ISSN: 0974-5823 Vol. 6 No. 3 December, 2021

International Journal of Mechanical Engineering

Conventional Forecasting of COVID 19 Pandemic

M. Nirmala, D. Angel, and M. Mallika

Department of Mathematics, Sathyabama Institute of Science and Technology, Chennai, India

Abstract

The novel corona virus COVID-19 is an ongoing threat worldwide. The World Health Organization (WHO) declared the COVID-19 caused by Severe Acute Respiratory Syndrome Corona Virus 2 (SARS CoV 2) as a pandemic on 11th March 2020. Originally, the outbreak of this infectious disease was identified in Wuhan, China in December 2019. Present study analyses the spread and aims to forecast the cumulative number of confirmed reported cases of the ongoing pandemic by applying the classical model, Auto Regressive Integrated Moving Average (ARIMA). As a case study, a dataset consisting of everyday report of confirmed cases for the 38 districts of the state, Tamilnadu, India was obtained for a period of 18th March to 5th June, 2020, from the department of Health and Family welfare, Tamilnadu state, India. The ARIMA model was identified by considering the diagnostic test for the most affected district Chennai and its neighbouring districts.

Keywords : *COVID 19, Forecasting, ARIMA, Mean Absolute Percentage Error*

1.Introduction

Infectious diseases like SARS, Spanish flu, MERS CoV, HIV/AIDS, Ebola, swine flu were the main cause of mortality in the world and has been even more important historically. Worldwide concern about infectious disease peaked with the onset of the Human Immunodeficiency Virus (HIV) that causes Acquired Immune Deficiency Syndrome (AIDS), the most feared diseases of the 20th century. Each year, millions of people worldwide die from infectious diseases such as measles, malaria, tuberculosis, HIV. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes. While there are many complicating factors, simple mathematical models can provide much insight into the dynamics of disease epidemics and help the officials to make decisions about public health policy. The practical use of epidemic models must rely heavily on the realism put into the models. Even the simple models will exhibit the about the underlying mechanisms of infection spread and possible means of control of the disease or epidemic.

2. Literature Survey

The human to the human spreading of the virus occurs due to close contact with an infected person, with the symptoms of coughing, sneezing, respiratory droplets or aerosols. These aerosols can penetrate into the lungs of the human body through the nose or mouth. In more severe cases, infection can cause pneumonia, severe acute respiratory syndrome, kidney failure, and death. According to current data, time from exposure to onset of symptoms is usually between two and 14 days, with an average of five days. The novel coronavirus originated from the seafood market at Wuhan, China where living things like bats, snakes, raccoon dogs, palm civets, and other animals are sold. Today it has been spread all over the world. There are no promising remedial medicines till now against human coronaviruses. However, the researchers are working to develop efficient therapeutic strategies to cope with the novel coronaviruses.

The best model for the number of COVID-19 cases in the countries, Turkey, Germany, Italy, Japan, Canada. Russia, United Kingdom and France (G8 countries) were identified [1] by applying ARIMA, Holt-Winters smoothing method and Exponential smoothing method apart from curve estimation models. The authors [2] have applied SEIQDR based method for estimation, which is different from the traditional SEIR method for a dataset which consists of one-month cumulative number of deaths, suspected number of cases, recovery people, death and in quarantine of Mainland, China. Apart from the mathematical modeling they have applied the time series analysis methods like exponential smoothing method, ARIMA and ARIMAX. The results were compared using R^2 statistic value. The authors [3] have forecasted by splitting the dataset into five groups and considering small groups of datasets using multiplicative exponential smoothing trend in Mainland, China made forecast. There is no remedial medicine for treating COVID 19 till data. The main objective of this research article is in modeling infectious diseases, where the major means of disease spread comes from the person-to-person interaction.

3.Study Area and Materials

Tamilnadu is the 10th largest state with an area 130058 sq. km. among the 28 states in the country, India with capital, Chennai. It stands 4th place in the country in population [4]. There are 38 districts in Tamilnadu and the most affected district is Chennai. The first case in the state Tamilnadu was reported on 7th March 2020. The Department of Health and Family Welfare has confirmed a total of 23,495 cases, including 184 deaths and 13,170 recoveries, as of 1 June 2020 [5].

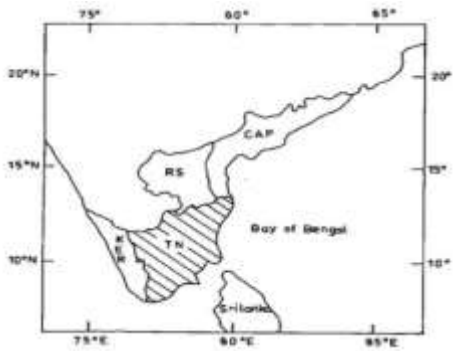

Figure 1 Geographical Location of Tamilnadu

Tamil Nadu has the second highest number of confirmed cases in India after Maharashtra. The state has 72 testing centres approved by Indian Council of Medical Research [6]. As on 5th June, the total number of confirmed cases in the state is 28694, the total number of active cases is 12697, the total number of deaths is 232 and total number of recovered cases is 15762. As per the Health Department, 88% of the patients are asymptomatic while 84% of deaths were among those with co-morbidities [7].

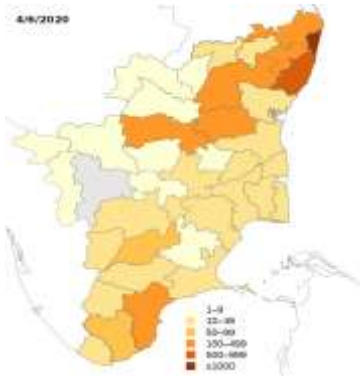

Figure 2 Map showing the districts of Tamilnadu with the number of affected cases

Table 1 Zonal classification of the Districts

Zone	Definition	Districts
Red	More than 80% of the cases in the state or with doubling rate less than four days	Chennai, Madurai, Villupuram, Ariyalur, Chengalpattu, Tiruvallur, Cuddalore, Perambalur, Tiruvannamalai, Tiruvarur, Ranipet, Virughunagar, Vellore, Kancheepuram
Orange	Districts without new cases in the last 14 days	Coimbatore, Tiruppur, Theni, Tenkasi, Nagapattinam, Thanjavur, Dindigul, Dharmapuri, Namakkal, Karur, Thoothukudi, Tiruchirappalli, Thirupathur, Kanyakumari, Ramanathapuram, Tirunelveli, Nilgris, Sivaganga, Pudukottai, Kallakurichi, Krishnagiri, Salem
Green	Non infected districts or no new cases in the last 28 days	Erode

Table 1 gives the zonal classification of the districts of Tamilnadu as per the criteria of Ministry of Health and Family welfare, Tamilnadu.

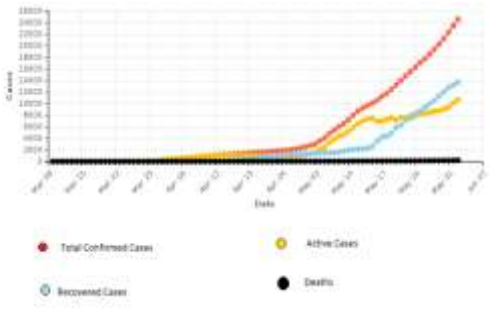

Figure 3 Time series plot showing the total confirmed cases, active cases, recovered cases and deaths.

4. Methodology

4.1 Auto Regressive Integrated Moving Average (ARIMA)

One of the advanced techniques in times series forecasting models is Auto Regressive Integrated Moving Average (ARIMA). This classical technique is based on autocorrelations in the data. This model is obtained by combining the differencing with autoregression and a moving average model. The ARIMA model of order (p, d, q) where p is the order of the autoregressive part, d is the order of differencing and q is the order of moving-average process is given by

$$\phi(B)(1-B)^d(Z_t-\mu)=\theta(B)a_t \qquad (1)$$

where Z_t denoted the observed value at time t, B represents the backward shift operator.

$\phi(B) and \theta(B)$ have the following representations:

$$\phi(B)=1-\phi_1 B-\phi_2 B^2-\ldots\ldots\ldots\ldots-\phi_p B^p \qquad (2)$$

$$\theta(B)=1-\theta_1 B-\theta_2 B^2-\ldots\ldots\ldots\ldots-\theta_q B^q \qquad (3)$$

The ARIMA model identification is made using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. In this study, the 14 red zone districts are considered. The autoregressive model and the moving average model are efficiently coupled to form a general and useful class of time series model called ARMA. This class of models can be extended as ARIMA model for non-stationary time series by differencing the data series [8]. Non-stationary data are unpredictable and cannot be modelled or forecasted. Based on the ACF and PACF plot, the general characteristics of the various models can be obtained by the following guidelines.

If the series is non-stationary, then the ACF plot remains significant for 6 or more lags instead of declining to zero quickly. In that case, the series must be differenced until it becomes stationary. Exponentially declining ACF and spikes in the first one or more lags in PACF indicate autoregressive model. Spikes in the first one or more lags in ACF and exponentially declining PACF specifies moving average

model. Exponential decline in both ACF and PACF stipulates mixed model namely autoregressive moving average model.

In most time series problems, data will be non-stationary. It will be converted into stationary by doing differences by an order of integration parameter 'd'. The first differences are given by the equation,

$$(z_t - z_{t-1}) = (1-B)z_t \quad (4)$$

The second differences, which are obtained by finding the differences of first differences, are given by the equation,

$$(z_t - 2z_{t-1} + z_{t-2}) = (1-B)^2 z_t \quad (5)$$

There are some general rules for detecting the values of 'p' and 'q' from the ACF and PACF plots which determines the order of the ARIMA model.

A gradual decay in the ACF plot and a sharp cut off in the PACF plot at lag 'k' suggests that the model is an autoregressive model of order 'k'. A gradual decay in the PACF plot and a sharp cut off in the ACF plot at lag 'k' implies that the model is a moving average model of order 'k'. A gradual decay in both ACF and PACF plots which starts after k_1^{th} and k_2^{th} lag for ACF and PACF respectively indicates that the model is an autoregressive moving average model of order (k_1, k_2).

Having made a good guess of the correct values of 'p' and 'q' using error measures, the correct model is fitted with these values. If the model is correct then the ACF and PACF plots of the residuals of the correctly identified model will not have any significant spikes (spikes are the vertical lines that extend beyond the horizontal dotted lines in the ACF and PACF plots) at the first few lags. After having found the correct model, it is fitted to a time series and has to be checked whether the model provides an adequate description of the data. If the values of 'p' and 'q' do not satisfy the stationarity conditions, then one has to identify a new model for which the new parameters are estimated and tested. This process of finding the parameters, testing the parameters and then finalizing the parameters will make up the required model [9]. Once the model is chosen forecast is done.

4.2 Performance Evaluation Criteria

To assess the quality of forecasting and to evaluate the consistency of the model, error measures are required. In this research work, the error measures Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used. The formula for the error measures is given below:

$$MAPE = \frac{1}{n}\Sigma\left|\frac{Actual - Forecast}{Actual}\right| * 100 \quad (6)$$

$$MAE = \frac{1}{n}\Sigma|Actual - Forecast| \quad (7)$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma(Actual - Forecast)^2} \quad (8)$$

5. Results and Discussions

The district Chennai has more than the half of active number of COVID 19 patients when compared to the remaining districts in Tamilnadu. This information gives the basic idea of this research article. The district Chennai, which is most affected district together with its neighbouring districts, Chengalpattu and Kancheepuram are considered for the analysis. For the identification of ARIMA model, the first preprocessing step is to check for stationarity using the AutoCorrelation function plot. The ACF plots fluctuate around an average value and tend to zero rapidly, the dataseries are stationary through the visual inspection of the plots for all the three districts. The ACF plots (Fig.4) after first order differencing and natural log transformation and the Partial AutoCorrelation function (PACF) (Fig.5) for the zones, Chennai, Chengalpattu, and Kancheepuram are plotted below for the identification of the order p and q of ARIMA model process.

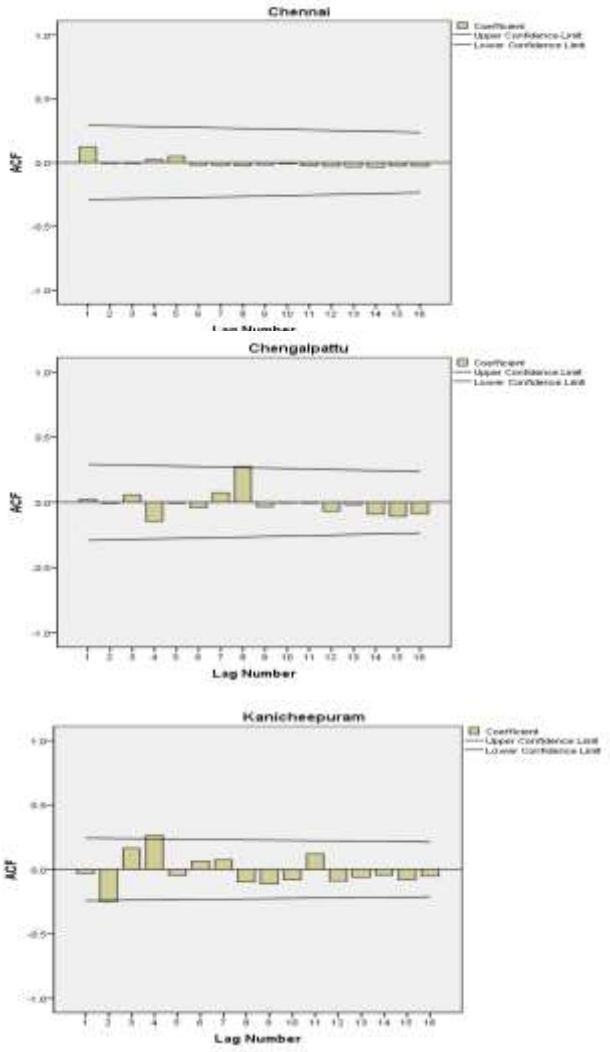

Figure 4 ACF plots of Chennai and its neighbouring districts, Chengalpattu and Kancheepuram

4.Main Text

Type your main text in 10-point Times New Roman, single-spaced and justified. Do not use double-spacing.

Figure and Table captions

Figure and Table captions should be 10-point times new roman and justified. The First words are in capital and all other words are in small letters. Figures and

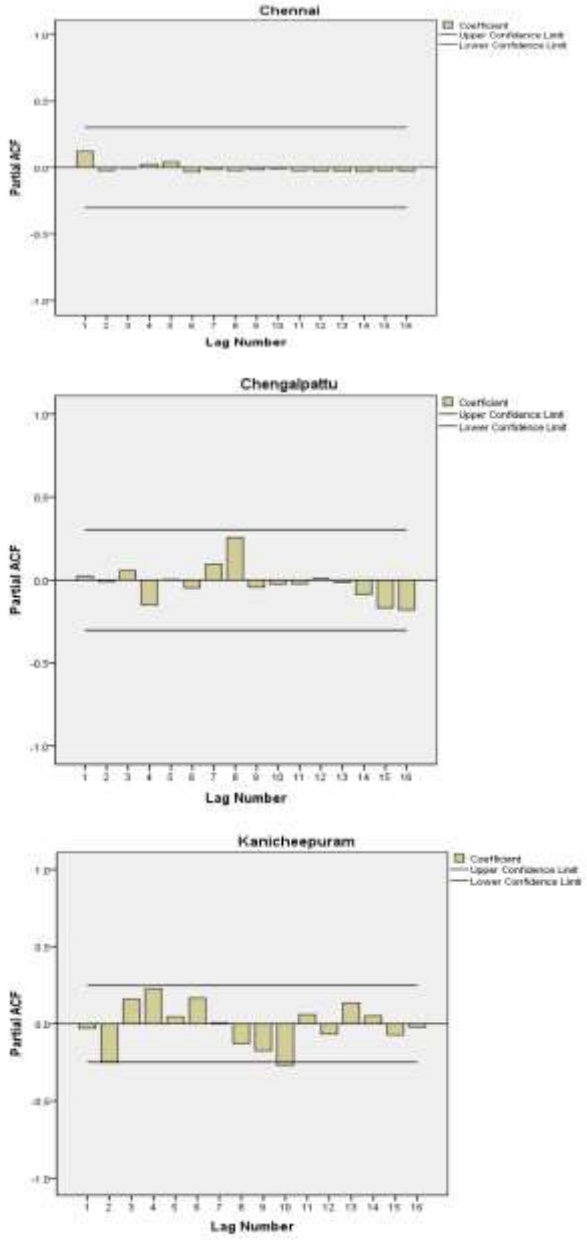

Figure 5 PACF plots of Chennai and its neighbouring districts, Chengalpattu and Kancheepuram

Table 2 ACF and PACF values for the first 6 lags for Chennai

Lag	Auto correlation	Std. Error	Box-Ljung statistic			Partial Auto correlation	Std. Error
			Value	df	Significant value		
1	0.124	0.146	0.719	1	0.397	0.124	0.151
2	-0.005	0.144	0.720	2	0.698	-0.021	0.151
3	-0.006	0.142	0.722	3	0.868	-0.003	0.151
4	0.024	0.141	0.751	4	0.945	0.025	0.151
5	0.049	0.139	0.873	5	0.972	0.043	0.151
6	-0.019	0.137	0.892	6	0.989	-0.030	0.151

Table 3 ACF and PACF values for the first 6 lags for Chengalpattu

Lag	Auto correlation	Std. Error	Box-Ljung statistic			Partial Auto correlation	Std. Error
			Value	Df	Significant value		
1	0.023	0.146	0.024	1	0.877	0.023	0.146
2	-0.007	0.144	0.027	2	0.987	-0.007	0.144
3	0.057	0.142	0.185	3	0.980	0.057	0.142
4	-0.146	0.141	1.267	4	0.867	-0.146	0.141
5	-0.006	0.139	1.269	5	0.938	-0.006	0.139
6	-0.040	0.137	1.353	6	0.969	-0.040	0.137

Lag	Auto correlation	Std. Error	Box-Ljung statistic			Partial Auto correlation	Std. Error
			Value	Df	Significant value		
1	-0.029	0.121	0.059	1	0.809	-0.029	0.124
2	-0.250	0.120	4.376	2	0.112	-0.251	0.124
3	0.165	0.119	6.286	3	0.098	0.158	0.124
4	0.264	0.118	11.251	4	0.024	0.226	0.124
5	-0.046	0.117	11.406	5	0.044	0.046	0.124
6	0.064	0.116	11.707	6	0.069	0.167	0.124

The Box – Ljung Statistic test is used to test whether any autocorrelation at different lags is different from zero. This test is mainly applied to test the overall randomness at different lags. Here the test statistic values shows that the data are not random and independently distributed.

Table 5 Ljung Box Q Statistic for the three districts

	Ljung Box Q Statisitic		
District	**Statistics**	**DF**	**Significant value**
Chennai	12.835	12	0.685
Chengalpattu	9.845	10	0.454
Kancheepuram	6.193	10	0.799

Based on the ACF and PACF plots, the tentative ARIMA model for forecasting the number of active COVID 19 cases in the three districts are identified as ARIMA (1,1,1), ARIMA (4,1,4) and ARIMA (4,1,4) for the three districts, Chennai, Chengalpattu and Kancheepuram respectively. Based on the Ljung Box Q statistic and the error measures, the ARIMA model gives better accuracy for the district Kancheepuram when compared to the other two districts.

Table 6 Error Measures obtained for ARIMA model for the three districts

		Error Measures		
District	**Model**	**MAPE**	**MAE**	**R Squared**
Chennai	ARIMA (1,1,1)	9.854	6.747	0.999
Chengalpattu	ARIMA (4,1,4)	11.298	9.387	0.986
Kancheepuram	ARIMA (4,1,4)	8.082	4.635	0.997

From the table 6, of error measures, the ARIMA model gives better accuracy for the district Kancheepuram when compared to the other two districts.

The district Chennai and its neighbouring districts, Chengalpattu and Kancheepuram are considered for forecasting the COVID 19 affected number of cases through conventional modeling Auto Regressive Integrated Moving Average (ARIMA) and the results obtained are discussed. The figure 6. shows that the model gives better accuracy for the district Kancheepuram.

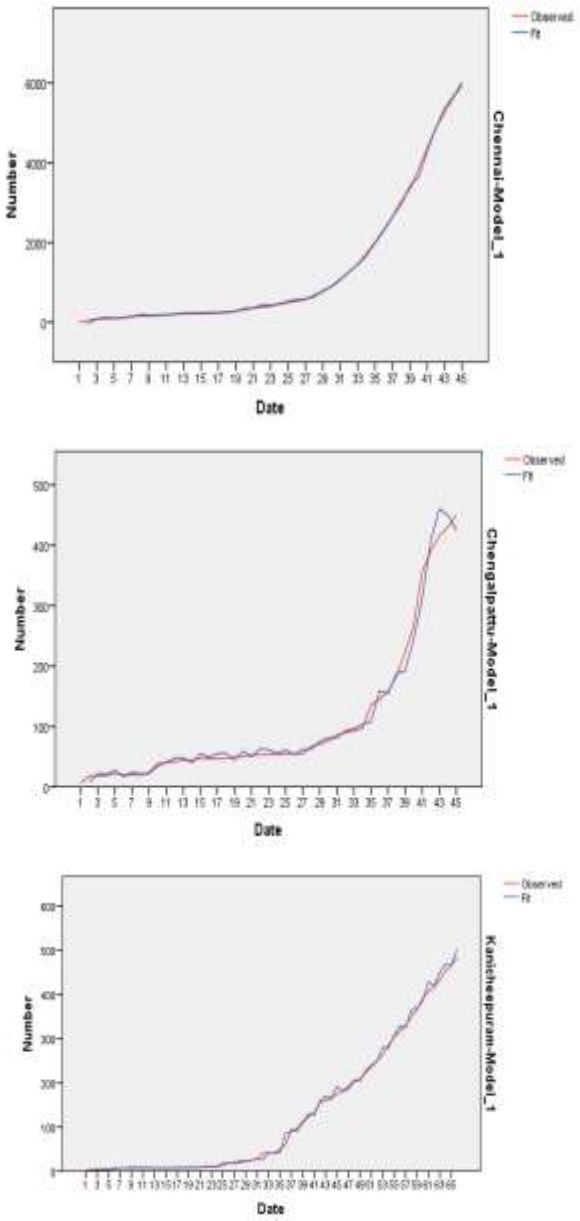

Figure 6 Observed vs Predicted plots of Chennai and its neighbouring districts, Chengalpattu and Kancheepuram

6. Conclusion

COVID 19 is an infectious disease caused by newly discovered Corona virus. There are no specific vaccines or treatments for COVID 19 till now. But there are many ongoing clinical trials evaluating potential treatments. The spread of Corona virus in a particular geographical location is based on many factors and one among them is densely populated area. Since the city Chennai and its neighbouring districts Chengalpattu and Kancheepuram are densely populated, these districts come under the red zone according to the Health and Family Welfare of the state, Tamilnadu. In this research article, forecasting models were identified for the cumulative number of active cases using the ARIMA model, based on Box-Jenkins. The results and discussions were made based on the diagnostic tools and performance accuracy criterion, MAPE. These models will play a major in preventing and controlling the spread of the disease. Classical and advanced sstatistical models are essential for understanding the relation between social and

biological mechanisms that influence the spread of COVID 19. These models can be used in the cost effectiveness evaluation of various control measures.

Conflicts of interest

The authors have no conflicts of interest to declare.

References

1. Yonar H, Yonar A, Tekindal MA, Tekindal M., Modeling and Forecasting for the number of cases of the COVID-19 pandemic with the Curve Estimation Models, the Box-Jenkins and Exponential Smoothing Methods. EJMO, 2020;4(2):160–165.

2. Li Y, Wang B, Peng R, Zhou C, Zhan Y,Liu Z et al., Mathematical Modeling and Epidemic prediction of COVID – 19 and its significance to Epidemic Prevention and Control Measures, Ann Infect Dis Epidemiol, 2020; 5(1):1052.

3. Petropoulos F, Makridakis S (2020) Forecasting the novel coronavirus COVID-19. PloS ONE 15(3): e0231236.

4. https://censusindia.gov.in/Census_And_You/area_and_population.aspx

5. https://en.wikipedia.org/wiki/COVID19_pandemic_in_Tamil_Nadu

6. https://www.mygov.in/covid-19

7. https://stopcorona.tn.gov.dhcp.in/daily-bulletin/

8. George E. P. Box, Gwilym M. Jenkins and Gregory C. Reinsel, Time Series Analysis – Forecasting and Control, 3rd edition, Pearson Education, Inc (1994).

9. Spyros Makridakis, Steven C. Wheelwright, and Victor E. McGee. Forecasting – Methods and Applications, 2nd edition, Wiley, 1983.