambridge, UK, 2012.
11. Motulsky, H.; Christopoulos, A. *Fitting Models to Biological Data Using Linear and Nonlinear Regression: A Practical Guide to Curve Fitting*; Oxford University Press: Oxford, UK, 2004.
12. Freund, R.J.; Wilson, W.J.; Sa, P. *Regression Analysis*; Elsevier: Amsterdam, The Netherlands, 2006.
13. Box, G.; Jenkins, G.M.; Reinsel, G.C. *Time Series Analysis: Forecasting and Control*, 3rd ed.; Prentice-Hall: Hoboken, NJ, USA, 1994; ISBN 0130607746.
14. Prophet Forecasting at Scale. Available online: https://facebook.github.io/prophet/ (accessed on 16 April 2019).
15. Pinho, A.; Costa, R.; Silva, H.; Furtado, P. Comparing Time Series Prediction Approaches for Telecom Analysis. In *International Conference on Time Series and Forecasting*; Springer: Cham, Switzerland, 2018; pp. 331–345.
16. Harko, T.; Lobo, F.S.; Mak, M.K. Exact analytical solutions of the Susceptible-Infected-Recovered (SIR) epidemic model and of the SIR model with equal death and birth rates. *Appl. Math. Comput.* **2014**, *236*, 184–194. [CrossRef]
17. Wang, M.; Wang, Y.; Wang, X.; Wei, Z. Forecast and Analyze the Telecom Income based on ARIMA Model. *Open Cybern. Syst. J.* **2015**, *9*, 2559–2564. [CrossRef]
18. Taylor, S.J.; Letham, B. Forecasting at scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
19. World-o-Meter Site with Worldwide and Per-Country Corona Virus Information. Available online: https://www.worldometers.info/coronavirus/ (accessed on 7 April 2020).