Time Series Analysis and Prediction of COVID-19 pandemic using Dynamic Harmonic Regression Models

Lei Wang[1,*]

[1]College of Science and Mathematics, Augusta University

Abstract

Rapidly spreading Covid-19 virus and its variants, especially in metropoli- tan areas around the world, became a major health public concern. The tendency of Covid-19 pandemic and statistical modelling represent an urgent challenge in the United States for which there are few solutions. In this paper, we demonstrate com- bining Fourier terms for capturing seasonality with ARIMA errors and other dynamicsin the data. Therefore, we have analyzed 156 weeks COVID-19 dataset on national level using Dynamic Harmonic Regression model, including simulation analysis and ac- curacy improvement from 2020 to 2023. Most importantly, we provide a new advanced pathways which may serve as targets for developing new solutions and approaches.

Research Article

Open Access &

Peer-Reviewed Article

DOI : 10.14302/issn.2643-2811.jmbr-23-4528

Corresponding author:

Lei Wang, College of Science and Mathematics, Augusta University.

Keywords:

Dynamic Harmonic Regression with ARIMA errors; COVID-19 pan- demic; Forecasting models; Time series Analysis; Weekly seasonality.

Received: Mar 18, 2023

Accepted: Apr 25, 2023

Published: April 26, 2023

Academic Editor:

Raul Isea, Fundación Instituto de Estudios Avanzados -IDEA.

Citation:

Lei Wang (2023) Time Series Analysis and Prediction of COVID-19 pandemic using Dynamic Harmonic Regression Models. Journal of Model Based Research - 2 (1):28-36. https://doi.org/10.14302/issn.2643-2811.jmbr-23-4528

Introduction

The COVID-19 pandemic has had a tremendous impact on the world for 3 years and spread to more than 200 countries worldwide, leading to more than 36 million confirmed cases as of October 10, 2020. Some well-respected organizations such as Johns Hopkins University, the Centers for Disease Control and Prevention, the World Health Organization and the United States Census Bureau are involved in the study and tracking of the Covid-19 pandemic [2].

To respond this urgent public health concern, we used 156 weekly time series datasets to evaluate the seasonal patterns of COVID-19 cases and mortality in the United States with the objective to determine the tendency of Covid-19 pandemic. Besides, the implantation of R and simulation analysis can improve the forecasting accuracy

Given my prospective research interest in Data Science, smart data analytics is giving profes- sionals and public more insight into the factors impacting than ever before. From assessing risks to analyzing evolving trends, we are now able to anticipate the success of a property more accurately thanks to the abundance of information available to academics and professionals. Our analysis canhelp in understanding the trends of the disease outbreak and provide suggestions and instructions of adopted countries.

Based on complex nature of virus transformation, traditional epidemic models such as Regressionand ARIMA methods have been applied for prediction of its spread. Particularly, Dynamic HarmonicRegression (DHR) approaches were used to predict the spreading trends of COVID-19, such as new cases and deaths. We reviewed studies that implemented these strategies [10].

Dynamic Harmonic Regression (DHR) is a nonstationary time-series analysis approach used to identify trends, seasonal, cyclical and irregular components within a state space framework. Many re- searchers studied about this forecasting methods. Dr.Kumar and Dr.Suan (2020) use ARIMA model and day level information of COVID-19 spread for cumulative cases from whole world and 10 mostlyaffected countries to forecast the impact of the virus in the affected countries and worldwide [1]. Also, Dr.Fuad Ahmed Chyon Md, Dr.Nazmul Hasan Suman employed ARIMA model to analyze the temporal dynamics of the worldwide spread of COVID-19 in the time window from January 22, 2020to April 7, 2020 [2]. Dr.Tandan, Dr.Acharya, Dr.Pokharel, Dr.Timilsina aimed to discover symptom patterns and overall symptom rules, including rules disaggregated by age, sex, chronic condition, andmortality status, among COVID-19 patients [12].

Methods

A Short Review of Covid-19 situations

- In early December 2019, an outbreak of coronavirus disease 2019 (COVID-19) caused by a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), occurred in Wuhan City, Hubei Province, China.
- On January 30, 2020 the World Health Organization declared the outbreak as a Public Health Emergency of International Concern (PHEIC).
- As of February 14, 2020, 49,053 laboratory-confirmed and 1,381 deaths have been reported globally.
- On March 2020, the Journal of the American Medical Association Ophthalmology reported that COVID-19 can be transmitted through the eye. One of the first warnings of the emergence of the SARS-CoV-2 virus came late in 2019 from a Chinese ophthalmologist, Li Wenliang, MD, who treated patients in Wuhan and later died at age 34 from COVID-19.
- On December 18, 2020, after demonstrating 94 percent efficacy, the NIH-Moderna vaccine was authorized by the U.S. Food and Drug Administration (FDA) for emergency use. Just days earlier, the similar Pfizer/BioNTech vaccine had become the first COVID-19 vaccine to beauthorized for use in the United States.
- In the late summer and fall of 2021, the delta variant was the dominate strain of COVID-19 inthe U.S.
- On 26 November 2021, WHO designated the variant B.1.1.529 a variant of concern, named Omicron.
- Director of the National Institute of Allergy and Infectious Diseases Anthony Fauci gave an update on the Omicron COVID-19 variant during the daily press briefing at the White House on December 1, 2021 in Washington, DC. He said that we will likely learn to live with COVID-19 like we do with the common cold and flu [10].
- Globally, as of 6:32pm CET, 27 January 2023, there have been 752,517,552 confirmed cases of COVID-19, including 6,804,491 deaths, reported to WHO. As of 24 January 2023, a total of13,156,047,747 vaccine doses have been administered.

Data Collection

The data for the ongoing Covid-19 outbreak in the United States is collected from the Centersfor

Disease Control and Prevention. The columns of this dataset include the Total number of weekly cases, Weekly Death and Weekly tests volume of Covid-19 patients accumulating all the states, on a weekly basis from 29th Jan 2020 to 18th Jan 2023. The total cases per 100,000, allow for comparisons between areas with different population sizes.

Weekly data is difficult to work with because the seasonal period (the number of weeks in a year)is both large and non-integer, like stock prices, employment numbers, or other economic indicators. The average number of weeks in a year is 52.18. Most of the methods we have considered require the seasonal period to be an integer. Even if we approximate it by 52, most of the methods will not handle such a large seasonal period efficiently.

So far, many publications and researchers have considered relatively simple seasonal patterns, such as quarterly and monthly data. However, higher frequency time series often exhibit more com- plicated seasonal patterns. For example, daily data may have a weekly pattern as well as an annual pattern. Hourly data usually has three types of seasonality: a daily pattern, a weekly pattern, andan annual pattern. Even weekly data can be challenging to forecast as it typically has an annual pattern with seasonal period of $365.25/7 \approx 52.179$ on average.

Exponential smoothing model didn't seem applicable, and ARIMA modelling is poor working with high integer seasonal periods (e.g. days/weeks rather than months/quarters), and also struggleswith a non-integer seasonal period (i.e. 52 weeks some years, 53 weeks other years).

Advanced Forecasting Model: Dynamic Harmonic Regression (DHR)

There are several methods for incorporating seasonality into a forecasting model. One common approach is to use time-series models such as SARIMA (Seasonal Autoregressive Integrated Moving Average) or Seasonal Exponential Smoothing. These models can capture the seasonal patterns in the data and adjust the forecast accordingly.

The time series processes are usually all stationary processes, but many applied time series, particularly those arising from economic and business areas are non-stationary. With respect to the class of covariance stationary processes, non-stationary time series can occur in many different ways. They could have non-constant means μ_t, time-varying second moments, such as non-constant variance σ^2, or both of these properties [9].

When applied to Covid-19 data, taking the natural logarithm of the number of cases or deaths can help stabilize the variance of the data and make the trend more apparent, especially in the earlystages of the pandemic when the growth was exponential. This can also help identify if there are anyunderlying patterns or seasonality in the data. After applying the log transformation, the resulting data will have a more linear trend and a constant variance, which makes it easier to model using standard statistical techniques such as linear regression or ARIMA models.

Many models used in practice are of the simple ARIMA type, which has a long history and was formalized in Box and Jenkins [6]. ARIMA stands for Autoregressive Integrated Moving Average and an ARIMA(p; d; q) model for an observed series, and 'I' stands for integration; where p is orderof autoregression, d is order of differencing, q is order of moving average [5].

Since we are also taking into account the seasonal pattern even if it is weak, we should also examine the seasonal ARIMA process. This model is built by adding seasonal terms in the non- seasonal ARIMA model we mentioned before. One shorthand notation for the model is

$$ARIMA(p,d,q)(P,D,Q)_m \qquad (3.1)$$

{(p, d, q)} : non-seasonal part

(P, D, Q)m}: seasonal part.

P = seasonal AR order,

D = seasonal differencing,

Q = seasonal MA order

m: the number of observations before the next year starts; seasonal period [12].

The seasonal parts have term non-seasonal components with backshifts of the seasonal period. For instance, we take {ARIMA(p, d, q)(P, D, Q)m} model for weekly data (m=52). Without differencing operations, this process can be formally written as:

$$\Phi(B^m)\phi(B)(x_t-\mu) = \Theta B^m\theta(B)(w_t) \qquad (3.2)$$

A seasonal ARIMA model inc*{(p, d, q)}* : non-seasonal part operates both non-seasonal and seasonal factors in a multiplicativefashion.

The time series models in ARIMA model and Exponential Smoothing model allow for the inclusion of information from past observations of a series, but not for the inclusion of other informationthat may also be relevant. For example, the effects of holidays, competitor activity, changes in thelaw, the wider economy, or other external variables may explain some of the historical variation andmay lead to more accurate forecasts. On the other hand, the regression models allow for the inclusionof a lot of relevant information from predictor variables but do not allow for the subtle time series dynamics that can be handled with ARIMA models.

An alternative approach uses a dynamic harmonic regression model. Next, we tried to extendARIMA models in order to allow other information to be included in the models. Firstly, we consid-ered regression model

$$y_t = T_t + C_t + S_t + \epsilon_t \qquad (3.3)$$

The system composed by four components: trend (T), sustained cyclical (C) with period differentto the seasonality, seasonal (S) and white noise (ϵ_t) [9].

The measured values of y are the output (observations) series of a system of stochastic state space equations, which can then be broken down to allow for estimation of the four components.

So for such time series, we prefer a harmonic regression approach where the seasonal pattern is modelled using Fourier terms with short-term time series dynamics handled by an ARIMA error.

In the following example, the number of Fourier terms was selected by minimising the AICc. The order of the ARIMA model is also selected by minimising the AICc although that is done within the auto.arima() function in R.

Dynamic harmonic regression is based on the principal that a combination of sine and cosine functions can approximate any periodic function.

$$y_t = b_t + \sum_{j=1}^{K}[\alpha_j sin\left(\frac{2\pi jt}{m}\right) + \beta_j cos\left(\frac{2\pi jt}{m}\right)] + \eta_t \qquad (3.4)$$

Where m is the seasonal period, α_j and β_j are regression coefficients, and η_t is modeled as a non-seasonal ARIMA process.

The fitted model has 18 pairs of Fourier terms and can be written as

$$y_t = b_t + \sum_{j=1}^{18}[\alpha_j sin\left(\frac{2\pi jt}{52.18}\right) + \beta_j cos\left(\frac{2\pi jt}{52.18}\right)] + \eta_t \tag{3.5}$$

Where $\eta_t$is an ARIMA(4,1,1) process. Because n_t is non-stationary, the model is actually esti- mated on the differences of the variables on both sides of this equation. There are 36 parameters tocapture the seasonality which is rather a lot but apparently required according to the AICc selection.The total number of degrees of freedom is 42 (the other six coming from the 4 AR parameters, 1 MAparameter, and the drift parameter)[4].

The advantages of this approach are :

Flexibility: DHR model can be used to model data with various levels of complexity, including data with multiple seasonal patterns, irregular patterns, and non-stationary patterns. It allows

any length seasonality; The short-term dynamics are easily handled with a simple ARIMA error. Especially, for data with more than one seasonal period, Fourier terms of different frequenciescan be included;

The smoothness of the seasonal pattern can be controlled by K, the number of Fourier sin andcos pairs – the seasonal pattern is smoother for smaller values of K ;

The only real disadvantage (compared to a seasonal ARIMA model) is that the seasonality is assumed to be fixed - the seasonal pattern is not allowed to change over time. But in practice, seasonality is usually remarkably constant so this is not a big disadvantage except for long time series.

Main Results

Forecasting Accuracy

Time series analysis and forecasting are an active research area over the last five decades. Thus, various kinds of forecasting models have been developed and researchers have relied on statistical techniques to predict time series data. The accuracy of time series forecasting is fundamental to many decisions processes, and hence the research for improving the performance of forecasting mod- els has never been stopped. However, the time series datasets are often nonlinear and irregular [3].An interdisciplinary approach afforded in the study of Data Science critically analyzes the relevant disciplinary insights and attempts to produce a more comprehensive understanding or purpose of a holistic solution.

The author measured forecasting performance by the mean absolute error (MAE), root mean square error (RMSE), root relative squared error (RSE), and mean absolute percentage error (MAPE).The MAE criterion is most appropriate when the cost of a forecast error rises proportionally with respect to the absolute size of the error. With RMSE, the cost of the error rises as the square of the error, and so large errors can be weighted far more than proportionally. Whether MAE or RMSEis most appropriate surely varies according to circumstances and individual institutions, and in any case we will find that the several measures pick the same model in all but several instances [8].

These measures were calculated by using the following Equations. P_t is the predicted value at time t, Z_t is the observed value at time t and N is the number of predictions.

$$ME = \frac{\sum_{i=1}^{N}(P_t - Z_t)}{N} \tag{4.1}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_t - Z_t| \tag{4.2}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}|\frac{P_t - Z_t}{Z_t}| \tag{4.3}$$

$$MPE = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Z_t - P_t}{Z_t}\right) \times 100\% \tag{4.4}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{N}(P_t - Z_t)^2}{N}} \tag{4.5}$$

$$AIC = -2ln(L) + 2k \tag{4.6}$$

$$AICc = AIC - \frac{2k(k+1)}{n-k-1} \tag{4.7}$$

where k is the number of parameters and n the number of samples.

It is important to note that these information criteria tend not to be good guides to selecting theappropriate order of differencing (d) of a model, but only for selecting the values of p and q. Thisis because the differencing changes the data on which the likelihood is computed, making the AIC values between models with different orders of differencing not comparable [4].

Conclusion

In this section, the focus is on statistical methodology and forecasting results on time series datasets regarding Covid-19 pandemic. The comparison table 1 below shows all the potential forecast-ing models. A given forecasting model may have a systematic positive or negative bias and do a poor job of tracking the actual mean of value changes, and measures such as RMSE and MAE could well miss this defect. Obviously, the Log Transformation DHR perform best among other models. Because we evaluated the different models with different criterion. The Log Transformation DHR minimize the RMSE, MAE and shows relatively better forecasting accuracy.

Collectively, these models are capable of identification of learning parameters that affect dissimilarities in COVID-19 spread across various regions or populations, combining numerous interventionmethods and implementing what-if scenarios by integrating data from diseases having analogous trends with COVID-19 pandemic [5]. (Figure 1)

As it was the case with the forecast in Table 2 and Table 3, the number of weekly cases and weekly deaths are projected to continue increase in the following weeks. It shows the noticeable increase in the future. However, weekly cases will decrease at the end of May 2023. However, the weekly deaths forecasting results shows the uncertainty and fluctuations until the end of 2023. The DHR show the smallest RMSE. Because it is a better model than $ARIMA(p, d, q)(P, D, Q)_m$ and dynamic harmonic

regression with ARIMA error. We can easily confirm from the above results that the transformation improves the accuracy if the time series have an unstabilized variance. It also shows that when there are long seasonal periods, a dynamic regression with Fourier terms is often better than other models we have considered from the raw datasets.

Table 1.Comparison Table for forecasting model

Model	ME	RMSE	MAE	MPE	MAPE	MASE	AICc
DHR with ARIMA(2,0,1) error	8447.324	148729.5	92906.71	43052.44	48766.5	0.1582	-18.38
ARIMA(2,1,0)(0,1,0)[52]	-4511.181	132336.8	57721.63	-3.0858	8.7082	0.0983	1711.99
Dynamic Regression	16520.74	162314.7	94507.58	0.1878	19.2053	0.1609	3105.5
with ARIMA (2,1,3)error							
Log transformation	0.00654	0.25964	0.18395	0.26225	1.8929	0.10279	-419.08
ARIMA (1,1,5)(0,0,1)[52]							
Log transformation DHR	0.01285	0.1753	0.13024	0.34485	1.4169	0.0728	15.88

Table 2. Forecasting results for weekly cases from regression with ARIMA (3,1,1) errors

Date	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2023.01.04	11.84703	2.16397924	21.53008	-2.9619173	26.65597
2023.01.11	11.67934	1.86601883	21.49266	-3.3288382	26.68751
2023.01.18	11.39147	1.44959306	21.33336	-3.813321	26.59627
2023.01.25	11.09728	1.02847775	21.16608	-4.3016243	26.49618
2023.02.01	11.01559	0.821447	21.20973	-4.5750067	26.60619
2023.02.08	11.27106	0.95309604	21.58902	-4.5089033	27.05102
2023.02.15	11.77707	1.33675702	22.21738	-4.1900106	27.74415
2023.02.22	12.34798	1.7867302	22.90922	-3.8040555	28.50001
2023.03.01	12.83167	2.15085746	23.51248	-3.5032215	29.16656
2023.03.08	13.12814	2.32908592	23.92719	-3.3875856	29.64386
2023.03.15	13.20719	2.29118114	24.1232	-3.487405	29.90179
2023.03.22	13.14645	2.11472479	24.17818	-3.7251195	30.01803
2023.03.29	13.05819	1.91194524	24.20444	-3.9885213	30.10491
2023.04.05	12.95955	1.69995251	24.21915	-4.2605198	30.17963
2023.04.12	12.79333	1.42150431	24.16515	-4.5983756	30.18503
2023.04.19	12.55773	1.07477713	24.04068	-5.0039298	30.11939
2023.04.26	12.31002	0.71700654	23.90303	-5.4199636	30.04
2023.05.03	12.06197	0.35992833	23.76401	-5.834757	29.95869
2023.05.10	11.79296	-0.01709568	23.60302	-6.2689634	29.85489
2023.05.17	11.55598	-0.36111708	23.47308	-6.669649	29.78162
2023.05.24	11.44662	-0.57657226	23.4698	-6.9412638	29.8345
2023.05.31	11.46867	-0.65967812	23.59702	-7.0800382	30.01738

Table 3. Forecasting results for weekly deaths with regression with ARIMA (4,0,1) errors

Date	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2023.01.04	7.881919	5.361387	10.402452	4.027098	11.736741
2023.01.11	7.231386	4.707106	9.755666	3.370833	11.091939
2023.01.18	7.014583	4.49027	9.538896	3.153979	10.875187
2023.01.25	7.316785	4.790167	9.843403	3.452656	11.180913
2023.02.01	7.972997	5.438274	10.50772	4.096473	11.849521
2023.02.08	8.628184	6.080713	11.175655	4.732163	12.524205
2023.02.15	8.983049	6.422724	11.543374	5.067369	12.898729
2023.02.22	8.973543	6.404637	11.54245	5.04474	12.902347
2023.03.01	8.738027	6.166017	11.310036	4.804477	12.671576
2023.03.08	8.455716	5.883601	11.02783	4.522006	12.389426
2023.03.15	8.228145	5.654814	10.801476	4.292575	12.163715
2023.03.22	8.086148	5.50775	10.664546	4.142829	12.029467
2023.03.29	8.056633	5.469725	10.643541	4.100298	12.012967
2023.04.05	8.171459	5.575543	10.767374	4.201348	12.141569
2023.04.12	8.408955	5.806693	11.011218	4.429138	12.388773
2023.04.19	8.675691	6.070895	11.280486	4.691999	12.659382
2023.04.26	8.875202	6.270236	11.480169	4.89125	12.859154
2023.05.03	8.969148	6.363566	11.57473	4.984254	12.954042
2023.05.10	8.95649	6.347756	11.565223	4.966776	12.946203
2023.05.17	8.833789	6.21938	11.448199	4.835395	12.832183
2023.05.24	8.605682	5.984966	11.226399	4.597643	12.613722
2023.05.31	8.304007	5.678611	10.929403	4.28881	12.319204

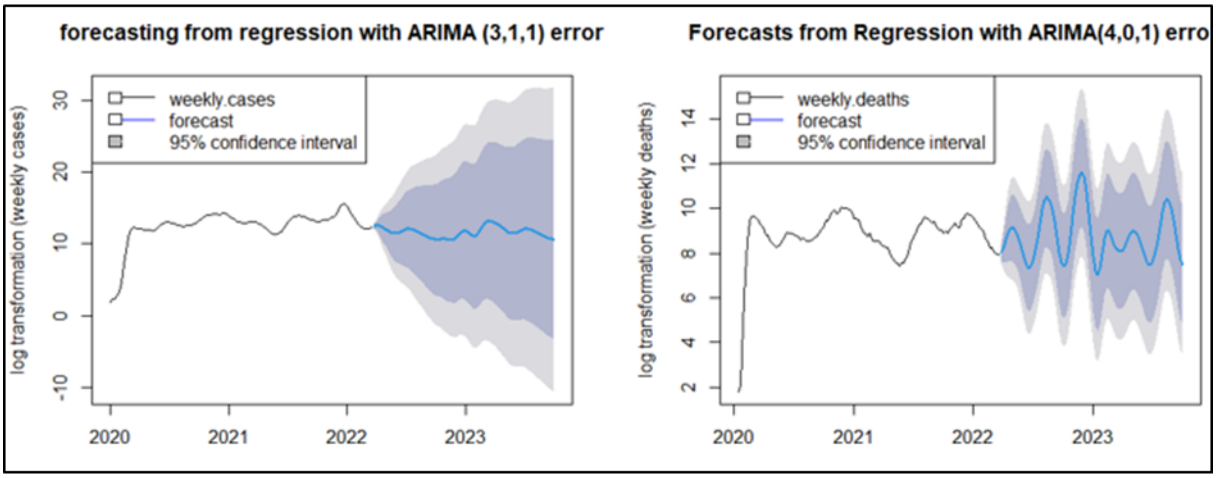

Figure 1. Forecasting results

The trend analysis shows unstable situation in the infected cases and weekly deaths and predic- tion study shows increase in the expected active and death cases nationally. However, the time series datasets are often nonlinear and irregular. This data has been used by researchers, policymakers, and others to better understand and respond to the effects of the pandemic.

The objective in providing crucial statistical techniques is to enable government and public to make informed decisions regarding Covid-19. Most importantly, we obtain how to add value to public health and apply skills in a real world environment. These models are essential for informing public health decision-making and resource allocation, as well as for predicting future trends in the spreadof the disease.

Acknowledgements

The author would like to thank some comments and constructive suggestions from Dr.Olusegun Michael Otunuga from the college of Science and Math and Dr.Hinton Romana from Writing Centerin Augusta University. Several stimulating discussions and comments allowed me to develop originalideas and improve my paper.

References

1. Naresh K; Seba S, (2020), COVID-19 Pandemic Prediction using Time Series Forecasting Models. The 11th ICCCNT 2020 conference
2. Saud S; Jaini G; Aishita J; Sunny A; Sagar J; Mani.R E (2021).Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting. 28-29 January 2021, 11th International Conference on Cloud Computing, Data Science & En- gineering.
3. Fotios P, Spyros M (2020). Forecasting the novel coronavirus COVID-19.Plos One 15(3): e0231236.https://doi.org/10.1371/journal.pone.0231236
4. Hyndman, R. J and Athanasopoulos G, (2014). Forecasting: Principles and Practice, OTexts, 2nd edition, ISBN 978-0-9875071-0-5.
5. RATNADIP A, (2013). An Introductory Study on Time Series Modeling and Forecasting , LAP Lambert Academic Pub-lishing, ISBN 10: 3659335088.
6. Box G. and Jenkins G, (1970) Time Series Analysis: Forecasting and Control, Holden-Day, San Francisco.
7. [Faraway, J. J., (2014). Linear Models with R, CRC Press, Taylor and Francis Group.
8. Brockwell P.J and Davis R.A, (2002). Introduction to Time Series and Forecasting, Second Edition, Springer, New York.
9. David A. M, Wlodzimierz T, (2019). Dynamic harmonic regression and irregular sampling; avoiding pre-processing and minimisingmodelling assumptions Environmental Modelling & Software Volume 121, November 2019, 104503