A Framework for Healthcare Management System Using Predictive Analysis of Diseases

Author

Khulood Nakhat

Regn Number

00000277633

Supervisor

Brig. Dr. Shoab Ahmed Khan

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

NOVEMBER, 2020

A Framework for Healthcare Management System Using Predictive Analysis of Diseases

Author

Khulood Nakhat

Regn Number

00000277633

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Computer Engineering

Thesis Supervisor:

Brig. Dr. Shoab Ahmed Khan

Thesis Supervisor's Signature:_____

DEPARTMENT OF COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

NOVEMBER, 2020

Declaration

I certify that this research work titled "*A Framework for Healthcare Management System Using Predictive Analysis of Diseases"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

_____

Signature of Student

(Khulood Nakhat)

2018-NUST-Ms-Comp-00000277633

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

_____

Signature of Student
(Khulood Nakhat)
Registration Number
00000277633

_____

Signature of Supervisor
(Brig. Dr. Shoab Ahmed Khan)

Copyright Statement

Acknowledgements

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Your will, so indeed none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life.

I would also like to express special thanks to my supervisor Brig. Dr. Shoab Ahmed Khan for his help throughout my thesis and also for Statistical Signal Modeling in Digital Signal Processing course which he has taught me. I can safely say that I haven't learned any other engineering domain in such depth than the ones which he has taught.

I would also like to pay special thanks to Miss Fatima Khalique for her tremendous support and cooperation. Each time I got stuck in something, she came up with the solution. Without her help I wouldn't have been able to complete my thesis. I appreciate her patience and guidance throughout the whole thesis.

I would also like to thank Dr. Ali Hassan and Dr. Urooj Fatima on my thesis guidance and evaluation committee and express my special Thanks to Miss Khansa Khan for her help.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

Dedicated to my exceptional parents and adored siblings whose tremendous support and cooperation led me to this wonderful accomplishment

Abstract

Disease outbreak detection is a major challenge in public health informatics. While big data in healthcare is constantly expanding, there still exists a need to effectively integrate and represent data in order to obtain useful information applicable in solving disease outbreak problems. This study presents a framework to analyze the disease outbreaks for a given population by performing predictive analytics on incidence data. This information is particularly useful for the decision-makers in the context of healthcare management to formulate intervention programs based on the results. None of the existing public health frameworks that support the integration of predictive analysis with decision making process for optimal resource planning and control. We present data acquisition and transmission framework with a predictive analytics on top to provide threshold based alert to decision-makers on disease incidence data. We use a temporal predictive Auto-Regressive Integrated Moving Averaging model (ARIMA) in combination with a minimum size moving window to forecast the disease incidences over a data collection and integration framework. We applied our model for predictive analysis of Hepatitis C incidences in Lahore and Vehari District of Punjab province in Pakistan. Model performance is evaluated based on Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The results of the analysis provide a sound reference for expanding capabilities of the disease management tools in healthcare management context.

Key Words: *Public health management; forecast; time series; ARIMA; predictive analysis; stochastic modeling*

Table of Contents

List of Figures

List of Tables

CHAPTER 1: INTRODUCTION

This chapter is about the introduction to our study and we divide it into four sections. The first section describes the background of predictive analytics of diseases and emphasizes the importance of its need. The second section contains the problem statement of this study and the third section describes the major contributions of this study. The fourth section describes the overall thesis structure.

1.1 Background and Motivation

Nowadays, healthcare is generating a diverse and rapidly growing massive amount of data. The variations and scales of such data, for example, medical and environmental data, etc. has increased the demand for smooth data access with the availability of reliable and efficient data analysis processes and services. On one hand, the availability of such a large amount of diverse data and on the other the availability of big data in health presents an opportunity to model an analytical framework that is able to extract information of interest from multiple underlying sources and perform analytics from a wide range of available techniques [1-3]. The healthcare data is obtained from various health sources including electronic health records, doctors' prescriptions, clinical reports, diagnostic reports, and medical images etc. This healthcare data is essential for analyzing disease prediction at the early stages of disease diagnosis and detection. Healthcare data analysis methods detect trends in data and facilitate appropriate treatment to improve people's lives in the early stages at a very low cost [4-7]. In the domain of healthcare, predictive analysis is crucial for early diagnosis and treatment and plays an important role in accurately diagnosing diseases, optimizing resource and cost allocation, enhancing patient care, and thus improving health outcomes [8]. It refers to the branch of analytics that extracts useful information from historical data to predict upcoming trends and outcomes and to make better decisions. It supports healthcare providers by applying a variety of techniques including modeling, data mining, machine learning, statistics, artificial intelligence and other related fields to investigate the present findings in order to predict certain future outcomes [9-10].

Time series analysis can be used to diagnose the management of multiple diseases and prediction is the most important aspect of this analysis. In this analysis, the properties of the data

are analyzed and useful information is obtained from it, and at the same time a predictive model is used which predicts the incidences of the upcoming days based on the past. It has been used for the past several decades by data analysts and statisticians. Different statistical time series models exist in the literature including exponential smoothing [11], fuzzy time series [12], linear regression [13], Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Moving Average (ARMA), Auto-Regressive Integrated Moving Average (ARIMA) [11], and correlation coefficient analysis [13]. Among the time series models, ARIMA is the only one that has a tendency to deal with changing trends, random disturbances, and periodic changes, which is why we use this model in our study for the purpose of disease incidence predictive analysis. In the ARIMA model, when all past disease incidences are used, then the structural changes that occur in the past in these incidences have an effect on the predictive incidences. The structural changes that appear in past disease incidences are not necessarily the same in the future. This has a bad effect on predicted incidences. That's why we've used the ARIMA model with a moving window for predictive analysis to overcome this so that we can move the window forward every time it's equal to the predicted incidences and reduce the impact of structural changes that have taken place in the past. We have also identified the minimum number of days of incidence to give a reference to the healthcare management system for the disease incidence so that the model can easily predict the incidence of the disease.

In public health context, predictive analysis of diseases analyzes the risk of disease outbreaks in a given population. The integration of this analysis with decision making tools will aid healthcare professionals to find the causes of the spread of the disease, allocate resources according to the expected number of patients, appoint more doctors in health centers, and conduct better targeted public awareness campaigns. There is, therefore, a need to define a framework within public health that integrates predictive analysis into the healthcare management system and facilitates the aforementioned features. A significant work has been done in field of predictive analysis given a considerable amount of data. The acquisition of the data is primarily dependent on surveillance programs that run independently of the analytical tools used to process information from raw data. For example, in many public health analytical dashboards, population or location based aggregates are used and raw point data is lost or stays unutilized. Therefore, there is a need to define an integrated approach that acquire, transmit and

analyze time-series data of public health interest in real time or near real time to allow detection of subtle changes as well as characterization of patterns for predictive analytics.

1.2 Problem Statement

None of the disease predictive analytics framework in healthcare is made in which predictive analysis is integrated with decision making process. ARIMA Model with a moving window is not used for disease incidence predictive analysis. This is the research gap that has been covered in this study.

1.3 Contributions

We present a novel framework called Disease Predictive Analysis (DPA) for public health management that integrates the processes from acquisition to analysis and informs public health officials about the population health status in order to make informed decisions. We focus on investigating period of measurement, also known as, window size in terms of number of incidences, for predictive analysis of diseases in a population. In particular, we use Auto-Regressive Integrated Moving Averaging Model (ARIMA) for predictive analysis with a moving window for potential resource, staff and policy planning. The identification of optimal window size allows predictive model applied to be valuable and effective. In order to validate the framework, we present a case study of Hepatitis C incidences consisting of Lahore and Vehari District in Punjab Province of Pakistan from 2015 to 2019 and decision makers in the Pakistan Ministry of Health.

The presented framework, in its implementation is unique in addition to the existing public healthcare frameworks through the following contributions

- It allows disease incidence data to be collected from multiple sources and integrated through a pathway provided by the framework
- It allows the predictive analysis of diseases in any given area by providing the required minimum number of incidences.
- It can be used for early warning triggers to alert decision makers during high and low risk periods when predictions cross a certain threshold level.

1.4 Thesis Structure

The rest of the thesis is organized as follows. Chapter 2 presents the related work on the existing frameworks of healthcare and the existing techniques. Chapter 3 presents the proposed framework in the context of the healthcare management system. This chapter also describes disease incidences of Lahore District of Punjab province in Pakistan as a case study for the implementation of our proposed framework. Chapter 4 presents the ARIMA model in detail. In chapter 5, we present the analysis of the results and discussion is presented in chapter 6, whereas, the conclusion and future work are presented in chapter 7.

CHAPTER 2: LITERATURE REVIEW

This chapter contains the literature review of our study and we divide it into three sections. The first section gives an overview of predictive modeling techniques and explains the reason for using the proposed modeling technique. The second section covers an overview of the existing healthcare frameworks and describes the gaps that exist in current literature. The third section describes the Healthcare Delivery System of Pakistan.

2.1 Predictive Modeling Techniques

Predictive modeling in healthcare is emphasized and focused after the evolution of EHRs, which leads to large scale production of a wide variety of data [14-16]. In the time series analysis, data sequences are recorded at the discrete-time intervals. A variety of statistical models have been used by researchers, including exponential smoothing method, fuzzy time series model, and grey model etc. In the exponential smoothing method, correlation is not found in successive time-series incidences. There is no correlation between the predicted errors and the spread is normally distributed with zero mean and constant variance. Practically the data is of non-stationary nature so AR, MA, and ARMA that deal with stationary nature data cannot be applied. There is a need to make non-stationary data stationary through some possible transformation. In this regard, Box Jenkins has created the ARIMA model and it is also capable of dealing with the univariate time series [11].The fuzzy algorithm provides a solution to any particular problem upon the execution of the defined fuzzy instructions. For the Taiwan's export values, the ARIMA model and the fuzzy-time series model have been compared, which suggests that the Fuzzy-time series model performs better for only a small number of data sample values [12].The ARIMA model and the grey model have been used to predict the mortality rate of diabetics, indicating that ARIMA performs better in dealing with more samples and correlative time series data [17]. Similarly, there has been significant research into the application of ARIMA to disease incidences including influenza [18], dengue outbreaks [19], Hepatitis-B [20], Hemorrhagic fever with renal syndrome [21], Inflammatory Bowel Disease [22], New Castle [23], Coronavirus disease (COVID-19) [24], and Tuberculosis [25] etc. When only the ARIMA model is used for prediction, some structural changes that occur in the past have an effect on predictions. To reduce the impact of these structural changes, a moving window is used with the

ARIMA model. Using the moving window, researchers have successfully developed predictive models, including smart meters [26], wind speed [27], vehicle speed [28], stock price [29], and multi-area power systems [30], etc.

2.2 Healthcare Frameworks

Several healthcare frameworks are proposed by researchers including [31-40]. To provide the ubiquitous healthcare services to each person during their exercise sessions smart healthcare framework based on Internet of Things (IoT) technology has been proposed. Real health-related conditions are also analyzed to forecast probabilistic health-related vulnerabilities, and layered architecture designs are made to meet different predefined tasks in a synchronized way [31]. Using a variety of emerging computing techniques, Smart Data Mining-based IoT (SMDIoT) framework for diabetes and cardiovascular patients in healthcare has been proposed, which provides treatment advice to patients with these diseases. Biosensors and IoT are used to collect and monitor patient data from time to time, and the sensitive patient information is obtained from sentiment analysis techniques and, ultimately, with the help of machine learning and data mining techniques aforementioned disease patients are classified from healthy patients [32]. To improve the digital technology in healthcare, the Secure Privacy Conserving Provable Data Possession (SPC-PDP) framework has been proposed. Simultaneous data storage, badge auditing, integrity, and dynamic data auditing is provided to conserve privacy in healthcare [33]. To secure the patients' EHR data in healthcare, a Sensitive and Energetic Access Control (SE-AC) framework has been proposed that facilitates access control in critical situations. Patient privacy is protected with the help of fast and secure encryption methods and in a very short time, the ciphertext EHR data is decrypted according to the identity of the applicant user and finally, permission is provided based on environmental conditions and the applicant's attributes [34]. Blockchain-based technology framework in healthcare has been developed to protect the confidentiality of patients' EHR data from disclosure that addresses access control challenges using blockchain immutability and built-in autonomy features. It uses an authorized blockchain that allows only authenticated and invited users to access data after verifying users' identities and cryptographic keys, and also guarantees accountability because all users are already verified [35]. To secure the patients' EHR data, a Keyless Signature Infrastructure with Blockchain technology (KSIBC) framework has been proposed in which patient data is provided with integrity, and authentication

is ensured [36]. The summary of the existing healthcare framework is presented in Table 2-1.

Table 2-1: Summary of the Healthcare Frameworks

Author	Year	Framework	Analytics Platform	Facilitation to Healthcare
Munish Bhatia and Sandeep K-Sood	2017	Healthcare Workout	Predictive	Provide ubiquitous healthcare services to person during his/her workout sessions
Sharan Srinivas and A.Ravi Ravindrun	2018	Framework for Optimizing Outpatient Appointment System	Prescriptive	Improve the performance of outpatient appointment system w.r.t patient satisfaction and resource utilization
Koumakis, L.etal	2017	Content Aware Analytics Framework	Predictive	Improve the precision in medicine at healthcare
M.Sharma etal	2018	Advance Conceptual Healthcare Diagnostic Framework for Diabetes and Cardiovascular Disorder	Sentiment	Collect and monitor the regular and periodic data of diabetic and heart patients
Mariadel Carmen Legaz-Garcia etal	2016	A semantic web based framework for interoperability of clinical models and EHR data	Not used	Leverages EHR and semantic web technologies for the interoperability and exploitation of archtypes, EHR data and ontologies
Our Proposed Framework	2020	Disease Predictive Analytics Framework	Predictive	Aids healthcare professionals to find the causes of the spread of the disease, to allocate resources according to the expected number of patients, to appoint more doctors in health centers, and to conduct better targeted public awareness campaigns

Using a collaborative filtering approach, a patient-centered healthcare framework has been proposed that achieves patient similarities and develops disease risk profiles for individuals. The individual patient's medical history is compared on the basis of similarity constraints determined by the medical history of all available patients. Based on these similarity constraints, a group of identical patients is selected and diseases are predicted [37]. To seamlessly integrate the EHR data and bridge the interoperability gap between clinical models and clinical records, a semantic web-based framework using Web Ontology Language (OWL) has been proposed. Patient data is retrieved from a relational database and finally, the resulting data is converted into an arch-type OWL to build an ontology so that the constructed ontology is used for data exploration [38]. Using machine learning algorithms, a prescriptive analytics framework has been proposed to improve patients' appointment system in healthcare that satisfies patients and makes better use of resources. Patient data is retrieved from EHR and publicly available datasets, and this data is refined through the detection and correction of missing values and finally, it is used as input to machine learning algorithms to predict patients' missed appointment risk [39]. To improve the precision in medicine at healthcare, a content-aware analytics framework has been proposed that consolidates healthcare data with analytics. Data is obtained from various health-related sources and it is seamlessly integrated with the help of semantics and standards and finally, this integrated information is anonymized and data analytics algorithms are applied to obtain statistical information [40].

To the best of our knowledge, no general framework has been proposed within healthcare that deals with predictive disease analysis and integrates predictive modeling of disease into decision making process to facilitate decision makers in the healthcare management system, and whose validation is carried out on a real-time data. This is the gap that exists in the current literature and has been covered in this research.

2.3 Healthcare Delivery System of Pakistan

Healthcare delivery system is a management to the population that provides better health services to any country with efficient and equitable allocation of resources, and distributions of funds over a systematized infrastructure to develop well. [41]. The healthcare sector is very crucial for any country as it directly influences on economies. If there would be better health there would be increased in the labor force as result productivity of country rises and economies

increase [42]. According to the constitution of Pakistan, it is stated: “to provide better health services to citizens is the obligation of the state". This obligation is shifted from the federal to the provincial governments after the 18th amendment [43, 44, 45]. The healthcare system of Pakistan is divided in to private, public and non-government health sector. Public health care sector comprises of federal and provincial governments, where the ministry of defense, ministry of Inter-Provincial Coordination (IPC), research institutes, ministry of health, and ministry of National Health Services Regulation and Coordination (NHSRC) lie under the control of federal governments and all provincial departments are under the responsibility of provincial governments. Primary, Secondary and Tertiary are the three levels of healthcare. Healthcare at the district level comprises primary and secondary that is surrounded through Basic Health Unit(BHUs), Maternal and Child Health Centers(MCHCs), Dispensaries, Rural Health Unit(RHUs), Tehsil Head Quarters (THQs), District Head Quarters (DHQs). All of the healthcare facilities are implemented at the tertiary level of healthcare where provincial governments directly handle it [46]. Data is aggregated at all levels of health care in the form of Electronic Health Records (EHR) [47].

CHAPTER 3: METHODOLOGY

This chapter presents our proposed Disease Predictive Analytics Framework. We divide this chapter into six sections and each section contains a one-layer description of this framework. The first section deals with the data collection layer, the second with the data cleansing layer, the third with the Predictive Analysis layer, the fourth with the Disease Threshold layer, the fifth with the Decision making process layer, and the sixth with the case study of validating the framework.

Disease Predictive Analytics Framework

Figure 3.1 shows the DPA framework layered architecture with different layers. In the data collection layer, data is collected from EHR systems installed in district hospitals. The data cleansing layer applies some operations for location separation of the disease incidences. The predictive analysis layer, predicts disease incidences based on disease incidence data selected from the data cleansing layer. The disease threshold layer compares defined disease thresholds with predictive incidences, and provide an alert to decision-makers of the healthcare management system in the decision making process layer.

The details about every layer of the framework is shown below:

3.1 Data Collection Layer

In this layer, data have collected from EHR systems installed in district hospitals at the provincial level. The information collected in this data includes the patient's hospital arrival date, his diagnosed disease, and the district. EHRs data is mainly available in the form of two data components namely structured EHRs and unstructured EHRs. The data components of EHRs which lie under the structured EHRs are demographics, laboratory results, prescriptions, vital signs, prescriptions, ICD9 codes and medications. The unstructured EHR data components are images, graphics, discharge summary, radiology reports, chief complaint and visit documentation [48].

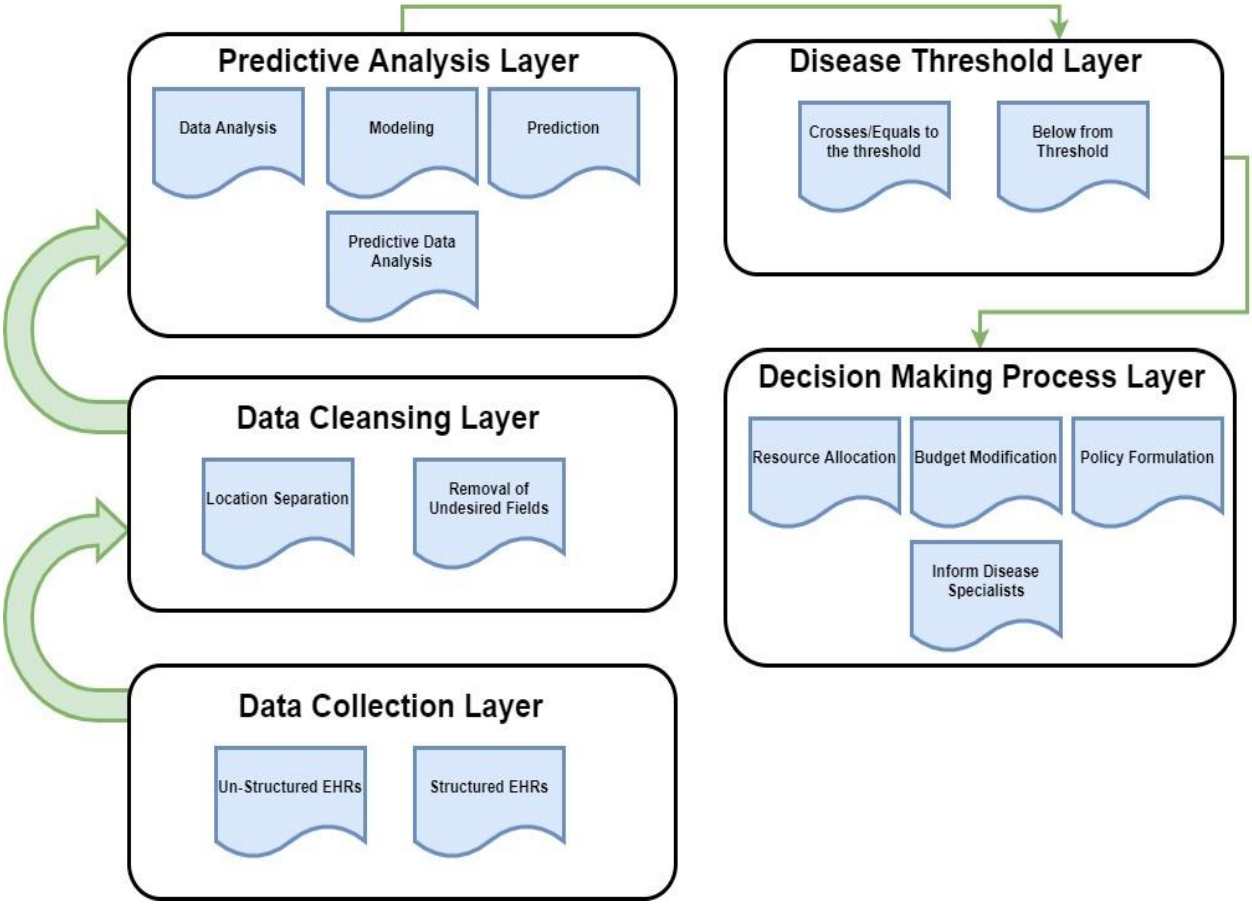

Figure 3.1: Disease predictive analysis framework for healthcare management system

3.2 Data Cleansing Layer

In this layer, inappropriate records are removed from the data samples to make desired predictions. Each location is separated from the data so that each location can be analyzed separately. In this study, we separate the disease incidence from the entire province to districts in order to analyze each district separately. We select one of the districts to analyze district incidences and uses the Moving Average (MA) filter to remove outliers from the incidences.

3.3 Predictive Analysis Layer

In this layer, required field data is analyzed in such a way that the data modeling can be performed on the basis of the analysis. After applying the model, upcoming disease incidences are predicted based on the analysis of data. The predictive data is then analyzed to provide a

health care management system decision-makers with warnings to prevent disease outbreak and to develop sound health policies.

For the predictive analysis of disease incidence, we use the ARIMA based predictive model with a moving window. Figure 3.2 depicts the main stages of our predicted model. Disease incidence data is retrieved from the data cleansing module and its time series is generated. The minimum number of incidences are determined by selecting the window size. Incidences that meet this window size are provided as input to Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF). Auto-Regressive (AR) terms are determined by PACF, while Moving Average (MA) terms are determined by ACF. As an input, these two terms are provided to the model. Disease incidences are predicted by this model and trends are determined by the interpretation of these predictable incidences. We then analyze the set trend and compare it with the disease threshold whether the trend has crossed the threshold of the disease or has fallen below from it. In both cases, a warning is provided to the decision-makers in the healthcare management system, to make better decisions.

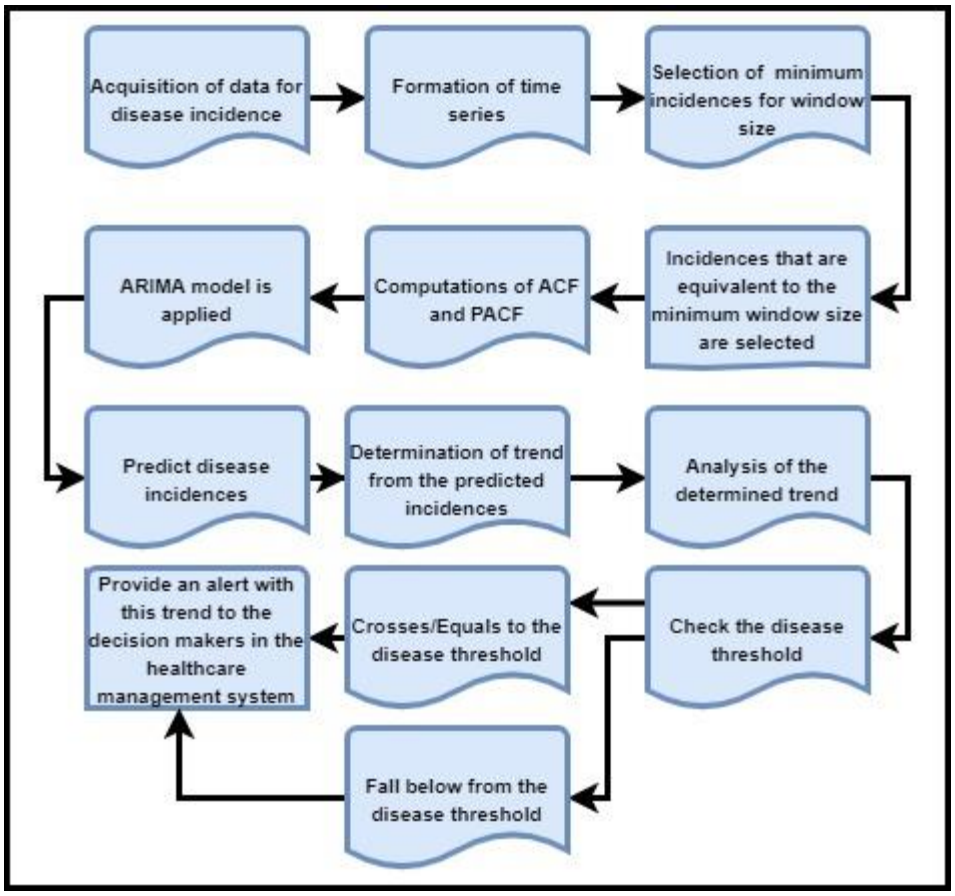

Figure 3.2: Block diagram depicts the main phases of the predictive model

Table 3-1: Window Selection process variable representation

Variable	Variable Representation
G_{e+f}	Total disease incidences
H_{e+f}	Window of total disease incidences
G_{e+1}	$e+1$ input incidences
G_{e+2}	$e+2$ input incidences

3.3.1 Window Selection Process

The variables used in equation 1, equation 2 and equation 3 in the window size selection process are represented in Table 3-1. Window size selection process for disease incidence is shown in Figure 3.3. Here, the disease incidences are the following:

$$G_{e+f} = G_{e+1}, G_{e+2}, \dots\dots, G_{e-1+f}, G_{e+f} \quad (1)$$
$$\ni \; e = 1, f = 1,2,3, \dots\dots f$$

$$H_{e+f} = H_{e+1}, H_{e+2}, \dots, H_{e-1+f}, H_{e+f} \quad (2)$$

$$G_{e+f} = H_{e+f} \quad (3)$$

This process begins with the selection of G_{e+1} incidences and the formation of their respective time series. The Augmented Ducky Fuller (ADF) test is used to check stationary in this configured time series and three cases are considered. In the first case, the probability value or p-value of the ADF test is more than 0.05, while in the second case its value is NA, and in the third case, it is less than 0.05. When it satisfies the first case, the relevant time series is differentiated and this process is continued till the p-value is less than 005. In the fulfillment of the second case, the selected incidences G_{e+1} are rejected and new incidences G_{e+2} are selected. On completion of the third case, the AR and MA terms of the relevant time series are computed from PACF and ACF. These terms are then provided as an input to the model for predicting incidences. These predictive incidences are then compared to the actual incidences and it is seen that both the values are close to each other, if it is, in that case, H_{e+1} window size is selected. But, if the value of the predicted incidences is negative out of range, in such case the window size H_{e+1} are rejected and incidences G_{e+2} are selected. This process continues for f iterations until the values of the predicted incidences approach to the actual incidences.

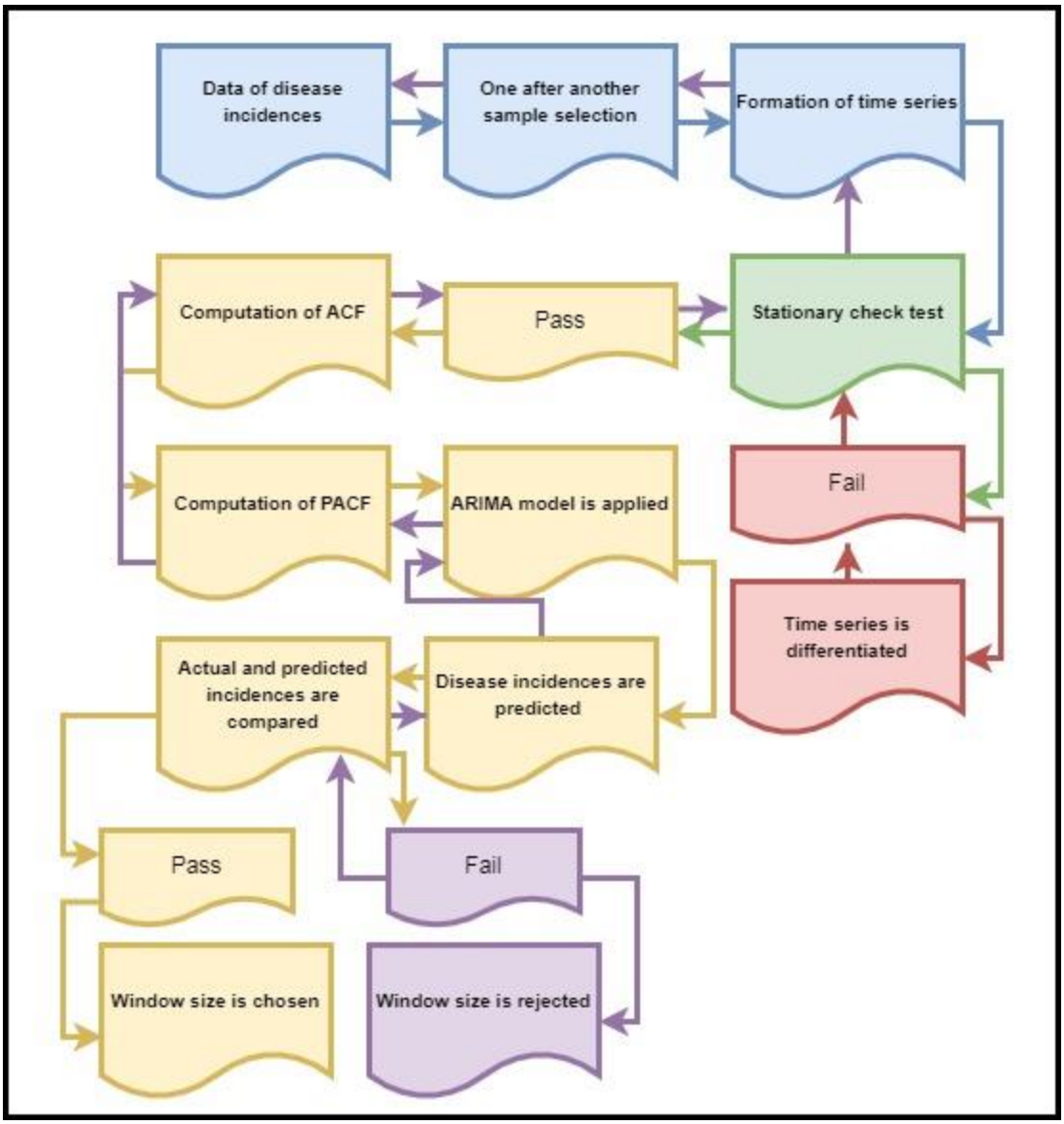

Figure 3.3: Window selection process

3.3.2 Trend Analysis

Figure 3.4 describes the process of determining the trend. In this process, after the selection of window size, the equivalent input incidences are selected and their time series is created. We use the ADF test to find stationary in this time series which checks this factor based on the p-value. If the p-value of this test is less than or equal to 0.05 then the element of stationary is present in this time series, and it is found due to the constant mean and variance. But if the p value is greater than 0.05, it indicates non-stationary and it is differentiated to make stationary till the p value is less than or equal to 0.05. In this way, after making the time series

stationary, the PACF and ACF of the corresponding time series are computed in order to find the AR and MA terms. Upcoming incidences are predicted by the ARIMA model. In the next iteration, the window is moved to the size of the predicted incidences, and this process is continued until the window covers all the input incidence. In this way, all the incidences except the input incidence that are selected in the first iteration are predicted by the model. The trend of diseases is determined by these predictable incidences and in its analysis, the threshold of diseases is compared with the predictable incidence and it is seen that the trend of these events has crossed the threshold of diseases or is below from it. In both cases, the decision-makers in the healthcare management system are made aware of this trend in order to make better policies.

Table 3-2: Trend analysis process variable representation

Variable	Variable Representation
B_1	AR terms' first coefficient
B_p	AR terms' p^{th} coefficient
d	Differenced time series' order
L_1	MA terms' first coefficient
L_q	MA terms' q^{th} coefficient
H_t	Actual disease incidences'
H_{t-p}	Actual disease incidences' p^{th} value
R_{t-q}	Error terms' q^{th} value
H_j	Actual incidences
H_{1j}	Predicted incidences
u	Total disease incidence's size

Table 3-2 shows the variable representation of the following ARIMA equations. With the help of equation 4, this model predicts the incidences of the stationary time series.

$$H_t = (B_1 H_{t-1} + \cdots + B_p H_{t-p})\left(1 + L_1 R_{t-1} \ldots M_q\, R_{t-q}\right) \quad (4)$$

To predict the disease incidences, t values of equation (4) are replaced with t+f values in equation (5).

$$H_{t+f} = (B_1 H_{t+f-1} + \cdots + B_p H_{t+f-p})\left(1 + L_1 R_{t-1} \ldots L_q\, R_{t+f-q}\right) \quad (5)$$

With the help of equation (6) and equation (7), non-stationary time series values are predicted.

$$H_t = d(1)^{d-1}H_{t-1} - \frac{d(d-1)}{2!}(1)^{d-2}H_{t-2} \ldots + d(1)H_{t-(d-1)} - H_{t-d} + B_1H_{t-1} - B_1e(1)^{e-1}H_{t-2} + B_1\frac{d(d-1)}{2!}(1)^{d-2}H_{t-3} \ldots - B_1d(1)H_{t-(d-1)} + B_1H_{t-(d+1)} \ldots . + B_pH_{t-p} - B_p\,d(1)^{d-1}H_{t-p} + B_p\frac{d(d-1)}{2!}(1)^{d-2}H_{t-(p+2)} \ldots - B_pd(1)H_{t-(d-1+p)} + B_pH_{t-(p+d)} + R_t + L_1R_{t-1} \ldots L_q\,R_{t-q} \quad (6)$$

To predict the f[th] disease incidence, t values in equation (6) are replaced with t+f values in equation (7).

$$H_{t+f} = d(1)^{d-1}H_{t+f-1} - \frac{d(d-1)}{2!}(1)^{d-2}H_{t+f-2} \ldots + d(1)H_{t+f-(d-1)} - H_{t+f-d} + B_1H_{t+f-1} - B_1d(1)^{d-1}H_{t+d-2}\frac{d(d-1)}{2!}(1)^{d-2}H_{t+f-3\ldots} - B_1d(1)H_{t+f-(d-1)} + B_1H_{t+f-(d+1)} \ldots + B_pH_{t+f-p} - B_p\,d(1)^{d-1}H_{t+f-p} + B_p\frac{d(d-1)}{2!}(1)^{d-2}H_{t+f-(p+2)} \ldots - B_pd(1)H_{t+f-(d-1+p)} + B_pH_{t+f-(p+d)} + R_t + L_1R_{t+f-1} \ldots L_q\,R_{t+n-q} \quad (7)$$

3.3.3 Performance Validation

The performance of the model is validated using the MAE and RMSE between the actual and predicted incidences. The computation of the Mean Absolute Error (MAE) between the actual and predicted incidences is given by equation (8).

$$MAE = \frac{1}{u}\sum_{j=1}^{u} |H_{1j} - H_j| \quad (8)$$

The computation of Root Mean Square Error (RMSE) between the actual and predicted incidences is given by equation (9).

$$RMSE = \sqrt{\sum_{j=1}^{u}\frac{(H_{1j} - H_j)}{u}} \quad (9)$$

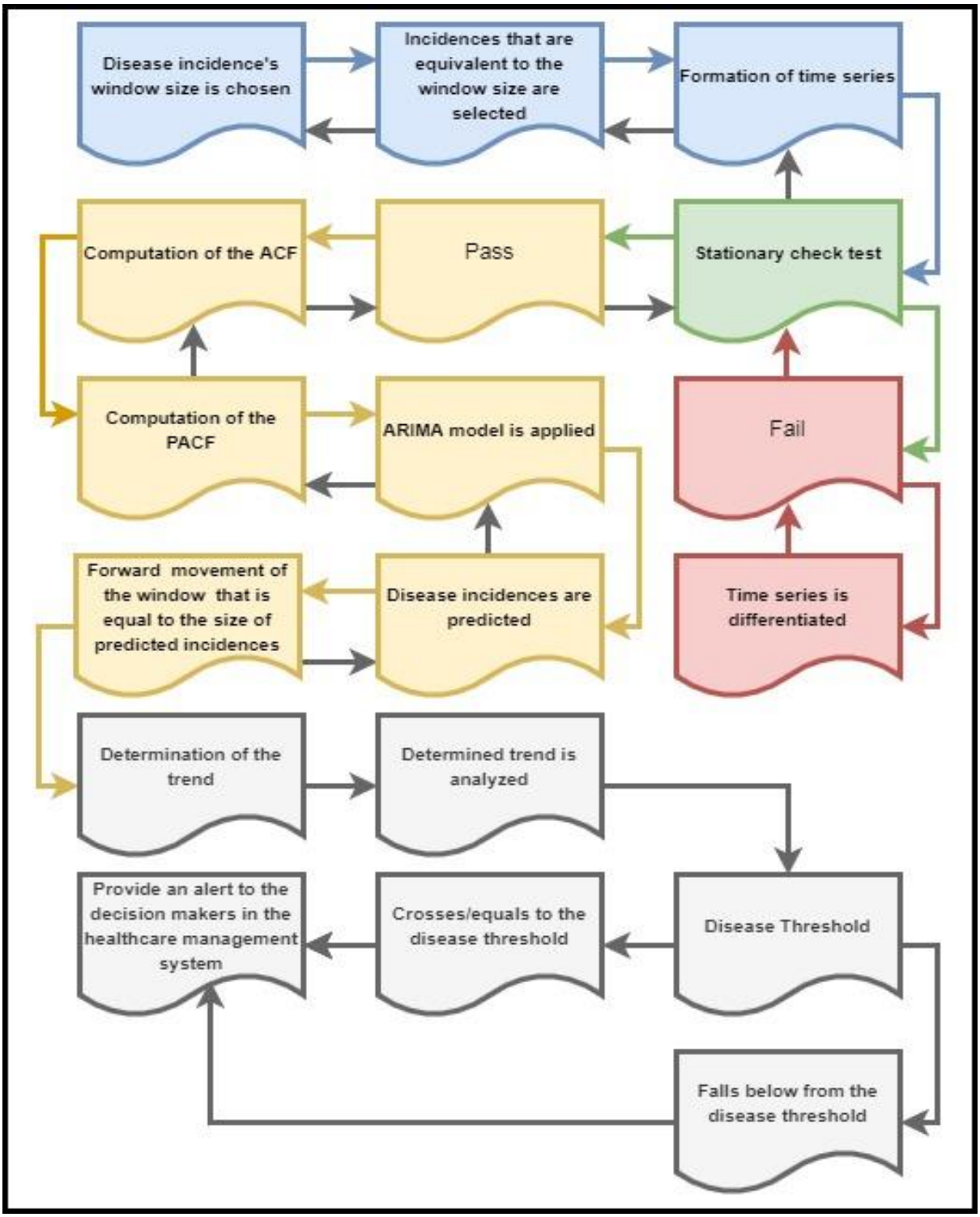

Figure 3.4: Trend determination process

3.4 Disease Threshold Layer

In this layer, disease thresholds are considered for predictive analysis so that decision-makers in the healthcare management system can take all necessary steps to prevent the spread of the disease. For every disease, this threshold is different. Two cases are considered in the incidence of predictive diseases, one is when the incidence of predictive diseases crosses or equals to the threshold and the other is when the incidence of predictive diseases falls below from the threshold. In both cases there is an alert to the decision-makers of the health care management system and at the same time a disease prediction trend is also sent so that they take into account this predictive trend.

The alert threshold varies for each disease, for example, occurrence of 1 incidence in 3 to 14 days for dengue fever, 1 incidence in 2 to 5 days for diphtheria, 3 to 6 incidences for hepatitis C in 2 to 6 weeks, 1 incidence in 24 to 28 hours for influenza and 1 incidence in 7 to 18 for measles, etc. [49].In this study, we do not compute alert thresholds for diseases, but use pre-computed alert thresholds.

3.5 Decision Making Process Layer

In this layer, the decision-makers in the healthcare management system receive the predictive trend when the predictive analysis of our predictive model fulfills either of the two cases described in the disease threshold module. In Case 1, they reallocate health resources and medical services. Increase health budgets. Inform disease specialists in a timely manner so they can find the cause of the spread of the disease. Develop new policies to protect people from the spread of the disease and to run health campaigns to protect themselves. On the other hand, allocate health resources and medical services according to the number of predicted incidences in case 2. Reduce health budgets to prevent the wastage of medical resources and the destruction of the national economy.

There are several departments in the Pakistan Healthcare Management System to support decision, including National Database and Registration Authority (NADRA), Drug Regulatory Authority of Pakistan (DRAP), National Health Emergency Preparedness and Response Network (NHEPRN), Pakistan Medical Commission (PMC), Pakistan Centre for Philanthropy (PCP), Pakistan Bureau of Statistics (PBS), National Institute of Health (NIH), and National Institute of Population Studies (NIPS). NADRA provides better solutions for e-Governance, to eliminate

identity theft and to protect and identify people's documents and interests. DRAP plays an important role in enforcing drug laws and regulating therapeutic goods in the country keeping in view the interest of the people of Pakistan [50]. The NHERN serves as a focal point for all aspects of disaster health care recovery, response, and preparedness, as well as creates coordination with international, national, and regional agencies and stakeholders [51]. PMC sets the standards for recognition of qualification, training, and education in the medical and dentistry profession [52]. PCP meets certification requirements in the areas of financial management, program delivery, and internal governance, as well as evaluates the performance of these areas by the FBR [53]. PBS is responsible for gathering, compilation, and distribution of consistent and appropriate statistical information to the researchers, policy makers, and planners [54]. NIH is actively involved in various public health related activities including laboratory diagnostics, antisera production, research and development, vaccines, and drug and food quality control [55]. NIPS is involved in producing research of high quality, data of evidence based, and information for utilization by different agencies in order to make strategic planning, formulation of policies, and for creating references in the ranges of development & population, demography, and health [56].

Director General Health Services (DGHs) is the central programmatic management, monitoring, and implementation arm of the provincial department of health and is responsible for controlling provision of secondary and primary services of the healthcare, and links with the health offices of the district in the entire province [57]. Concerned DG takes appropriate action and forward the information of predictive trend to Prime Minister, Health Minister, and Health Secretary. When the communicable diseases, Non opt Medical Equipment and test facility, and HR shortages crossed the threshold in such a case the information of the predictive trend is sent to DG Health in order to take appropriate actions. DG Health access the current situation and recommend way forward to Prime Minister, Health Minister, and Health Secretary. Conduct analysis for long term measures. Select appropriate actions for HCDU Management. Identify the intervention to curb/mitigate the negative impact. When the healthcare funding and allergy vaccinations is reduced from the threshold in such a case an information of the predictive trend is sent to the DG Finance Division, and Executive Director NIH. DG Finance Division and Executive Director NIH access the current situation and recommend way forward to Prime Minister, Health Minister, and Health Secretary. When the fake medicines is crossed the

threshold in such a case an information of the predictive trend is sent to the DG DRAP. DG DRAP access the current situation and recommend way forward to Prime Minister, Health Minister, and Health Secretary. Conduct analysis for long term measures. Select appropriate actions for HCDU Management. Identify the intervention to curb/mitigate the negative impact.

3.6 Case Study

We show the incidences in the Lahore and Vehari district of Punjab province and the Pakistan Ministry of Health as a case study in Figure 3.5. In this case study, data is collected on Hepatitis C cases in Punjab Province between the years 2015-2019 and is separated into Lahore District incidences by pre-processing it in Excel 2013. After analyzing and predicting the events of the upcoming days from these events, we compare them with the threshold of Hepatitis C disease and in case of crossing the threshold or below it, the information is sent to the dashboard of the ministry of health, and at the same time, an alert through SMS is sent to DG Health about the spread of the disease. DG Health takes appropriate action to address health issues and also informs the Prime Minister, Health Ministry, and Health Secretary about the situation. R Studio is used for the implementation purpose of predictive analysis.

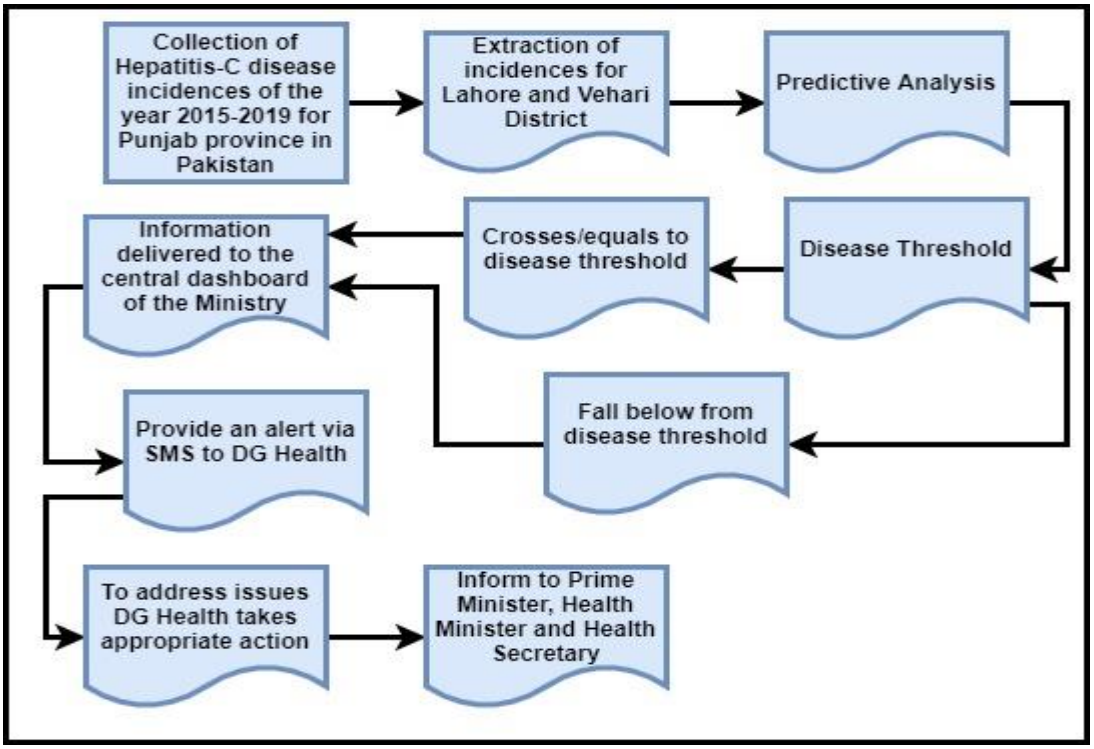

Figure 3.5: Case study for Lahore district Hepatitis-C incidences and Pakistan Ministry of Health

CHAPTER 4: ARIMA MODELING

4.1 Introduction

Auto Regressive Integrated Moving Average (ARIMA) is a class of statistical analysis model that uses historical time-series data to predict and analyze the upcoming trends. Data that consists of actual incidences have a non-stationary nature, so it is made stationary to get a better prediction from the data in the ARIMA model.

4.2 Components of ARIMA Model

The components of the ARIMA model are described below:

Auto Regressive (AR): The model shows a dependent relationship between the current and prior lagged incidences.

Integrated (I): The term integrate represents the process of taking the difference in time series in order to maintain the stationary factor.

Moving Average (MA): The model shows a dependent relationship between the error terms of the current and prior lagged incidences.

The standard notation used for ARIMA model is ARIMA (p, d, q).

Where p represents the order of AR terms, d represents the order of difference terms, and q represents the order MA terms of the ARIMA model.

4.3 Steps of the ARIMA Model

The steps of the ARIMA model building process are described below**:**

4.3.1 Model Identification

In this step time series disease incidences are plotted. ADF test is applied to check the stationary in time series. Stationary in time series is necessary to produce constant mean and finite variance. ARIMA model predicts the present value on the basis of past values and error terms of past values. Constant mean and finite variance remain the same in the future as it is in the past so the ARIMA model easily predicts the present value based on the past value and error terms.

ADF test checks the stationary in time series by this procedure.

Two types of hypotheses are made. NULL Hypothesis 'H0' and Alternative Hypothesis 'H1'.H0 is accepted when non-stationary lie in time series plot and the p-value is greater than 0.05. H1 is accepted when stationary in time series is present and the p-value is less than 0.05. Non-stationary time series are made stationary by differentiating the time series. The order of differencing of the time series is started from lower then move to higher.

The first difference of non-stationary time series is illustrated by the following equation.

$$\Delta H_t = H_t - H_{t-1} \tag{10}$$

Where H_t is the actual value of time series in (10) and H_{t-1} is the value time series at t-1 lags.

The nth difference of non-stationary time series is illustrated by this equation.

$$\Delta H_t = H_t \dots\dots\dots\dots\dots\dots, -H_{t-(n-1)} - H_{t-n} \tag{11}$$

Where H_t is the actual value of time series in (11) and H_{t-n} is the value time series at nth lag.

Non stationary time series are made stationary by diff function in the Rstudio.

4.3.2 Parameter Estimation

In this step, ACF and PACF of actual and the differenced time series incidences are plotted. PACF computes the AR terms and ACF computes the MA terms. The term Autocorrelation is used due to the presence of one variable. Correlation is computed between the first lag value and the second lag value in the case of ACF. In PACF, correlation is computed between the first lag value and third lag value by skipping the second lag value. Those terms which lie above the threshold value in both ACF and PACF are considered significant. The threshold value is 0.05. It is shown by the blue line in Rstudio. ACF and PACF graphs lie within the range of -1 to +1. In the case of stationary time series, ACF and PACF plots of the actual time series are considered. For non-stationary timeseries ACF and PACF plots of the differenced time series are considered. With the help of equation (12) ACF computes the values of MA terms.

$$Correlation(H_t, H_{t-s}) = Covariance(H_t, H_{t-s})/(\sqrt{variance(H_t)})(\sqrt{variance(H_{t-1})}) \tag{12}$$

Where in (12) H_t is the present value of time series and Ht-s is the value of time series which are at s lag.

PACF computes the values of AR terms with the help of equation (13).

$$Correlation(H_t, H_{t-s} | H_{t-s+1}, \ldots\ldots\ldots\ldots H_{t-1}) \quad (13)$$
$$= Covariance(H_t, H_{t-s} | H_{t-s+1})$$
$$/(\sqrt{variance(H_t | H_{t-s+1})})(\sqrt{variance(H_{t-s} | H_{t-s+1})})$$

Where in (13) H_t is the present value of time series, H_{t-s} is the value of time series which are at s lag and H_{t-s+1} is the value of intermediate lag between H_t and H_{t-s}.

To compute ACF and PACF acf and pacf functions are used in R studio.

AR terms are computed from significant values of PACF, MA terms are computed from significant values of ACF, d terms are computed from the order of difference terms in time series.

4.3.3 Model Checking

In this step, AR, MA and d terms which are computed in the above steps are used to fit ARIMA (p, d, q), model. AR terms are used to set the parameters of p, MA terms are used to set the parameters of q value and d terms are used to set the parameters of d. Time series data is given to ARIMA (p, d, q) model and upcoming data is predicted from ARIMA (p, d, q) model. This model is applied to time series data with the help 'arima' function in R studio.

With the help of equation (14), this model predicts the incidences of the stationary time series.

$$H_t = (B_1 H_{t-1} + \cdots + B_p H_{t-p})\left(1 + L_1 R_{t-1} \ldots M_q R_{t-q}\right) \quad (10)$$

To predict the disease incidences, t values of equation (14) are replaced with t+f values in equation (15).

$$H_{t+f} = (B_1 H_{t+f-1} + \cdots + B_p H_{t+f-p})\left(1 + L_1 R_{t-1} \ldots L_q R_{t+f-q}\right) \quad (11)$$

With the help of equation (16) and equation (17), non-stationary time series values are predicted.

$$H_t = d(1)^{d-1} H_{t-1} - \frac{d(d-1)}{2!}(1)^{d-2} H_{t-2} \ldots + d(1) H_{t-(d-1)} -$$
$$H_{t-d} + B_1 H_{t-1} - B_1 e(1)^{e-1} H_{t-2} + B_1 \frac{d(d-1)}{2!}(1)^{d-2} H_{t-3} \ldots - B_1 d(1) H_{t-(d-1)} +$$
$$B_1 H_{t-(d+1)} \ldots . + B_p H_{t-p} - B_p\, d(1)^{d-1} H_{t-p} +$$
$$B_p \frac{d(d-1)}{2!}(1)^{d-2} H_{t-(p+2)} \ldots - B_p d(1) H_{t-(d-1+p)} + B_p H_{t-(p+d)} + R_t + L_1 R_{t-1} \ldots L_q R_{t-q}$$

(12)

To predict the f[th] disease incidence, t values in equation (16) are replaced with t+f values in equation (17).(7)

$$H_{t+f} = d(1)^{d-1}H_{t+f-1} - \frac{d(d-1)}{2!}(1)^{d-2}H_{t+f-2} \ldots + d(1)H_{t+f-(d-1)} \quad (13)$$
$$- H_{t+f-d} + B_1 H_{t+f-1} - B_1 d(1)^{d-1} H_{t+d-2} \frac{d(d-1)}{2!}(1)^{d-2}H_{t+f-3\ldots}$$
$$- B_1 d(1)H_{t+f-(d-1)} + B_1 H_{t+f-(d+1)} \ldots + B_p H_{t+f-p}$$
$$- B_p\, d(1)^{d-1}H_{t+f-p}$$
$$+ B_p \frac{d(d-1)}{2!}(1)^{d-2}H_{t+f-(p+2)} \ldots - B_p d(1)H_{t+f-(d-1+p)}$$
$$+ B_p H_{t+f-(p+d)} + R_t + L_1 R_{t+f-1} \ldots L_q R_{t+n-q}$$

CHAPTER 5: RESULTS AND ANALYSIS

This chapter contains the results of our proposed Disease Predictive Analytics Framework. In this chapter, we present the results of the implementation of incidences in Lahore and Vehari districts and compare the predictive analysis to the pre-computed disease threshold and analyze how the decision-makers take appropriate steps to prevent the disease in the Pakistan healthcare management system.

We have presented a disease incidences of Lahore and Vehari District of Punjab province of Pakistan and the decision makers in the Pakistan Ministry of Health as a case study to validate our proposed predictive analytics framework.

We have selected 25, the minimum number of days of HCV incidence for the Lahore District. During this selection process, we do not use the incidences of the first three days due to the absence of variance within it. The incidences of the day 4,5,6,7,8,9,10,11,12,18,22 are rejected due to non-maintenance of the stationary element while the incidence at 13,14,15,16,17,19,20,21,23,24 days are rejected based on negative out of range prediction.

The plot of Figure 5.1 shows the incidence of first 25 days of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 6, 11 to 18, and 20 to 21 show an increase in the trend while the incidences on the day 7 and 19 to 20 show a decrease in the trend. It also shows the non-stationary element in the time series.

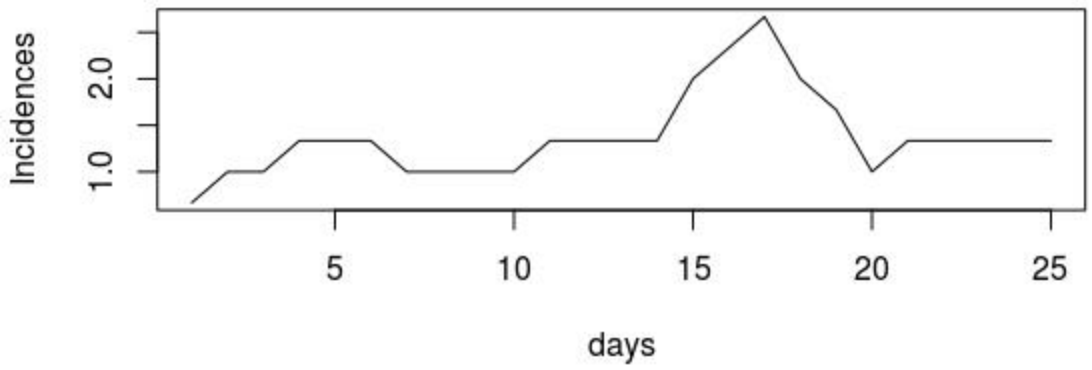

Figure 5.1: Plot of first 25 day's incidences of the Lahore District

In the case of this non-stationary time series, the p value of the ADF test is 0.2093. It is made in to stationary by differentiating it once and in this case the p-value of the ADF test is 0.01837 which indicates the stationary presence. The plot of the stationary time series of 25-day incidences after making a difference once is shown in Figure 5.2. This plot also shows the constant mean and variance in this time series.

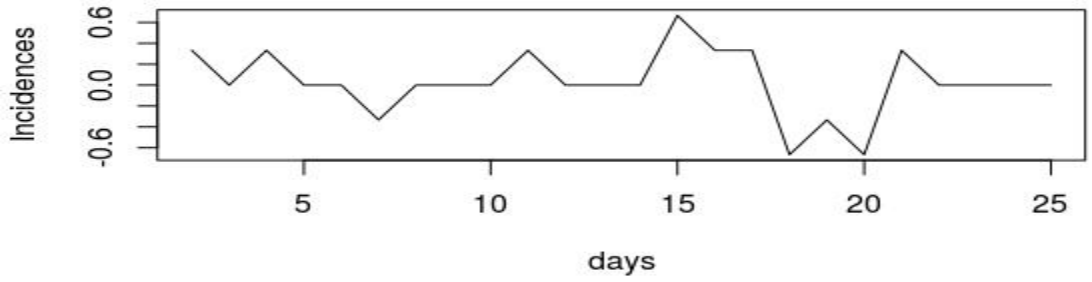

Figure 5.2: Stationary time series plot of 25 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.3. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by only one coefficient of the AR terms so its order is one and the value of this coefficient is -0.6118. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

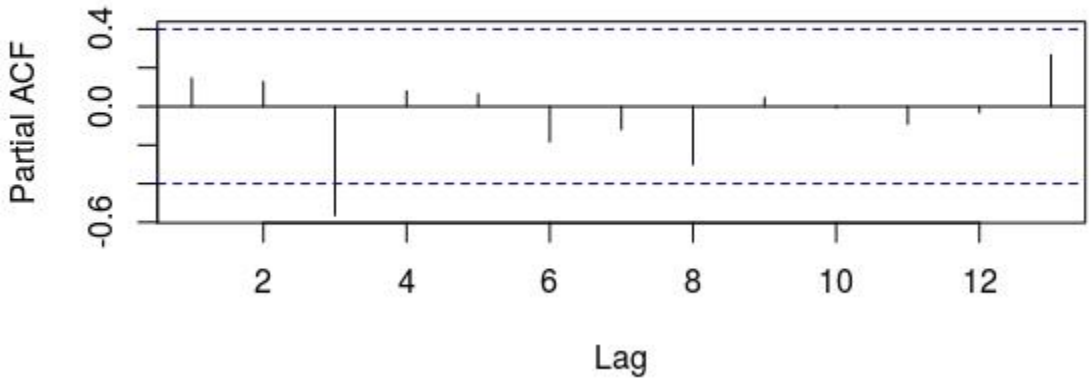

Figure 5.3: PACF plot of Lahore District

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.4. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 1.1921 and 0.9999.

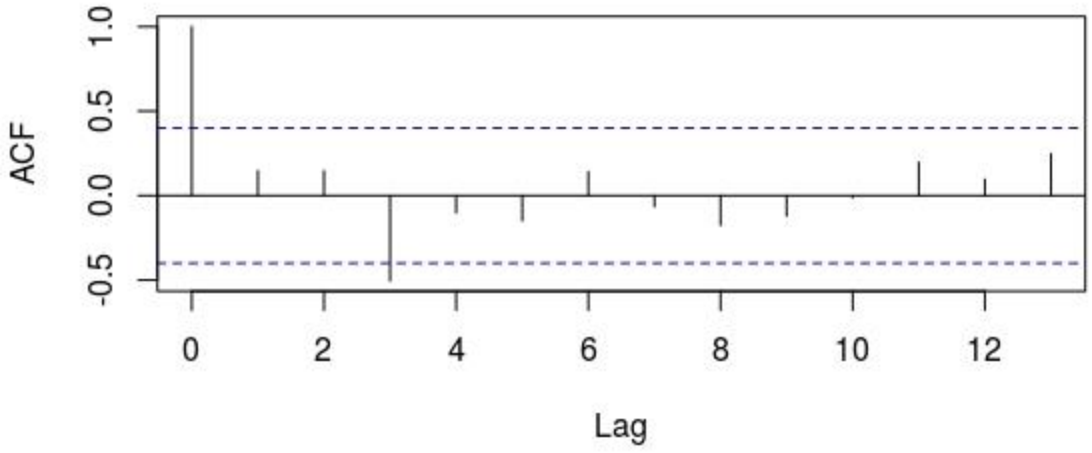

Figure 5.4: ACF plot of the Lahore District

As the order 1 of AR terms, order 2 of MA terms, and order 1 of differential terms are computed, so the model that fits best is ARIMA (1,1,2) on these incidences and it is shown in equation (19).

$$H_t = (1 - B_1)H_{t-1} - B_1 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (14)$$

After the substitution of B_1, L_1, and L_2 coefficient values, equation (19) is replaced with equation (20).

$$H_t = 1.6118H_{t-1} + 0.6118H_{t-2} + R_t + 1.1921R_{t-1} + 0.9999R_{t-2} \quad (15)$$

The above equations (20) are used to predict 26, 27, 28, 29, 30 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.5 shows the incidence of 6 to 30 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 9 to 12, 15 to 16, 20 to 21, and 22 show an increase in

the trend while the incidences on the day 2, 13 to 15, and 22 show a decrease in the trend. It also shows the non-stationary element in the time series.

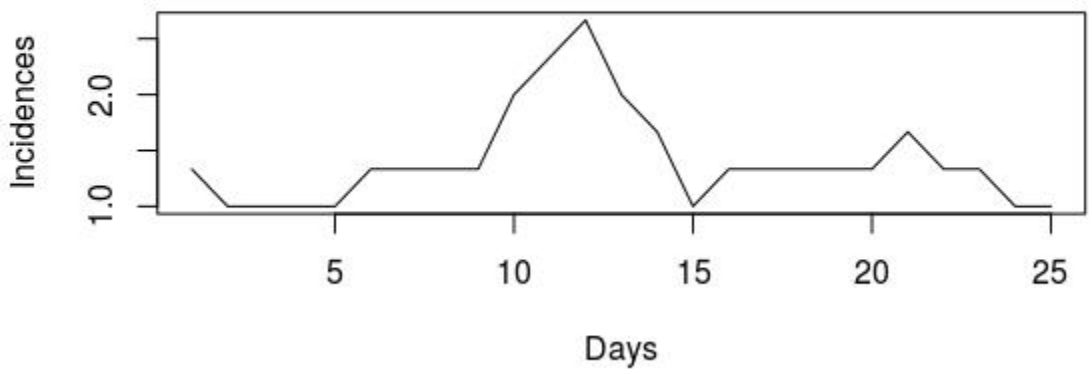

Figure 5.5: Plot of 6 to 30 day's incidences of the Lahore District

In the case of this non-stationary time series, the p value of the ADF test is 0.3072. It is made in to stationary by differentiating it once and in this case the p-value of the ADF test is 0.02663 which indicates the stationary presence. The plot of the stationary time series of 25-day incidences after making a difference once is shown in Figure 5.6. This plot also shows the constant mean and variance in this time series.

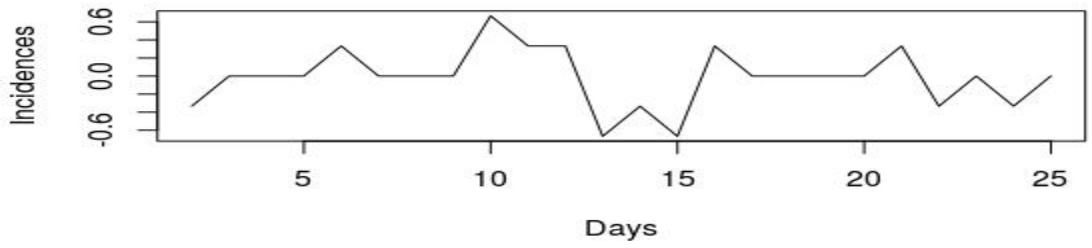

Figure 5.6: Stationary time series plot of the 6 to 30 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.7. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by only one coefficient of the AR terms so its order is one and the value of this coefficient is -0.6586. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

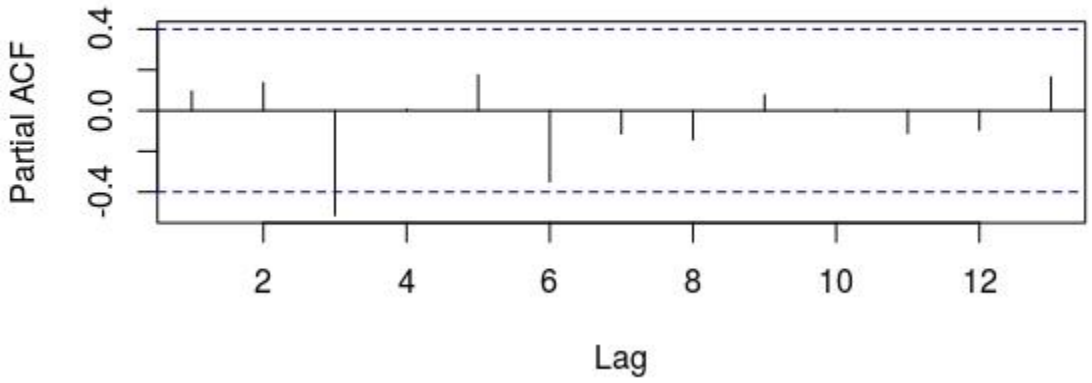

Figure 5.7: PACF plot of Lahore District for days 6 to 30

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.8. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 1.1984 and 1.0000.

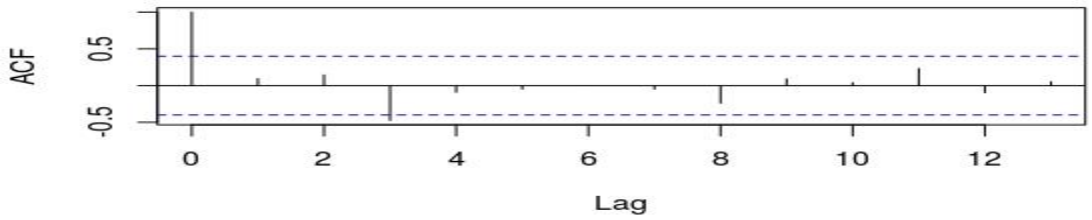

Figure 5.8: ACF plot of Lahore District for days 6 to 30

As the order 1 of AR terms, order 2 of MA terms, and order 1 of differential terms are computed, so the model that fits best is ARIMA (1,1,2) on these incidences and it is shown in equation (21).

$$H_t = (1 - B_1)H_{t-1} - B_1 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (16)$$

After the substitution of B_1, L_1, and L_2 coefficient values, equation (21) is replaced with equation (22).

$$H_t = 1.6586H_{t-1} + 0.6586H_{t-2} + R_t + 1.1984R_{t-1} + 1.0000R_{t-2} \quad (17)$$

The above equations (22) are used to predict 31, 32, 33, 34, 35 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.9 shows the incidence of 11 to 35 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 4 to 8, 16, and 25 show an increase in the trend while the incidences on the day 9 to 10, and 17 to 24 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

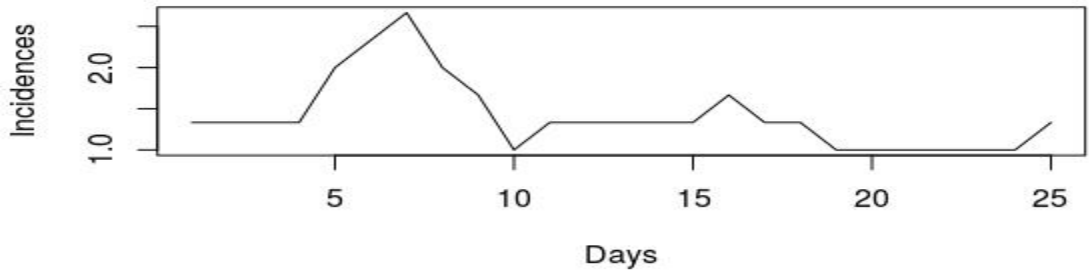

Figure 5.9: Plot of 11 to 35 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 15. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by only one coefficient of the AR terms so its order is one and the value of this coefficient is 0.4930. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

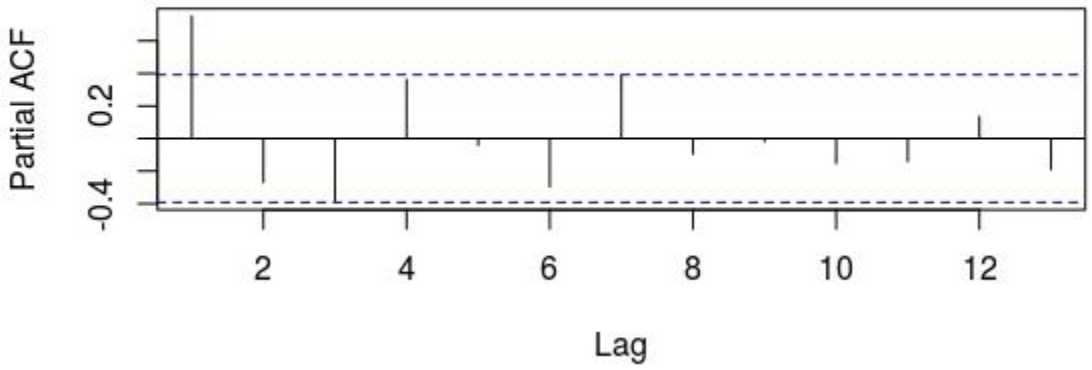

Figure 5.10: PACF plot of Lahore District for days 11 to 35

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.11. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are 0.7351, 0.6981, and -0.2935.

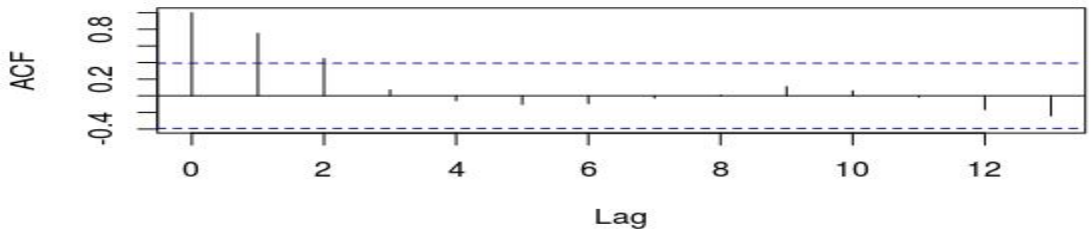

Figure 5.11: ACF plot of Lahore District for days 11 to 35

As the order 1 of AR terms, order 3 of MA terms, and order 1 of differential terms are computed, so the model that fits best is ARIMA (1, 0, 3) on these incidences and it is shown in equation (23).

$$H_t = B_1 H_{t-1} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 R_{t-3} \quad (18)$$

After the substitution of B_1, L_1, and L_2 coefficient values, equation (23) is replaced with equation (24).

$$H_t = 0.4930 H_{t-1} + R_t + 0.7351 R_{t-1} + 0.6981 R_{t-2} - 0.2935 R_{t-2} \quad (19)$$

The above equations 24 are used to predict 36, 37, 38, 39, 40 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.12 shows the incidence of 16 to 40 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, 10 to 11, and 19 to 24 show an increase in the trend while the incidences on the day 3 to 5, and 25 show a decrease in the trend. It also shows the non-stationary element in the time series.

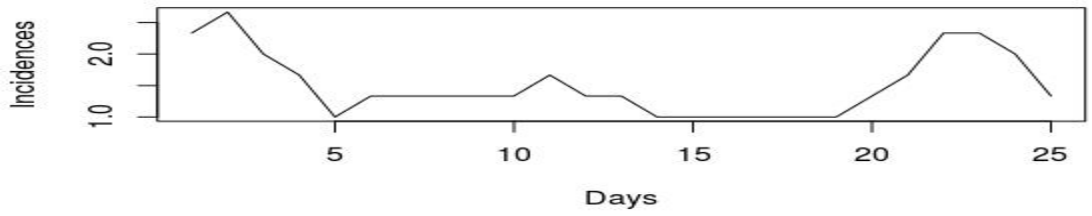

Figure 5.12: Plot of 16 to 40 day's incidences of the Lahore District

In the case of this non-stationary time series, the p value of the ADF test is `0.05229.` It is made in to stationary by differentiating it thrice and in this case the p-value of the ADF test is 0.01279 which indicates the stationary presence. The plot of the stationary time series of 16 to 40

day incidences after making a difference three times is shown in Figure 5.13. This plot also shows the constant mean and variance in this time series.

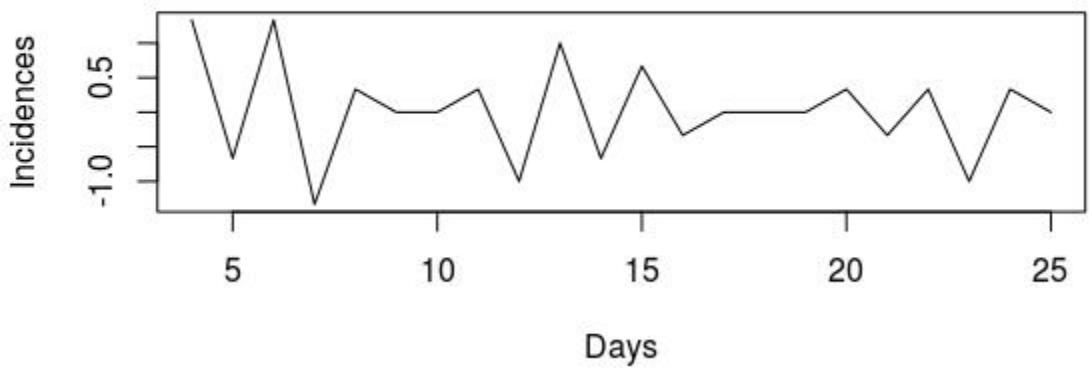

Figure 5.13: Stationary time series plot of 16 to 40 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.14. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by only one coefficient of the AR terms so its order is one and the value of this coefficient is -0.6816. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

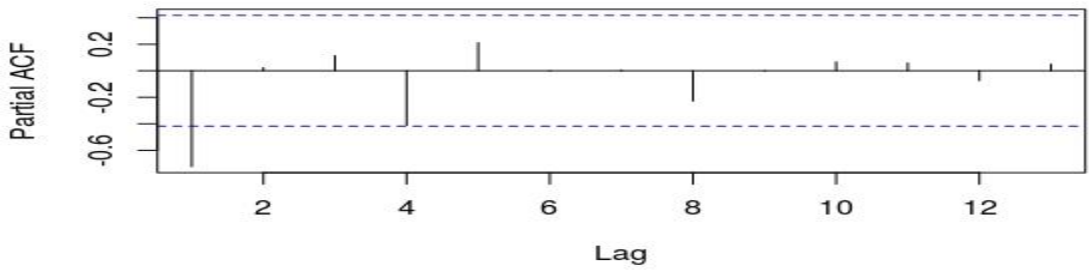

Figure 5.14: PACF plot of Lahore District for days 16 to 40

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.15. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by six coefficients of the MA terms so its order is six and the values of these coefficients are -0.6139, -0.3490, -0.8-47, 0.5325, 0.1765, and 0.1595.

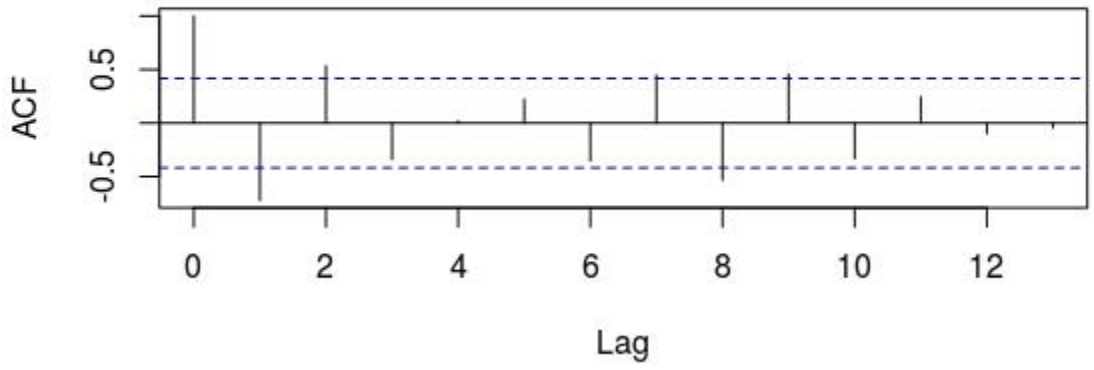

Figure 5.15: ACF plot of Lahore District for days 16 to 40

As the order 1 of AR terms, order 6 of MA terms, and order 3 of differential terms are computed, so the model that fits best is ARIMA (1, 3, 6) on these incidences and it is shown in equation (25).

$$H_t = -(-3 - B_1)H_{t-1} - (-3 + 3B_1)H_{t-2} - (-1 - 3B_1)H_{t-3} - B_1H_{t-4} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} + L_4R_{t-4} + L_5R_{t-5} + L_6R_{t-6} \quad (20)$$

After the substitution of B_1, L_1, L_2, L_3, L_4, L_5, and L_6 coefficient values, equation (25) is replaced with equation (26).

$$H_t = 0.23184H_{t-1} - 0.9552H_{t-2} - 1.0448H_{t-3} + 0.6816H_{t-4} + R_t - 0.6139R_{t-1} - 0.3490R_{t-2} - 0.8047R_{t-3} + 0.5325R_{t-4} + 0.1765R_{t-5} + 0.1595R_{t-6} \quad (21)$$

The above equations (26) are used to predict 41, 42, 43, 44, 45 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.16 shows the incidence of 21 to 45 days of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on day 6, and 15 to 18 show an increase in the trend while the incidences on the day 7 to 14, and 19 to 25 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this non-stationary time series, the p value of the ADF test is `0.04423.` This plot also shows the constant mean and variance in the time series.

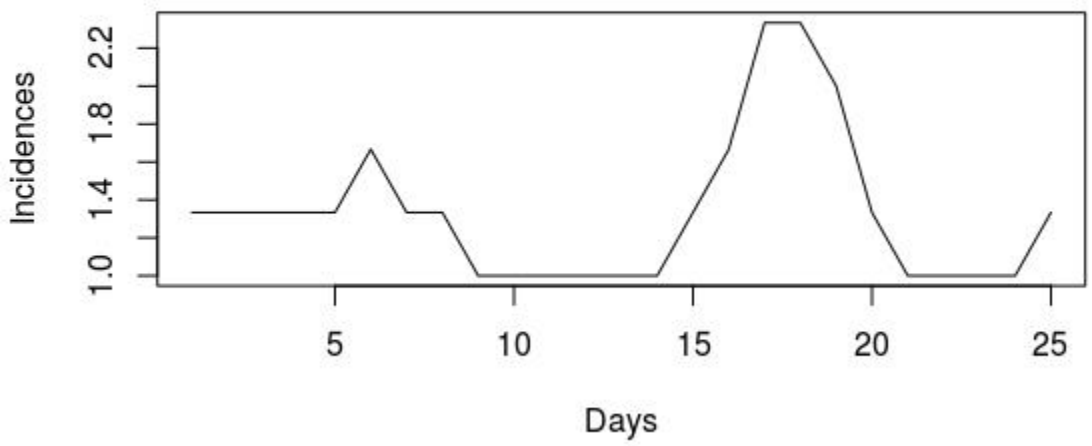

Figure 5.16: Plot of 21 to 45 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.17. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are 1.4572, and -0.6812. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

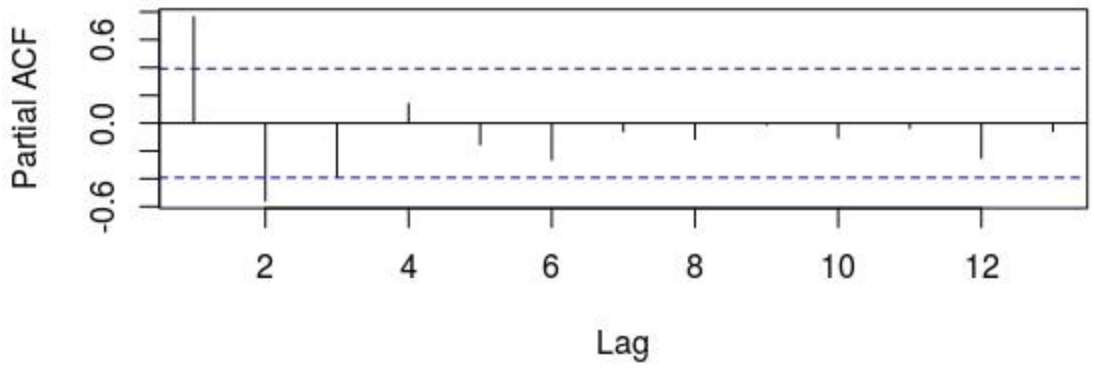

Figure 5.17: PACF plot of Lahore District for days 21 to 45

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.18. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by five coefficients of the MA terms so its order is five and the values of these coefficients are -0.5210, 0.1249, -0.8282, 0.2135, and 0.0119.

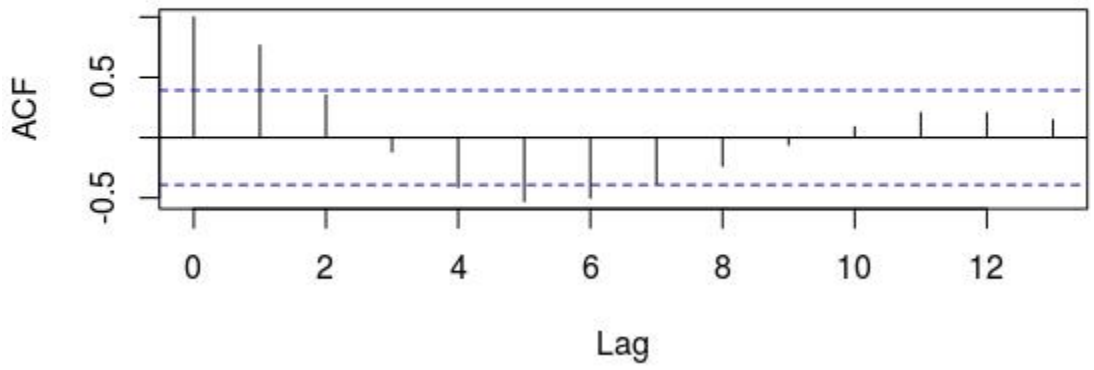

Figure 5.18: ACF plot of Lahore District for days 21 to 45

As the order 2 of AR terms, and order 5 of MA terms are computed, so the model that fits best is ARIMA (2, 0, 5) on these incidences and it is shown in equation (27).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 R_{t-3} + L_4 R_{t-4} + L_5 R_{t-5} \quad (22)$$

After the substitution of B_1, B_2, L_1, L_2, L_3, L_4, and L_5 coefficient values, equation (27) is replaced with equation (28).

$$H_t = 1.4572 H_{t-1} - 0.6812 H_{t-2} + R_t - 0.5210 R_{t-1} + 0.1249 R_{t-2} - 0.8287 R_{t-3} + 0.2135 R_{t-4} + 0.0119 R_{t-5} \quad (23)$$

The above equations (28) are used to predict 46, 47, 48, 49, 50 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.19 shows the incidence of 26 to 50 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 10 to 13 show an increase in the trend while the incidences on the day 1 to 9, and 14 to 25 show a decrease in the trend. It also shows the non-stationary element in the time series.

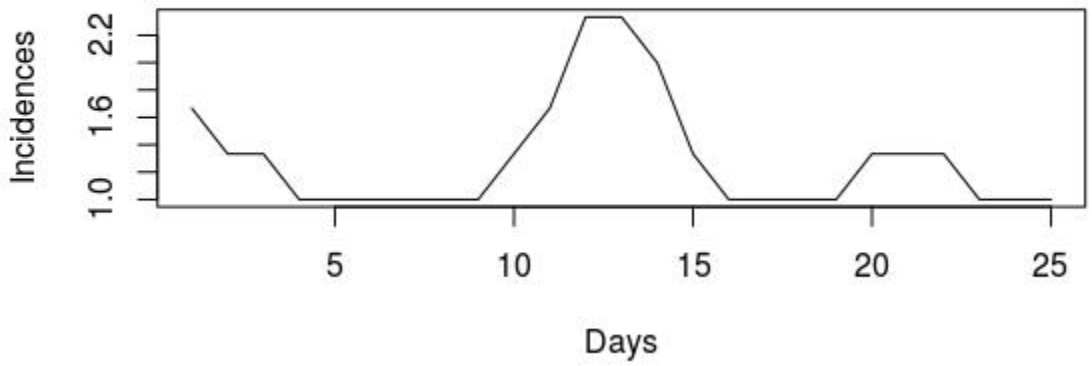

Figure 5.19: Plot of 26 to 50 day's incidences of the Lahore District

In the case of this non-stationary time series, the p value of the ADF test is `0.09334.` It is made in to stationary by differentiating it once and in this case the p-value of the ADF test is 0.02627 which indicates the stationary presence. The plot of the stationary time series of 26 to 50

day incidences after making a difference two times is shown in Figure 5.20. This plot also shows the constant mean and variance in this time series.

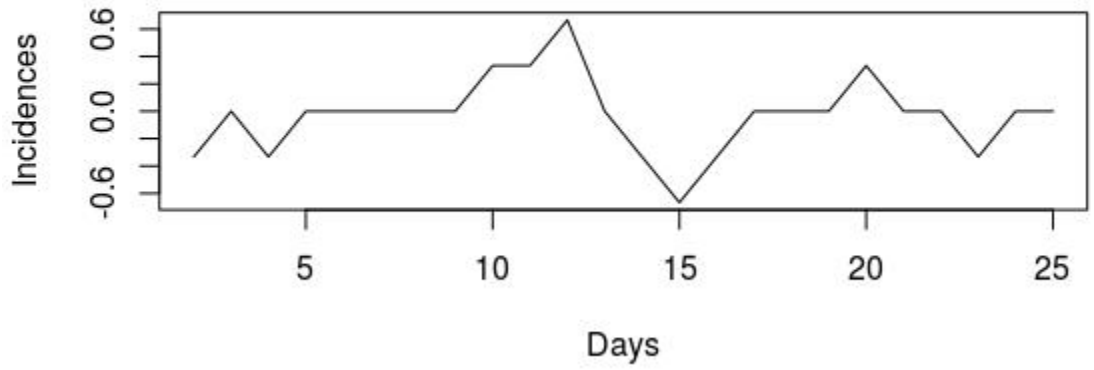

Figure 5.20: Stationary time series plot of 26 to 50 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.21. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are -0.2223, and -0.0882. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

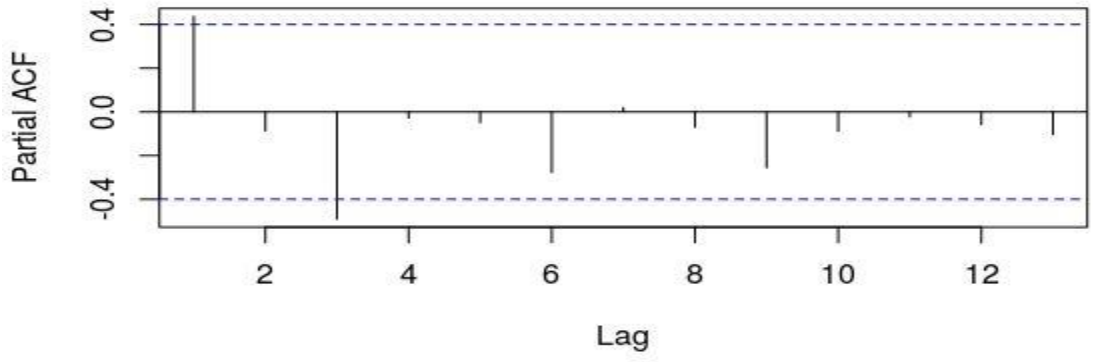

Figure 5.21: PACF plot of Lahore District for days 26 to 50

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.22. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 0.8745, and 1.000.

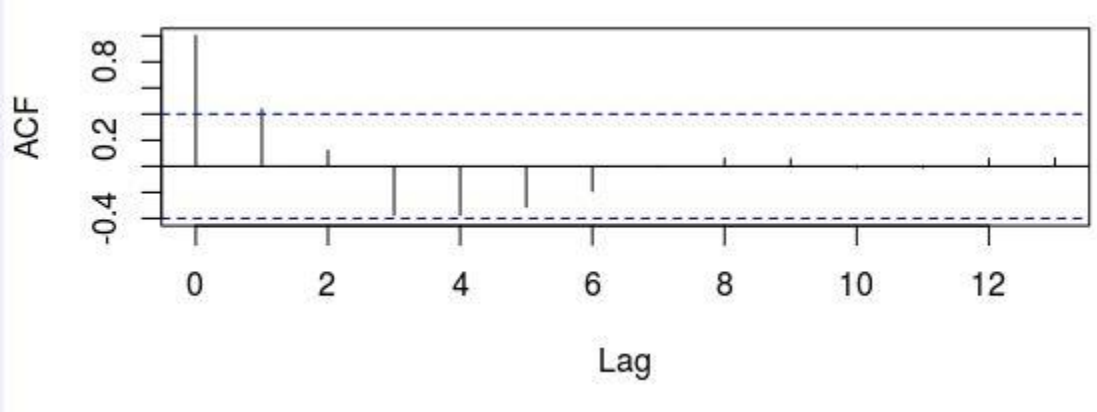

Figure 5.22: ACF plot of Lahore District for days 26 to 50

As the order 2 of AR terms, order 2 of MA terms, and order 1 of the difference terms are computed, so the model that fits best is ARIMA (2, 1, 2) on these incidences and it is shown in equation (29).

$$H_t = -(-B_1 - 1)H_{t-1} - (-B_2 + B_1)H_{t-2} + B_2 H_{t-3} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \qquad (24)$$

After the substitution of B_1, B_2, L_1, and L_2 coefficient values, equation (29) is replaced with equation (30).

$$H_t = 0.7777H_{t-1} + 0.1341H_{t-2} - 0.0882H_{t-3} + R_t + 0.8745R_{t-1} + 1.0000R_{t-2} \qquad (25)$$

The above equation (30) is used to predict 51, 52, 53, 54, and 55 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.23 shows the incidence of 31 to 55 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 8 show an increase in the trend while the incidences on the day 9 to 25 show a decrease in the trend. It also shows the stationary element

in the time series. In the case of this stationary time series, the p value of the ADF test is 0.02075. This plot also shows the constant mean and variance in this time series.

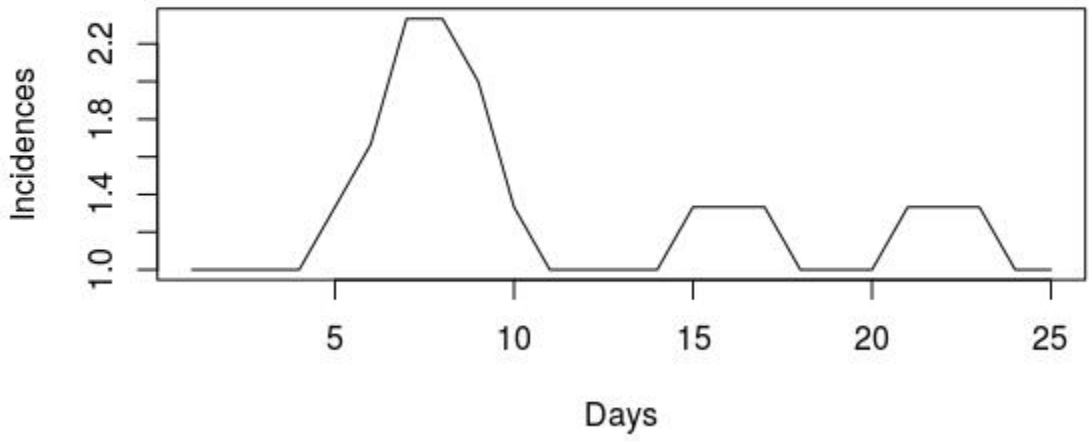

Figure 5.23: Plot of 31 to 55 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.24. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are 1.3687, and -0.5573. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

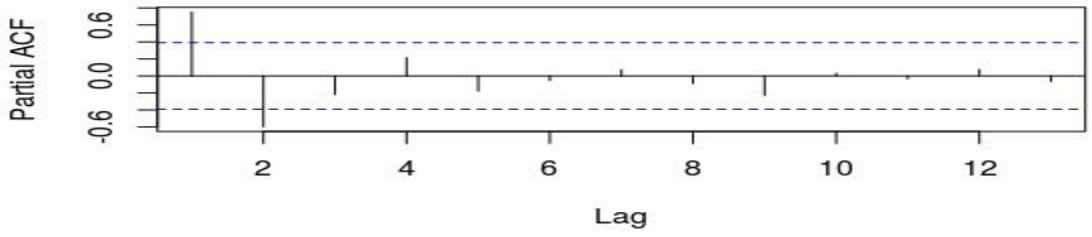

Figure 5.24: PACF plot of Lahore District for days 31 to 55

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.25. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are -0.1638, 0.0184, and -0.8545.

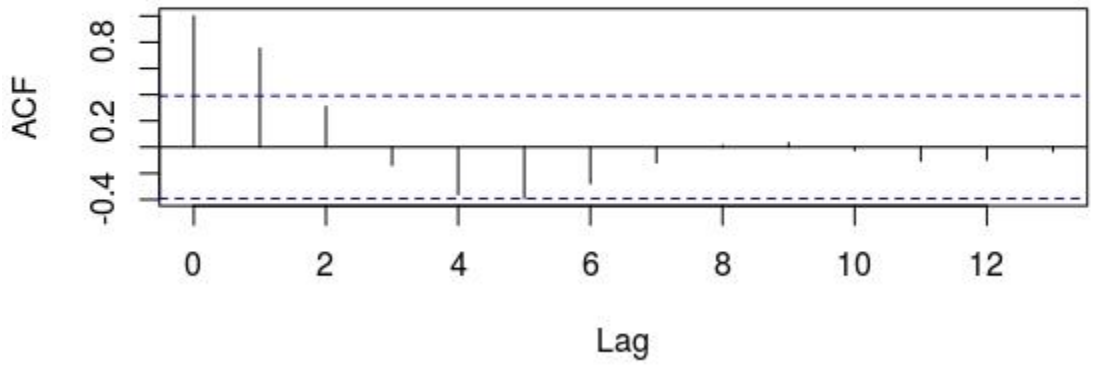

Figure 5.25: ACF plot of Lahore District for days 31 to 55

As the order 2 of AR terms, and order 3 of MA terms are computed, so the model that fits best is ARIMA (2, 0, 3) on these incidences and it is shown in equation (31).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 R_{t-3} \quad (26)$$

After the substitution of B_1, B_2, L_1, L_2, and L_3 coefficient values, equation (31) is replaced with equation (32).

$$H_t = 1.3687 H_{t-1} - 0.5573 H_{t-2} + R_t - 0.1638 R_{t-1} + 0.0184 R_{t-2} - 0.8545 R_{t-3} \quad (27)$$

The above equation (32) is used to predict 56, 57, 58, 59, and 60 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.26 shows the incidence of 36 to 60 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be

seen that the incidences that occur on days 1 to 4, 10 to 12, 15 to 18, and 23 to 34 show an increase in the trend while the incidences on the day 18 to 19, and 12 to 13 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

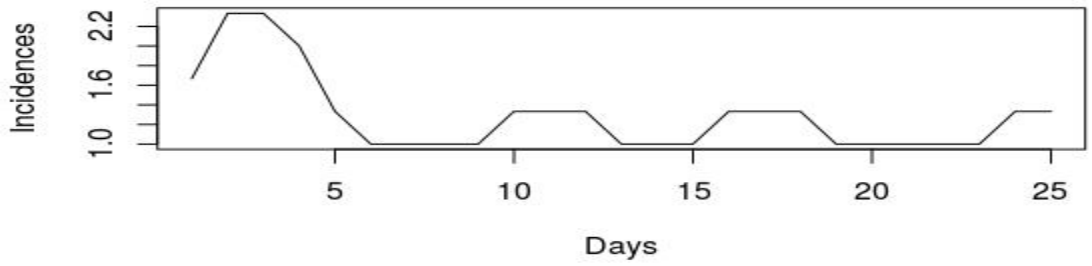

Figure 5.26: Plot of 36 to 60 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.27. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are 0.4039, and -0.1073. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

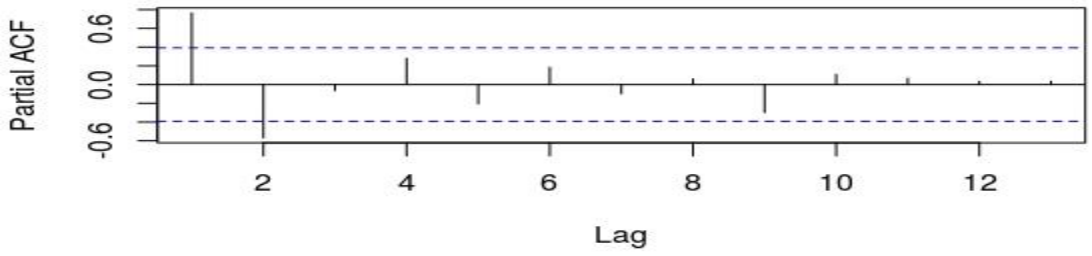

Figure 5.27: PACF plot of Lahore District for days 36 to 60

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.28. In this plot, lag values are represented by the x-axis while the coefficients of MA

terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 0.9184, and 1.0000.

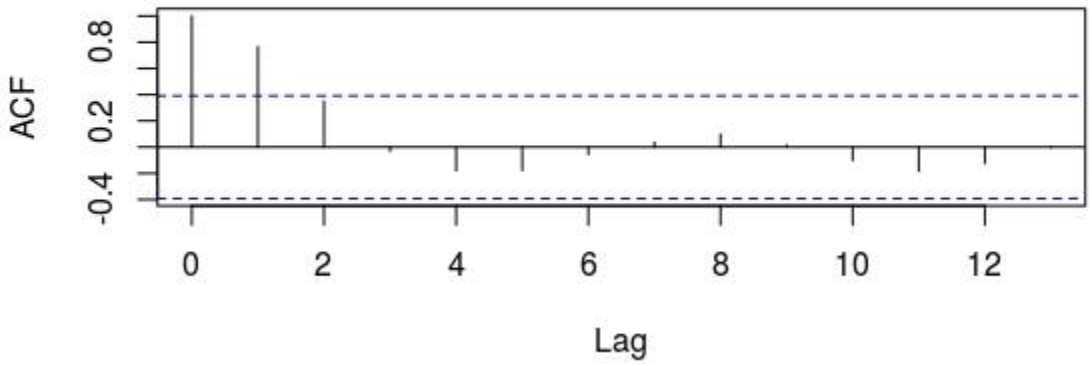

Figure 5.28: ACF plot of Lahore District for days 36 to 60

As the order 2 of AR terms, and order 2 of MA terms are computed, so the model that fits best is ARIMA (2, 0, 2) on these incidences and it is shown in equation (33).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (28)$$

After the substitution of B_1, B_2, L_1, and L_2 coefficient values, equation (33) is replaced with equation (34).

$$H_t = 0.4039 H_{t-1} - 0.1073 H_{t-2} + R_t + 0.9184 R_{t-1} + 1.0000 R_{t-2} \quad (29)$$

The above equation (34) is used to predict 61, 62, 63, 64, and 65 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.29 shows the incidence of 41 to 65 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 4 to 7, 11 to 13, 19 to 21, and 24 to 25 show an increase in the trend while the incidences on the day 8 to 10, 14 to 18, and 22 to 23 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this

stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

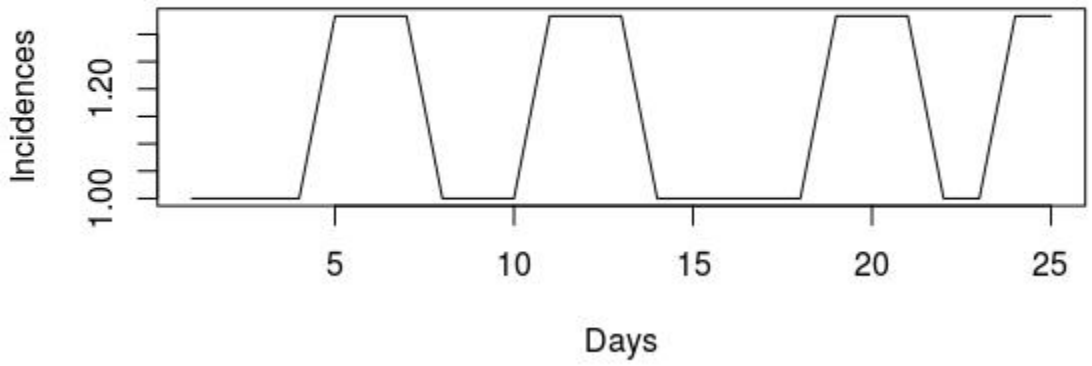

Figure 5.29: Plot of 41 to 65 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.30. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are 0.6361, and -0.5085. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

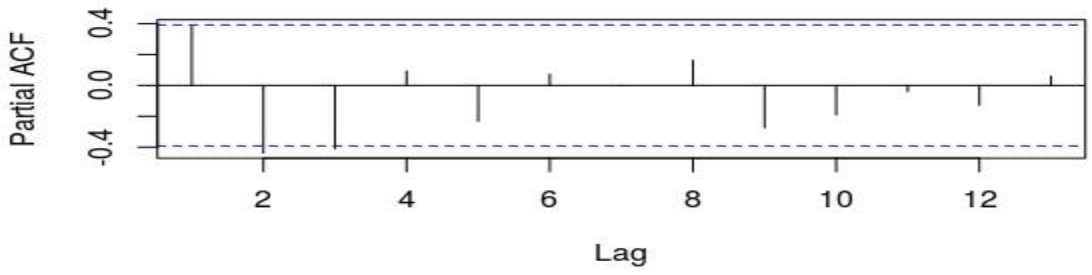

Figure 5.30: PACF plot of Lahore District for days 41 to 65

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.31. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are -0.3528, 0.1344 and -0.7816.

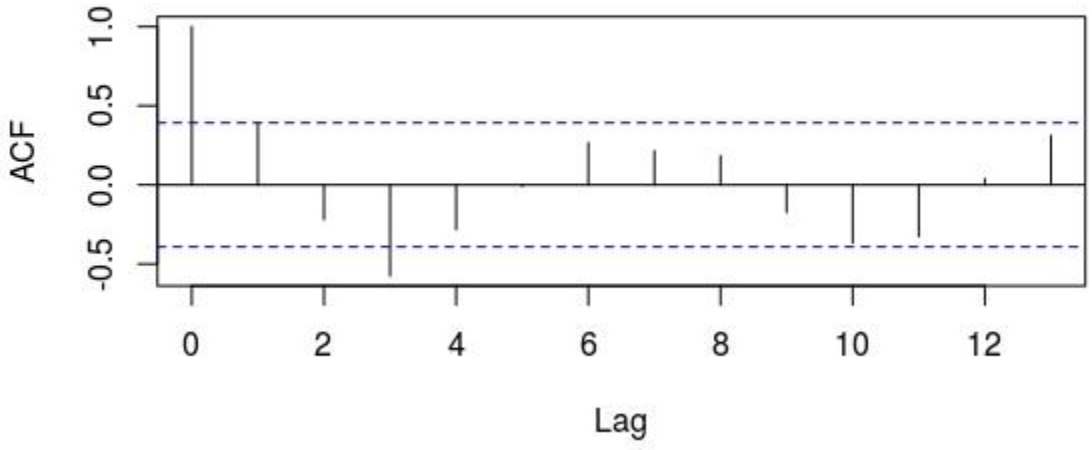

Figure 5.31: ACF plot of Lahore District for days 41 to 65

As the order 2 of AR terms, and order 3 of MA terms are computed, so the model that fits best is ARIMA (2, 0, 3) on these incidences and it is shown in equation (35).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 R_{t-3} \quad (30)$$

After the substitution of B_1, B_2, L_1, L_2, and L_3 coefficient values, equation (35) is replaced with equation (36).

$$H_t = 0.6361 H_{t-1} - 0.5085 H_{t-2} + R_t - 0.3528 R_{t-1} + 0.01344 R_{t-2} - 0.7816 R_{t-3} \quad (31)$$

The above equation (36) is used to predict 66, 67, 68, 69, and 70 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.32 shows the incidence of 46 to 70 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen

that the incidences that occur on days 1 to 2, 6 to 8, 14 to 16, 19 to 21 and 24 to 25 show an increase in the trend while the incidences on the day 3 to 5, 9 to 13, and 17 to 18 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

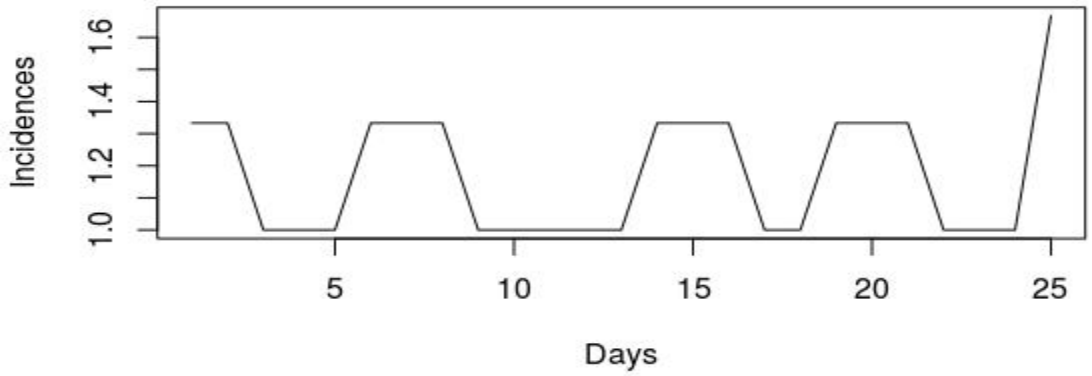

Figure 5.32: Plot of 46 to 70 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.33. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by one coefficients of the AR terms so its order is one and the value of the coefficient is 0.4411. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

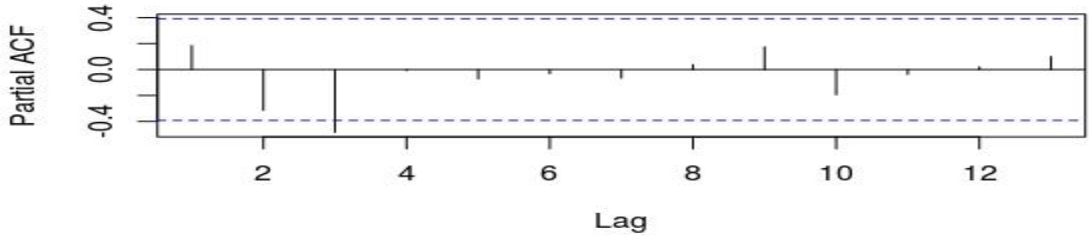

Figure 5.33: PACF plot of Lahore District for days 46 to 70

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.34. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are -0.5672, and -0.4328.

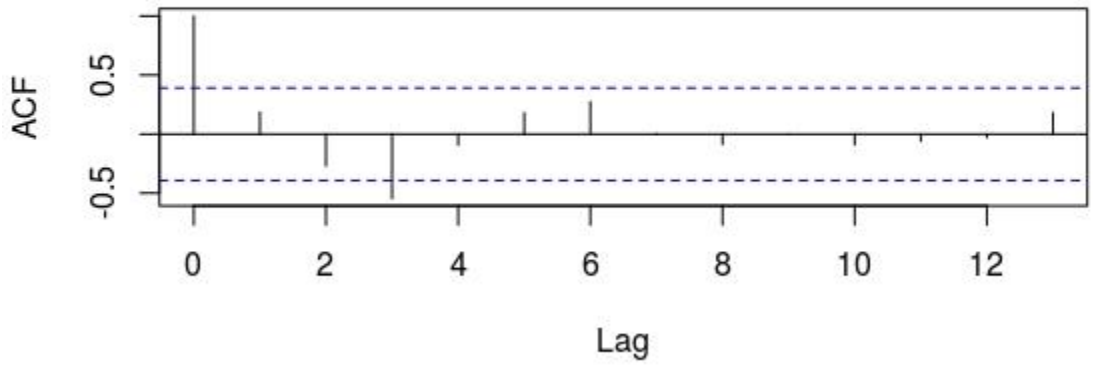

Figure 5.34: ACF plot of Lahore District for days 46 to 70

As the order 1 of AR terms, and order 2 of MA terms are computed, so the model that fits best is ARIMA (1, 0, 2) on these incidences and it is shown in equation 37.

$$H_t = B_1 H_{t-1} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (32)$$

After the substitution of B_1, L_1, and L_2 coefficient values, equation (37) is replaced with equation (38).

$$H_t = 0.4511 H_{t-1} + R_t - 0.5672 R_{t-1} - 0.5328 R_{t-2} \quad (33)$$

The above equation (38) is used to predict 71, 72, 73, 74, and 75 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.35 shows the incidence of 51 to 75 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 3, to 11, 14 to 16, 20 to 22, and 25 show an

increase in the trend while the incidences on the day 4 to 8, 12 to 13, 17 to 19, and 23 to 24 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

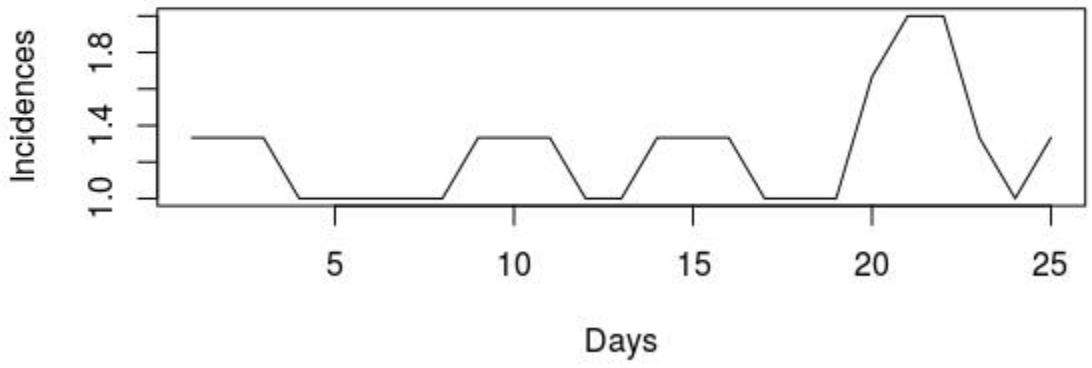

Figure 5.35: Plot of 51 to 75 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.36. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are 0.4048, and -0.5803. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

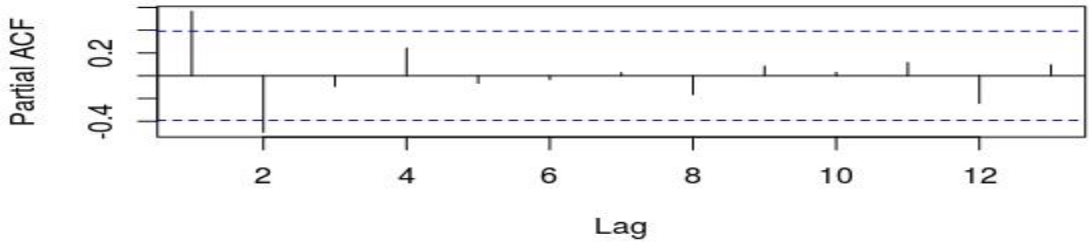

Figure 5.36: PACF plot of Lahore District for days 51 to 75

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.37. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 0.6838, and 1.0000.

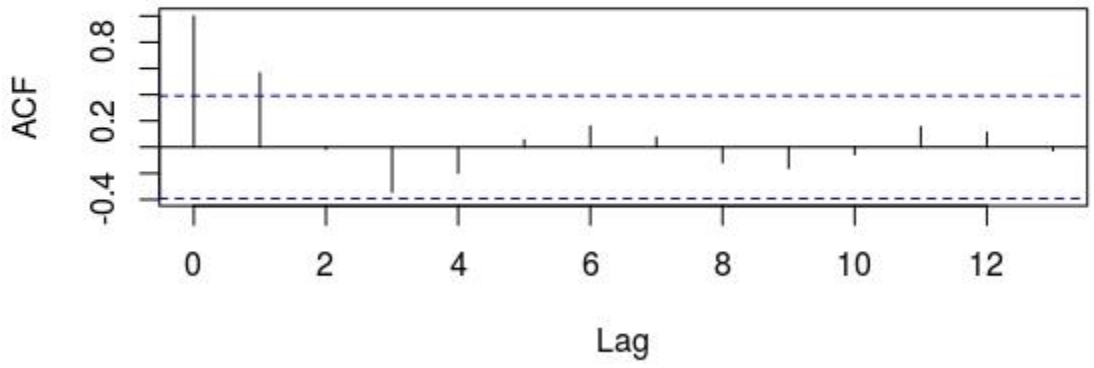

Figure 5.37: ACF plot of Lahore District for days 51 to 75

As the order 2 of AR terms, and order 2 of MA terms are computed, so the model that fits best is ARIMA (2, 0, 2) on these incidences and it is shown in equation (39).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (34)$$

After the substitution of B_1, B_2, L_1, and L_2 coefficient values, equation (39) is replaced with equation (40).

$$H_t = 0.4048 H_{t-1} - 0.5803 H_{t-2} + R_t + 0.6838 R_{t-1} + 1.0000 R_{t-2} \quad (35)$$

The above equation (40) is used to predict 76, 77, 78, 79, and 80 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.38 shows the incidence of 56 to 80 day's of the HCV patients in the Lahore District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 3 to 6, 9 to 11, 14 to 17, and 20 to 22 show an

increase in the trend while the incidences on the day 7 to 8, 18 to 19, and 23 to 25 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series

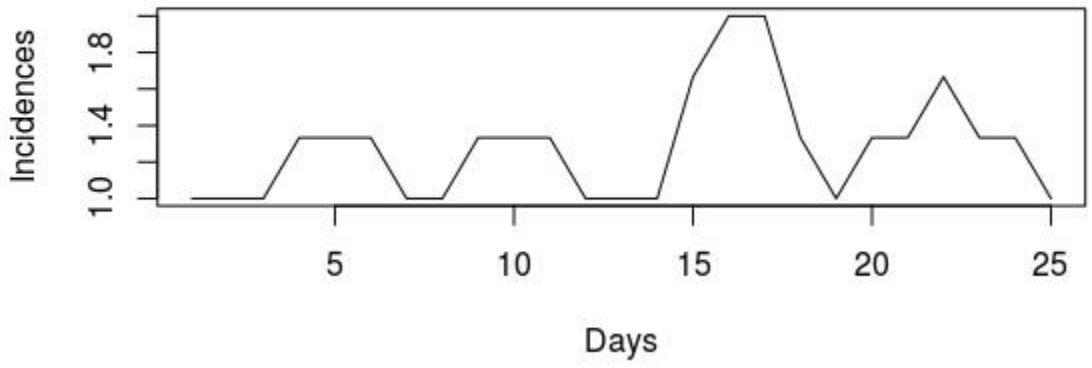

Figure 5.38: Plot of 56 to 80 day's incidences of the Lahore District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.39. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by three coefficients of the AR terms so its order is three and the value of these coefficients are 0.6835, -0.7488, and -0.2641. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

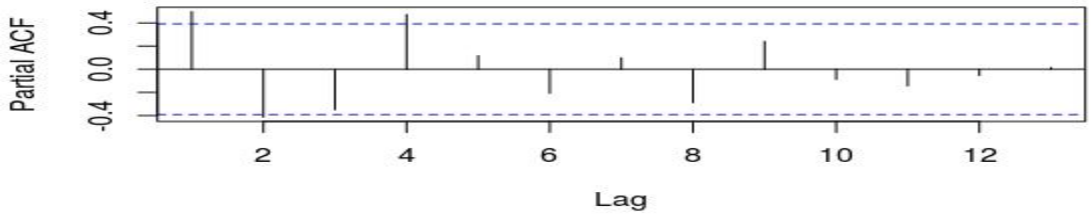

Figure 5.39: PACF plot of Lahore District for days 56 to 80

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.40. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by four coefficients of the MA terms so its order is four and the values of these coefficients are 0.0914, 1.0669, 0.0848, and 0.9925.

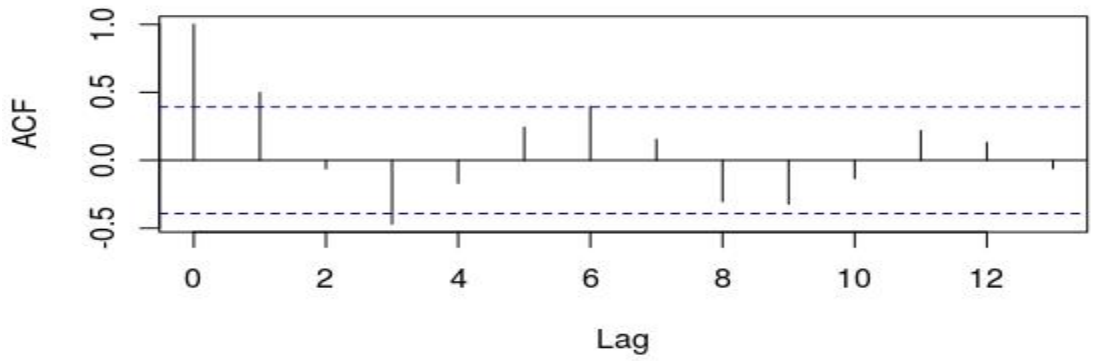

Figure 5.40: ACF plot of Lahore District for days 56 to 80

As the order 3 of AR terms, and order 4 of MA terms are computed, so the model that fits best is ARIMA (3, 0, 4) on these incidences and it is shown in equation (41).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + B_3 H_{t-3} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 H_{t-3} + L_4 H_{t-4} \quad (36)$$

After the substitution of B_1, B_2, B_3, L_1, L_2, and L_3 coefficient values, equation (41) is replaced with equation (42).

$$H_t = 0.6835 H_{t-1} - 0.7488 H_{t-2} - 0.2641 H_{t-3} + R_t + 0.0914 R_{t-1} + 1.0669 R_{t-2} + 0.0848 H_{t-3} + 0.9925 H_{t-4} \quad (37)$$

The above equation (42) is used to predict 81, 82, 83, and 84 day incidences using the ARIMA model by adding +1 to t value each time.

Figure 5.41 shows the plot of actual and predicted incidences from 26 to 84 days. In this plot, x-axis shows the number of the days while y-axis shows the Hepatitis C incidences that took place in Lahore district. This plot indicates that when the trend increase in real incidences is

provided as an input to the model, the trend of the incidences that are predicted also increases. On the other hand, giving the decreasing trend to model as input also reduces the trend of predicted incidences. This is because the predictive incidences trend in this model depends on the incidences that are provided to it as input. As the hepatitis C threshold is 3 to 6 incidences in 2 to 6 weeks. This plot shows the incidences of two weeks from day 26 to 39 while from day 40 to 81, six weeks incidences are shown. During this period, the number of incidences predicted by our model is 68.1134659. In this plot, predicted incidences have crossed the threshold so an alert is sent to the DG Health in the form of an SMS that Hepatitis C incidences have crossed the threshold in Lahore District of Punjab Province and also this predicted trend is sent to the central Dashboard of the health ministry. The DG Health analyzes this predictive trend and also assesses the current health situation, including the budget allocations for health facilities so far. The spread of the disease can be controlled in time or not, suggests increasing and decreasing the budget according to the current situation and also informs the Prime Minister, Health Minister and Health Secretary about these conditions.

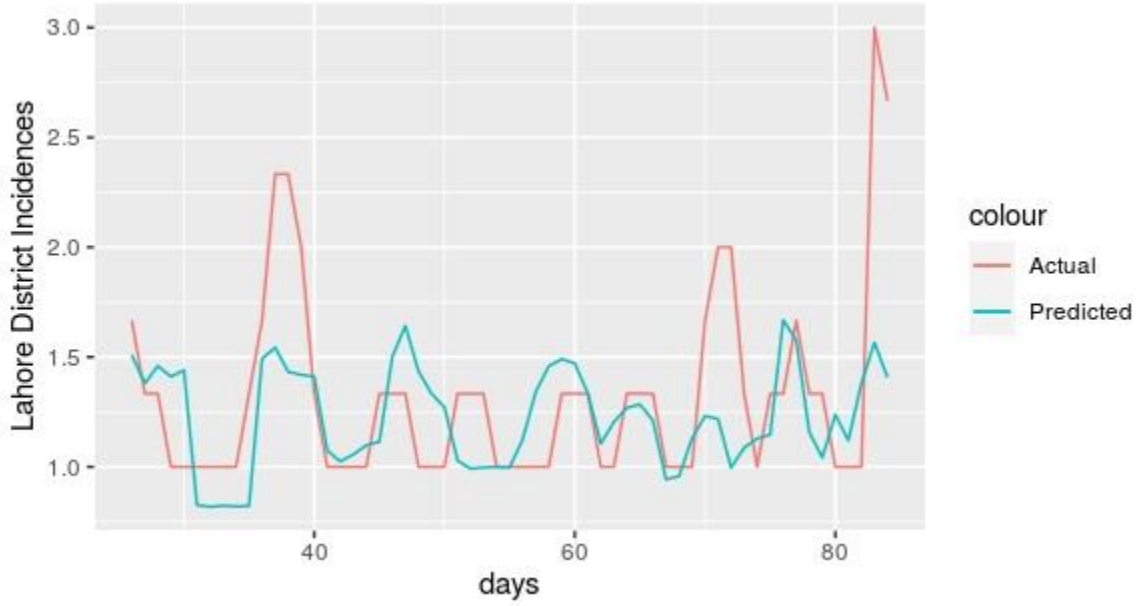

Figure 5.41: Lahore District actual and predicted incidences

The errors found in the 12 iterations in the actual and predicted incidences are shown in Figure 5.42, the error values are mentioned in Table 3, while the mapping that is done on the

model variables of the predicted day incidences is shown in Table 4. This plot in Figure 6 indicates that MAE and RMSE values in all iterations within the 0-2 range. Therefore, the predictions made by this model can be used to prevent the spread of the disease.

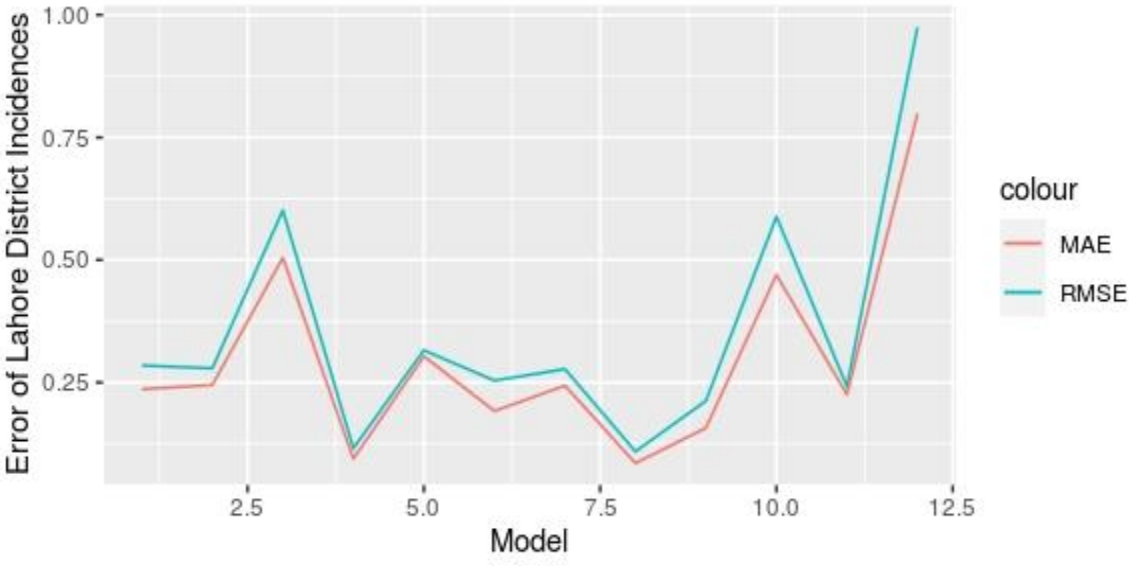

Figure 5.42: Error of Lahore District incidences

Table 5-1: Error values of Lahore District incidences

Model	MAE	RMSE
1	0.236198	0.284768
2	0.244899	0.278796
3	0.504621	0.601339
4	0.094456	0.115636
5	0.303483	0.315654
6	0.191684	0.253856
7	0.243591	0.277371

8	0.085374	0.109038
9	0.157218	0.212449
10	0.46942	0.588511
11	0.225521	0.241543
12	0.798496	0.974826

Table 5-2: Mapping of predicted day incidences to model variable

Predicted day incidences	**Model**
26-30	1
31-35	2
36-40	3
41-45	4
46-50	5
51-55	6
56-60	7
61-65	8
66-70	9
71-75	10
76-80	11

81-84	12

We have selected 28, the minimum number of days of HCV incidence for the Vehari District. During this selection process, we do not use the incidences of the first three days due to the absence of variance within it. The incidences of the day 4,5,6,7,8,9,10,11,12,14,19 are rejected due to non-maintenance of the stationary element while the incidence at 11,13,15,17,18,20,21,23,24,25 days are rejected based on negative out of range prediction.

The plot of Figure 5.43 shows the incidence of 1 to 28 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 6, 10 to 15, and 24 to 27 show an increase in the trend while the incidences on the day 7 to 9, 16 to 23, and 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

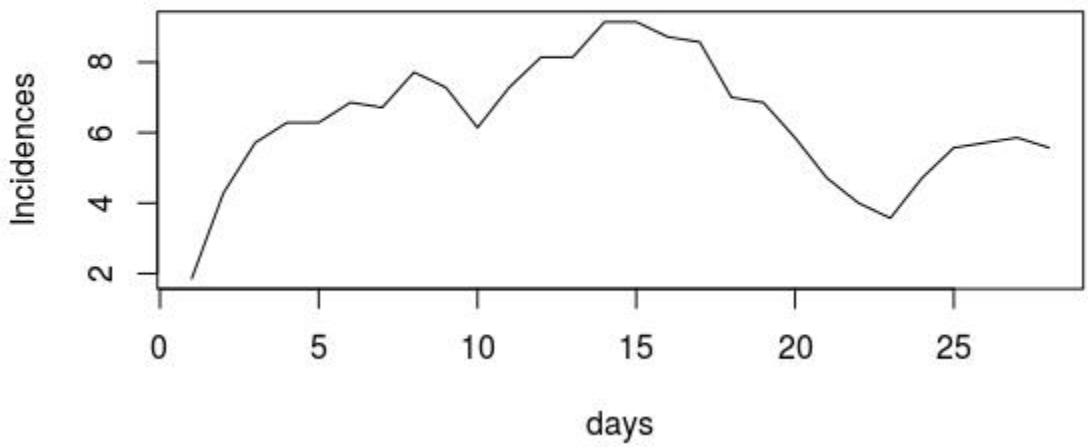

Figure 5.43: Plot of 1 to 28 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.6077`. It is made in to stationary by differentiating it thrice and in this case the p-value of the ADF test is 0.02041 which indicates the stationary presence. The plot of the stationary time series of 1 to 28

day incidences after making a difference three times is shown in Figure 5.44. This plot also shows the constant mean and variance in this time series.

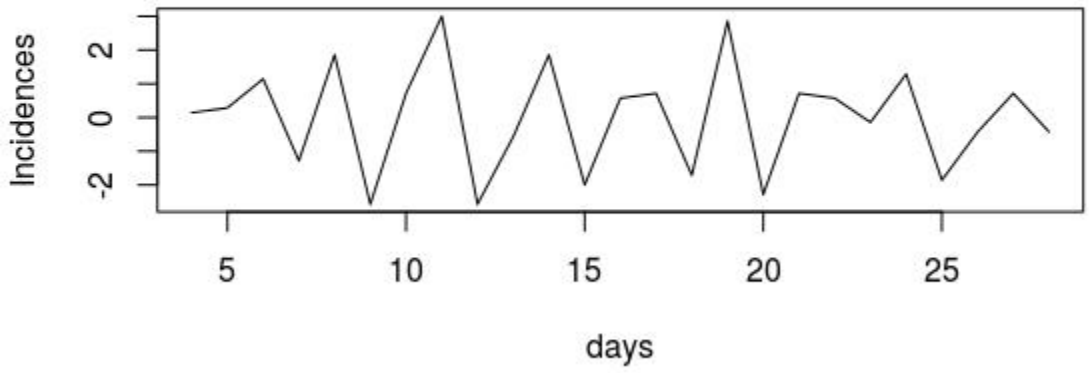

Figure 5.44: Stationary time Series plot of 1 to 28 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.45. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are -0.0023, and 0.5007. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

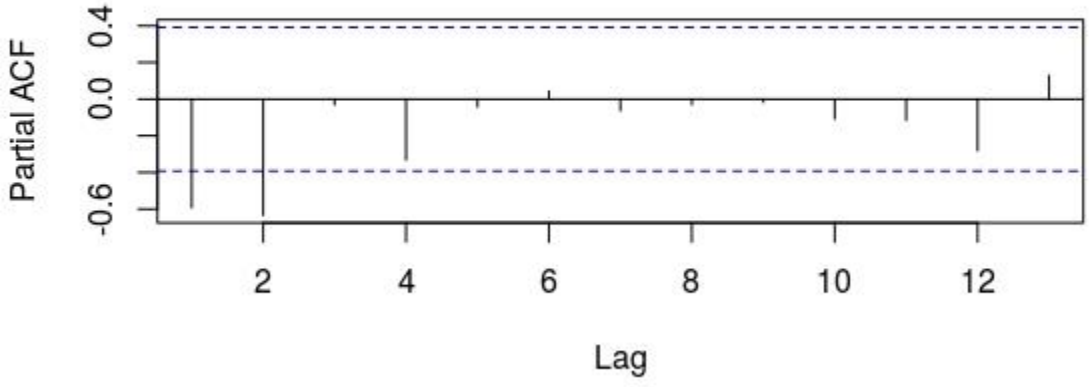

Figure 5.45: PACF plot of Vehari District for days 1 to 28

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.46. In this plot, lag values are represented by the x-axis while the coefficients of MA

terms are represented by the y-axis. The threshold is crossed by four coefficients of the MA terms so its order is four and the values of these coefficients are -1.3014, -0.6355, 1.3265, and -0.3394.

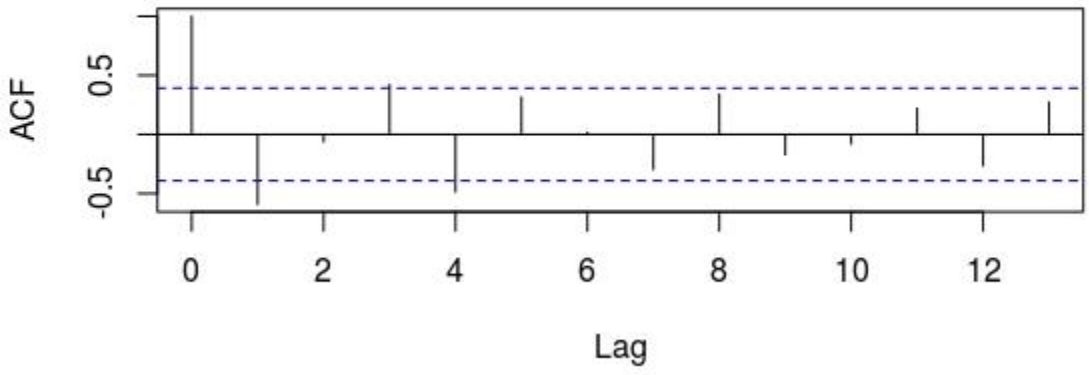

Figure 5.46: ACF plot of Vehari District for days 1 to 28

As the order 2 of AR terms, order 4 of MA terms, and order 3 of difference terms are computed, so the model that fits best is ARIMA (2, 3, 4) on these incidences and it is shown in equation (43).

$$H_t = -(-3 - B_1)H_{t-1} - (-3 + 3B_1 - 3B_2)H_{t-2} - (-1 - 3B_1 + 3B_2)H_{t-3} - (B_1 - 3B_2)H_{t-4} - B_2H_{t-5} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} + L_4R_{t-4} \quad (38)$$

After the substitution of B_1, B_2, L_1, L_2, L_3, and L_4 coefficient values, equation (43) is replaced with equation (44).

$$H_t = 2.9977H_{t-1} - 1.491H_{t-2} - 0.5044H_{t-3} + 1.5044H_{t-4} - 0.5007H_{t-5} + R_t - 1.3014R_{t-1} - 0.6355R_{t-2} + 1.3265R_{t-3} - 0.3394R_{t-4} \quad (39)$$

The above equation (44) is used to predict 29, 30, 31, 32 and 33 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.47 shows the incidence of 6 to 33 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 3, 6 to 10, and 19 to 22 show an increase in the

trend while the incidences on the day 4 to 5, 11 to 18, and 23 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

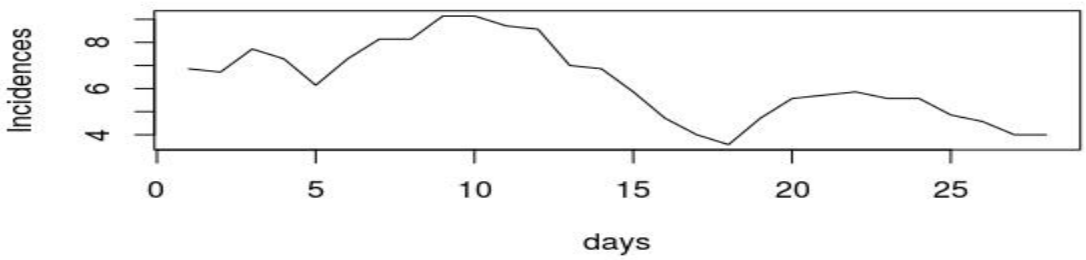

Figure 5.47: Plot of 6 to 33 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.1045`. It is made in to stationary by differentiating it thrice and in this case the p-value of the ADF test is 0.01709 which indicates the stationary presence. The plot of the stationary time series of 6 to 33 day incidences after making a difference three times is shown in Figure 5.48. This plot also shows the constant mean and variance in this time series.

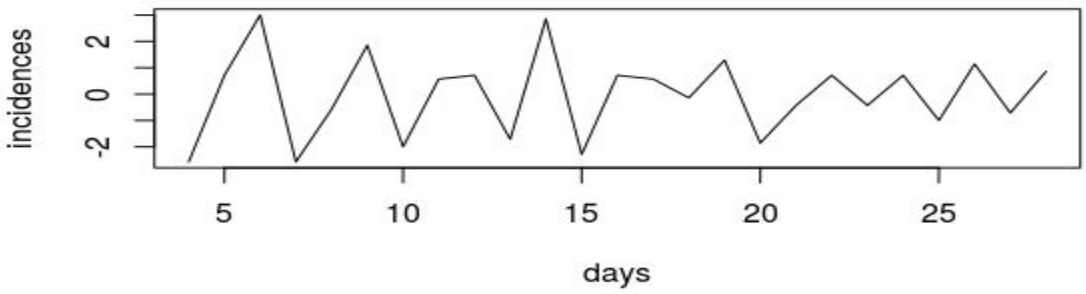

Figure 5.48: Stationary time series plot of 6 to 33 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.49. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are -1.0434, and -0.5461. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

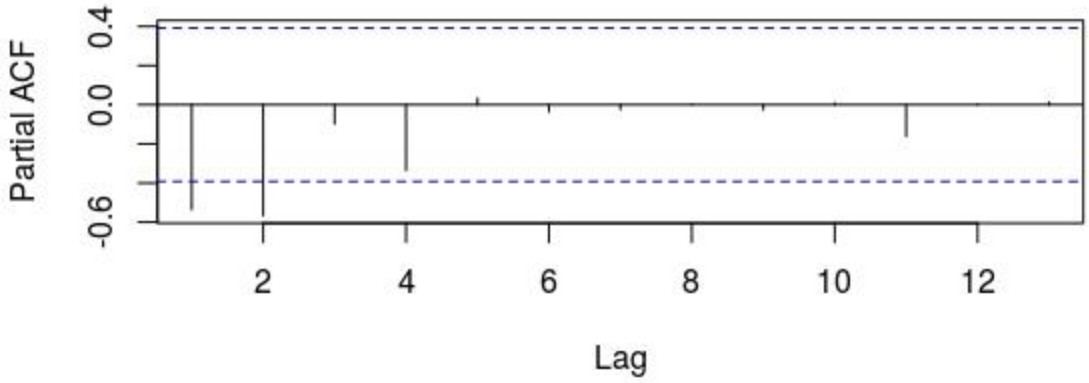

Figure 5.49: PACF plot of Vehari District for days 6 to 33

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.50. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are -0.3401, and -0.6599.

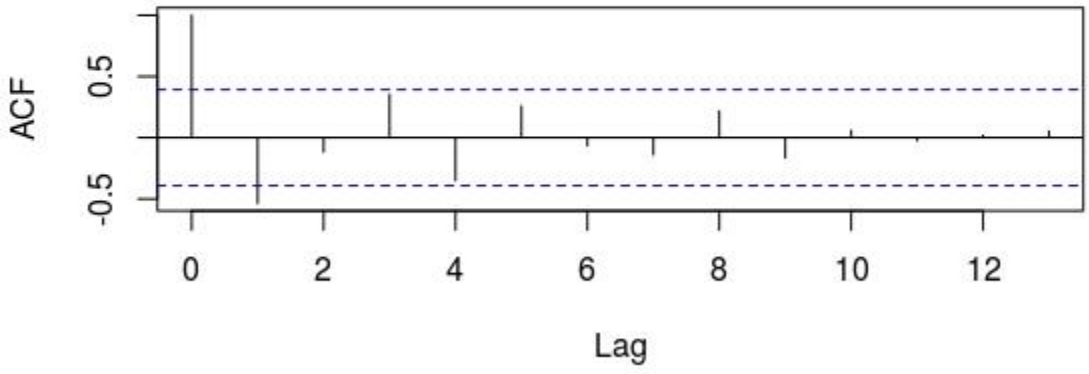

Figure 5.50: ACF plot of Vehari District for days 6 to 33

As the order 2 of AR terms, order 2 of MA terms, and order 3 of difference terms are computed, so the model that fits best is ARIMA (2, 3, 2) on these incidences and it is shown in equation (45).

$$H_t = -(-3 - B_1)H_{t-1} - (-3 + 3B_1 - 3B_2)H_{t-2} - (-1 - 3B_1 + 3B_2)H_{t-3} \quad (40)$$
$$- (B_1 - 3B_2)H_{t-4} - B_2H_{t-5} + R_t + L_1R_{t-1} + L_2R_{t-2}$$

After the substitution of B_1, B_2, L_1, and L_2 coefficient values, equation (45) is replaced with equation (46).

$$H_t = 1.9566H_{t-1} - 1.5081H_{t-2} - 0.4919H_{t-3} - 0.5949H_{t-4} + 0.5461H_{t-5} + R_t \quad (41)$$
$$- 0.3401R_{t-1} - 0.6599R_{t-2}$$

The above equation (46) is used to predict 34, 35, 36, 37 and 38 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.51 shows the incidence of 11 to 38 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 5, 14 to 17, and 24 to 27 show an increase in the trend while the incidences on the day 6 to 13, 18 to 23, and 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

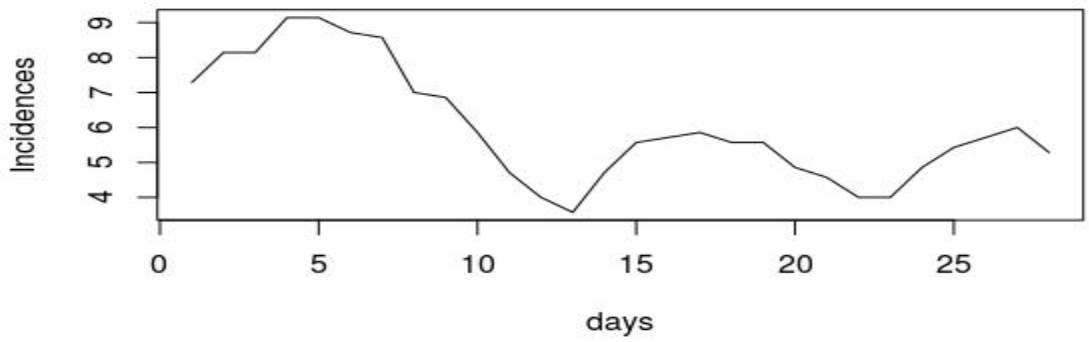

Figure 5.51: Plot of 11 to 38 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.2574.` It is made in to stationary by differentiating it thrice and in this case the p-value of the ADF test is 0.03834 which indicates the stationary presence. The plot of the stationary time series of 11 to 38 day incidences after making a difference three times is shown in Figure 5.52. This plot also shows the constant mean and variance in this time series.

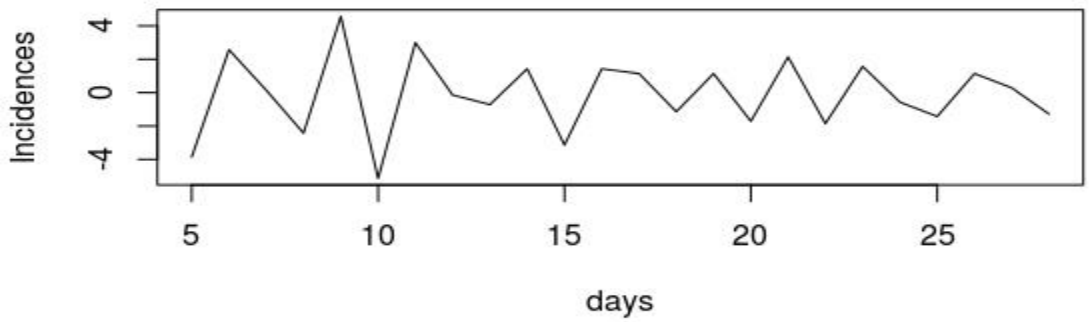

Figure 5.52: Stationary time series plot of 11 to 38 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.53. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are -1.6235, and -0.9847. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

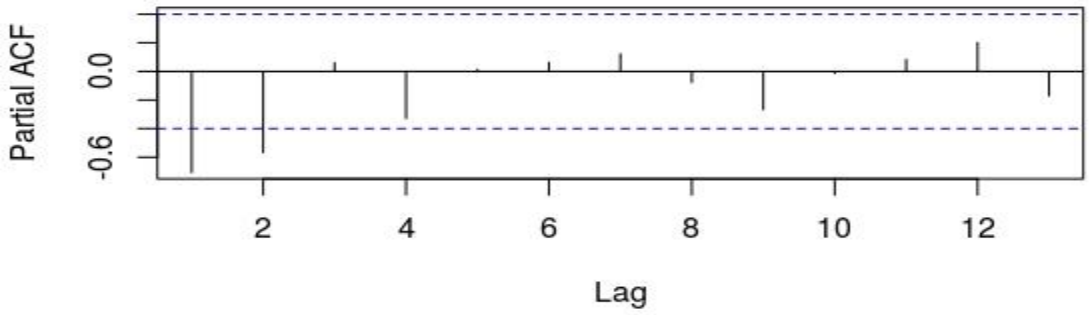

Figure 5.53: PACF plot of Vehari District for days 11 to 38

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.54. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by four coefficients of the MA terms so its order is four and the values of these coefficients are -0.3254, -1.0217, -0.3028, and 0.6838.

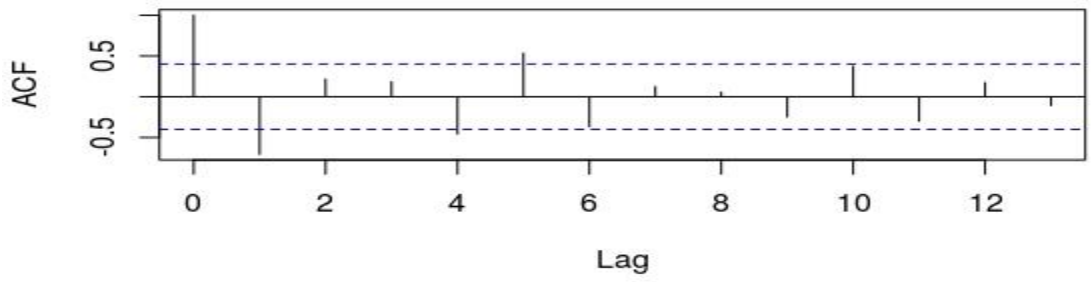

Figure 5.54: ACF plot of Vehari District for days 11 to 38

As the order 2 of AR terms, order 4 of MA terms, and order 3 of difference terms are computed, so the model that fits best is ARIMA (2, 3, 4) on these incidences and it is shown in equation (47).

$$H_t = -(-3 - B_1)H_{t-1} - (-3 + 3B_1 - 3B_2)H_{t-2} - (-1 - 3B_1 + 3B_2)H_{t-3} \quad (42)$$
$$- (B_1 - 3B_2)H_{t-4} - B_2H_{t-5} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3}$$
$$+ L_4R_{t-4}$$

After the substitution of B_1, B_2, L_1, L_2, L_3, and L_4 coefficient values, equation (47) is replaced with equation (48).

$$H_t = 1.3765H_{t-1} + 4.9164H_{t-2} - 0.9164H_{t-3} - 4.5776H_{t-4} + 0.9847H_{t-5} + R_t \quad (43)$$
$$- 0.3254R_{t-1} - 1.0217R_{t-2} - 0.3028R_{t-3} + 0.6838R_{t-4}$$

The above equation (48) is used to predict 39, 40, 41, 42 and 43 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.55 shows the incidence of 16 to 43 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 9 to 12, and 19 to 22 show an increase in the trend while the incidences on the day 1 to 8, and 13 to 18, and 24 to 28 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

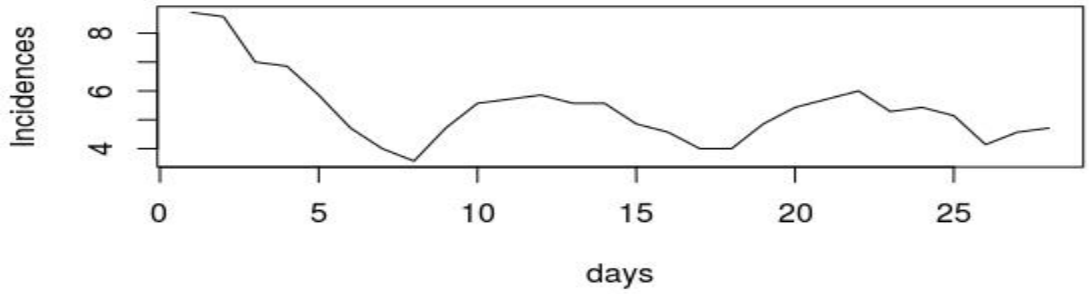

Figure 5.55: Plot of 16 to 43 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.56. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by one coefficient of the AR terms so its order is one and the value of this coefficient is 0.0825. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

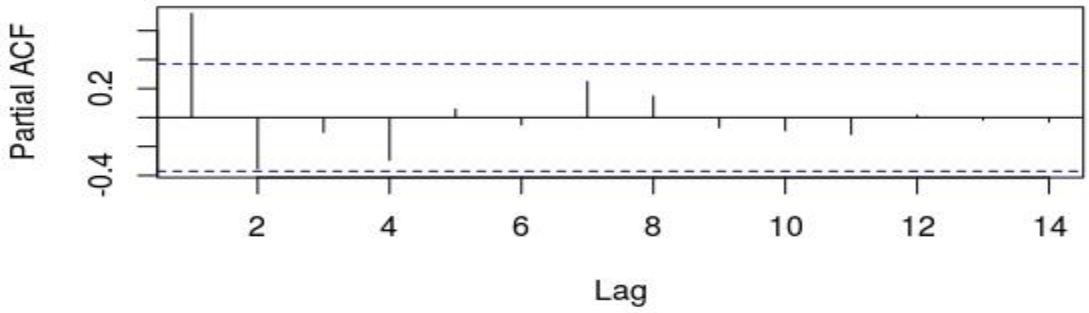

Figure 5.56: PACF plot of Vehari District for days 16 to 43

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.57. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are 1.2927, 1.3888, and 0.8158.

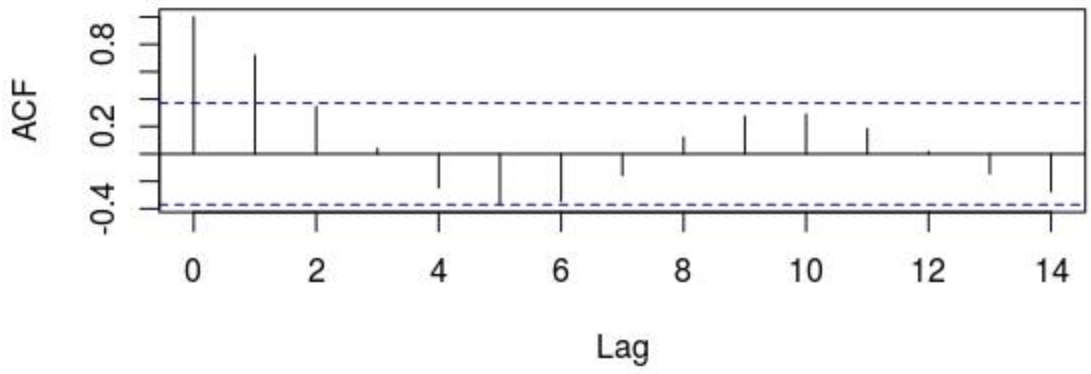

Figure 5.57: ACF plot of Vehari District for days 16 to 43

As the order 1 of AR terms, and order 3 of MA terms are computed, so the model that fits best is ARIMA (1, 0, 3) on these incidences and it is shown in equation (49).

$$H_t = B_1 H_{t-1} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 H_{t-3} \quad (44)$$

After the substitution of B_1, L_1, L_2, and L_3 coefficient values, equation (49) is replaced with equation (50).

$$H_t = 0.0825 H_{t-1} + R_t + 1.2927 R_{t-1} + 1.3888 R_{t-2} + 0.8158 H_{t-3} \quad (45)$$

The above equation (50) is used to predict 44, 45, 46, 47 and 48 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.58 shows the incidence of 21 to 48 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 4 to 7, 14 to 19, and 22 to 23 show an increase in the trend while the incidences on the day 1 to 3, 8 to 13, 20 to 21, and 25 to 28 show a decrease in the trend. It also shows the stationary element in the time series. In the case of this stationary time series, the p value of the ADF test is 0.01. This plot also shows the constant mean and variance in this time series.

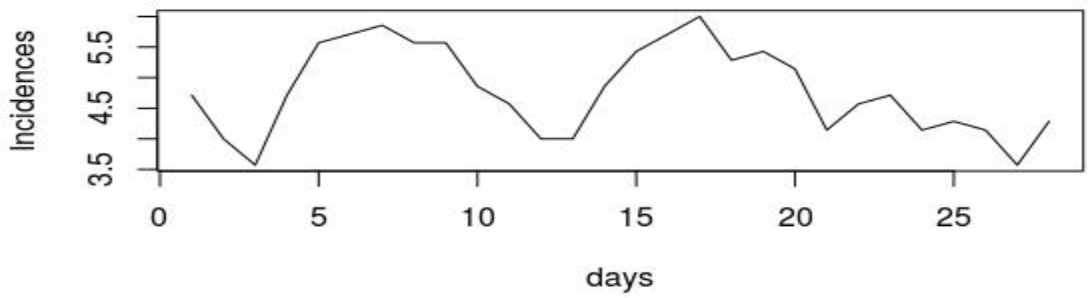

Figure 5.58: Plot of 21 to 48 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.59. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by three coefficients of the AR terms so its order is three and the value of these coefficients are 1.3484, -0.7916, and -0.0740. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

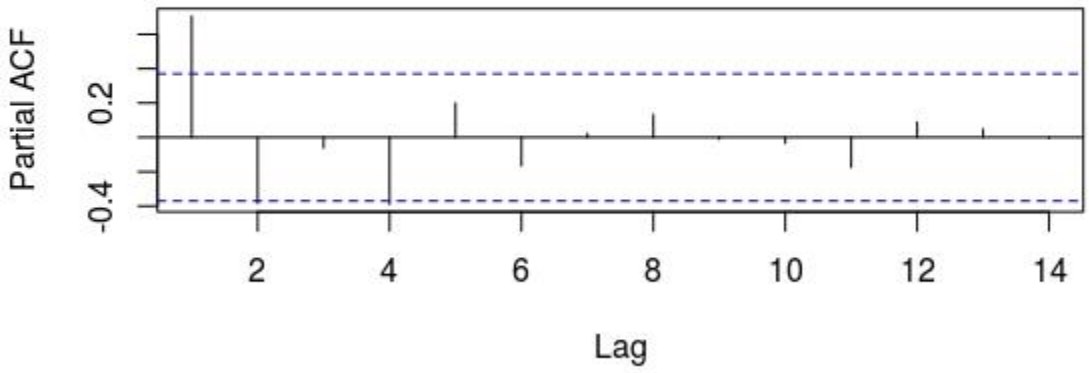

Figure 5.59: PACF plot of Vehari District for days 21 to 48

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.60. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are -0.5286, 0.0479, and 0.7465.

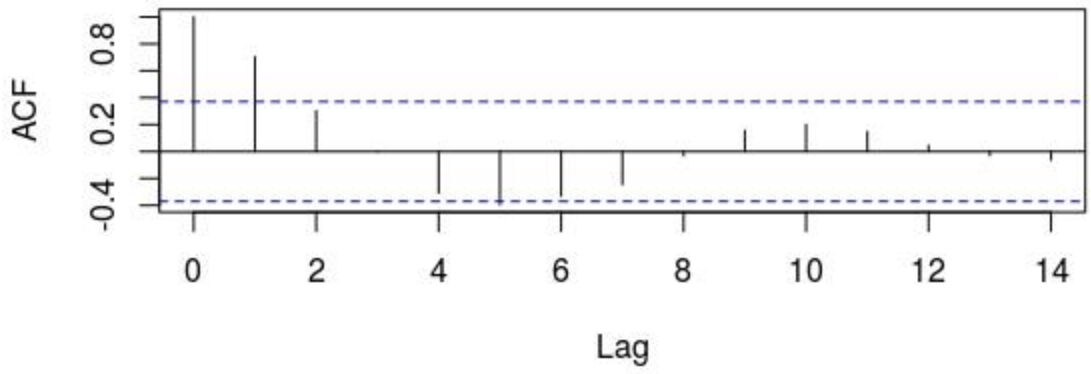

Figure 5.60: ACF plot of Vehari District for days 21 to 48

As the order 3 of AR terms, and order 3 of MA terms are computed, so the model that fits best is ARIMA (3, 0, 3) on these incidences and it is shown in equation (51).

$$H_t = B_1 H_{t-1} + B_2 H_{t-2} + B_3 H_{t-3} + R_t + L_1 R_{t-1} + L_2 R_{t-2} + L_3 H_{t-3} \quad (46)$$

After the substitution of B_1, B_2, B_3, L_1, L_2, and L_3 coefficient values, equation (51) is replaced with equation (52).

$$H_t = 1.3484 H_{t-1} - 0.7916 H_{t-2} - 0.0740 H_{t-3} + R_t - 0.5286 R_{t-1} + 0.0479 R_{t-2} + 0.7465 H_{t-3} \quad (47)$$

The above equation (52) is used to predict 49, 50, 51, 52 and 53 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.61 shows the incidence of 26 to 53 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, 9 to 14, and 17 to 18 show an increase in the trend while the incidences on the day 3 to 8, 16, and 19 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

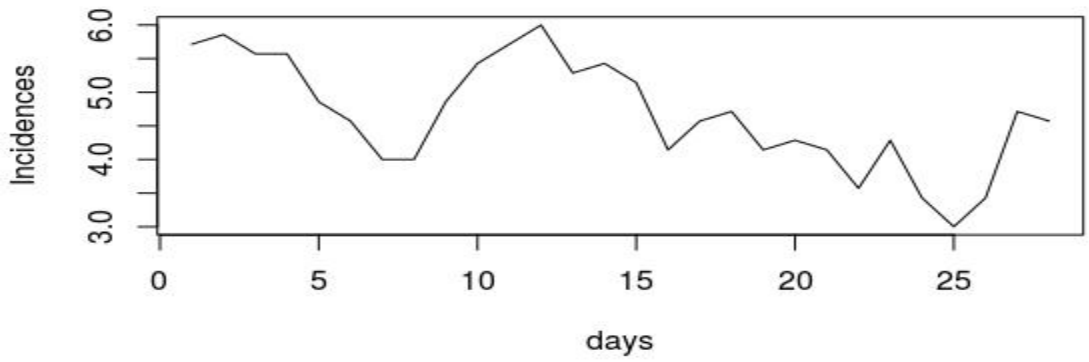

Figure 5.61: Plot of 26 to 53 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.1831.` It is made in to stationary by differentiating it twice and in this case the p-value of the ADF test is 0.0463 which indicates the stationary presence. The plot of the stationary time series of 26 to 53 day incidences after making a difference two times is shown in Figure 5.62. This plot also shows the constant mean and variance in this time series.

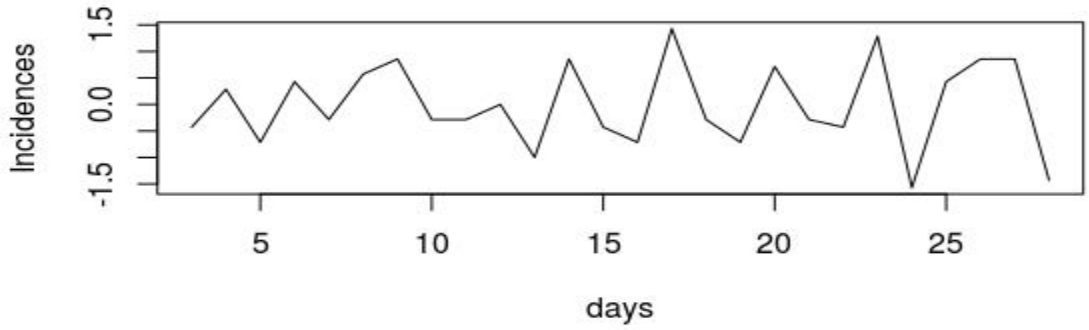

Figure 5.62: Stationary time series plot of 26 to 53 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.63. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is crossed by two coefficients of the AR terms so its order is two and the value of these coefficients are -0.458, and -0.1367. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

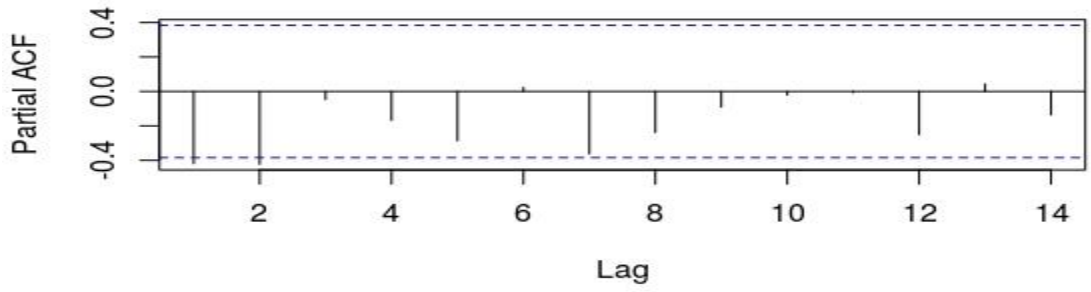

Figure 5.63: PACF plot of Vehari District for days 26 to 53

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.64. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are -0.4860, and -0.5139.

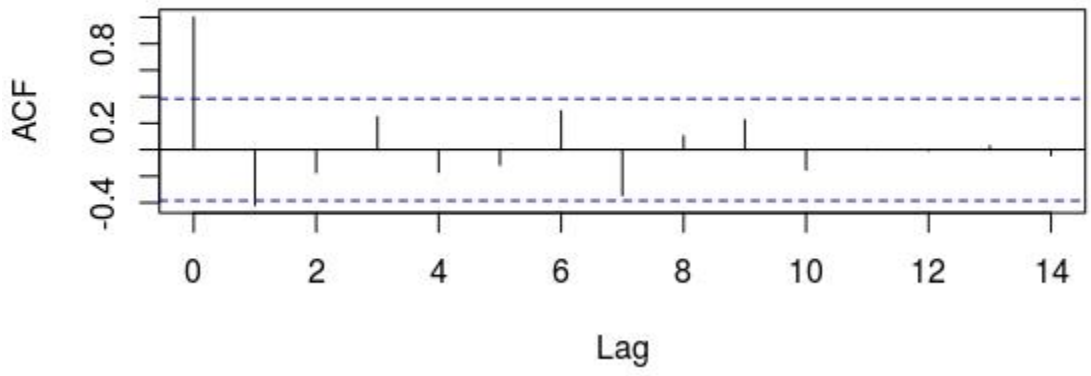

Figure 5.64: ACF plot of Vehari District for days 26 to 53

As the order 2 of AR terms, order 2 of MA terms, and order 2 of the difference terms are computed, so the model that fits best is ARIMA (2, 2, 2) on these incidences and it is shown in equation (53).

$$H_t = -(-2 - B_1)H_{t-1} - (1 + 2B_1 - B_2)H_{t-2} - (-B_1 + 2B_2)H_{t-3} + B_2H_{t-4} + R_t + L_1R_{t-1} + L_2R_{t-2} \quad (48)$$

After the substitution of B_1, B_2, L_1, and L_2 coefficient values, equation (44) is replaced with equation (54).

$$H_t = 1.542H_{t-1} - 2.0527H_{t-2} - 0.1846H_{t-3} - 0.1367H_{t-4} + R_t - 0.4860R_{t-1} - 0.5139R_{t-2} \quad (49)$$

The above equation (54) is used to predict 54, 55, 56, 57 and 58 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.65 shows the incidence of 31 to 58 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 4 to 7 show an increase in the trend while the incidences on the day 1 to 3, and 10 to 26 show a decrease in the trend. It also shows the non-stationary element in the time series.

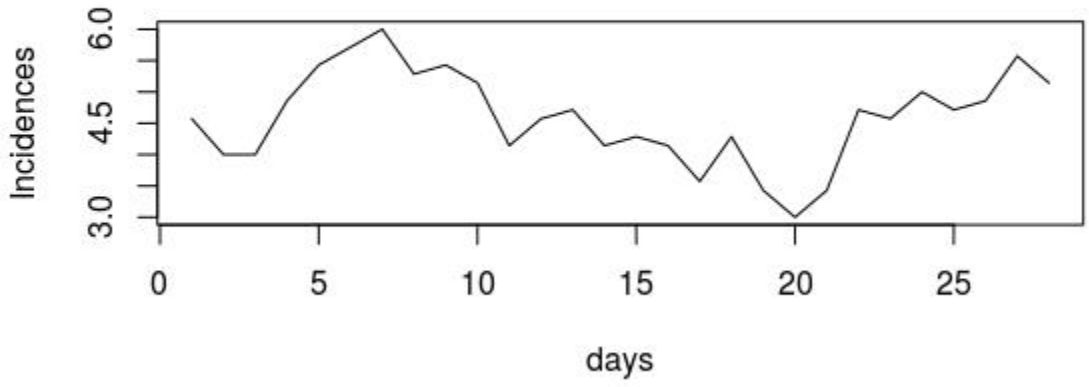

Figure 5.65: Plot of 31 to 58 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.5963.` It is made in to stationary by differentiating it twice and in this case the p-value of the ADF test is 0.01264 which indicates the stationary presence. The plot of the stationary time series of 31 to 58 day incidences after making a difference two times is shown in Figure 5.66. This plot also shows the constant mean and variance in this time series.

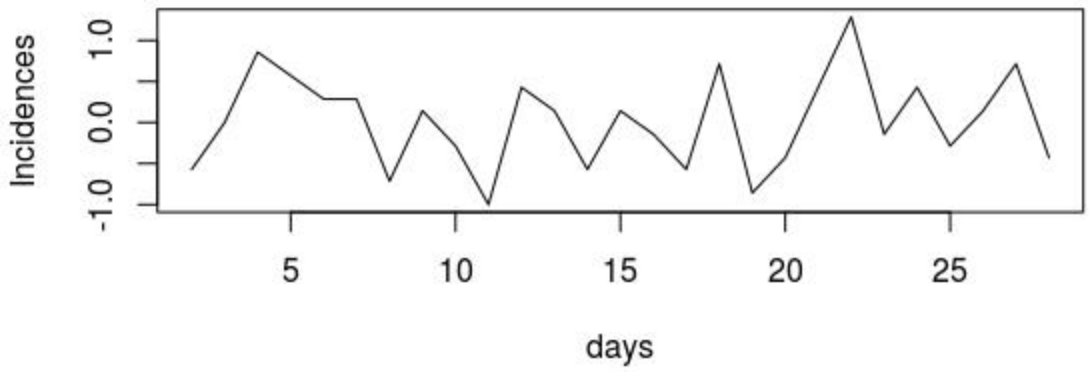

Figure 5.66: Stationary time series plot of 31 to 58 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.67. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold is not crossed by any coefficients of the AR terms so its order is zero. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

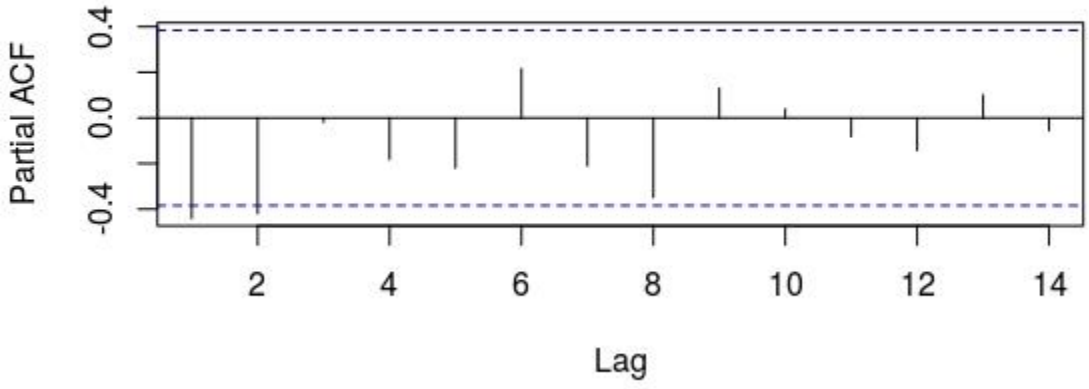

Figure 5.67: PACF plot of Vehari District for days 31 to 58

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.68. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are -1.0209, -0.1325, and 0.1534.

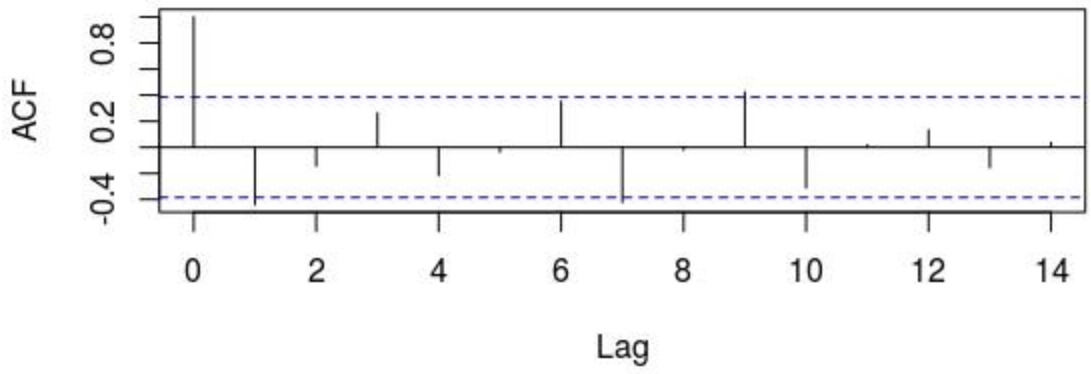

Figure 5.68: ACF plot of Vehari District for days 31 to 58

As the order 0 of AR terms, order 3 of MA terms, and order 2 of the difference terms are computed, so the model that fits best is ARIMA (0, 2, 3) on these incidences and it is shown in equation (55).

$$H_t = 2H_{t-1} - H_{t-2} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} \quad (50)$$

After the substitution of L_1, L_2, and L_3 coefficient values, equation (55) is replaced with equation (56).

$$H_t = 2H_{t-1} - H_{t-2} + R_t - 1.0209R_{t-1} - 0.1325R_{t-2} + 0.1534R_{t-3} \quad (51)$$

The above equation (56) is used to predict 59, 60, 61, 62 and 63 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.69 shows the incidence of 36 to 63 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, 3 to 4, 7 to 8, 9 to 10, 16 to 17, 18 to 19, and 20 to 22 show an increase in the trend while the incidences on the day 5 to 6, 11 to 12, 13 to 15, 23 to 24, 25 to 26, and 27 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

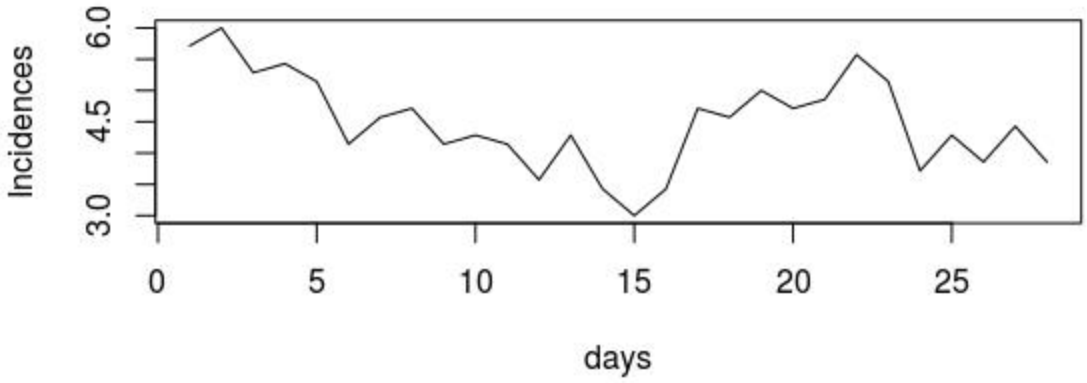

Figure 5.69: Plot of 36 to 63 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.4915.` It is made in to stationary by differentiating it twice and in this case the p-value of the ADF test is 0.01 which indicates the stationary presence. The plot of the stationary time series of 36 to 63 day incidences after making a difference two times is shown in Figure 5.70. This plot also shows the constant mean and variance in this time series.

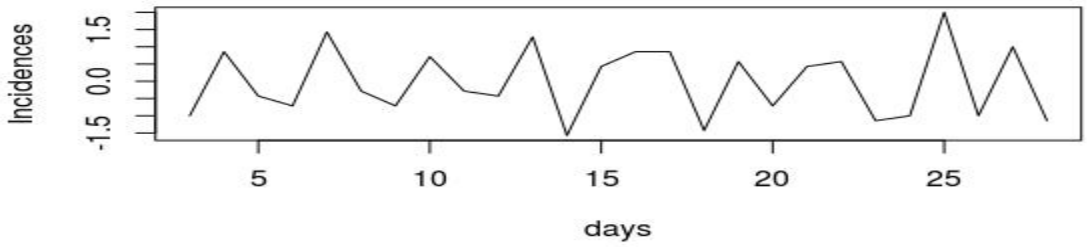

Figure 5.70: Stationary time series plot of 36 to 63 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.71. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold crossed by two coefficients of the AR terms so its order is two and the values of these coefficients are -1.5434, and -0.5809. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

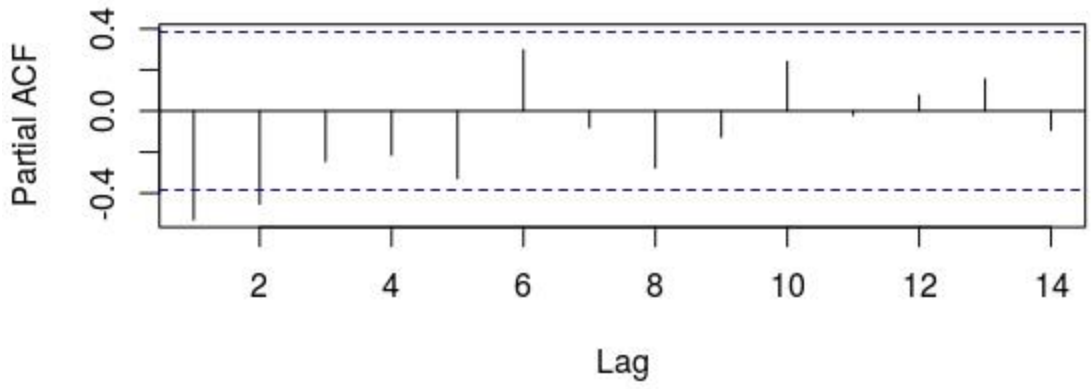

Figure 5.71: PACF plot of Vehari District for days 36 to 63

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.72. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by four coefficients of the MA terms so its order is four and the values of these coefficients are 0.5348, -1.3013, -0.6303 and 0.3968.

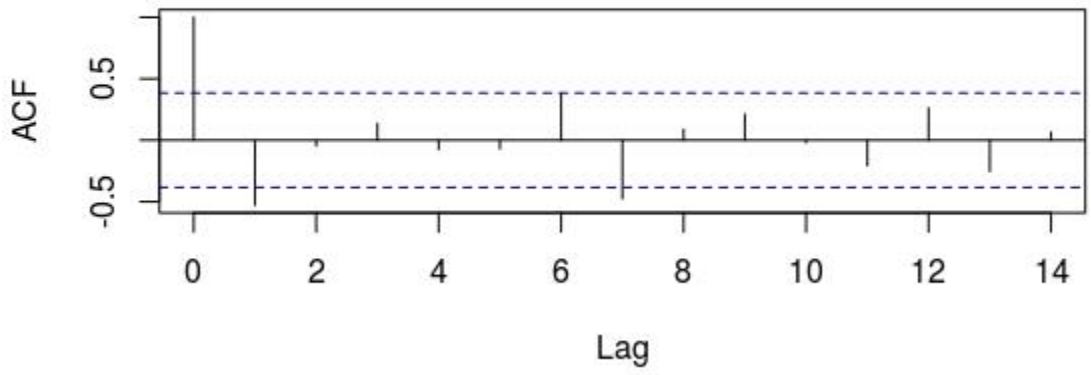

Figure 5.72: ACF plot of Vehari District for days 36 to 63

As the order 2 of AR terms, order 4 of MA terms, and order 2 of the difference terms are computed, so the model that fits best is ARIMA (2, 2, 4) on these incidences and it is shown in equation (57).

$$H_t = -(-2 - B_1)H_{t-1} - (1 + 2B_1 - B_2)H_{t-2} - (-B_1 + 2B_2)H_{t-3} + B_2H_{t-4} + R_t \quad (52)$$
$$+ L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} + L_4R_{t-4}$$

After the substitution of B_1, B_2, L_1, L_2, L_3, and L_4 coefficient values, equation (57) is replaced with equation (58).

$$H_t = 0.4566H_{t-1} + 1.5059H_{t-2} - 0.3816H_{t-3} - 0.5809H_{t-4} + R_t + 0.5348R_{t-1} \quad (53)$$
$$- 1.3013R_{t-2} - 0.6303R_{t-3} + 0.3968R_{t-4}$$

The above equation (58) is used to predict 64, 65, 66, 67 and 68 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.73 shows the incidence of 41 to 68 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, 3 to 4, 7 to 8, 9 to 10, 16 to 17, 18 to 19, and 20 to 22 show an increase in the trend while the incidences on the day 5 to 6, 11 to 12, 13 to 15, 23 to 24, 25 to 26, and 27 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

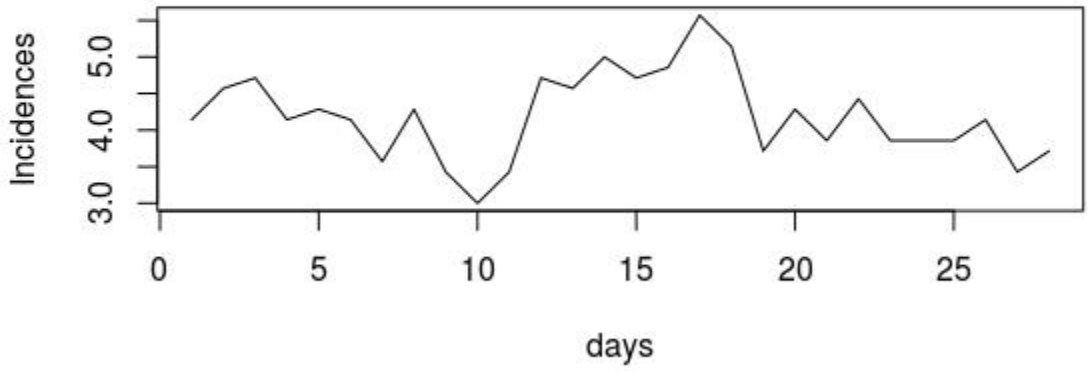

Figure 5.73: Plot of 41 to 68 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.716.` It is made in to stationary by differentiating it twice and in this case the p-value of the ADF test is 0.01 which indicates the stationary presence. The plot of the stationary time series of 41 to 68 day incidences after making a difference two times is shown in Figure 5.74. This plot also shows the constant mean and variance in this time series.

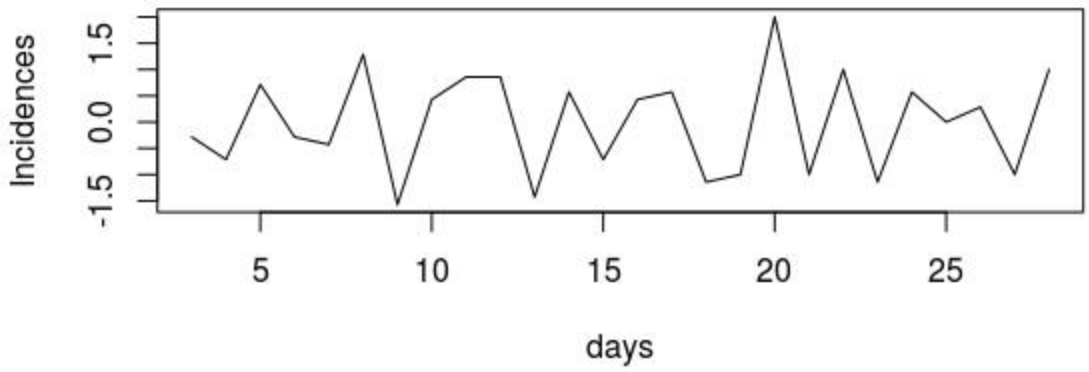

Figure 5.74: Stationary time series plot of 41 to 68 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.75. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold crossed by only one coefficient of the AR terms so its order is one and the values of these coefficients are 0.8019. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

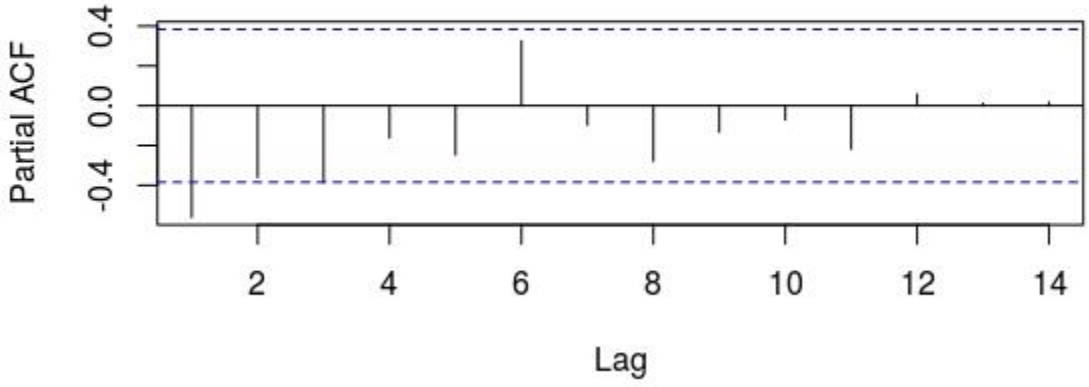

Figure 5.75: PACF plot of Vehari District for days 41 to 68

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.76. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by three coefficients of the MA terms so its order is three and the values of these coefficients are -2.1798, 1.3711, and -0.1862.

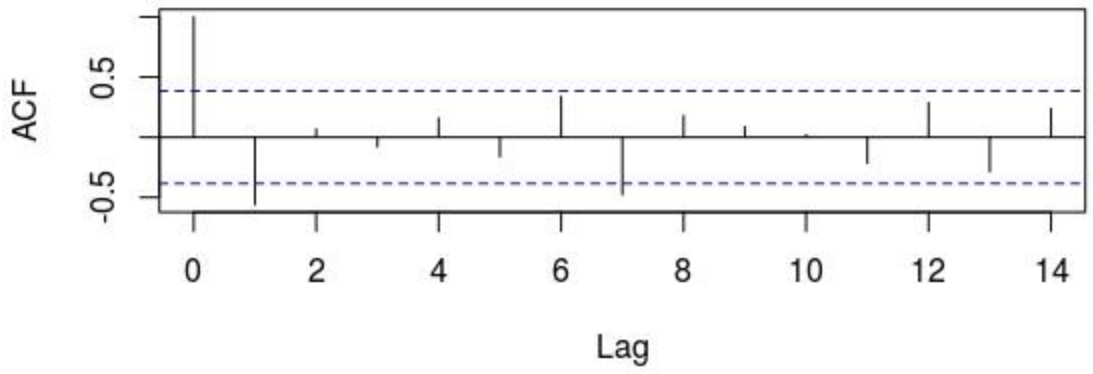

Figure 5.76: ACF plot of Vehari District for days 41 to 68

As the order 1 of AR terms, order 3 of MA terms, and order 2 of the difference terms are computed, so the model that fits best is ARIMA (1, 2, 3) on these incidences and it is shown in equation (59).

$$H_t = -(-2 + B_1)H_{t-1} - (1 - 2B_1)H_{t-2} - B_1H_{t-3} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} \quad (54)$$

After the substitution of B_1, L_1, L_2, and L_3 coefficient values, equation (59) is replaced with equation (60).

$$H_t = 1.1981H_{t-1} + 0.6038H_{t-2} - 0.8019H_{t-3} + R_t - 2.1798R_{t-1} + 1.3711R_{t-2} - 0.1862R_{t-3} \quad (55)$$

The above equation (60) is used to predict 69, 70, 71, 72 and 73 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.77 shows the incidence of 46 to 73 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 6 to 7, 8 to 9, and 10 to 11 show an increase in the trend while the incidences on the day 1 to 2, 3 to 5, 12 to 14, 15 to 16, 17 to 20, 21 to 22, and 23 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

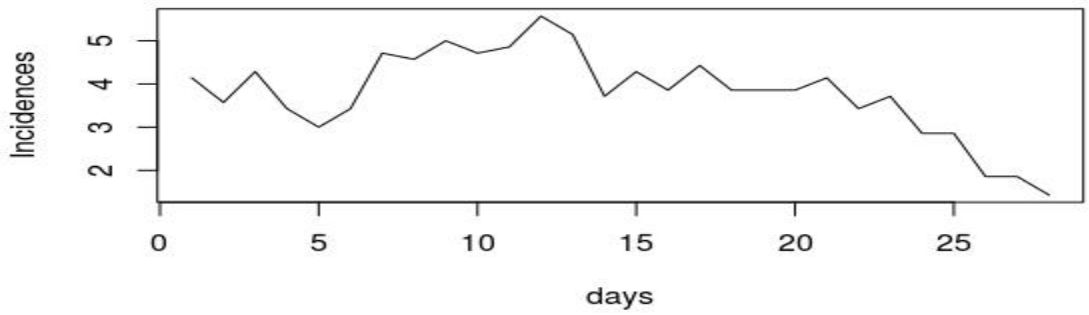

Figure 5.77: Plot of 46 to 73 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.9617.` It is made in to stationary by differentiating it once and in this case the p-value of the ADF test is 0.01745 which indicates the stationary presence. The plot of the stationary time series of 46 to 73 day incidences after making a difference one time is shown in Figure 5.78. This plot also shows the constant mean and variance in this time series.

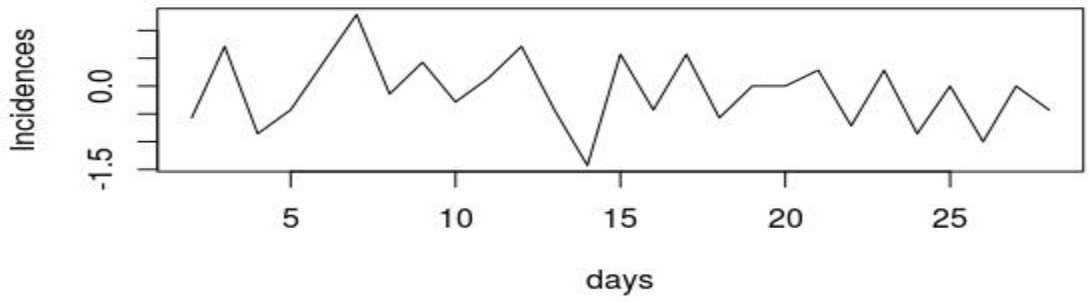

Figure 5.78: Stationary time series plot of 46 to 73 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.79. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold crossed by only one coefficient of the AR terms so its order is one and the values of these coefficients are -0.9998. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

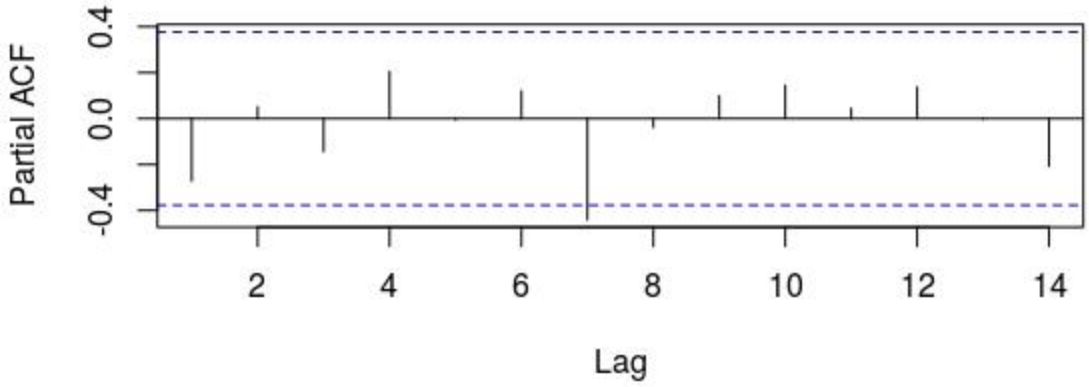

Figure 5.79: PACF plot of Vehari District for days 46 to 73

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.80. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 1.0714, and 0.083.

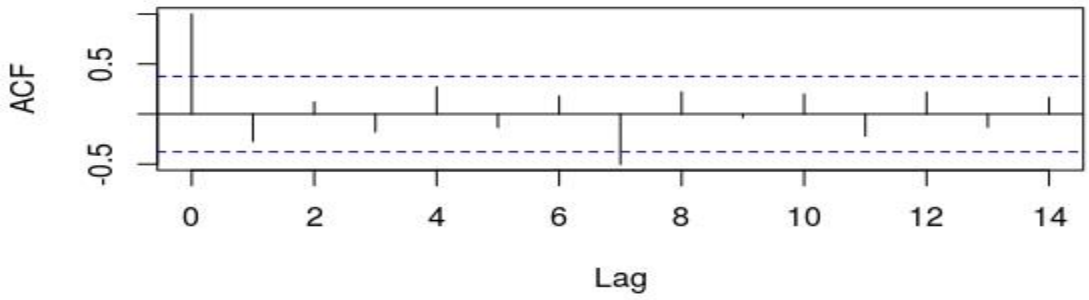

Figure 5.80: ACF plot of Vehari District for days 46 to 73

As the order 1 of AR terms, order 2 of MA terms, and order 1 of the difference terms are computed, so the model that fits best is ARIMA (1, 1, 2) on these incidences and it is shown in equation (61).

$$H_t = -(-1 - B_1)H_{t-1} - B_1 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (56)$$

After the substitution of B_1, L_1, and L_2 coefficient values, equation (61) is replaced with equation (62).

$$H_t = 0.0002H_{t-1} + 0.9998H_{t-2} + R_t + 1.0714R_{t-1} + 0.083R_{t-2} \quad (57)$$

The above equation (62) is used to predict 74, 75, 76, 77 and 78 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.81 shows the incidence of 51 to 78 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, 3 to 4, and 5 to 6 show an increase in the trend while the incidences on the day 7 to 9, 10 to 11, 12 to 15, 16 to 17, 18 to 22, and 22 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

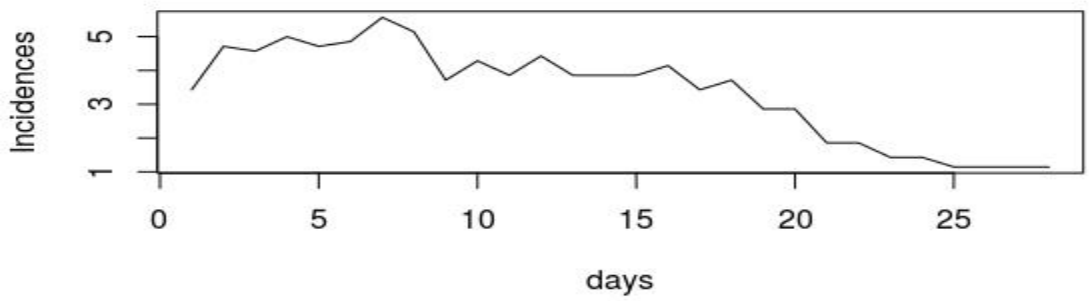

Figure 5.81: Plot of 51 to 78 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.4658.` It is made in to stationary by differentiating it once and in this case the p-value of the ADF test is 0.057 which indicates the stationary presence. The plot of the stationary time series of 51 to 78 day incidences after making a difference one time is shown in Figure 5.82. This plot also shows the constant mean and variance in this time series.

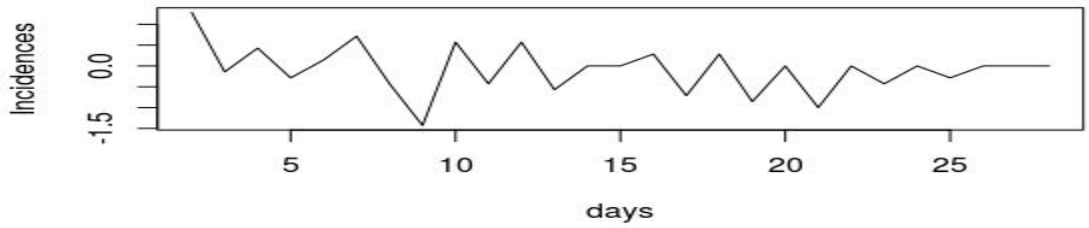

Figure 5.82: Stationary time series plot of 51 to 78 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.83. In this plot, lag values are represented by the x-axis while the coefficients of AR

terms are represented by the y-axis. The threshold crossed by only one coefficient of the AR terms so its order is one and the values of these coefficients are -0.9998. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

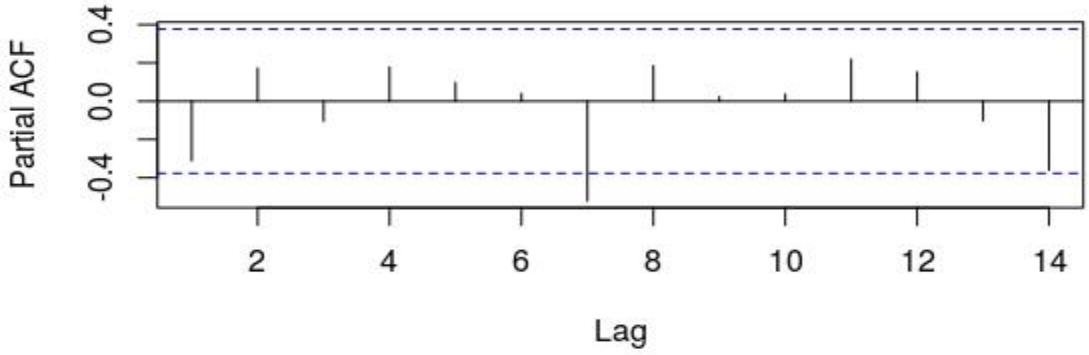

Figure 5.83: PACF plot of Vehari District for days 51 to 78

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.84. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by two coefficients of the MA terms so its order is two and the values of these coefficients are 0.9737, and -0.0128.

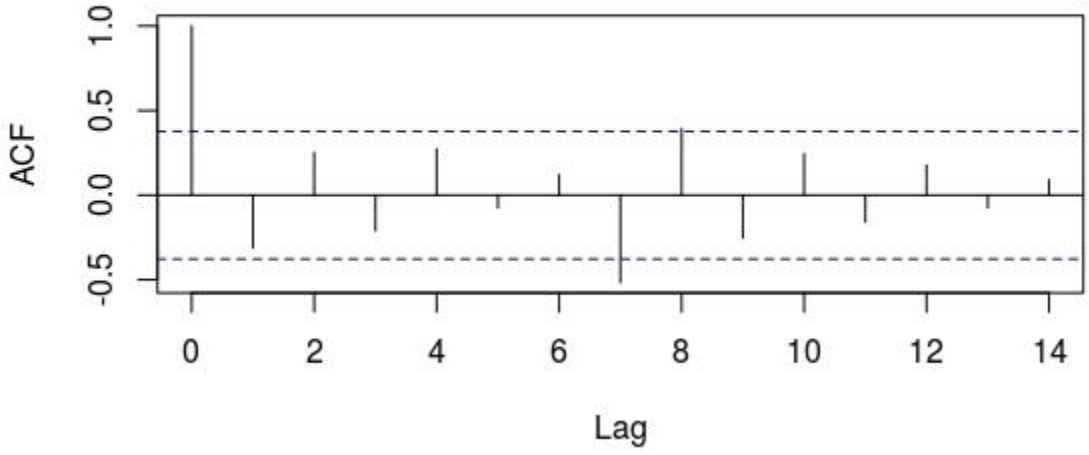

Figure 5.84: ACF plot of Vehari District for days 51 to 78

As the order 1 of AR terms, order 2 of MA terms, and order 1 of the difference terms are computed, so the model that fits best is ARIMA (1, 1, 2) on these incidences and it is shown in equation (63).

$$H_t = -(-1 - B_1)H_{t-1} - B_1 H_{t-2} + R_t + L_1 R_{t-1} + L_2 R_{t-2} \quad (58)$$

After the substitution of B_1, L_1, and L_2 coefficient values, equation (63) is replaced with equation (64).

$$H_t = 0.0002H_{t-1} + 0.9998H_{t-2} + R_t + 0.9737R_{t-1} - 0.0128R_{t-2} \quad (59)$$

The above equation (64) is used to predict 79, 80, 81, 82 and 83 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.85 shows the incidence of 56 to 83 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, and 28 show an increase in the trend while the incidences on the day 3 to 4, 5 to 6, 7 to 10, 11 to 12, and 13 to 27 show a decrease in the trend. It also shows the non-stationary element in the time series.

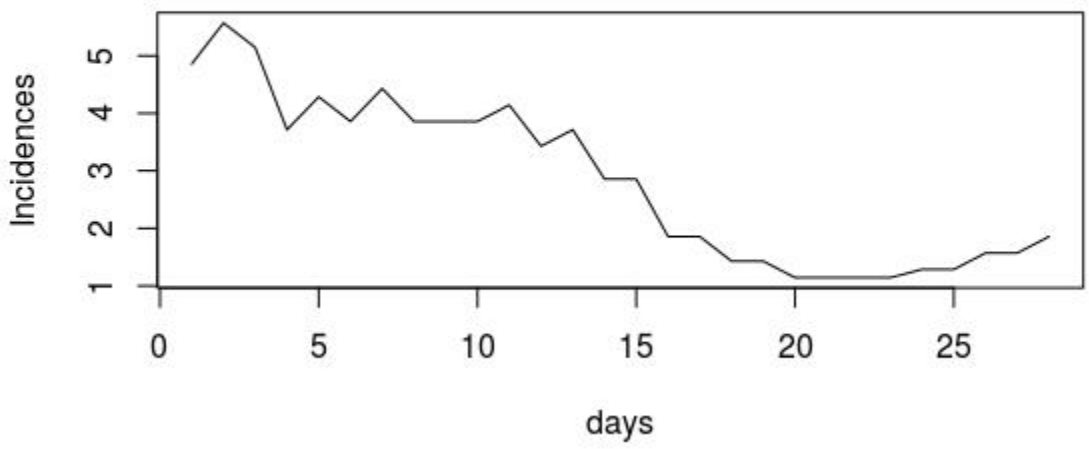

Figure 5.85: Plot of 56 to 83 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.9129.` It is made in to stationary by differentiating it twice and in this case the p-value of the ADF test is 0.01 which indicates the stationary presence. The plot of the stationary time series of 56 to 83

day incidences after making a difference two times is shown in Figure 5.86. This plot also shows the constant mean and variance in this time series.

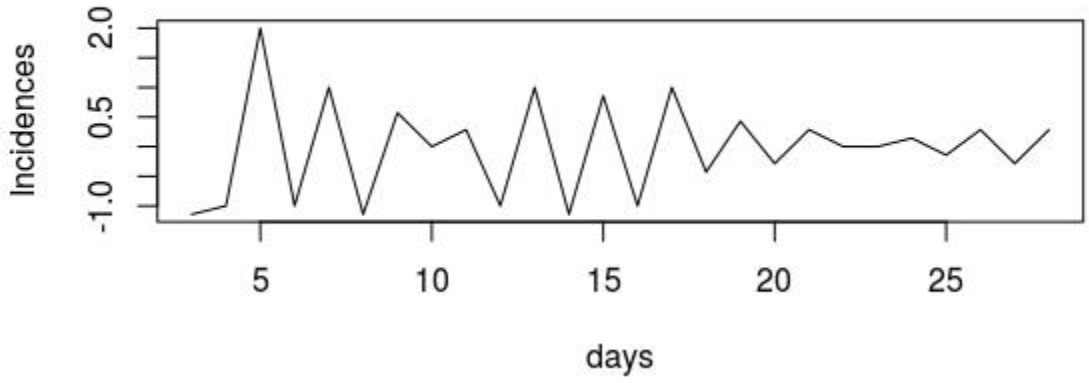

Figure 5.86: Stationary time series plot of 56 to 83 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.87. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold crossed by only one coefficient of the AR terms so its order is one and the values of these coefficients are 0.1082. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

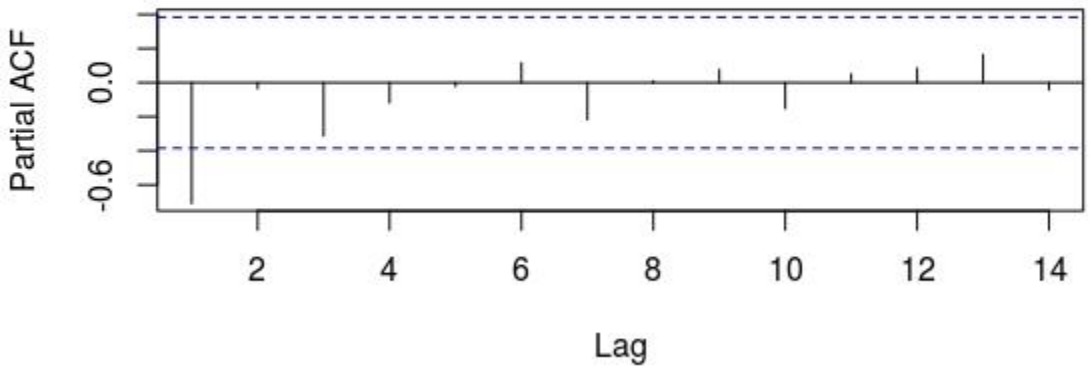

Figure 5.87: PACF plot of Vehari District for days 56 to 83

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.88. In this plot, lag values are represented by the x-axis while the coefficients of MA

terms are represented by the y-axis. The threshold is crossed by five coefficients of the MA terms so its order is five and the values of these coefficients are -1.5527, 1.4398, -1.4398, 1.5527, and -1.0000.

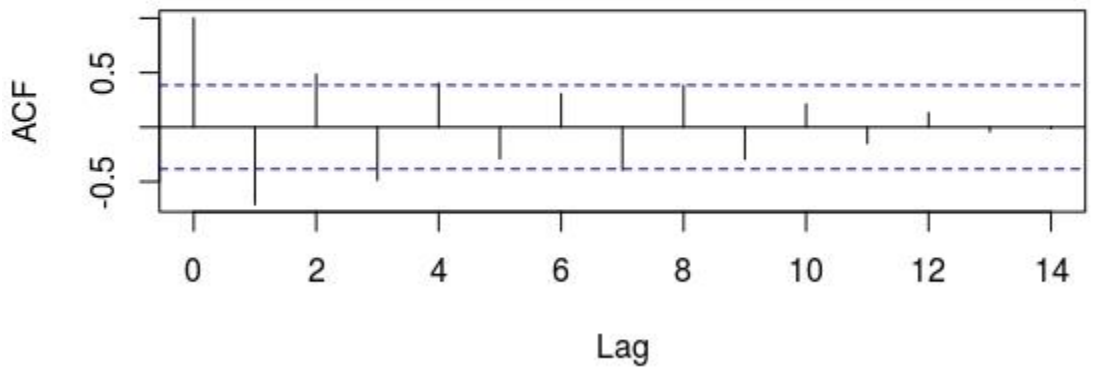

Figure 5.88: ACF plot of Vehari District for days 56 to 83

As the order 1 of AR terms, order 5 of MA terms, and order 2 of the difference terms are computed, so the model that fits best is ARIMA (1, 2, 5) on these incidences and it is shown in equation (65).

$$H_t = -(-2 + B_1)H_{t-1} - (1 - 2B_1)H_{t-2} - B_1H_{t-3} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} + L_4R_{t-4} + L_5R_{t-5} \quad (60)$$

After the substitution of B_1, L_1, L_2, L_3, L_4, and L_5 coefficient values, equation (65) is replaced with equation (66).

$$H_t = 1.8918H_{t-1} + 0.7836H_{t-2} - 0.1082H_{t-3} + R_t - 1.5527R_{t-1} + 1.4398R_{t-2} - 1.4398R_{t-3} + 1.5527R_{t-4} - 1.0000R_{t-5} \quad (61)$$

The above equation (66) is used to predict 84, 85, 86, 87 and 88 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.89 shows the incidence of 61 to 88 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 1 to 2, 5 to 6, 7 to 8, and 21 to 27 show an increase in

the trend while the incidences on the day 9 to 20, and 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

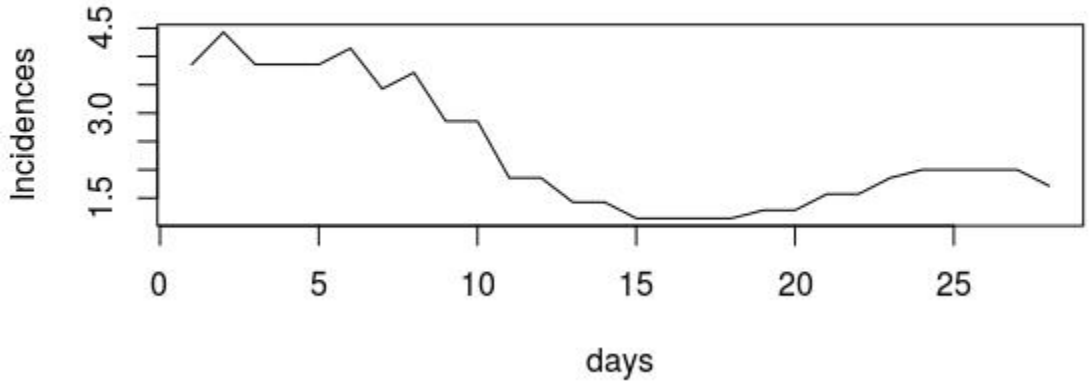

Figure 5.89: Plot of 61 to 88 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.5138.` It is made in to stationary by differentiating it twice and in this case the p-value of the ADF test is 0.04785 which indicates the stationary presence. The plot of the stationary time series of 61 to 88 day incidences after making a difference two times is shown in Figure 5.90. This plot also shows the constant mean and variance in this time series.

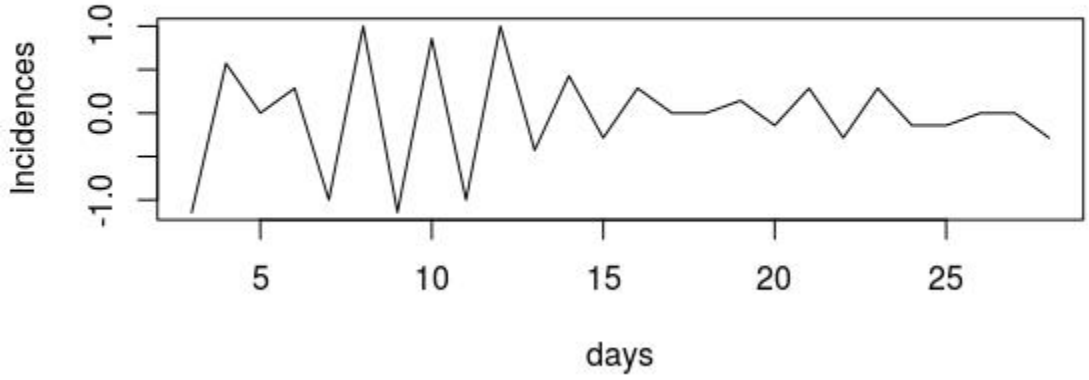

Figure 5.90: Stationary time series plot of 61 to 88 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.91. In this plot, lag values are represented by the x-axis while the coefficients of AR

terms are represented by the y-axis. The threshold crossed by only one coefficient of the AR terms so its order is one and the values of these coefficients are -0.8985. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

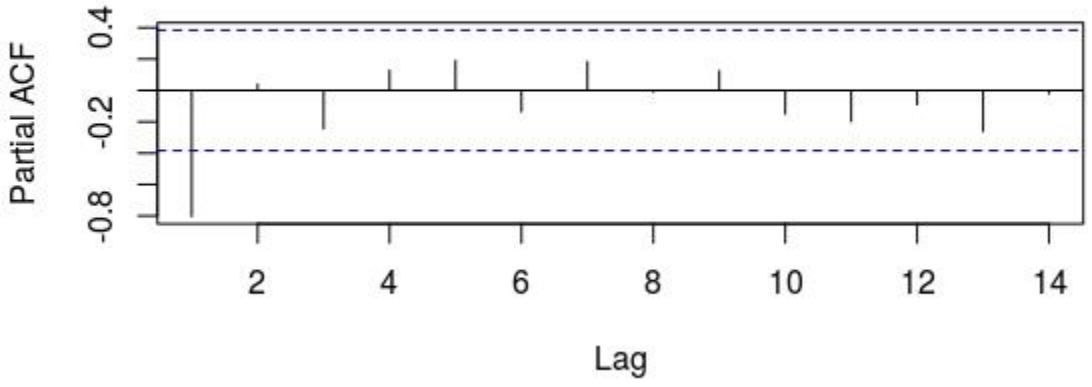

Figure 5.91: PACF plot of Vehari District for days 61 to 88

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.92. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by six coefficients of the MA terms so its order is six and the values of these coefficients are -0.3583, -0.0138, 0.3046, -0.4232, -0.5617, and 0.0540.

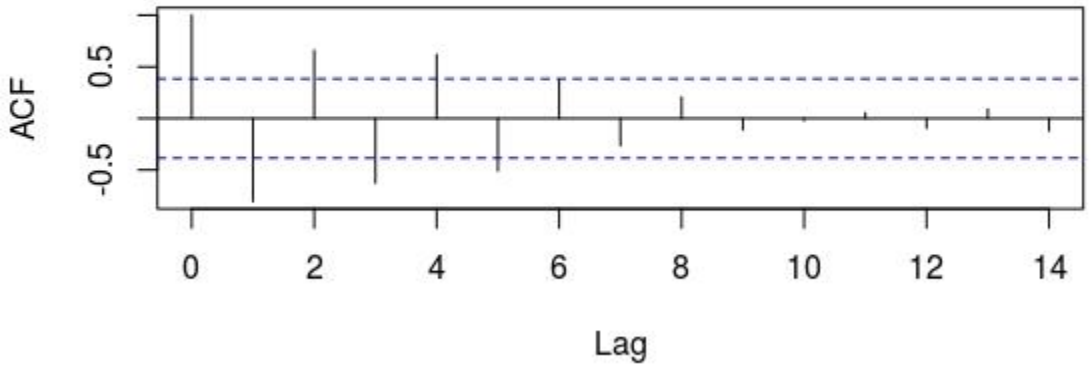

Figure 5.92: ACF plot of Vehari District for days 61 to 88

As the order 1 of AR terms, order 6 of MA terms, and order 2 of the difference terms are computed, so the model that fits best is ARIMA (1, 2, 6) on these incidences and it is shown in equation (67).

$$H_t = -(-2 + B_1)H_{t-1} - (1 - 2B_1)H_{t-2} - B_1H_{t-3} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} + L_4R_{t-4} + L_5R_{t-5} + L_6R_{t-6} \quad (62)$$

After the substitution of B_1, L_1, L_2, L_3, L_4, L_5, and L_6 coefficient values, equation (67) is replaced with equation (68).

$$H_t = 2.8985H_{t-1} - 2.797H_{t-2} + 0.8985H_{t-3} + R_t - 0.3583R_{t-1} - 0.0128R_{t-2} + 0.3046R_{t-3} - 0.4232R_{t-4} - 0.5617R_{t-5} + 0.0540R_{t-6} \quad (63)$$

The above equation (68) is used to predict 89, 90, 91, 92 and 93 day incidences using the ARIMA model by adding +1 to t value each time.

The plot of Figure 5.93 shows the incidence of 66 to 93 day's of the HCV patients in the Vehari District of Punjab Province in Pakistan. In this plot, x-axis shows the incidence of HCV infected patients while y-axis shows the days in which incidence occurs. From this plot, it can be seen that the incidences that occur on days 16 to 22 show an increase in the trend while the incidences on the day 1 to 2, 3 to 7, 8 to 15, and 23 to 28 show a decrease in the trend. It also shows the non-stationary element in the time series.

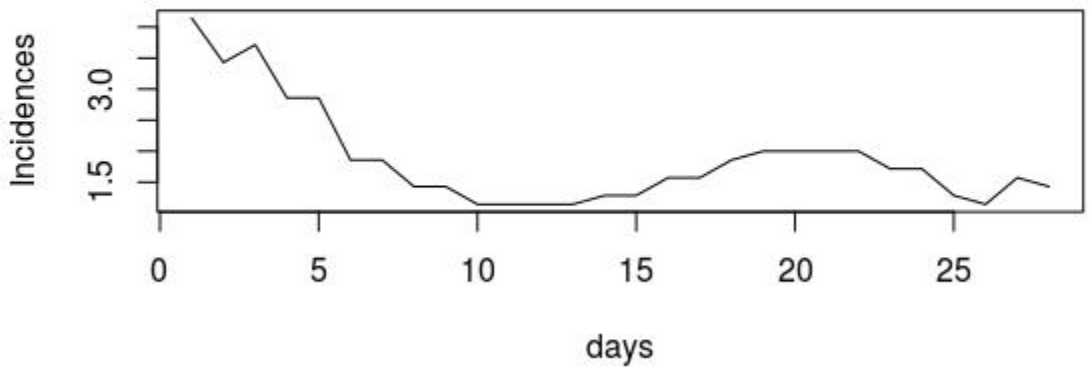

Figure 5.93: Plot of 66 to 93 day's incidences of the Vehari District

In the case of this non-stationary time series, the p value of the ADF test is `0.07401.` It is made in to stationary by differentiating it thrice and in this case the p-value of the ADF test is 0.01 which indicates the stationary presence. The plot of the stationary time series of 66 to 93

day incidences after making a difference three times is shown in Figure 5.94. This plot also shows the constant mean and variance in this time series.

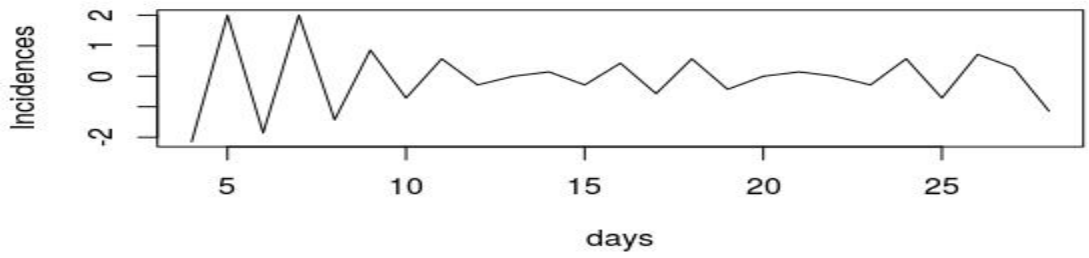

Figure 5.94: Stationary time series plot of 66 to 93 day's incidences of the Vehari District

PACF gives us the order of AR terms of the ARIMA model and this plot is shown in Figure 5.95. In this plot, lag values are represented by the x-axis while the coefficients of AR terms are represented by the y-axis. The threshold crossed by only one coefficient of the AR terms so its order is one and the values of these coefficients are -0.6807. This threshold is shown in the form of blue dots and this line is drawn at 5% confidence interval.

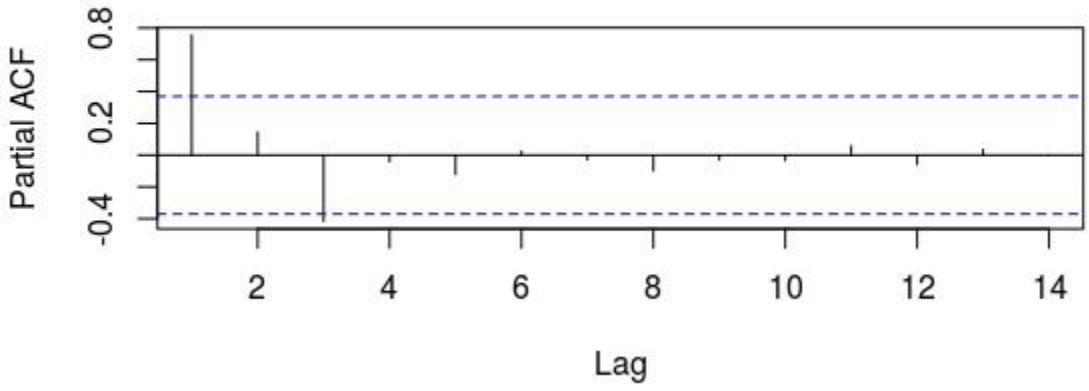

Figure 5.95: PACF plot of Vehari District for days 66 to 93

ACF gives us the order of MA terms of the ARIMA model and this plot is shown in Figure 5.96. In this plot, lag values are represented by the x-axis while the coefficients of MA terms are represented by the y-axis. The threshold is crossed by four coefficients of the MA terms so its order is four and the values of these coefficients are -1.2185, 0.6908, -1.1341, and 0.6619.

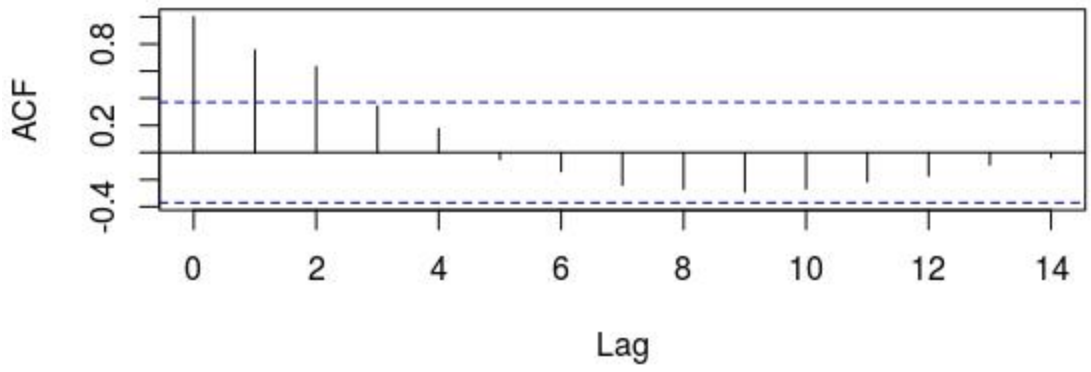

Figure 5.96: ACF plot of Vehari District for days 66 to 93

As the order 1 of AR terms, order 4 of MA terms, and order 3 of the difference terms are computed, so the model that fits best is ARIMA (1, 3, 4) on these incidences and it is shown in equation (69).

$$H_t = -(-3 - B_1)H_{t-1} - (-3 + 3B_1)H_{t-2} - (-1 - 3B_1)H_{t-3} - B_1H_{t-4} + R_t + L_1R_{t-1} + L_2R_{t-2} + L_3R_{t-3} + L_4R_{t-4} \quad (64)$$

After the substitution of B_1, L_1, L_2, L_3, and L_4 coefficient values, equation (69) is replaced with equation (70).

$$H_t = 2.3193H_{t-1} - 0.9579H_{t-2} - 1.0421H_{t-3} - 0.6807H_{t-4} + R_t - 1.2185R_{t-1} + 0.6908R_{t-2} - 1.1341R_{t-3} + 0.6619R_{t-4} \quad (65)$$

The above equation (70) is used to predict 94, 95, 96, 97 and 98 day incidences using the ARIMA model by adding +1 to t value each time.

CHAPTER 6: DISCUSSION

Incidence prediction is playing a crucial rule for disease prevention and control before its epidemic. It will help decision makers in improving the accuracy of diagnostics and treatment of patient, operational efficiency, precision in medicines for healthcare. It will also help in reducing the fraud, abuse and waste for healthcare. Statistical Methods that have rarely been used for subpopulation at district level is used in this study to construct and validate a model and will aid decision makers in health care management system in prevention and control of disease incidences. Surveillance of infectious diseases in Epidemiology is prevalent, forecasting from model will help in efficient utilization of surveillance data [58, 59]. Statistical models are playing a substantial rule in forecasting incidence of communicable disease, and it is very important for sanitation departments to early recognize the behavior of epidemic [60]. ARIMA was initially considered for economics but now it has been used for infectious disease predictive analysis [61, 62].

We have presented a disease incidences of Lahore and Vehari District of Punjab province of Pakistan and the decision makers in the Pakistan Ministry of Health as a case study to validate our proposed predictive analytics framework.

The analysis of the results of the predictive model indicate that when an increase in trend occurs, it is the provision of an early warning for decision makers in healthcare management system before the disease reaches its peak level. This aids in public disease management including better policies for disease prevention and control, raising awareness among people, allocation of appropriate budget for production of medicines, increasing production of protective medicines and protective vaccines. Similarly, the decrease in trend would allow decision makers in healthcare management system to make corresponding decisions such as utilizing the budget for production of medicines to appropriate disease.

The availability of vast amount of data available in the domain of public health is of countless significance for prevention and control of diseases [63]. Time series analysis on the data of infectious diseases is valuable to suggest novel hypothesis, forecast trends, to enhance the prevention and control systems. In this study the predictive model that is constructed and is validated for Lahore and Vehari District of HCV incidences and the Pakistan Ministry of Health will aid decision makers of healthcare management system to impose early warning of diseases

before its outbreak, to improve the awareness of public health and to efficiently allocate required resources, thus assisting decision makers in healthcare management system.

6.1 Limitations

Limitations of the study may include that the model takes the data samples on the basis of defined minimum value that is set as a threshold. This limits the intake of the data samples at first place.

CHAPTER 7: CONCLUSION AND FUTURE WORK

7.1 Conclusion

Disease predictive analytics framework is successfully made in this time frame of study and is validated on case study of HCV incidences of Lahore and Vehari District for Punjab Province and the Ministry of Health of Pakistan. The research gap of proposing a generalized predictive model of disease analysis and its integration with the decision making process in healthcare management system is successfully covered. 12 and 14 ARIMA models are best fitted in this study. The combined MAE and RMSE of all of the data samples of Lahore and Vehari that are forecast from model are 0.2881705, and 0.4113594, and 0.9295707, and 1.180788. The model however, is capable of aiding in finding trend of any disease and also helps timely decision making i.e. appointing doctors to respective places to attend the patients suffering from that particular disease, production of medicines and vaccination schemes etc.

7.2 Future Work

In future work this proposed framework can be connected with the workflow of Ministry of Health and other stochastic modeling techniques can be studied and applied to yield more accuracy in results of determining disease trends purpose.

REFERENCES

[1] Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association, 25(10), 1419-1428.

[2] Ancker, J. S., Kern, L. M., Edwards, A., Nosal, S., Stein, D. M., Hauser, D., ... & with the HITEC Investigators. (2014). How is the electronic health record being used? Use of EHR data to assess physician-level variability in technology use. Journal of the American Medical Informatics Association, 21(6), 1001-1008.

[3] Kim, S. K., & Huh, J. H. (2020, June). Consistency of Medical Data Using Intelligent Neuron Faster R-CNN Algorithm for Smart Health Care Application. In Healthcare (Vol. 8, No. 2, p. 185). Multidisciplinary Digital Publishing Institute.

[4] Heart, T., Ben-Assuli, O., & Shabtai, I. (2017). A review of PHR, EMR and EHR integration: A more personalized healthcare and public health policy. Health Policy and Technology, 6(1), 20-25.

[5] Hossain, A., Quaresma, R., & Rahman, H. (2019). Investigating factors influencing the physicians' adoption of electronic health record (EHR) in healthcare system of Bangladesh: An empirical study. International Journal of Information Management, 44, 76-87.

[6] Al-Rayes, S. A., Alumran, A., & AlFayez, W. (2019). The Adoption of the electronic health record by physicians. Methods of information in medicine, 58(02/03), 063-070.

[7] Campanella, P., Lovato, E., Marone, C., Fallacara, L., Mancuso, A., Ricciardi, W., & Specchia, M. L. (2016). The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. The European Journal of Public Health, 26(1), 60-64.

[8] Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208.

[9] Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: how can we know it works?. Journal of the American Medical Informatics Association, 26(12), 1651-1654.

[10] Liu, V. X., Bates, D. W., Wiens, J., & Shah, N. H. (2019). The number needed to benefit: estimating the value of predictive analytics in healthcare. Journal of the American Medical Informatics Association, 26(12), 1655-1659.

[11] Farooqi, A. (2014). ARIMA model building and forecasting on Imports and Exports of Pakistan. Pakistan Journal of Statistics and Operation Research, 157-168.

[12] Udom, P., & Phumchusri, N. (2014). A comparison study between time series model and ARIMA model for sales forecasting of distributor in plastic industry. IOSR Journal of Engineering, 4(2), 32-38.

[13] Wang, Y. J., Zhao, T. Q., Wang, P., Li, S. Q., Huang, Z., Yang, G. Q., ... & Liu, B. (2006). Applying linear regression statistical method to predict the epidemic of hemorrhagic fever with renal syndrome. Chinese Journal of Vector Biology and Control, 17(4), 333-4.

[14] Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3-13.

[15] Reddy, B. K., Delen, D., & Agrawal, R. K. (2019). Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. Health informatics journal, 25(4), 1201-1218.

[16] Ng, K., Ghoting, A., Steinhubl, S. R., Stewart, W. F., Malin, B., & Sun, J. (2014). PARAMO: A PARAllel predictive MOdeling platform for healthcare analytic research using electronic health records. Journal of biomedical informatics, 48, 160-170.

[17] Pan, Y., Zhang, M., Chen, Z., Zhou, M., & Zhang, Z. (2016, June). An ARIMA based model for forecasting the patient number of epidemic disease. In 2016 13th International Conference on Service Systems and Service Management (ICSSSM) (pp. 1-4). IEEE.

[18] Soebiyanto, R. P., Adimi, F., & Kiang, R. K. (2010). Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters. PloS one, 5(3), e9450.

[19] Dom, N. C., Hassan, A. A., Abd Latif, Z., & Ismail, R. (2013). Generating temporal model using climate variables for the prediction of dengue cases in Subang Jaya, Malaysia. Asian Pacific journal of tropical disease, 3(5), 352-361.

[20] SHEN, Z. Z., WANG, Y. W., MA, S., ZHAO, P. Y., YU, K., YAN, B. H., & JIANG, Y. (2018). Prediction on the number of hepatitis B by AIMRA model in China. Chinese Journal of Disease Control & Prevention, (3), 19.

[21] Liu, Q., Liu, X., Jiang, B., & Yang, W. (2011). Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC infectious diseases, 11(1), 218.

[22] Jones, G. R., Lyons, M., Plevris, N., Jenkinson, P. W., Bisset, C., Burgess, C., ... & Kirkwood, K. (2019). IBD prevalence in Lothian, Scotland, derived by capture–recapture methodology. Gut, 68(11), 1953-1960.

[23] WANG, P., CHENG, B., & SUN, J. (2016). Risk Prediction Model for Newcastle Disease Based on ARIMA. Journal of Qingdao Agricultural University (Natural Science), (1), 18.

[24] Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, 105340.

[25] Moosazadeh, M., Nasehi, M., Bahrampour, A., Khanjani, N., Sharafi, S., & Ahmadi, S. (2014). Forecasting tuberculosis incidence in iran using box-jenkins models. Iranian Red Crescent Medical Journal, 16(5).

[26] Alberg, D., & Last, M. (2018). Short-term load forecasting in smart meters with sliding window-based ARIMA algorithms. Vietnam Journal of Computer Science, 5(3-4), 241-249.

[27] Vafaeipour, M., Rahbari, O., Rosen, M. A., Fazelpour, F., & Ansarirad, P. (2014). Application of sliding window technique for prediction of wind velocity time series. International Journal of Energy and Environmental Engineering, 5(2-3), 105.

[28] Mozaffari, L., Mozaffari, A., & Azad, N. L. (2015). Vehicle speed prediction via a sliding-window time series analysis and an evolutionary least learning machine: A case study on San Francisco urban roads. Engineering science and technology, an international journal, 18(2), 150-162.

[29] Khan, I. A., Akber, A., & Xu, Y. (2019, May). Sliding window regression based short-term load forecasting of a multi-area power system. In 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) (pp. 1-5). IEEE.

[30] Hota, H. S., Handa, R., & Shrivas, A. K. (2017). Time series data prediction using sliding window based rbf neural network. International Journal of Computational Intelligence Research, 13(5), 1145-1156.

[31] Bhatia, M., & Sood, S. K. (2017). A comprehensive health assessment framework to facilitate IoT-assisted smart workouts: A predictive healthcare perspective. Computers in Industry, 92, 50-66.

[32] Sharma, M., Singh, G., & Singh, R. (2019). An advanced conceptual diagnostic healthcare framework for diabetes and cardiovascular disorders. arXiv preprint arXiv:1901.10530.

[33] Jayaraman, I., & Mohammed, M. (2019). Secure Privacy Conserving Provable Data Possession (SPC-PDP) framework. Information Systems and e-Business Management, 1-27.

[34] Riad, K., Hamza, R., & Yan, H. (2019). Sensitive and energetic IoT access control for managing cloud electronic health records. IEEE Access, 7, 86384-86393.

[35] Xia, Q., Sifah, E. B., Smahi, A., Amofa, S., & Zhang, X. (2017). BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. Information, 8(2), 44.

[36] Nagasubramanian, G., Sakthivel, R. K., Patan, R., Gandomi, A. H., Sankayya, M., & Balusamy, B. (2020). Securing e-health records using keyless signature infrastructure Blockchain technology in the cloud. Neural Computing and Applications, 32(3), 639-647.

[37] Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: a patient-centered framework. Journal of general internal medicine, 28(3), 660-665.

[38] del Carmen Legaz-García, M., Martínez-Costa, C., Menárguez-Tortosa, M., & Fernández-Breis, J. T. (2016). A semantic web based framework for the interoperability and exploitation of clinical models and EHR data. Knowledge-Based Systems, 105, 175-189.

[39] Srinivas, S., & Ravindran, A. R. (2018). Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: a prescriptive analytics framework. Expert Systems with Applications, 102, 245-261.

[40] Koumakis, L., Kondylakis, H., Katehakis, D. G., Iatraki, G., Argyropaidas, P., Hatzimina, M., & Marias, K. (2017, November). A content-aware analytics framework for open health data. In International Conference on Biomedical and Health Informatics (pp. 59-64). Springer, Singapore.

[41] Meghani, S. T., Sehar, S., & Punjani, N. S. (2014). Comparison and analysis of health care delivery system: Pakistan versus China. Int J Endorsing Health Sci Res, 2(1), 46-50.

[42] Hassan, A., Mahmood, K., & Bukhsh, H. A. (2017). Healthcare system of Pakistan. IJARP, 1(4), 170-3.

[43] Javed, S. A., Liu, S., Mahmoudi, A., & Nawaz, M. (2019). Patients' satisfaction and public and private sectors' health care service quality in Pakistan: Application of grey decision analysis approaches. The International journal of health planning and management, 34(1), e168-e182.

[44] Javed, S. A., & Ilyas, F. (2018). Service quality and satisfaction in healthcare sector of Pakistan—the patients' expectations. International Journal of Health Care Quality Assurance.

[45] Shafiq, M., Naeem, M. A., Munawar, Z., & Fatima, I. (2017). Service quality assessment of hospitals in Asian context: An empirical evidence from Pakistan. INQUIRY: The Journal of Health Care Organization, Provision, and Financing, 54, 0046958017714664.

[46] Khalique, F., Khan, S. A., & Nosheen, I. (2019). A Framework for Public Health Monitoring, Analytics and Research. IEEE Access, 7, 101309-101326.

[47] Hripcsak, G., & Albers, D. J. (2013). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20(1), 117-121.

[48] Ehrenstein, V., Kharrazi, H., Lehmann, H., & Taylor, C. O. (2019). Obtaining Data from Electronic Health Records. In Tools and Technologies for Registry Interoperability, Registries for Evaluating Patient Outcomes: A User's Guide, 3rd Edition, Addendum 2 [Internet]. Agency for Healthcare Research and Quality (US).

[49] (2020, July 21). Disease surveillance Thresholds. Retrieved from https://emergency.unhcr.org/entry/38802/disease-surveillance-thresholds

[50] (2020, October 20). National Database and Registration Authority (NADRA). Retrieved from https://www.nadra.gov.pk/

[51] (2020, October 20). National Health Emergency Preparedness and Response Network (NHEPRN). Retrieved from https://www.nheprn.gov.pk/

[52] (2020, October 20). Drug Regulatory Authority of Pakistan (DRAP). Retrieved from https://www.drap.gov.pk/

[53] (2020, October 20). Pakistan Medical Commission (PMC). Retrieved from https://www.pmc.gov.pk/

[54] (2020, October 20). Pakistan Centre for Philanthropy (PCP). Retrieved from https://www.pcp.gov.pk/

[55] (2020, October 20). Pakistan Bureau of Statistics (PBS). Retrieved from https://www.pbs.gov.pk/

[56] (2020, October 20). National Institute of Health (NIH). Retrieved from https://www.nih.gov.pk/

[57] (2020, October 20). National Institute of Population Studies (NIPS). Retrieved from https://www.nips.gov.pk/

[58] Jamison, D.T., Breman, J.G., Measham, A.R., Alleyne, G., Claeson, M., Evans, D.B., Jha, P., Mills, A., Musgrove, P.: Disease control priorities in developing countries. The World Bank, (2006)

[59] Liu, Q., Liu, X., Jiang, B., Yang, W.: Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC infectious diseases 11(1), 218 (2011).

[60] Earnest, A., Chen, M.I., Ng, D., Sin, L.Y.: Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. BMC Health Services Research 5(1), 36 (2005).

[61] García-Díaz, J.C.: Monitoring and forecasting nitrate concentration in the groundwater using statistical process control and time series analysis: a case study. Stochastic Environmental Research and Risk Assessment 25(3), 331-339 (2011).
[62] McGee, V.E., Jenkins, E., Rawnsley, H.M.: Statistical forecasting in a hospital clinical laboratory. Journal of medical systems 3(3-4), 161-174 (1979).
[63] Dolley, S.: Big data's role in precision public health. Frontiers in public health 6, 68 (2018).