分 类 号_____ 学号 I201422204

学校代码 10487 密级_____

华中科技大学

博士学位论文

Time Series Analysis for Thalassemia disease in Maysan Province

学位申请人： RANA SABEEH ABBOOD

学 科 专 业： 概率论与数理统计

指 导 教 师： 刘继成 教授

答 辩 日 期： 2017 年 5 月 日

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Science

Time Series Analysis for Thalassemia disease in Maysan Province

Ph.D Candidate: Rana Sabeeh Abbood Al-Sudani

Major: Probability Theory and Mathematical Statistics

Supervisor: Prof. Liu Jicheng

Huazhong University of Science and Technology

Wuhan 430074, P. R. China

May 2017

独创性声明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究 成果。尽我所知，除文中已经标明引用的内容外，本论文不包含任何其他个人或集体 已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以 明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期： 2017 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ，在_____ 年解密后适用本授权书。

本论文属于

不保密 □ 。

（请在以上方框内打“ ✓”）

学位论文作者签名： 指导教师签名

日期： 年 月 日 日期： 年 月 日

Abstract

Thalassemia is one of the most famous genetic diseases in Iraq that lead to severe anemia and other complications in the long term.one of major diseases, spread throughout Iraq and other Middle East countries. Despite a prevention program, there has been no decrease in the prevalence of the disease, due to a lack of awareness. Thalassemia is disease an inherited autosomal recessive blood disorder characterized by the underproduction of globin chains because of globin gene defects, resulting in malfunctioning red blood cells and oxygen transport. In this paper, we use three time series models to study thalassemia from the database from Maysan Health Center specific for Thalassemia the Maysan Provence, Iraq.

Firstly, we used Box and Jenkins methodology to build an ARIMA model to forecast the number of people with Thalassemia, for the period from 2016-2018. After the model selection, the best model for forecasting was ARIMA (0, 1, 1), The results showed our data and charts, there is an increase in cases with Thalassemia in the coming years from 2016 to 2018, where the number of patients monthly will be between (7-11) patients and these numbers are high compared to previous years.

Secondly，also used in this study Co-integration (Engle-granger) model to analyses data to find out the relationship between times of blood transfusion and iron overload in thalassemia patients as well as studying the effect of some physiological factors, such as age, gender, blood type and the type of thalassemia Our results demonstrated that there was a positive relationship between both the number of blood transfusions and blood iron levels, the more the number of blood transfusion increased, the more blood iron level and the males was mostly infected than females and children between 1-4 years were the age category with the highest

level of infection. The group with blood type O+ was the most infected group and, finally, thalassemia major beta was the highest registered type.

Finally，we used in this study the vector autoregression model (VAR) for forecasting the number of deaths in patients with thalassemia and also addressed the causes of these deaths, There was a strong relationship between mortality in thalassemia patients and an increase in the proportion of iron and the highest number of deaths was recorded for patients who had a very high proportion of iron. It was the most important cause of mortality，such as Cardiac disease, infections, the liver, the spleen.

Keywords: Time Series, Thalassemia, ARIMA model, VAR model, physiological factors, mortality.

摘　要

地中海贫血是伊拉克最著名的疾病之一，会长期导致严重的贫血和其他并发症，广泛分布在伊拉克和其他中东国家。它是一种遗传性常染色体隐性血液疾病，因基因缺陷造成球蛋白链的产生不足而导致红细胞氧运输产生故障。尽管有预防计划，但由于缺乏防范意识这种疾病的流行率并没有下降。本论文中，我们用三个时间序列的模型来研究地中海贫血疾病，数据来源于伊拉克 Maysan 省地中海贫血健康中心的数据库。

首先，我们用 Box 和 Jenkins 方法建立了 ARIMA 模型，来预测 2016 - 2018 年间伊拉克地中海贫血的人数。通过选择模型，得到预测的最佳模型是 ARIMA（0,1,1）。预测结果表明，在 2016 年至 2018 年的未来几年，地中海贫血病例将有所增加，每个月将有 7-11 名患者，该数字高于前几年。

其次，用 Co-integration 模型来处理数据，分析地中海贫血患者的铁过量和输血时间之间的关系，以及一些生理因素对铁过量的影响，如年龄，性别，血型和地中海贫血的类型等。结果表明，输血次数和血液铁含量之间存在正相关，输血次数越多，血液中铁含量越高，以及男性感染多于女性，1-4 岁之间的儿童是具有最高感染水平的年龄类别，血型 O+的组是最易感染组，并且 Beta 型地中海贫血是最高的类型。

最后，用矢量自回归模型（VAR），用于预测地中海贫血患者的死亡人数，并揭示了这些死亡的原因。结果是，血液中铁过量是地中海贫血患者的死亡率最重要原因，铁含量高的患者死亡率最高。死亡的重要病因是心脏病，感染，肝脏病，脾脏病。

关键词：时间序列分析，地中海贫血，ARIMA 模型，VAR 模型，生理因素，死亡率

Table of Contents

LIST OF FIGURE

LIST OF TABLE

Declaration

I declare that except where explicit reference is made to the contribution of others that this dissertation is the result of my own work and has not been submitted for any other degree at Huazhong University of Science and Technology or any other institution.

1 Introduction

1.1 Background

The time series is a group views sorted by time (and often time periods equal and successive periods vary according to the nature of this phenomenon)[1], and time series have important applications in many areas, including economic, trade and population statistics[2]. As time-series, models typically used to predict the variable values. If the variable to be studied is known determinants, and the factors that affect it, is also used in the case of variable is subject to the expectations of its clients, which is reflected in the future based on what happened in the past.

Mathematically: we say that the independent variable time (t) and the corresponding values him dependent variable (y) and that each value at time t corresponding values of y variable y is a function of time t in which: y = F(t) [3].

The time series analysis of statistical methods task method, which has evolved a lot and it, was possible to use it for the purpose of expectation for the future supply and demand for a commodity or service. Moreover, supports time-series analysis to track the phenomenon style (or variable) over a certain time (several years, for example), then expect for the future based on different values that have emerged in the time series and the pattern of growth in values. This is superior to the conventional method, since the methods traditional calculate the difference in value between the only two date ranges of the time series and builds future expectation. Without taking into account the general style of the series or of the rise and decline that occurs to the values of the time continuum [4].

The variable x_t it dealt with as a random variable. The measurements taken during an of the event in a time series are arranged in a proper based on chronological order[5].

A time series containing records of a single variable is termed as invert. However, if records of more than one variable are considered, it is termed as multivariate. A time series can be continuous or discrete. In a continuous time series notes it is measured in each case of time, while a discrete time series includes observations measured at discrete points of time[6].

For example, temperature readings, flow of a river, concentration of a chemical process etc. can record as a continuous time in the form of a series. On the other hand population of a particular city, production of a company, exchange rates between two different currencies may represent discrete time series[7], the variable being observed in the a discrete time series it is supposed to be measured as a continuous variable using the real number measurement. in addition to a continuous time series can be easily turn to a discrete one by merging data together during a certain period of time. Time series a fumble of measurements taken from one variable or for a number of variables according to the time they occur. In other words, statistical data collected or recorded on the phenomenon of the time periods specified consecutive and equal, and these phenomena are (chemical, economic, social, natural, engineering, environmental ... etc.) and these vary periods based on the nature of the phenomenon may be a day or a week or a year or any other period of time.

Time series analysis has substantial beginnings in together the physical and social sciences. Basic concepts it has appeared in each subject and they made their way to the other to consequent transfer of technology. Historical researchers important in the development of the field include: Thiele, Hooker, Einstein, Wiener, Yule, Fisher, Tukey, Whittle and Bartlett [8-10]Time series models are very useful models when they're serially correlated data.

Majority business houses work on time series data to analyze a sales number for the next year, website traffic competition position and much more, in this study, we used

three models of time series (ARIMA, Co-Integration, vector auto regression model(VAR)) for the purposes of prediction and correlation relationship between two variables. Autoregressive integrated moving average (ARIMA) models are mathematical models of persistence, or autocorrelation, in a time series. Box and Jenkins (1970) introduced it. ARIMA models enable us to not only to uncover the hidden patterns in the data, but also for generate forecasts and predict a variable's future value from its past values. One of the more common model among forecasters is the autoregressive integrated moving average (ARIMA) model[11]. An ARIMA model consists of three parts.

An autoregressive component (AR) pointing the number of lags of the dependent variable that is to be included, a component order of integration (I) and a moving average (MA) component that captures the effect of lagged values of the error term. The early work of among others Yule (1927) that lay the foundation for the development of AR and MA models. Box and Jenkins (1970) integrated the earlier work in this field and developed a three-stage approach for identifying, estimating and verifying ARIMA models. The Box-Jenkins method, as it has come to know, is still widely used today. Box-Jenkins (ARIMA) is an important forecasting method that can yield highly accurate forecasts types of data.

ARIMA models initially generated a lot of excitement in the academic community, due mostly to their theoretical underpinnings which proved that if certain assumptions were met, the models would yield optimal forecasts[12].

The second model used is the Co- integration model, Since the mid-eighties, co-integration techniques have become increasingly popular, along with a remarkable amount of work in the time series Econometrics, fundamentally, Granger (1986) identified that regression constructed with no stationary time series on the other non-stationary series, generates a spurious regression. Nevertheless, Engle and Granger (1987) emphasize a situation that a regression did not yield spurious relationship as two

series co-integrated in the latter work. For the first condition of co integration, we have to determine the integration level of series and the most useful and common way to determine the integration order of the series is unit root tests[13].

The third model used is the VAR models, pioneered by Chris Sims about 25 years ago, have acquired a permanent place in the toolkit of applied macroeconomists both to summarize the information contained in the data and to conduct certain types of policy experiments. Since the critique of Sims (1980) in the early eighties of the last century, multivariate data analysis in the context of vector autoregressive models (henceforth: VAR) has evolved as a standard instrument in econometrics, because statistical tests are frequently used in determining inter-dependencies and dynamic relationships between variables, this methodology was soon enriched by incorporating non-statistical a priori information [14].

Thalassemia was first recognized in 1925 by a Detroit physician, Cooley and Lee, who described a series of infants who became profoundly anemic and developed splenomegaly and bone change over the first year of life[15]. George and William (1932), described the pathological changes of the condition for the first time, recognized that many of their patients came from the Mediterranean region, and hence invented the word thalassemia from the Greek words (“thalassa”: meaning sea and (“aima”: meaning blood) [16]. It was only after 1940 that the true genetic character of this disorder was fully appreciated[17].

Thalassemia is described as a heterogeneous group of inherited anemia’s characterized by a reduced or absent amount of hemoglobin. Thalassemia can classify according to the deficient globin chain, alpha or beta thalassemia, the symptoms vary from relatively mild anemia to life threatening. The signs of the disease included anemia, a large spleen, and characteristic bone deformities. Despite several therapeutic attempts all the children died within a few months of diagnosis[15].

Almost at the same, time an Italian pediatrician, Antonio Maccanti.

1.2 Problem statement

The genetic blood diseases, a group of diseases that pose a major threat to the generations to transition from parents to their children, which is witnessing an increase in the number of cases of those chronic diseases that afflict one blood components such as red and white blood cells or platelets, which constitutes a major threat to the generations. Thalassemia is the pathological case resulting from the defect or disorder in one gene or more can each move from generation to generation and mostly affects the individual during embryonic life, as well as that of those diseases related to the occurrence of disability in people infected if they do not receive proper treatment and appropriate[18]. The increase in the number of births thalassemia because thalassemia trait spread at high rates in some areas and communities, because marriage is often of the same environment, the thalassemia campaign so far unknown by society, ignorant of everything about the situation, and do not have any means of guidance in matters of marriage. Pregnancy so the lack of special centers in thalassemia treatment, study and guidance, and then the lack of control over the new deliveries will increase the problem is a very sharp increase in the coming years [19].

1.3 Study objective

The main objective of the study is

(1) Study time series to determine the best and most efficient statistical model for using it to predict the numbers with Thalassemia for the years 2016-2018.

(2) Find a relationship between the number doses of blood obtained by the patient and the amount of iron using co-integration model, and the study of some physiological factors associated with Thalassemia.

(3) Predict the numbers of deaths in patients with thalassemia (2016-2020) and find

out the complications of the caused by disease the most effect.

1.4 Research Questions

(1) Is there an increase in the number infected with Thalassemia in the coming years 2016-2018?

(2) Is there a relationship between the blood doses which given for thalassemia patients and increase the proportion of iron?

(3) What are the physiological factors that affect patients with thalassemia?

(4) Is there an increase in the number of deaths in patients with thalassemia in the coming years 2016-2020?

(5) Which diseases caused by Thalassemia?

(6) What are the most diseases that cause mortality for patients with thalassemia?

1.5 Literature Review

1.5.1 District under study

Maysan is an Iraqi city and center Maysan province, about 320 kilometers south east of the capital Baghdad. This city is located on the banks of the Tigris River, and lies about 50 km from the Iran-Iraqi border, an estimated population of 340 thousand people in 2002 and 420 thousand inhabitants in 2005, 459,216 thousand people in 2009, 977.34 thousand in 2015.

The characterized by of Maysan province clan character where marriages are often between relatives, which led to the high rate of genetic diseases caused inbreeding. If continued for several generations, will the accumulation of genetic traits, where most of the scientific studies confirm for common diseases genetic, notably hemoglobin blood diseases "hemoglobin "congenital metabolic defects and diseases common single gene, they are the main cause of many diseases and disabilities in children. According to some studies, the inbreeding inherited 82 a disease, such as recurrent miscarriage, multiple

disabilities, multiple vesicles disease , thalassemia, a disease of iron overload in blood, atrophy of facial and shoulder muscles of the disease, a disease multiple colon tumors, birth weight is less than the marriages of non-relatives , and other diseases.

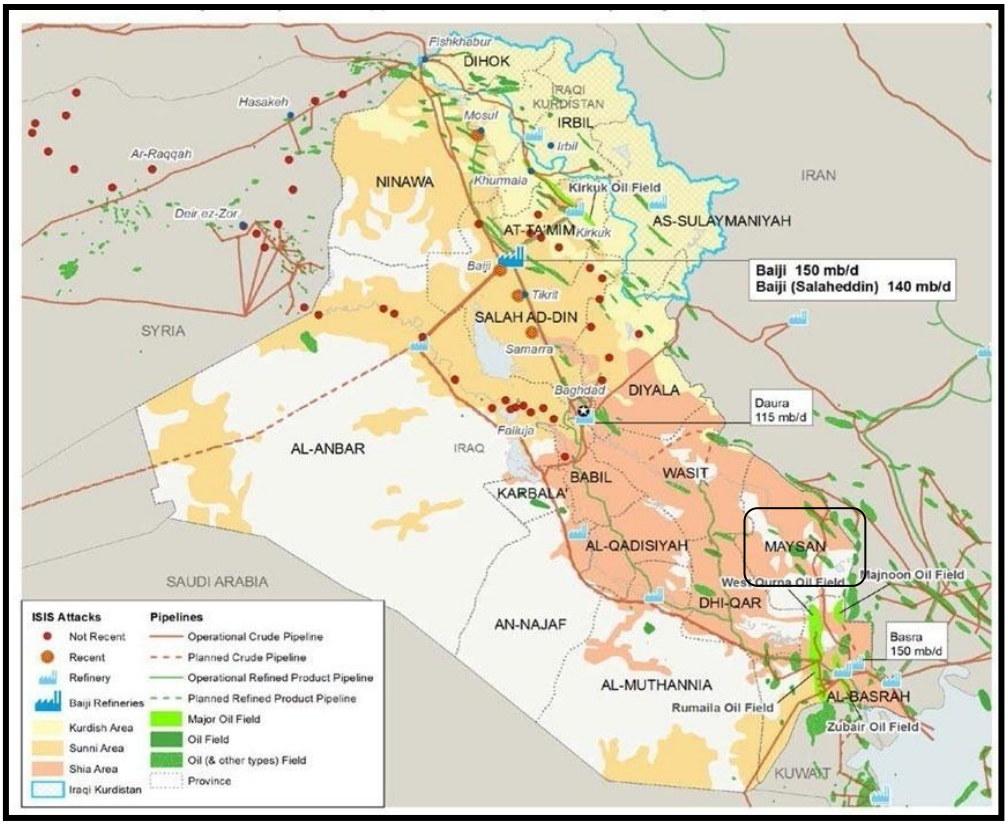


Figure 1-1 map of Iraq

1.5.2 Thalassemia Disease

Thalassemia is a heritable blood disease where they become shape abnormal blood cells of hemoglobin. Hemoglobin is the protein molecule in red blood cells that bearing oxygen. In addition, trouble produces in excessive damage of red blood cells, which lead to anemia.

Anemia the body does not have sufficient of healthy red blood cells in a state. Thalassemia disease get, meaning that at least one of parents there must be a carrier of the disease. It's due to either a genetic change or a deletion of the basic gene parts[20], gene for Thalassemia it is passed from parents to their children through genes, just as eye color and hair color are. If the couples are carriers of the disease, they are likely to be child

infected. In adding to, is a genetic condition, you cannot control it and carriers will have not become ill because of it.

There are two forms of Thalassemia [21]:

- Alpha Thalassemia
- Beta Thalassemia

(1) Alpha thalassemia

- Patients in this type, hemoglobin does not produce sufficient alpha protein. It affects the production of normal hemoglobin -a key constituent of human red blood cells. This type of thalassemia is commonly found in Africa, the Middle East, India, Southeast Asia, southern China, and occasionally the Mediterranean region[22]. There are different types of alpha thalassemia that range from mild to severe:

- **Alpha Thalassemia Trait or Mild Alpha Thalassemia:** In this condition, there is lack of alpha protein and Patients suffering has smaller red blood cells and a mild anemia, the condition is symptoms free [23].

- **Hemoglobin H Disease:** In this condition, there is lack of alpha protein, which causes severe anemia and serious health problems such as bone deformities, fatigue, and enlarged spleen. Abnormal hemoglobin H destroys red blood cells and causes anemia[24].

- **Hemoglobin H-Constant Spring**: This is more case acute than hemoglobin H diseases. patients suffering this disease tend to be more severe anemia and suffer more often than not from enlargement of the spleen and viral infections [25].

- **Alpha Thalassemia Major:** In this case, there are no alpha genes in-patient DNA, which lead the gamma hemoglobin produced by the fetus to form abnormal hemoglobin that called hemoglobin Brats. It is the most severe α-thalassemia, the homozygous case for α 0-thalassemia. None of four α-globin genes is functioning and no α-chains produced. It causes severe anemia leading to the death of the fetus[24]. Patients who suffer from this case dying before or after a birth. In some very rare

cases where the case is discovered before birth, in intrauterine blood transfusions it has allowed her the birth of children with hydrops fetal is that need to medical care over the life and blood transfusions continuous[26].

(2) Beta thalassemia:

It affects product of normal hemoglobin, shalt not make enough beta protein. It found in people of assets Mediterranean, such as Italians and Greeks, and as well found in the Arabian Peninsula, Africa, Iran, Southern China and Southeast Asia. There are three types of beta thalassemia disease ranging from mild to severe according to their effects on the body[27].

- **Beta-Thalassemia Minor (BTMi) or Thalassemia Trait :**

  In this case, suffers patient the lack of beta protein causes no problems in the normal functioning of the hemoglobin. A person with this suffers simply carry genetic trait for thalassemia with any problems other than a prospect mild anemia[28].

- **Beta-Thalassemia Intermedia (BTI):**

  It is a case central among the major and minor forms. In this type, the not available beta protein in the hemoglobin leads to severe anemia and major health problems, including splenomegaly and bone deformities [29], there are wide ranges in the clinical severity of this case, and the borderline between thalassemia middle and the most severe form, thalassemia major, can be confusing. Affected individuals can often manage a normal life but may need blood transfusions that are at times of illness or pregnancy, depending upon the severity of their anemia [27].

- **Thalassemia Major (TM) or Cooley's anemia:**

  TM or β-thalassemia happens when similar from a genetic defects affect production of the beta globin protein [30]. This is the more severe form of beta thalassemia in which there is complete absence of beta protein in the hemoglobin, which cause life-threatening anemia, which requires regularity blood transfusions and continuing medical care.

  These , lifelong blood transfusions lead to high iron concentration which must be treated with chelation therapy in order to prevent early death from organ failure [31].

The low hemoglobin concentration lowers the oxygen level in the blood stream. This serious condition increases the risk of congestive heart and is fatal if not treated. For unknown reasons, iron absorption in the gastro-intestinal tract often enhanced in individuals with TM. This may lead to iron overload and subsequent organ damage.[32]

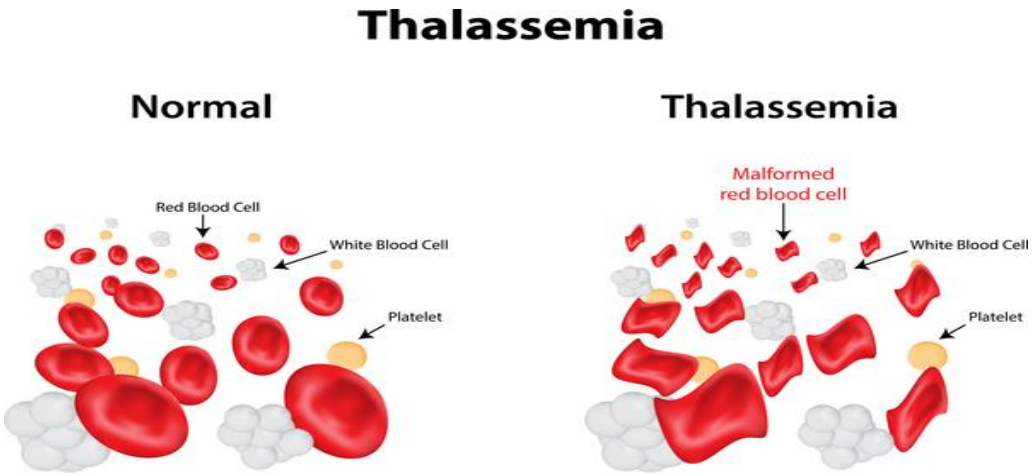


Figure 1-2 Shape of red blood cells for patients with thalassemia

About 60,000 individuals are born annually with thalassemia.[33]. It is a severe inherited anemia arising from the failure of hemoglobin synthesis.[34]. Symptoms of thalassemia do not appear at birth immediately, because the appearance of a child with thalassemia is not different to that of other newborns, but over time, especially after the first six months, the child begins to suffer from anemia, and prominence in the face bones appears, and the child becomes susceptible to infections. We will talk about that in Chapter 3.

Thalassemia disease requires regular blood transfusion every three to four weeks, accompanied by pain to the patients and great suffering to their families. On top of this, the transfer of continuous blood also leads to the accumulation of iron in the vital organs of the body, such as the liver and heart, which leads to serious complications. We will talk about that in Chapter 4, the result of all complications caused by thalassemia that lead to the increasing number of deaths that will remember in the chapter 5.

1.5.3 Overview of the thesis

In order to be able to learn the aim of research, the thesis divided into four distinct parts as shown in figure 1. 3

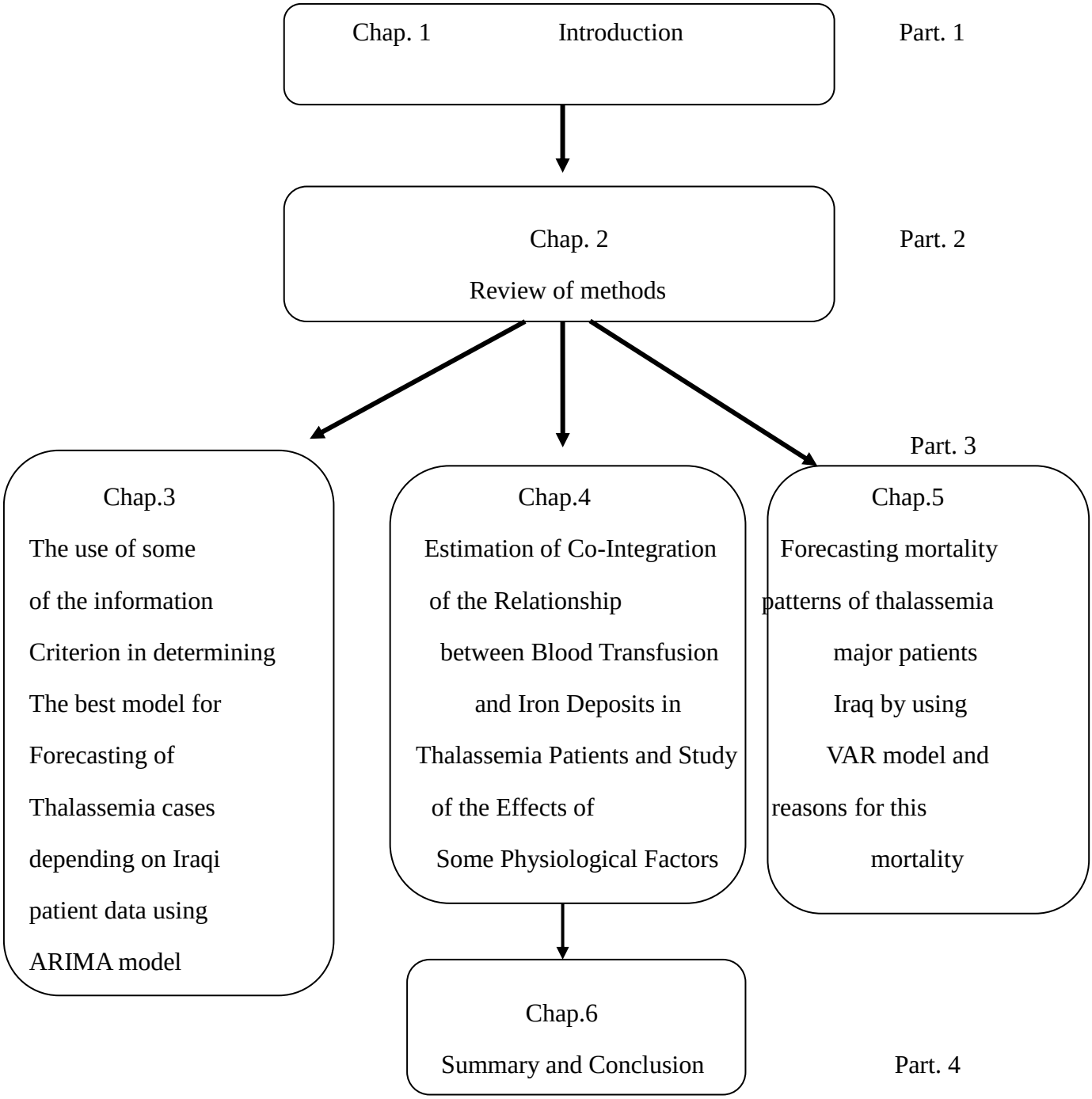


Figure 1-3 Overview of the thesis

Part 1 (Chapter 1) frames the present thesis by introducing the health related issues problem statement rational of the study, study objective research questions and literature review.

Part 2 (Chapter 2) outlines the theoretical framework for the present thesis. The aim is to present the reader a deeper understanding of the time series methods, and how it has evolved through time, give an idea of the models used.

Part 3 (Chapter 3) Used in this chapter model Box Jenkins (ARIMA) to forecasting the thalassemia patients for the years 2016 to 2018 in Chapter 4. This chapter dealt with the use of co-integration model to find the relationship between the blood dosing, which given for thalassemia patients and increase the proportion of iron, and the effects of some physiological factors on the incidence of the disease. Chapter 5 forecasting the number of deaths in patients with thalassemia for the years 2016-2020 using the VAR model, and find out the causes of death.

Finally, Part 4 (Chapter 6) will conclude on the finding in part 3 and reflect on possible future research and extensions.

2 REVIEW OF METHODS

2.1 Introduction

In this review the theoretical basis for the time-series models and the types of time series and the goal of time series analysis and definition of some of the concepts used in subsequent chapters.

2.2 Definition of Time Series

Time is the most important factor, which ensures success in all a business. It is hard to keep up with of time. However, technology developed some more powerful methods that we can see things early time ahead of time.

A time series is a sequence of data points, typically consisting of successive measurements or observations on quantifiable variables, made over a time interval [35]. Usually the observations are chronological and taken at regular intervals (days, months, years), but the sampling could also be irregular. According to [35], time series can be represented as a set of observations x_T, each one being recorded at a specific time T; written as:

$$\{x_1, x_2, x_3, \dots\dots\dots, x_t\}.$$

If a time series has a regular pattern i.e. trend, then a value of the series should be a function of previous values. If x is the target value that is to be model and predicted, and x_t is the value of x at time t, then the goal is to create a model of the form:

$$x_t = f(x_{t-1}, x_{t-2}, x_{t-3}, \dots\dots x_{t-n}) + e_t, \tag{2.1}$$

where $x_{t-1}$is the value of x for the previous observation, x_{t-2} is the value two observations ago, etc., and e_t represents noise that does not follow a predictable pattern (this is called a *random shock*). Values of variables occurring prior to the current observation called lag values.

If a time series follows a repeating pattern, then the value of x_t is usual high correlated with x_t-cycle where cycle is the number of observations in the regular cycle. For example [36], monthly observations with an annual cycle often modeled by:

$$x_t = f(x_{t-12}). \quad (2.2)$$

Time series analysis and related research methods will represent the sophisticated leap in the ability to data analyze longitudinal data collected about topics or units. in Early time series designs, especially as it used is within all science, heavily dependent on graphs analysis to describe and the interpretation of results. While graphical methods are useful and still availability important additional information to the understanding of a time series process, the ability to achieve a sophisticated statistical methodology to influence on this class of data has revolutionized the area of research [37].

Currently, Time series are used in statistics, and signal processing, learn patterns mathematical finance, econometrics, , weather forecasting, earthquake prediction, astronomy, control engineering, communications engineering, medicine science and significantly in any areas of applied science and engineering which involves temporal measurements [36].

2.3 Types of time series:

There are two types of time series: single variable and multivariate. Single variable time series measured variable is only one over time, but the multiple time series are those, measured more than a variable in the same time.

2.4 Objectives of time series analysis

There are two main goals for the analysis of time series:

1. To understand the basic structure of the time series by breaking it down into its components,

2. To fit a mathematical model and then proceed to predict the future[38].

Basically, there are two ways for the analysis of time series, which is associated with the time area (the trend component) or the frequency domain (periodic component).It represents a time-domain approach to the time series as a function of time[39].

The main concern is explore whether the time series have a tendency (high or low) and, if so, to fit with the prediction model. The approach is based on the frequency domain on the assumption that most of the ordinary, and therefore it is likely that the league can be unpredictable, and the behavior of the time series[40].

Thus, the main concern of this approach is to determine the rotating components is an integral part of the time series. Choose between frequency domain and time domain mainly depends on the types of questions, which raised in the various fields of study.

2.4.1 Components of a time series [41]

A time series is composed of the following four elements:

2.4.1.1 Trend *Components*

The trend can usually it would be detected by the inspection of the time series. It may be upward, downward or constant, depending on a slope of the trend-line. The trend-line equation of the line is the equation of the regression line of y (t) on t.

2.4.1.2 Seasonality *Components*

The seasonal factor it can easily be detected it from the graph of the time series data. Peaks and troughs that occur at regular time intervals, suggesting the variable attains maxima and minima, usually represent it. interval between any two successive peaks or troughs is known as the period [42].

2.4.1.3 Cycle *Components*

It is the cycle resembles a season except that its period usually much longer period. Cycles occur as result of changes of qualitative nature, that is, the changes in taste, trend and fashion for example. A cycle is very difficult to detect visually from a time series graph and subsequently mostly assumed to be do not remember, especially for short-term data[43].

2.4.1.4 Residuals

Residuals also known as errors that put on the account of unpredictable external factors known as freaks of nature. They are the differences between the forecast and

observed values of the variables. Theoretical values are the combination (additions or multiplication) of trend, seasonality and cycle, it is assumed to residuals usually distributed to the scope and a long range of time, they cancel one another in such a way their sum is zero [44].

2.5 Definitions and concepts:

2.5.1 Stationary and Non-stationary Series

Stationary is a critical assumption in time series models. Stationary mean homogeneity the sense that the chain behaves in a similar regardless of the time, which means that the characteristics of statistical time series, no change over time and remain fixed. Words finer, stationary mean that the distribution of probability is of fixed over time, Non-stationary series have systematic trends, such as linear, quadratic, and other. A non-stationary series that made stationary by differencing called non- stationary in the homogenous sense. However, practically, a much weaker definition of stationary, often referred to as weak stationary is used [45].

The condition of weak stationary requires that the elements of the time series should have a common finite expected value and that the auto covariance of two elements should depend only on their temporal separation. Mathematically, these conditions are:

Mean = μ and Variance = σ^2 are constant for all values of t

Covariance $\gamma(x_s, x_r)$ is a function of $(s - r)$ only.

There are three of the basic criterions for a series, through which classified as stationary series.

1. The mean of the series should not be a function of time instead of that should be a constant. The graph below has the left hand graph satisfying the condition whereas the graph other has a time dependent mean [46]

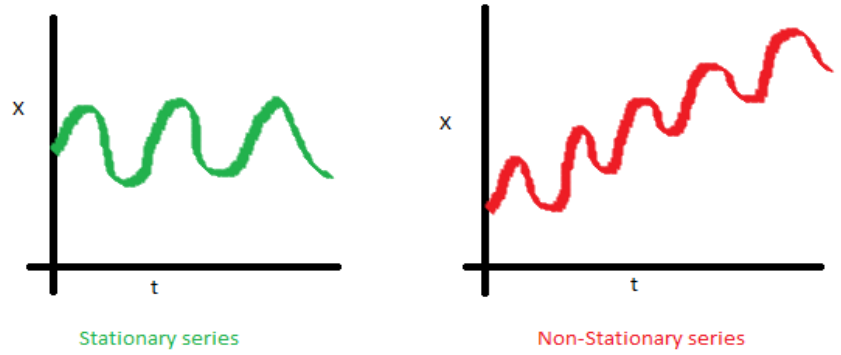


Figure 2-1: Time series, one stationary and the other is non-stationary in Mean

2. The variance of the series should not be a function of time, this property known as homoscedasticity.

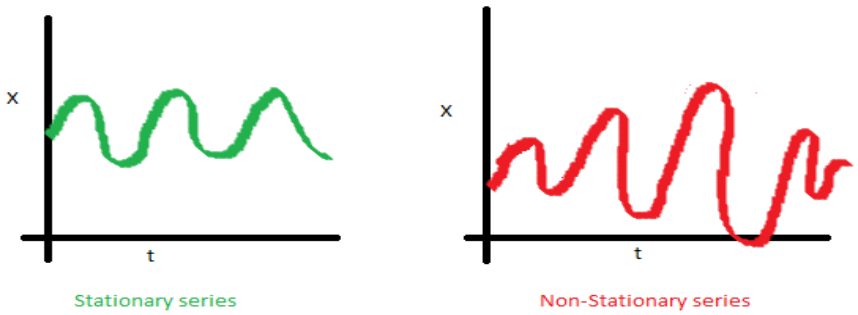


Figure 2-2 Time series, one stationary and the other is non-stationary in variance

3. The covariance of the i-th term and the (i + m)-th term should not be a function of time. In the following graph, you will notice the spread becomes closer as the time increases[46].

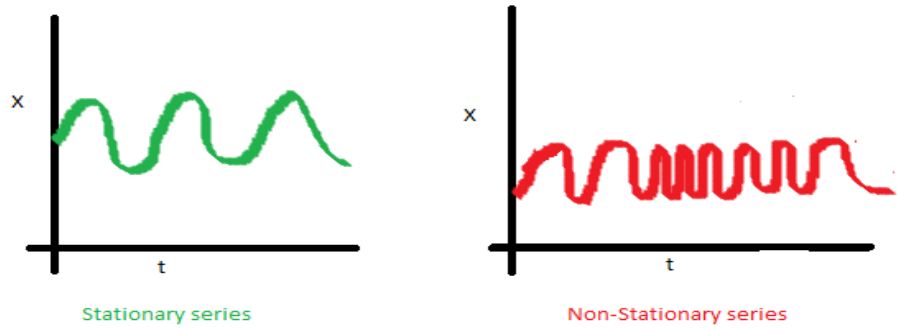


Figure 2-3 Time series, one stationary and the other is non-stationary in covariance

The first stage of building time series model is to identify whether the variable, which is forecasting, is stationary in the time-series or not. By stationary, with auto covariance functions, we can define the covariance stationarity, or weak stationarity. In the literature, usually stationarity means weak stationarity, unless otherwise specified. The time series $(x_{t,}t\epsilon z)$ [47], where z is the integer set is said to be stationarity if

$$var(x_t) < \infty, t \in z,$$

$$EX_t = \mu, \ \forall t \in Z,$$

$$\gamma x(s,t) = \gamma x(s+h, t+h), \ \forall s,t,h \in Z. \tag{2.3}$$

The time plot of the $\{x_t\}$ must have three features: finite variation, constant first moment, and that the second moment $\gamma x(s,t)$ only depends on (t - s) and not depends on s or t In light of the last point, we can rewrite the auto covariance function of a stationary process as

$$\gamma x(h) = Cov(x_t, x_{t+h}), \ t,h \in Z. \tag{2.4}$$

Also, when $x_t$is stationary, we have

$$\gamma x(h) = \gamma x(-h),$$

When h=0, $\gamma x = (0) = cov(x_t, x_t)$ is the covariance of x_t, so the autocorrelation function for stationary time series $x_t$is defined to be [48]

$$px(h) = \frac{\gamma x(h)}{\gamma x(0)}.$$

2.5.2 Differencing

Differencing this simply means subtracting the value of an previous observation from the value of a later observation, calculated differences among pairs of observations at some lag that make a non-stationary series stationary [49].

There are potential shifts in both the mean and the dispersion by the time for this series. The mean may be edging upwards, and the variability may be increasing. If the mean is changing, the trend removed by differencing once or twice, if the variability it is changing, the process may made stationary by logarithmic transformation. Difference the scores this is the easiest way to make a no stationary mean stationary, the number of times you have to find a difference to make a more stationary sets the value of d. d=0 be a model is originally stationary and it has no trend. When the series is difference once, d=1 and linear trend is removed. When the difference is then difference, d=2 and both linear and quadratic trend be removed, for no stationary series, d values of 1 or 2 are usually it would be adequate to make the mean stationary [50].

If the time series data analyzed exhibits a deterministic trend, the spurious results can avoid by trend. Sometimes the non-stationary series may combine a stochastic and deterministic trend at the same time and to avoid obtaining misleading results both differencing and trend should be applied, as differencing will remove the trend in the variance and trend will remove the deterministic trend[51].

2.5.2.1 Dickey–Fuller test

In statistics, the Dickey–Fuller test, tests the null hypothesis of whether a unit root is present in an autoregressive model. The alternative hypothesis is different depending on which version of the test is used, but is usually stationary or trend-stationary. It is named after the statisticians David Dickey and Wayne Fuller, who developed the test in 1979 [52].

A simple model is

$$Y_t = py_{t-1} + U_t, \tag{2.5}$$

where y_t is the variable of interest, t is the time index p is a coefficient, and u_t is the error term. A unit root is present if p=1.The model would be non-stationary in this case. The regression model can write as

$$\nabla y_t = (p-1)y_{t-1} + u_t = \delta y_{t-1} + u_t, \tag{2.6}$$

where ∇ is the first difference operator. This model can estimate and testing for a unit root is equivalent to testing $\delta = 0$ (where$\delta \equiv p - 1$). Since the test done over the residual term rather than raw data, it is not possible to use standard t-distribution to provide critical values. There are three main versions of the test:

Test for a unit root:

$$\nabla y_t = \delta y_{t-1} + u_t, \tag{2.7}$$

Test for a unit root with drift

$$\nabla y_t = a_0 + \delta y_{t-1} + u_t, \tag{2.8}$$

Test for a unit root with drift and deterministic time trend

$$\nabla y_t = a_0 + a_1 t + \delta y_{t-1} + u_t. \tag{2.9}$$

Each version of the test has its own, critical value, which depends on the size of the sample. In each case, the null hypothesis is that there is a unit root $\delta = 0$. The tests have low statistical in that they often cannot distinguish between true unit-root processes ($\delta = 0$) and near unit-root processes (δ is close to zero). This called the "near observation equivalence" problem. The intuition behind the test is as follows. If the series y is stationary (or trend stationary), then it has a tendency to return to a constant (or deterministically trending) mean.

Therefore large values will tend to be followed by smaller values (negative changes), and small values by larger values (positive changes). Accordingly, the level of the series will be a significant predictor of the next period's change, and will have a negative coefficient. If, on the other hand, the series is integrated, then positive changes and

negative changes will occur with probabilities that do not depend on the current level of the series, in a random walk [53]

2.5.3 The Trend (d)

The trend it is simply the underlying long -term behavior or pattern of the data or series. Defined trend behalf of the 'long term' movement in a time series without calendar related and irregular the effects of, and is a reflection of the statute level. It is the result of effect such as price inflation, population growth and general economic changes. A model with two trend terms (dC2) has to be difference twice to make it stationary. The first difference removes linear trend, the second difference removes quadratic trend, and so on [54].

2.5.4 Random time series

A time series, in it the observations wiggle around a constant mean, have a constant variance and is statistically independent, called a random time series, or other words, the time series does not exhibit any pattern:

- The observations do not trend up or down
- The variance does not increase or decrease with the passage of time
- The observations do not tend to be greater in some periods than it was in other periods[38].

A random model can write as

$$x_t = \mu + e_t, \tag{2.10}$$

μ is a constant, the average x_t, e_t is the residual (or error) term which is assumed to have a zero mean, a constant variance, and e_t's statistically independent [55]. We can test if the time series random or not random by the following procedures:

1. Visually, if it was time series plot shows any trend or not.

2. By looking at the correlogram of the time series.

3. Statistically, by conducting statistical tests to see whether the random series.

2.5.5 Autocorrelation:

The most noticeable of time series is that successive values are not independent; they are correlated with each another, i.e. they are serially correlated or auto correlated Auto covariance and autocorrelation functions be important tools in the description the serial (or temporal) dependence structure of a univariate time series[3].

Autocorrelation it is a measure of the dependence of time series values at a specific time on the values at another time. It is the Pearson correlation between all a pair of points in the time series with a given separation in time or lag. Positively auto correlated series are sometimes referred to as persistent because high values tend to follow high values and low values tend to follow low values. Negatively auto correlated series are characterized by reversals from high to low values from one time segment to the next, and vice versa[44].

The first order correlation (i.e. lag = 1) is the correlation coefficient of the first N −1 observations

[x_t : t = 1, 2, ...,N − 1] and the next N − 1 observations [x_t : t = 2, 3, ...,N]. It given by the following formula:

$$r_1 = \frac{\sum_{t=1}^{n-1}(x_t-\bar{x}_{(1)})(x_{t+1}-\bar{x}_{(2)})}{\sqrt{\sum_{t=1}^{n-1}(x_t-\bar{x}_{(1)}}\ ^2\sqrt{\sum_{t=1}^{n-1}(x_{t+1}-\bar{x}_{(2)})^2}}, \tag{2.11}$$

where $x_{(1)}$ is the mean of the first N − 1 observations and $x_{(2)}$ is the mean of the last N – 1 observations[56].

2.5.6 Partial Autocorrelation Function [57]

Partial autocorrelation function measures the degree of association between Y_t and Y_{t+k} when the effect of other time lags on Y are held constant. The partial autocorrelation function PACF denoted by the set of partial autocorrelations at various lags k are defined by

$$r_{kk} = \frac{rk-\sum_{j=1}^{k-1} rk-1,jrk-1}{1-\sum_{j=1}^{k-1} rk-1,jrj}\ , \tag{2.12}$$

where $r_{k,j} = r_{k-1,j} - r_{kk}r_{k-1,k-1}$, j= 1, 2, 3….k-1. Particular, partial autocorrelations it is useful in identifying ranking of an autoregressive model.

3 The information criterion in determining the best model for Forecasting of Thalassemia

3.1 Introduction

One of the most important elements of building health is the prevention of all diseases, including serious diseases, such as Thalassemia, that cause a high percentage of deaths compared to other diseases, due to the increasing number of people infected with this disease in recent times. This study conducted in order to forecast the prevalence of this disease in future, which has been increasing in all areas of Iraq, especially in the province of Maysan.

This study based on monthly data for the patients with Thalassemia for the period between 2005 and 2015, and used data of patients as a series of time for the purpose of analysis for optimal modeling using the ARIMA Methodology, the forecasting to predict the numbers of people with this disease in subsequent periods in order to take the necessary measures and substitutions to reduce morbidity in the future.

A time series typically consists of a set of observations of a variable taken at evenly spaced intervals of time [58].The most comprehensive of all popular and widely known statistical models which have been used in the last four decades for time series forecasting are the Box-Jenkins method. However, the ARIMA model is only a class of linear model and, thus, it can only capture linear feature of data time series [59].

Many standard determining the rank of models have proposed by researchers [60-62]. There are different models of precision that can juggle in time series analysis to clarify the given set of data, not be easy to choose the better model in many cases. It has developed several criteria to compare models in the process of selecting the rank of model and derive the importance of selecting the rank of the model from the fact that choosing the lowest rank of the actual rank of the model leads to inconsistency of the

model parameters, while choosing a higher rank than the actual rank of the form to increase the model variance. This leads to a loss of accuracy due to the increase in the number of model parameters chosen.

3.2 Method

For realizing the forecast of the analyses time-series we use modern methods, such as ARIMA models, because they are among the models that can analyses large time-series data and forecast future cases.

3.2.1 Models of Box & Jenkins:

The pioneers in this area were Box and Jenkins, who popularized an approach that combines the moving average and the autoregressive models (1971). An ARMA (*p, q*) model is a combination of AR (*p*) and MA (*q*) models, and is suitable for univariate time -series modeling. In an AR (*p*) model, the future value of a variable assumed to be a linear combination of *p* past observations and a random error, together with a constant term. Mathematically, the AR (*p*) model can be expressed as[7]:

$$Y_t = c + \sum_{i=1}^{p} \varphi_1 y_{t-i} + \varepsilon_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t, \quad (3.1)$$

Here y_t and ε_t are respectively the actual value and random error (or random shock) at time period t,φ_i (i= 1, 2... p) are model parameters and c is a constant, the integer constant p known as the order of the model. Sometimes the constant term omitted for simplicity.

Usually, for estimating parameters of an AR process using the given time series, the Yule-Walker equations are used. Just as an AR (p) model regresses against past values of the series, an MA (q) model uses past errors as the explanatory variables. The MA (q) model is given by[63]

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots . + \theta_q \varepsilon_{t-q} + \varepsilon_t, \quad (3.2)$$

Here μ is the mean of the series θ_j (j =1, 2... *q*) are the model parameters and *q* is the

order of the model. The random shocks assumed to be a white noise process, i.e. a sequence of independent and identically distributed (i.i.d.) random variables with zero mean and a constant variance σ^2 . Generally, the random shocks assumed to follow the typical normal distribution. Thus, conceptually, a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations. Fitting an MA model to a time series is more complicated than fitting an AR model because in the case of the former the random error terms are not fore seeable. Autoregressive (AR) and moving average (MA) models can be effectively combined together to form a general and useful class of time series models, known as the ARMA models. Mathematically an ARMA (p, q) model represented as

$$y_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_1 y_{t-1} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \quad (3.3)$$

Usually ARMA models manipulated by using the lag operator notation, the lag or Backshift operator is defined as $Ly_t = y_{t-1}$ .Polynomials of lag operator or lag polynomials are used to represent ARMA models as follows:[63]

$$\text{AR } (p) \text{ model: } \boldsymbol{\varepsilon_t = \varphi(L) y_t}, \quad (3.4)$$

$$\text{MA (q) model: } \boldsymbol{y_t = \theta(L) \varepsilon_t}, \quad (3.5)$$

$$\text{ARIMA (p, q) model: } \boldsymbol{\varphi(L) y_t = \theta(L) \varepsilon_t},$$

(3.6) Here $\varphi(L) = 1 - \sum_{i=1}^{p} \varphi_1 L^i$ and $\theta(L) = 1 + \sum_{J=1}^{Q} \theta_j L_j$. It shown in that an important property of AR (p) process is invertible, i.e. an AR (p) process can always write in terms of an MA (∞) process. Whereas for an MA (q) process to be invertible, all the roots of the equation θ (L) = 0 must lie outside the unit circle, this Condition known as the inevitability condition for an MA process.

3.2.2 ARIMA model:

In both statistics and econometrics, time series analysis of an autoregressive integrated moving average, an ARIMA model is the integration of an autoregressive moving average (ARMA) model. These models fitted to time-series data, either to better

understanding the data or to forecast future points in the series (forecasting). It is applied, in some cases where the figures show proof that they are not stationary, where an initial differencing step (corresponding to the integrated fraction of the model) can be applied to reduce the non-stationarity[64].

Non-periodical ARIMA models which are generally denoted by ARIMA (p,d,q) where parameters p, d, and q are non-negative integers, p is the order of the Autoregressive model, d is the degree of differencing, and q is to arrange of the Moving-average model. ARIMA models form is an important part of the Box-Jenkins approach to time-series modeling [65].

ARIMA models can see as a chain of two models. The first is not fixed:

$$x_t = \theta_0 + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots \phi_p x_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots \theta e_t - q, \qquad 3.6)$$

where x_t and e_t are the actual values and random error at time t, respectively, ϕ_i (i = 1,2,…,p) and θ_j (j = 1,2,…,q) are model parameters. The p and q are integers and often referred to as orders of autoregressive and moving average polynomials respectively.

3.2.3 Steps of ARIMA Methodology [9]:

- Analysis of the series: The first step in the process of modelling is to check for the stationary of the time series data.

- Identification of the model: This step aimed to detect periodically and to identify the order of seasonal autoregressive terms and seasonal moving average terms. This stage includes calculation of the estimated autocorrelation function (ACF) and estimation of partial autocorrelation function (PACF) these functions measure the statistical dependence between observations of data outputs.

- Estimation of ARIMA parameters: The estimation of ARIMA parameters achieved by the nonlinear least squares method. The values of the model coefficients are determined in relation to a particular criterion; one of these may be the maximum likelihood criterion.

It can show that the likelihood function associated with a correct ARIMA model, used to determine the estimates of maximum likelihood of the parameters, contains all the useful information from the data series about the model's parameters.

- **Diagnostic checking:** In this stage it is assumed that the errors represent a stationary process and the residues are white noise (or independent if the distribution is normal), a normal distribution with mean and variance stable. The tests used to validate the model based on the estimated residues. It checked that the components of this vector are auto correlated. If there is autocorrelation, the checked model do not correctly specified. In this case, the dependencies between the components series specified in an incomplete manner, and we have to return to the model identification step and try another model. Otherwise, the model is good and can be used to make predictions for a given time horizon.

- **Forecasting:** stage forecast is the final stage that you can find forecasting is the process of making predictions of the future，based on past and present data and most commonly by analysis of trends.

3.3 Criteria for selection of the rank of the model:

3.3.1 Bayesian Information Criterions (BIC) [61]

$$BIC = n\,ln\hat{\sigma}_a^2 - (n-M)\ln\left(1-\frac{M}{n}\right) + M\ln(n) + M\ln\left[\frac{\left(\frac{\hat{\sigma}_Y^2}{\hat{\sigma}_a^2}1\right)}{M}\right], \tag{3.7}$$

where

P: model rank

n: Views

M: The number of parameters

$\hat{\sigma}_Y^2$: Estimator series variance

$\hat{\sigma}_a^2$: Estimator error variance

$$\hat{\sigma}_a^2 = \sum_{t=1}^{n}(y_t - \hat{y}_t)^2/(n-p). \quad (3.8)$$

3.3.2 Akaike Information Criterion[61]

$$AIC(M) = n\, ln\, \hat{\sigma}_a^2 + 2M, \quad (3.9)$$

Or

$$AIC(p,q) = ln\, \hat{\sigma}_a^2 + 2(p+q)/n, \quad (3.10)$$

where

M: p + q, p, q: model rank, n: views, $\hat{\sigma}_a^2$: estimator error variance.

3.4 The application of data:

This study based on the time-series data provided by the hereditary blood disease center in Iraq (Maysan) from people diagnosed with Thalassemia for the period 2005-2015.

3.4.1 Analysis of the Series

The first step in the process of modelling is to check for the stationary of the time series data. This is done by observing the graph of the data or autocorrelation and the partial autocorrelation functions[38]. It notes, through a graph of the time series, that there are high rates of Thalassemia as compared with previous years (Figure 3-1)

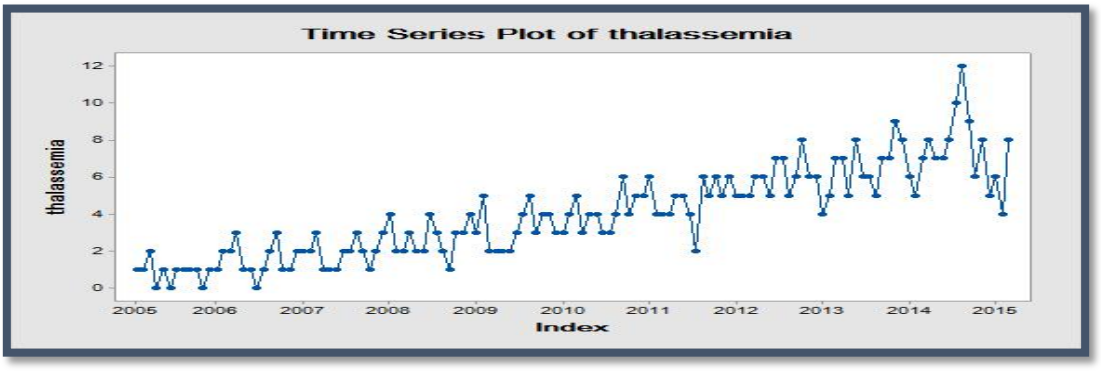


Figure 3-1 **graph of original series, increase the number of patients suffering from thalassemia year 2005-2015 in southern Iraq, time series before differencing shows the variability of the series appears to be changing with time. Therefore the mean and variance are not constant, suggesting that the series is not stationary**

3.4.2 Stationary:

In Figure 3-2 the thalassemia data, clearly shows that the data is not stationary (actually, it shows an increasing trend in time series). The ARIMA model cannot be built until we make this series stationary. We first have to differentiate the time series 'd' times, to obtain a stationary series in order to have an ARIMA (p, d, q) model with 'd' as the order of differencing. Care should be taken in differencing, as over differencing will tend to an increase in the standard deviation, rather than a reduction. The best idea is to start with differencing of the lowest order (of first order, d=1) and test the data for unit root problems [59]

$$\textit{First Difference:} \quad Z_t = y_{t-1}, t = 2, 3 \ldots n$$

$$\textit{Second Difference:} \quad Z_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}), \textit{ where } t = 3, 4 \ldots n$$

As a result we obtained a time-series of first order differencing and Figure 3-2, below, is the line plot of the first order differenced Thalassemia data. It can easily infer from the graph that the time series appears to be stationary both in its mean and in variance. Moreover, the time series data subjected to Dickey-Fuller test to check the number of differenced time-series data for stationarity.

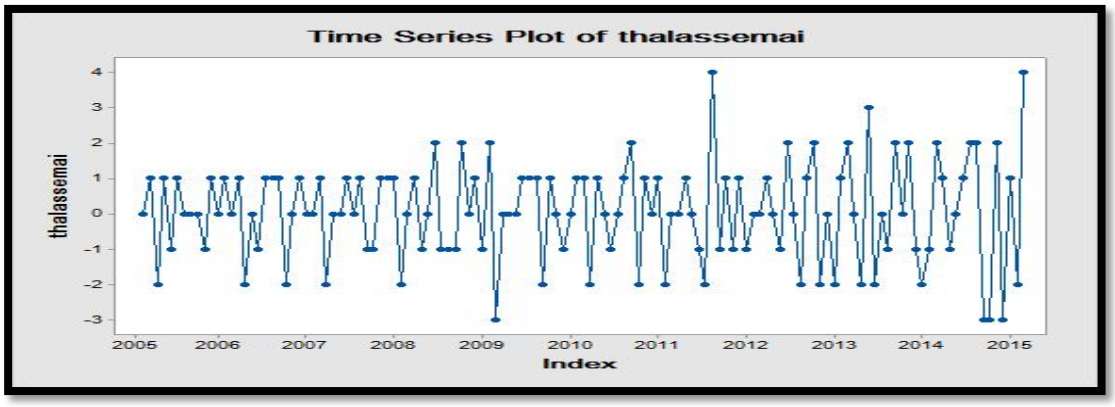


Figure 3-2 Time series for after first difference d = 1 become stationary

3.4.2.1 Using the adjusted (ADF) test[53]:

Our null hypothesis (H_0) in the test is that the time series data is non-stationary; while the alternative hypothesis (H_a) is that, the series is stationary. The hypothesis then tested by performing appropriate differencing of the data in d-th order and applying the ADF tests to the differenced time series data. First order differencing (d=1) means we generate a table of differenced data of current and immediately previous time $\Delta x_t = x_t - x_{t-1}$. The ADF test result, as obtained upon application, shown below (using Eviews program) in Table 3-1.

We, therefore, fail to accept H_0 and, hence, can conclude that the alternative hypothesis is true i.e. the series is stationary in its mean and variance. Thus, there is no need for further differencing of the time series and we adopt d = 1 for our ARIMA (p, d, q) model. This test enables us to go further in steps for ARIMA model development i.e. to find suitable values of p in AR and q in MA in our model. For that, we need to examine the correlogram and partial correlogram of the stationary (first order differenced) time series.

Table 3-1: The Augmented Dickey-Fuller test results

Augmented Dickey –Fuller test statistic		t-statistic -11.38261	Prob. 0.0000	
Test critical values	1% level 5% level 10% level	-3.482035 -2.884109 -2.578884		
Augmented Dickey-Fuller Test Equation Dependent variable :D(DTHALASSEMIA) Method Least Squares Sample(adjusted)2005 M05 2015M12				
Variable	Coefficient	Std. error	t-statistic	Prob.
DTHALASSEMIA(-1)	-2.176864	0.191245	-11.38261	0.0000
D(DTHALASSEMIA(-1)	0.699137	0.145316	4.811158	0.0000
DTHALASSEMIA(-2)	0.381127	0.085537	4.455672	0.0000
C	0.094297	0.109530	0.860923	0.3909
R- sequard	0.719809	Mean dependent VAR		0.046875

Adjusted R- sequard	0.713030	S.D dependent var	2.310058
S.E. of regression	1.237487	Akaike info criterion	3.294795
Sum squared resid	189.8905	Schwarz criterion	3.383920
Log likelihood	206.8669	Hannan –Quinn criter	3.331007
F- Statistic	106.1852	Dubin- Watson stat	1.998406

3.4.3 Correlogram and Partial Correlogram

Figure 3-3, below, represents the plot of correlogram (auto-correlation function, ACF) from lags 1 to 20 of the first order differenced time-series of the Thalassemia patients (Figure 3-3). The above correlogram infers that the auto-correlation and partial autocorrelation between lag 1 and 20 does not exceed the significance limits and auto-correlations tail off to zero the autocorrelation at rest Figure (3- 4).

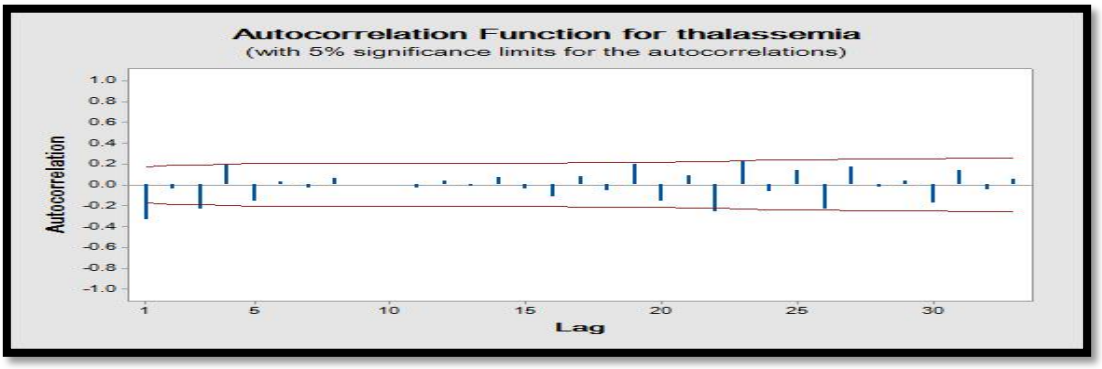


Figure 3-3: **Plot of the Autocorrelation functions of the differenced Series**

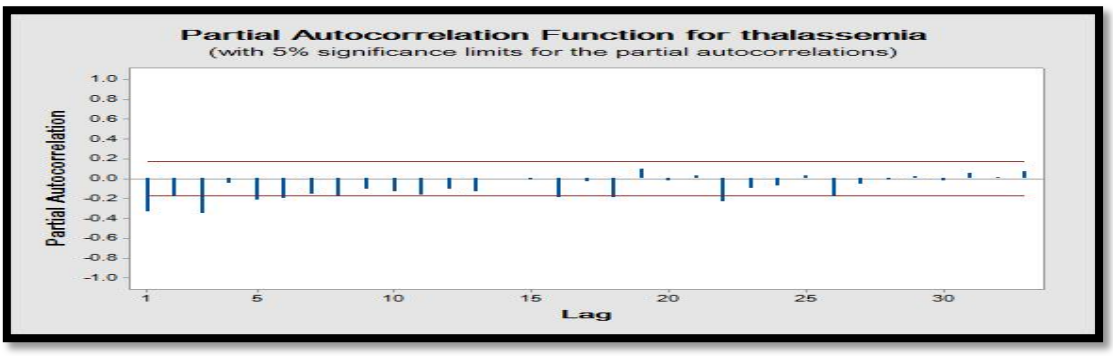


Figure 3-4: Plot of the Partial Autocorrelation functions of the differenced Series

3.4.4 Selecting the rank of model:

Table 3-2 shows the results of information criterion for various models of the time series

Models	AIC	BIC
ARIMA(1, 1, 0)	89.7921	-65.0102
ARIMA(0, 1, 1)	38.5633	-122.7428

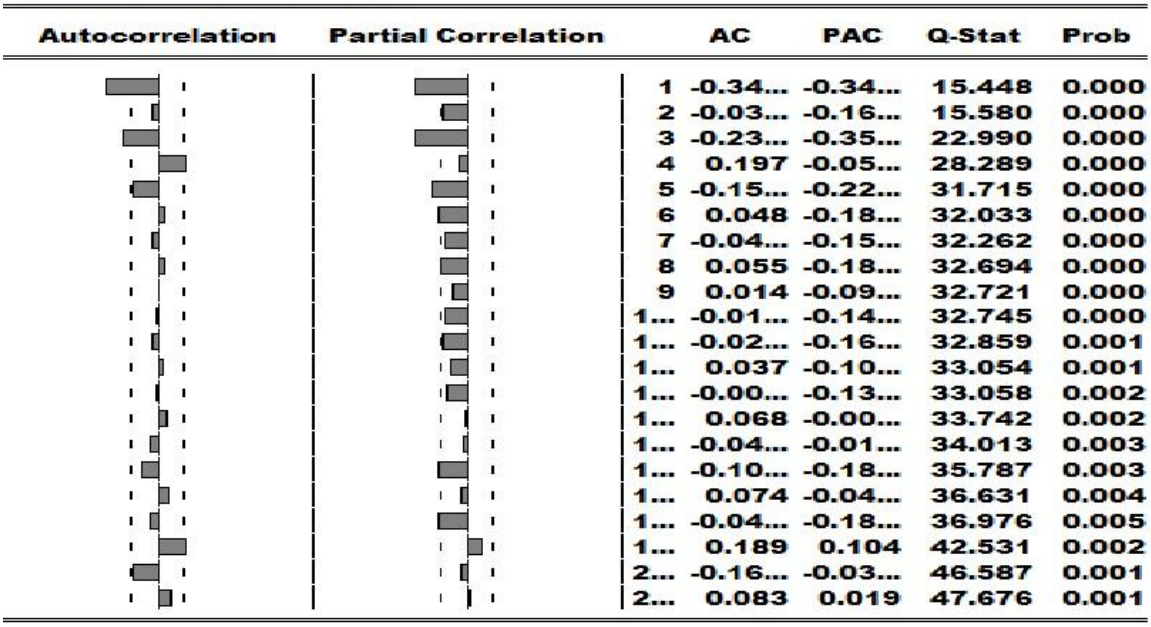
Seen from the table (3-2) that the best model for this series is ARIMA (0, 1, 1) for having the lowest values for the standards of information. This is perhaps the most commonly used model in forecasting. It is the exponential smoothing model. The general form of the model is:

$$Z_t - Z_{t-1} = c + a_t - \theta a_{t-1}, |\theta| < 1. \quad (3.11)$$

Following the common practice, we shall assume $c = 0$. Since the model is invertible, the π weights are $\pi_i = \theta_i - 1(1-\theta)$, for $i \geq 1$. Thus

$$Z_t - Z_{t-1} = \quad 0.05577677 + 0.98928205\, _{t-1}$$

Date: 02/16/17 Time: 10:42
Sample: 2005M01 2015M12
Included observations: 131



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	-0.34...	-0.34...	15.448	0.000
		2	-0.03...	-0.16...	15.580	0.000
		3	-0.23...	-0.35...	22.990	0.000
		4	0.197	-0.05...	28.289	0.000
		5	-0.15...	-0.22...	31.715	0.000
		6	0.048	-0.18...	32.033	0.000
		7	-0.04...	-0.15...	32.262	0.000
		8	0.055	-0.18...	32.694	0.000
		9	0.014	-0.09...	32.721	0.000
		1...	-0.01...	-0.14...	32.745	0.000
		1...	-0.02...	-0.16...	32.859	0.001
		1...	0.037	-0.10...	33.054	0.001
		1...	-0.00...	-0.13...	33.058	0.002
		1...	0.068	-0.00...	33.742	0.002
		1...	-0.04...	-0.01...	34.013	0.003
		1...	-0.10...	-0.18...	35.787	0.003
		1...	0.074	-0.04...	36.631	0.004
		1...	-0.04...	-0.18...	36.976	0.005
		1...	0.189	0.104	42.531	0.002
		2...	-0.16...	-0.03...	46.587	0.001
		2...	0.083	0.019	47.676	0.001

Figure 3-5 plot of the autocorrelation and partial autocorrelation function of the differenced series.

3.4.5 The forecasting

The results showed that after we apply (eviews , gretl, minitab) specific statistical programs on our data and charts, there is an increase in cases with Thalassemia in the coming years from 2016 to 2018. According to the figure(3-6) figure (3-7) and table (3-3), where the number of patients monthly will be between (7-11) patients and these numbers are high compared to previous years, where highest forecasting for the patients is during June of 2018. Some months during the period of 2005-2015 did not record any cases while in the forecasting period 2016-2018 every month accepted to have some patients.

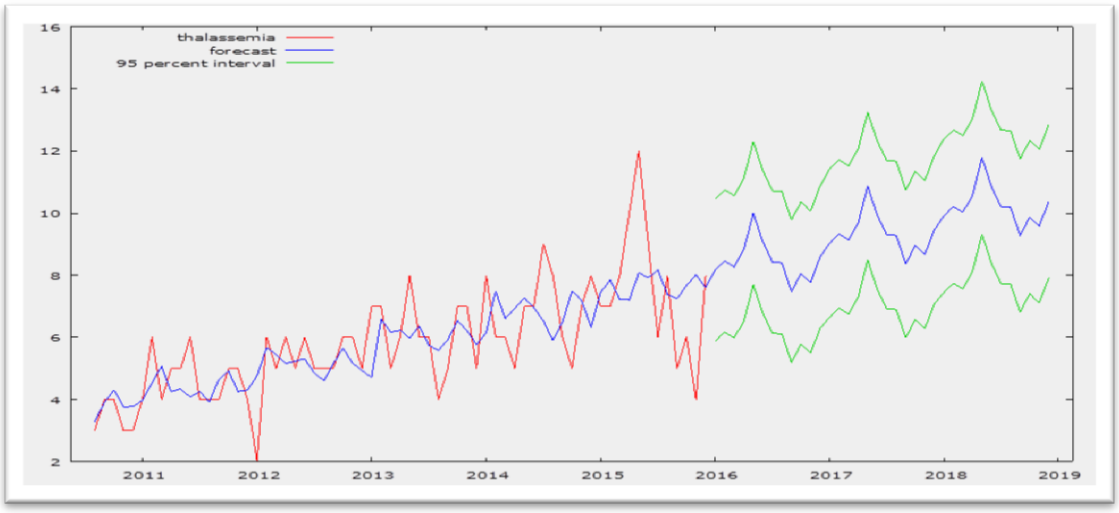


Figure 3-6 forecasting the expected number of patients with thalassemia years (2016-2018) using a form ARIMA (0, 1, 1)

Figure3-7 forecasting the expected number of patients with thalassemia years (2016-2018) using a form ARIMA (1, 1, 0).

Table 3-3: Year Forecasting for thalassemia patients, the number of people forecaster their thalassemia in 2016-2018 For 95% confidence intervals, z (0.025) = 1.96.

Obs	Forecast ARIMA(1,1,0)	Forecast ARIMA(0,1,1)	obs	Forecast ARIMA(1,1,0)	Forecast ARIMA(0,1,1)
2016:01	6.94	8.17	2017:07	6.82	9.30
2016:02	6.94	8.46	2017:08	8.08	9.28
2016:03	6.82	8.28	2017:09	5.32	8.37
2016:04	7.41	8.82	2017:10	5.82	8.96
2016:05	9.27	10.00	2017:11	4.79	8.67
2016:06	8.01	9.07	2017:12	8.05	9.46
2016:07	7.77	8.43	2018:01	6.99	9.92
2016:08	8.10	8.40	2018:02	6.99	10.22
2016:09	5.65	7.48	2018:03	7.14	10.04
2016:10	5.53	8.07	2018:04	8.07	10.58
2016:11	5.76	7.78	2018:05	9.96	11.78
2016:12	8.08	8.58	2018:06	8.26	10.85
2017:01	7.04	9.03	2018:07	7.33	10.21
2017:02	7.04	9.32	2018:08	8.06	10.19
2017:03	7.54	9.14	2018:09	5.47	9.28
2017:04	8.91	9.68	2018:10	5.62	9.87
2017:05	10.85	10.87	2018:11	5.29	10.39
2017:06	8.61	9.94	2018:12	8.02	9.969

3.4.5.1 final prediction error (FPE)[62]:

Been using the final prediction error (FPE**)** a good estimate of prediction error for model with n parameters is given by the final prediction error:

$$FPE = \sigma_r^2(N,\hat{\beta})\frac{N+n+1}{N-n-1}, \quad (3.12)$$

σ_r^2 =variance of the residuals,

N is the number of values in the estimation data set.

Table 3-4 the values of criterion of final prediction error

Model	FPE
ARIMA(0 ,1,1)	0.03256
ARIMA(1, 1, 0)	0.06542

From Table (3-4) that the model ARIMA (0,1,1) has the lowest value of the criterion of (FPE).

Conclusions:

Cases of Thalassemia will increase within coming years, which means that, currently, no serious efforts offered to solve or treat this disease in Iraq.

Recommendations:

(1) Doing more studies that are similar using data from other provinces of Iraq to overcome the disease.

(2) Doing pre-marriage tests to detect the carriers to limit the future incidence.

(3) The bone marrow transplant is the most suitable treatment of the disease but it is so costly for patients that governmental support will be very helpful.

(4) Should take care to sanitary ware industry pain less and more safety during blood transfusions to relieve pain and prevent infection from bacterial and viral diseases.

4 Estimation of Co-Integration of the Relationship between Blood Transfusion and Iron Deposits

4.1 Introduction

Blood transfusion can prevent the transfer of some of the most serious growth, skeletal and nerve complications of thalassemia major. However, once it has started, comp

Locations of blood transfusion become a major source of morbidity. We must set standards and maintain them to ensure a safe and reasonable approach to the use of blood transfusions in the management of these rare disorders [66]. Chronic iron overload is a serious complication of potentially life-saving blood transfusion, which can lead to excess iron deposits in various tissues of the body, especially the liver, heart and endocrine organs [67].

Once storage capacity exceeded in the body, free iron stimulates the formation of highly reactive hydroxyl radical roots, which leads to membrane damage and denaturation of proteins. This process leads to tissue damage and, ultimately, to disease and large mortality rates [68]. In fact, device failure due to chronic iron overload is the main cause of death in patients with Mrdy-β thalassemia major who receives regular blood transfusions without the proper treatment of chelation. Within 1-2 years from the start of regular blood transfusions, evidence of iron overload is evident as high iron concentration in the liver (LIC) values and high levels of serum ferritin.

There is an increased risk of heart disease caused by iron in thalassemia patients with LIC values above 15 mg iron / g dry weight (dry), and patients with ferritin serum values above 2500 mg/L8.Patients with a number of other congenital issues and acquired anemia. Who may receive frequent blood transfusions are also prone to the negative effects of iron [69].

This study aimed determine the causal relationship between the number of blood transfusions for thalassemia patients and the content of high iron, as well as the effects and relationship of some factors on thalassemia (blood type, age, gender, thalassemia type) with the focus on thalassemia patients in Southern Iraq (Maysan). These data have obtained from the blood Centre of genetic disease and thalassemia in the province. The Engle Granger method used to provide a causal analysis of data.

The (Engle Granger) and Co-integration are of great importance in the field of applied research and statistical analysis, which many researchers had addressed in various areas of research, but it's used in statistical and medical research was scarce and almost non-existent. As thalassemia poses a threat to the lives of people in general, and the seriousness of iron accumulation, because of taking blood doses by people suffering from thalassemia is particularly threatening. The study aimed to determine the long-term causal relationship between high iron levels and a dose of blood given per month for thalassemia patients taking in consideration blood type, gender, age and thalassemia type as important factors that may have an effect on thalassemia morbidity.

The study uses common integration (Engle Granger) and other traditional methods of statistical analysis to measure the relationship between variables.

4.2 Methods and materials

4.2.1 Causal relationship between giving blood doses and high iron:

Here we investigate the co-integration analysis, the study of Granger causality tests to find the direction of the causal relationship between giving blood doses and high iron levels in thalassemia patients. The purpose was studying the relationship between the number of doses of blood given and the increase in the proportion of iron for thalassemia patients. We used annual data for -100 patients-, from 2015, taken from the - Genetic Disease / Thalassemia Centre in Maysan province.

This study provides an overview of an important and relatively recent approach to estimate long-run relationship between blood and iron using 'Co-integration', a technique becoming widely used in macroeconomic modelling.

Before estimating the Co-integration and VAR, it is required to examine the stationarity of the variables. Stationarity means that the mean and variance of the series are constant through time and the auto covariance of the series is not time

varying[70].Therefore, the first step is to test the order of integration (I) of the variables. Integration means that past shocks, remaining undiluted, affects the realizations of the series forever and a series has theoretically infinite variance and a time-dependent mean. For this study, we used tests proposed by Dickey and Fuller [53].Phillips and Perron [71], and Kwiatkowski, Phillips, Schmidt and Shin [72].

For testing the properties of unit root for all variables used, if all of the series are non-stationary in levels, it should be stationary in first difference with the same level of lags. For appropriate lag lengths, we use the Akaike .Information Criterion (AIC) and Schwartz Bayesian Criterion (SBC).The Dickey and Fuller test (ADF) takes the following form:

$$\Delta y_t = \boldsymbol{\alpha}_0 + \boldsymbol{\delta} T + \boldsymbol{\beta} y_{t-1} + \sum_{i=1}^{p} \boldsymbol{\theta}_1 \Delta y_{t-1} + \boldsymbol{\mu}_t. \quad (4.1)$$

ADF regression tests help to root units in YT, which is the logarithm of the number of the patient's blood and takes a year of ratios of iron deposited in the body doses. Where T denotes the deterministic time trend and ΔYt-i is the lagged first differences to accommodate a serial correlation in the error μ and t. α,δ,β, and θ are the parameters to be estimated.

Meanwhile, the Phillips-Peron (PP) test shown by the equation below:

$$\Delta \boldsymbol{y_t} = \boldsymbol{\mu} + \boldsymbol{p y_{t-1}} + \boldsymbol{\varepsilon_t}. \quad (4.2)$$

The PP test is used because it will make a correction to the t-statistics of the coefficient from the AR (1) regression to account for the serial correlation. The PP test is a test of the

hypothesis P=1 in equation 2. However, unlike the ADF test, there are no lagged difference terms. Instead, the equation estimated by OLS and then the t-statistics of the P coefficient corrected for serial correlation in ε_t.

The Co-integration allows the analysis to clarify the true relationship between two variables, by searching for co- integration factor and removing its influence when necessary. Nevertheless, the basic time series integrated at the same class variables, which are first class for the purposes of this study. We used a linear regression model to determine the nature of the relationship between Iron and Blood as follows:

$$\mathrm{BLOOD}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}.\,\mathrm{IRON}_t + \boldsymbol{\varepsilon}_t. \tag{4.3}$$

Blood refers to the number of doses given to the patient within one year, and the proportion of iron: is the iron deposited in the patient's body.

4.3 Results

Statistical study of the data for our results has shown, it can be divided into several components, listed in the following order:

4.3.1 Results of the statistical analysis of the two series time periods:

The first step in the two-time series analysis is drawing views variables to determine the general direction of the two, where it represents the figure (4-1) time series for each of the doses of iron and blood given to thalassemia patients within months of 2015. We have observed that there is a growing trend where the more portions given to the patient, the greater the proportion of iron deposits

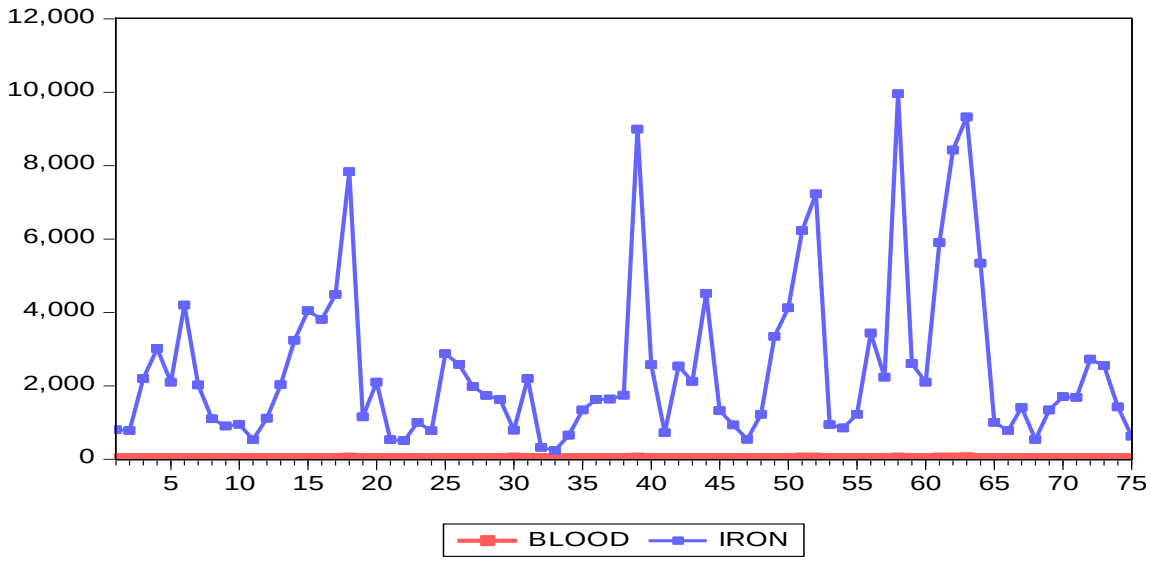


Figure 4-1 the general trend of a series of blood and iron

4.3.2 Results test Stationarity time series:

Preprocessing testing aims to examine the properties of time series for each of the iron-rate and portions of blood given to thalassemia patients during the year 2015, and to ascertain the extent of stationarity, and determine the rank of the integration of each variable separately. We use (unit root tests).

Despite the multiplicity of these tests, however, we have adopted two tests for this study, namely: the test Dickey Fuller expanded testing (Previously mentioned), and Phillips- Perron to test the null hypothesis theory of the existence of unit root of the instability of the time series, (Phillips-Perron).

For testing Philip-Perron's estimation is based on the same model's Dickey Fuller (DF) that it takes into account the variation is a homogenized error by correcting the non-parametric statistics Dickey Fuller process [73].

Table 4-1 shows the statistical results that obtained by the application of two tests at all levels, and includes the critical values for each test at the moral level of 5%. Through the results of previous tests, it turns out that the two strings are static, not containing a unitary root, as the calculated values are larger than the critical values (Mackinnon), We

note that the two strings oscillate around the middle of our constant, with no variation relationship in time. This means that there is a possibility for a joint integration between the blood and increasing doses of iron. To verify this, we will use the method of Engle - Granger co-integration.

Table 4-1 Results of the unit root stillness time series tests

Augmented Dickey Fuller (ADF)Test						
Variable	Constant			Trend		
	Level	Value	Conclusion	level	value	Conclusion
BLOOD	-3.521579	-5.311300	I(1)	-4.086877	-5.317518	I(1)
IRON	-3.524233	-4.340150	I(1)	-4.090602	-4.382865	I(1)
Phillip-perron (pp)Test						
Variable	Constant			Trend		
	Level	Value	Conclusion	level	value	Conclusion
BLOOD	-3.521579	-5.369641	I(1)	-4.086877	-5.381007	I(1)
IRON	-3.524233	-5.474117	I(1)	-4.086877	-5.491696	I(1)

4.3.3 The results of the co- integration tests:

Through the root of the previous unit test, it became clear that each variable on an integrated unit of the zero class. The focus of common integration theory on time-series analysis, where each of the Angel and Granger refers to the possibility of generating a linear combination characterized by stationary of time series, if possible, to generate this linear mix static, these static time series in this case is the integrated versions of the same rank. Thus, they can used as variables in the regression level and the gradient is, in this case, false and described the relationship equilibrium in the end. The formation of a linear combination of the study model is as follows:

$$\varepsilon_t = BLOOD_t - \alpha - \beta.IRON_t. \quad (4.4)$$

4.3.3.1 The results of the co-integration of analysis in a manner Engle- Granger:

The Co-integration that has been developed by Engle Granger's 1983 analysis Engle Granger, the year 1987 is when many economists recognized this as one of the most important new concepts in the field of econometrics, as well as for the analysis of time

series. This method requires two-step, the first estimate concerned the relationship in a way (least squares) where we get the regression equation of the joint integration, then get on the estimated regression residuals (ε), Mix Linear generated from the decline of long-term equilibrium relationship. The second test the stationarity residuum obtained from the first steps in accordance with the following:

$$\Delta\hat{\varepsilon} = \alpha + \delta\hat{\varepsilon}_{t-1} + \Delta\hat{\varepsilon}_{t-1} + \varepsilon_{t-1,} \tag{4.5}$$

$$e_t \sim IN(0).$$

If the statistical (τ) to (ε_{t-1}) is significant, we reject the null hypothesis ($\Delta\boldsymbol{\varepsilon_{t-1}} \sim I(1)$), The existence of the root of the units in the residuum and accept the alternative hypothesis static residuum or($\Delta\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{t}} \sim I(0)$) [38].

The application of the ordinary least squares method and a gradient between the number of doses of blood and iron we got the estimated relationship table (2) as shown by the following estimation:

$$\hat{BLOOD} = 5.611 + 0.002 * \hat{IRON}. \tag{4.6}$$

After obtaining the leftover regression, several statistical tools where used to test the stationarity residuum, in addition to estimating equation (4-6) to test the unit root, in order to confirm their findings. We research the possibility of a long-term equilibrium relationship; through the implementation of Co-integration between the studied variables would be so out of leftover appreciation ($\widehat{\boldsymbol{\varepsilon}_{\boldsymbol{t}}}$). We have to make sure that the latter is stable, and for this purpose we examined residuals estimated equation, as well as the autocorrelation of residuals transactions and, in the last test, we used Dickey Fuller expanded and Philip–Perron, in order to enhance the results obtained it. To examine the regression residuals co- integration for this purpose, we have equation Graph values leftover appreciation, see figure (4-2), which shows that the residuals series.

Regression equation Co-integration was stable. The "string is if stable fluctuated around the middle of our constant; with the variation in time has nothing to do [74].

To be sure, we examined transactions where the residuum series is stable, if the function linked transactions (pk) not different from zero for (k > 0) and Table (4-2) shows the self and the partial series residuum link function and can be seen from this table that the residuum string that represents the process of jamming. White as a function autocorrelation series residuum was found that all the affiliated transactions of gaps (k) generally not different from zero and within the confidence interval [75].

Table 4-2 Estimate the relationship between blood and iron in a manner least squares

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.611229	0.449608	12.48027	0.0000
IRON	0.002038	0.000134	15.19910	0.0000
R-squared	0.759878	Mean dependent var	4.000000	
Adjusted R-squared	0.756589	S.D. dependent var	2.398718	
S.E. of regression	2.603048	Akaike info criterion	3.476488	
Sum squared reside	494.6376	Schwarz criterion	3.520384	
Log likelihood	-177.1580	Hannan-Quinn Criter.	3.494325	
F-statistic	231.0127	Durbin-Watson Stat	2.408372	
Prob(F-statistic)	0.000000			

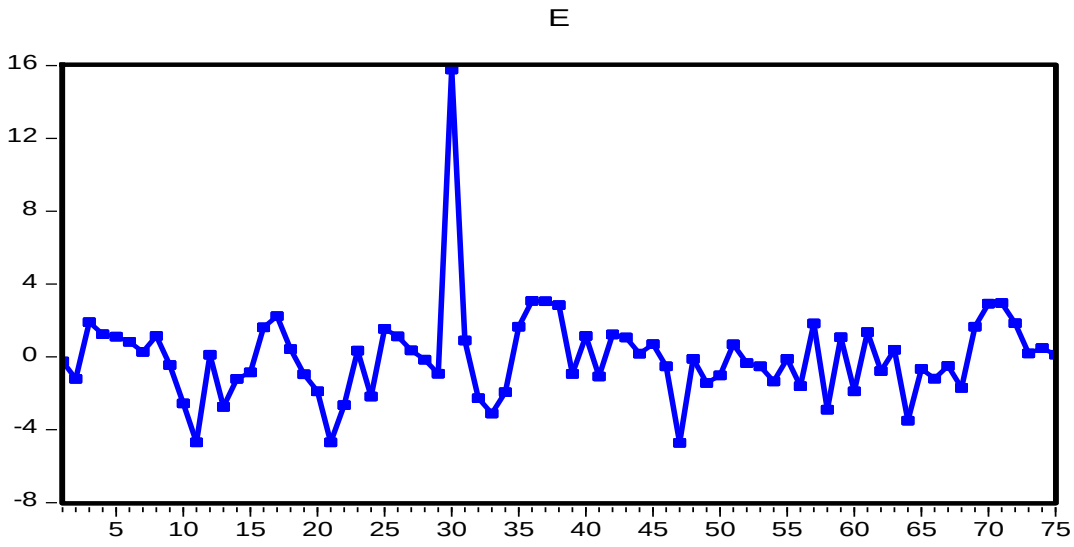


Figure 4-2 Leftover downhill Co- integration equation

Table 4-3 Autocorrelation function and partial residuals estimate the slope of the Co-integration equation

	AC	PAC	Q-Stat	Prob		AC	PAC	Q-Stat	Prob
1	0.152	0.152	1.7985	0.180	17	-0.271	-0.199	21.120	0.221
2	0.052	0.029	2.0086	0.366	18	-0.036	0.006	21.248	0.267
3	-0.149	-0.165	3.7906	0.285	19	-0.134	-0.170	23.090	0.233
4	-0.085	-0.042	4.3762	0.357	20	-0.085	-0.167	23.857	0.249
5	-0.015	0.021	4.3948	0.494	21	0.024	0.041	23.919	0.297
6	-0.006	-0.025	4.3982	0.623	22	-0.012	-0.036	23.934	0.351
7	0.033	0.018	4.4895	0.722	23	-0.030	-0.107	24.034	0.402
8	0.053	0.048	4.7355	0.785	24	-0.105	-0.097	25.278	0.391
9	-0.138	-0.169	6.4049	0.699	25	-0.005	-0.033	25.281	0.447
10	0.009	0.058	6.4126	0.779	26	-0.005	-0.113	25.284	0.503
11	0.009	0.041	6.4199	0.844	27	0.078	0.111	26.023	0.517
12	0.064	0.011	6.7986	0.871	28	-0.136	-0.224	28.282	0.450
13	0.152	0.141	8.9483	0.777	29	-0.014	0.020	28.308	0.501
14	0.029	-0.006	9.0256	0.829	30	-0.079	0.036	29.112	0.512
15	-0.055	-0.084	9.3172	0.860	31	0.052	-0.016	29.462	0.545
16	-0.214	-0.156	13.808	0.613	32	0.049	0.018	29.780	0.579

4.3.3.2 Dickey Fuller test results expanded Philip -Perron:

To confirm previous results, we conducted Dickey Fuller expanded testing (ADF) and Philip -Perron (PP), the leftover tests shown in Table (4-4), which shows the results of testing the stability leftover downhill Co-integration equation.

Table 4-4 the results of unit root for residuals export tests

Model type	Model(1) Without constant or trend		Model (2) With constant		Model (3) With constant and trend	
Type of test	ADF	PP	ADF	PP	ADF	PP
Calculated value	-7.332237	-7.350155	-7.281841	-7.300486	-7.235920	-7.255285

The critical value	-2.596586	-2.596586	-3.521579	-3.521579	-4.086877	-4.086877
Probability embarrassment	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

4.3.4 Results of the study of the causal relationship between blood and iron:

Granger demonstrated that the existence of a Co- integration between two variables means that there is a causal relationship in at least one direction. Therefore, we conclude that the lack of a common integration between two variables means there is no causal relationship between them. According to Granger, if we had two time series chains, we can talk about the evolution of two different phenomena over time (t) and two in this study (blood and iron) If the series (blood) containing the information by which they can improve the outlook forecasting series (iron).

In this case, we say that the variable (blood) causes the variable (iron). One of the problems that exist in this case is that the time series data for a variable are often associated, i.e., there is an auto-correlation between one variable's values over time and excluding the impact of this auto-correlation. If any, the inclusion of the same variables values for a number of time gaps as explanatory variables in the causal relationship to measure. That requires a causality test Granger estimate vector regression model self (VAR), which describes the behavior of the two variables (blood) and (iron):

$$BLOOD_t = \alpha_O + \sum_{i=1}^{p} \beta_i \ . BLOOD_{t-i} + \sum_{i=1}^{p} \ . IRON_{t-i} + \mu_t \ , \qquad (4.7)$$

where (μ_t) represents the model residuals. However, before determining the causal relationship between the two variables, we must specify the number of time gaps (p) and the accurate form VAR (p). If it is a smaller number of (p) then this leads to an error in the description, and if the number is greater than (p), this leads to a lack of full exploitation of information time series. It also reduces the degrees of freedom are usually determine the number of time gaps based on a standard (AIC) (SC) table (4-5) the steps Granger causality test:

Table 4-5 **Valuable Akaike and Schwarz**

Slow	**1**	**2**	**3**	**4**	**5**	**6**
AIC	**11,51**	**11,61**	**11,83**	**11,89**	**11,60**	**11,52**
SC	**11,76**	**12,04**	**12,43**	**12,66**	**12,55**	**12,67**

4.3.4.1 Estimate restrictive formula

$$\mathrm{BLOOD}_t = \boldsymbol{\alpha}_O + \sum_{i=1}^{p} \boldsymbol{\beta}_i \,.\, \mathrm{IRON}_{t-i} + \boldsymbol{\varepsilon}_t. \qquad (4.8)$$

We assume to be $\sum_{i=1}^{p} \beta_i = 0$ in equation. Meaning that the variable (blood) does not affect the variable (iron) and then we get the total estimated residuum squares recovered from the restricted equation (4.8)

-Estimation formula is restricted:

That the equation (4.8), and then we can get the total estimated residuum squares recovered from the formula unrestricted equation.

-Testing the imposition of the following null-hypothesis

$H_0 \sum_{i=1}^{p} \beta_i = 0$, for that, we must calculate statistical Fisher FC:

$$\mathrm{Fc}= \frac{\mathrm{RSS}_1 - \mathrm{RSS}_2 \mathrm{lm}}{\mathrm{RSS}_2/(n-k)} \quad \text{and} \quad \mathrm{F}\sim(m, n-k), \qquad (4.9)$$

where, m is number of lags; k is number of parameters involved in the model; and n is the sample size. The test is to reject the null hypothesis of non-causality between blood and iron, it uses test (f) to decide the existence of a causal relationship or not between the variables in the following form:

If (f) is greater than the calculated tabular, it will reject the null hypothesis testing the hypothesis H_0: $\sum_{i=1}^{p} \beta_i = 0$,

$$\mathrm{F_c} = \frac{\mathrm{RSS}_1 - \mathrm{RSS}_2/m}{(\mathrm{RSS}_2)/(n-k)} = 4.4 \text{ is greater than } \mathrm{F_c}= 3.96.$$

Hence, reject the null hypothesis and accept the alternative hypothesis that there is a long-term causal relationship between the number of doses of blood given and the high

proportion of iron deposited in the organs of the body.

4.4 Relationships of some factors on the Thalassemia:

Our research included several possible factors that may have effects on thalassemia, among these factors:

4.4.1 *The type of thalassemia*:

The two main types of thalassemia are alpha thalassemia and beta thalassemia. (The alpha and beta refer to which hemoglobin gene is affected, and which of the hemoglobin chains is faulty.) There are some rare types too. Each type of thalassemia (alpha and beta) then classified into subtypes, according to how severe the condition is. This mainly depends on how many thalassemia genes are involved. The mildest types called thalassemia trait (or thalassemia minor). The more severe beta types are beta thalassemia major (BTM) and beta thalassemia intermedia (BTI). The more severe alpha forms are Hb Barts (very severe) and HbH disease (moderate). There are also some rare types of thalassemia such as delta beta thalassemia, or combinations of a beta-thalassemia gene with another abnormal hemoglobin gene, such as HbE[76]. Among the data analyzed in our research, results showed that thalassemia beta major was the most common type in the patients table see Table (4-6)

Table 4-6 the relationship between age, gender, type of thalassemia

Age	1-4		5-9		10-14		15-24		25-34		35-44		Sum	
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F
Great	13	7	9	5	10	11	6	5	1	1	-	-	39	29
middle	9	3	1	3	1	1	1	2	-	-	-	-	12	9
minor	2	1	1	2	1	-	1	1	1	1	-	-	6	5
Sum	24	11	11	10	12	12	8	8	2	2	-	-	57	43
percentage	35%		21%		24%		16%		4%		-			

4.4.2 Blood type:

Patients included in this study were of different blood types and Rh factor (A+, A- ,

B+, B- , AB+, AB-, O+, O-) but our results demonstrated that the blood type O+ was the most common type affected see Table (4-7).

Table 4-7 the relationship between age, gender and type of blood

Age	1-4		5-9		10-14		15-24		25-34		35-44		Sum	
Gender	M	F	M	F	M	F	M	F	M	F	M	F	M	F
A+	3	1	1	1	1	1	1	1	-	1	-	-	6	5
A-	1	1	1	1	1	1	-	1	-	-	-	-	3	4
B+	4	2	2	1	1	1	1	2	1	1	-	-	9	7
B-	1	-	1	1	1	-	1	1	-	-	-	-	4	2
O+	9	4	3	3	5	5	1	1	1	-	-	-	19	13
O-	4	1	1	-	2	1	2	1	-	-	-	-	9	3
AB+	1	1	2	2	1	2	1	1	-	-	-	-	5	6
AB-	1	1	-	1	-	1	1	-	-	-	-	-	2	3
Sum	24	11	11	10	12	12	8	8	2	2	-	-	57	43

4.4.3 The age and gender:

Ages varied among people infected with Thalassemia and ranged between (1- 44) and they were subdivided in to categories (1-4), (5-9), (10-14), (15-24), (25-34) and (35-44) year's. Results showed that the most infected age groups is the category (1-4). Males more affected than females. We also noticed that data were absent for patients from category (35-44), which urges us and other researchers to do a more detailed study on that age groups of patients (see Table 4-7).

Recommendations:

(1) Do more research on this subject focusing on finding other relationships, such as the relationship between age and an increased demand for blood transfusions?

(2) Try to educate patients with Thalassemia to use specific medicines that relive the symptoms of iron overload and make these medicines available and cheap as much as possible.

(3) Increase awareness of patients and their families to avoid foods that contain high level of iron such as beef, green leaves and beans.

(4) Regular analysis for iron levels will greatly improve the control of iron overload cases.

(5) Patients that need blood transfusion must regulate their intake times as much as possible.

(6) As we noticed from our research, children between 1-4 years were the most affected, and children at that age, as we know, are in a high rate of growth, therefore, multivitamins and growth factors given regularly under medical supervision to reduce the effect of thalassemia on their growth as much as possible.

5 Forecasting mortality patterns by using VAR model and reasons for this mortality

5.1 Introduction

Iron overload is the primary cause of mortality and morbidity in thalassemia major patients, despite advances in chelation therapy, accumulating levels of iron in the body generated from hundreds of blood transfusions, together with increased absorption of iron from the diet, cause critical organs and glands in patients with thalassemia to experience serious iron-induced toxicity and early death [77].

Several publications provide evidence that the heart is unquestionably the most critical organ affected by the iron, jeopardizing the survival of thalassemia patients; 36% of patients between the ages of 15 and 18 showed detectable cardiac iron toxicity [78]. The life of patients with thalassemia has improved both duration and quality in industrialized countries. but complications are still common and include heart disease, growth retardation, pallor, jaundice, poor musculature, hepatosplenomegaly, leg ulcers, and the development of masses from extra medullary hematopoiesis, and skeletal changes (due to bone marrow expansion) are found in untreated or poorly transfused individuals with thalassemia major [79].

This study aimed to develop a standard model for forecasting the number of deaths in patients with thalassemia from Maysan province of southern Iraq in the coming years (2016-2020). The study concluded that the vector auto regression (VAR) model is one of the most successful, flexible, and easy to use models for the analysis of multivariate time series. It is a natural extension of the univariate autoregressive model for dynamic multivariate time series.

The VAR model has proven to be especially useful for describing the dynamic behavior of economic and financial time series and for forecasting. It often provides

superior forecasts to those from univariate models of time series and elaborate theory-based simultaneous equation models. Forecasts of VAR models are quite flexible because they made conditional on the potential future paths of specified variables in the model. Vector autoregressive (VAR) models have a long tradition as tools for multiple time series analysis[80].

Being linear models, they are relatively easy to work in both theory and practice. Although the related computations are relatively straightforward, they are sufficiently involved to make applied work cumbersome before powerful computers were in widespread use. VAR models became popular for economic analysis when Sims (1980) advocated them as alternatives to simultaneous equation models.

The later models have used extensively since the 1950s. The availability of longer and more frequent observed time series emphasized the need for models that focused on the dynamic structure of the variables, however. Sims also criticized the erogeneity assumptions for some of the variables in simultaneous equation models as ad hoc and often not backed by fully developed theories. In contrast, in VAR models often all observed variables treated as a priori endogenous, statistical procedures rather than subject matter theory used for imposing restrictions. Our study also included a precise scanning of our data for most complications caused by thalassemia by selecting the most causative one for mortality.

5.2 Methodology: VAR model :(Vector Autoregressive)

Proposed as a model by Sims in 1981 [81], vector auto regression (VAR) is an econometric model used to capture the linear interdependencies among multiple time series. VAR models generalize the univariate autoregressive model (AR model) by allowing for more than one evolving variable.

All variables in a VAR enter the model in the same way: each variable has an

equation explaining its evolution based on its own lags and the lags of the other model variables. VAR modelling does not require as much knowledge about the forces influencing a variable, as do structural models with simultaneous equations: the only prior knowledge required is a list of variables that hypothesize to affect each other. General models write as:

$$\phi(B)Y_t = \varepsilon_t, \tag{5.1}$$

where:

Y_t : a random context dimensioned n stable,

$\phi(B)$: polynomial matrix, grade P, slow time factor B

$$\phi(B) = \phi_0 - B_{\phi_1} - B^2\phi_2 - \cdots - B^p\phi_p, \tag{5.2}$$

$\emptyset_0$: matrix monocrystalline grade n,

ε_t : white noise.

VAR model can also write as follows: [Shumway RH and Stoffer, 2006]

$$Y_{1t} = \phi_{11}^{(1)} y_{1,t-1} + \cdots + \phi_{11}^{(p)} y_{1,t-p} + \cdots + \phi_{1n}^{(1)} y_{n,t-1} + \cdots + \phi_{1n}^{(p)} y_{n,t-p} + \varepsilon_{1,t}, \tag{5.3}$$

$$Y_{nt} = \phi_{n1}^{(1)} y_{1,t-1} + \cdots + \phi_{n1}^{(p)} y_{1,t-p} + \cdots + \phi_{nn}^{(1)} y_{n,t-1} + \cdots + \phi_{nn}^{(p)} y_{n,t-p} + \varepsilon_{n,t}. \tag{5.4}$$

5.3 Constructing the model (VAR):

1. Using a stationary time series, this does not contain the root of unity.
2. Specify the number of periods.
3. Study the causal relationship between the variables. Equations [82].

5.3.1 Stationary: this study, we used the test Dickey Fuller expanded

5.3.2 Determine the periods a slowdown:

When Sims gave, his model did not give any limitation with respect to the length of the period of a slowdown that applied. There are several quantitative criteria used to determine the periods in a slowdown:

5.3.2.1 Final Predictor Error Criterion (FPE) [83]:

Akaike's Final Prediction Error (FPE) criterion provides a measure of model quality

by simulating the situation where the model tested on a different data set. After computing several different models, you can compare them using this criterion. According to Akaike's theory, the most accurate model has the smallest FPE. If you use the same data set for both model estimation and validation, the fit always improves as you increase the model order and, therefore, the flexibility of the model structure. Akaike's Final Prediction Error (FPE) defined by the following equation:

$$FPE = \det\left(\frac{1}{n}\sum_{1}^{N} e(t,\theta_N)\right)(e(t,\theta_N))^T)\left(\frac{1+{}^{d}/_{N}}{1-{}^{d}/_{N}}\right), \tag{5.5}$$

where

N is the number of values in the estimation data set,

e(t) is a ny-by -1 vector of prediction errors,

θ_N Represents the estimated parameters.

5.3.2.2 AKAIKE Information Criterion (AIC)[83]

This computes the following relationship:

$$AIC(P) = Log\left(det\Omega(p) + 2\frac{n^2 p}{N}\right), \tag{5.6}$$

where

Ω: Matrix variations and changes of the estimated remainders,

n: The number of internal variables,

N: Total views.

After that, we choose p_0 that achieve the equation

$$AIC(P_0) = \underset{p=1}{\overset{k}{Min}}\ AIC(P). \tag{5.7}$$

5.3.2.3 Bayesian Information Criterion (BIC)[84]:

Computes the following relationship:

$$BIC(P) = Log\left(de + \Omega(p)\right) + \left(\frac{n^2.p.logN}{N}\right). \tag{5.8}$$

Of the calculate the number of extended slowdown

$$BIC(P_0) = \underset{p=1}{\overset{k}{Min}}\ BIC(P). \tag{5.9}$$

5.3.2.4 Hannan & Quinn Information Criterion (HQIC)[85]

It calculated from the following relationship:

$$HQIC(P) = Log\left(det\Omega(p)\right) + \left(2n^2 Pc\,\frac{log.logN}{N}\right), \tag{5.10}$$

where C: strength indicator benchmark and we consider it equal to 2 in the practical application.

Of the calculate the number of extended slowdown

$$HQIC(P_0) = \underset{p=1}{\overset{k}{Min}}\ HQIC(P). \tag{5.11}$$

Note: we can apply the different outcomes of these criteria in practice in this case we select the time slowdown for which we got the largest number of criteria.

5.3.3 Causality test[86]:

Estimates the following equation by using the least squares method:

$$Y_t = \phi_1(B).Y_t + \phi_2(B).X_t + \varepsilon_t, \tag{5.12}$$

where

$$\phi_1(B) = \sum_{i=1}^{P} \emptyset_{1i}.B^i \text{ and } \phi_2(B) = \sum_{i=1}^{q} \emptyset_{2i}.B^i.$$

We calculate the total deviations of actual values about estimated symbolized her SCR1. Estimate the following equation

$$Y_t = \phi_1(B).Y_t + \varepsilon_t. \tag{5.13}$$

We calculate the total deviations of actual values about estimated symbolized her SCR2. Calculate statistically the test Fc of the relationship

$$F_C = \frac{(SCR2-SCR1)/P}{SCR1/(M-N)}, \tag{5.14}$$

where M=T-Max (p,q) and N=p+q+2, T: number of views, P: the number of time slowdowns of internal variables, q: the number of time slowdowns of external variables.

5.4 Data analysis:

We obtained time series stretching from 2005 to 2015, showing the number of deaths in patients with thalassemia.

5.4.1 Showing time series:

Figure (5-1) shows clearly the instability of the series, but does not indicate whether the instability is due to the presence of unit root or not? Therefore, we must test the unit root.

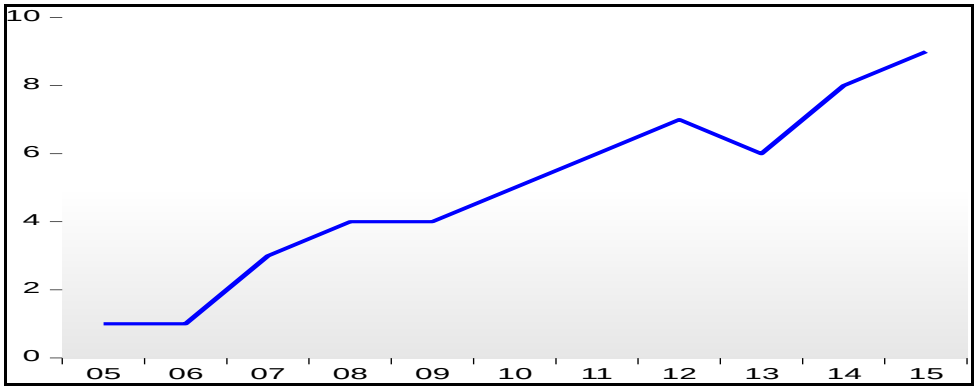


Figure 5-1Time series of mortality for patients with thalassemia

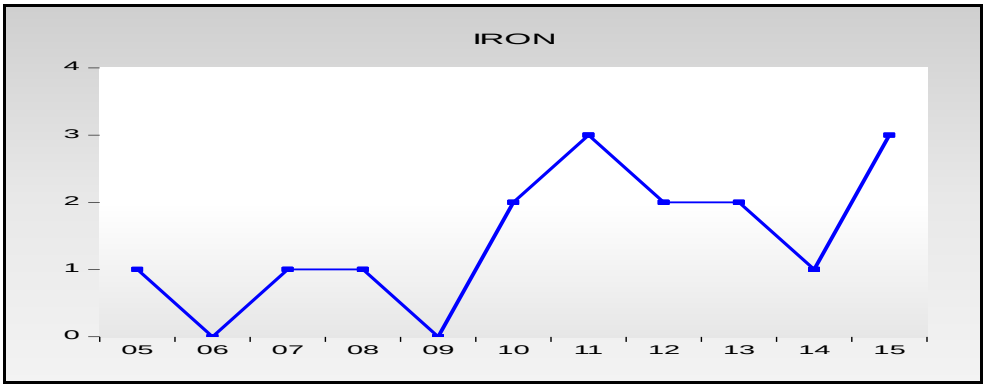


Figure 5-2 Time series for the preparation of cases of death due to high iron for the years 2005-2015

5.4.2 Stationarity test[87]:

5.4.2.1 Augmented Dickey & Fuller (ADF)

Because the instability in time series is often due to the existence of unit root, the Dickey–Fuller test has been proposed to detect whether unit root is present or not (Previously mentioned)

Table 5-1 Expanded Dickey Fuller test

Augmented Dickey –Fuller test statistic		t-statistic	Prob.
		-0.438295	0.8662
	1% level	-4.297073	
Test critical values	5% level	-3.212696	
	10% level	-2.747676	

H_0:φ = 0 Acceptance of this hypothesis implies acceptance that there is unit root instability

H_0: φ < 1 Accepting this implies acceptance of the hypothesis of stability and the lack of unit root [53]. We have the test result shows us that the time series is not stable and contains the root of the unit we convert it to the first application of a stable candidate variances series Δ= (1 − B) and After which test the resulting string[55]

Table 5-2 Expanded Dickey Fuller test after the first time series difference

Augmented Dickey –Fuller test statistic		t-statistic	Prob.
		-3.785640	0.0273
	1% level	-4.582648	
Test critical values	5% level	-3.320969	
	10% level	-2.801384	

To get rid of the unit root we apply again the first candidate difference to the first series of difference and we can obtain a stable series as shown in Table (5-3).

Table 5-3 Expanded Dickey Fuller test after the second difference of the time series

Augmented Dickey –Fuller test statistic		t-statistic	Prob.
		-3.845741	0.0024
	1% level	-2.937216	
Test critical values	5% level	-2.006292	
	10% level	-1.598068	

Table 5-4 Dickey-Fuller test enlarged Iron series

Augmented Dickey –Fuller test statistic		t-statistic	Prob.
		-1.784308	0.3657
	1% level	-4.297073	
Test critical values	5% level	-3.212696	
	10% level	-2.747676	

We find from Table (5-4) that the absolute value of (DAF) is calculated (1.784308) smaller than the absolute values at different levels of significance. If we accept the null hypothesis, we accept the existence of unit root in the time series. Applying the unit root test again to the first series of difference, we find that the results confirm the series in Table (5-5):

Table 5-5 Dickey-Fuller test enlarged Iron series, after taking the first differences of the series

Augmented Dickey –Fuller test statistic		t-statistic	Prob.
		-3.632122	0.0296
	1% level	-4.420595	
Test critical values	5% level	-3.259808	
	10% level	-2.771129	

To get rid of the unit root we apply again the first candidate difference to the first series of difference and we can obtain a stable series as shown in Table (5-6).

Table 5-6 Dickey-Fuller test enlarged Iron series, after taking the second differences of the series

Augmented Dickey –Fuller test statistic		t-statistic	Prob.
		-3.132381	0.0072
	1% level	-2.937216	
Test critical values	5% level	-2.006292	
	10% level	-1.598068	

5.4.3 Specify the amount of extended slowdown time

To specify the amount of extended slowdown time we use the criteria that we have advanced in advance and that appear in the following Table (5-7).

Table 5-7Criteria determine the number of periods of slowdown of time

VAR Lag Order Selection Criteria

Endogenous variables: DTHALASSEMIA DIRON

Exogenous variables: C

Sample: 2005 2015

Lag	LogL	LR	FPE	AIC	SC	HQ
0	-24.44796	NA*	2.557079	6.611991	6.631851	6.478041
1	-18.62741	7.275690	1.747198	6.156853	6.216434	5.755002
2	-10.99232	5.726322	1.005012*	5.248079*	5.347381*	4.578327*

* indicates lag order selected by the criterion

LR: sequential modified LR test statistic (each test at 5% level)

FPE: Final prediction error

AIC: Akaike information criterion

SC: Schwarz information criterion

HQ: Hannan-Quinn information criterion

We find from Table 5-6 that the three criteria (FPE), (AIC), (SC) and (HQ) refers to taking 2 lag.

5.4.4 Causality test

Table (5-8) shows the iron variable causes deaths for patients with thalassemia with 2 time lags at a level of significance of 5%.

Table 5-8 Granger causality test

Pairwise Granger causality tests

Sample:2005-2015

Lags :2

Null Hypothesis	Obs	F –statistic	Prob
D iron does not Granger cause D thalassemia	8	4.31177	0.1311
D thalassemia does not Granger cause D iron		0.23802	0.8018

Table 5-8 shows that the variable iron causes a thalassemia mortality variable at the level of 5%.

5.4.5 Estimates VAR model:

Given the causality test results in Table (5-8), and values of slowdown extended standards in Table (5-7), in order to reconcile these we choose 2 lags when estimating the model (VAR), see Table (5-9).

Table 5-9 the transactions estimates of the model

Vector Auto regression Estimates

Sample (adjusted): 2008 2015

Included observations: 8 after adjustments

	D thalassemia	D iron
	-0.395164	0.277190
DTHALASSEMIA(-1)	(0.33018)	(0.76185)
	[-1.19681]	[0.36384]
	0.926277	0.049635
D thalassemia(-2)	(0.24567)	(0.56685)
	[3.77041	[0.08756]
	0.784854	0.018613
D iron (-1)	(0.33239)	(0.76695)
	[2.36124]	[0.02427]
	-0.421989	-0.415146
Diron (-2)	(0.27399)	(0.63219)
	[-1.54017]	[-0.65668]
	3.012226	0.553650
C	(0.58070)	(1.33988)
	[5.18728]	[0.41321]
R –squared	0.958945	0.275973
Adj .R-squared	0.904205	-0.689396
Sum sq.resids	0.815967	4.344161
S .E.equation	0.521526	1203351
F –statistic	17.51820	0.285873

Log likelihood	-2.220217	-8.909072
Akaike AIC	1.805054	3.477268
Schwarz SC	1.854705	3.526919
Mean dependent	5.375000	1.500000
S.D. dependent	1.685018	0.925820
Determinant resid covariance (dof adj)		0.380596
Determinant resid covariance		0.053521
Log likelihood		-10.99232
Akaike information criterion		5.248079
Schwarz criterion		5.347381

VAR (2) model estimate

DTHALASSEMIA = - 0.395164233577*DTHALASSEMIA (-1) + 0.926277372263 *DTHALASSEMIA (-2) + 0.784854014599*DIRON (-1) - 0.421989051095*DIRON (-2) + 3.01222627737

DIRON= 0.277189781022*DTHALASSEMIA (-1) + 0.0496350364964*DTHALASSEMIA (-2) + 0.0186131386861*DIRON (-1) - 0.415145985401*DIRON (-2) + 0.553649635036

Written by Layout the matrix:

$$\begin{bmatrix} \mathrm{D(THALASSEMIA)_t} \\ \mathrm{D(IRON)_t} \end{bmatrix} = \begin{bmatrix} -0.3951 & 0.7848 \\ 0.2771 & 0.0186 \end{bmatrix} \cdot \begin{bmatrix} \mathrm{D(THALASSEMIA)_{t-1}} \\ \mathrm{D(IRON)_{t-1}} \end{bmatrix} + \begin{bmatrix} 0.9262 & -0.4219 \\ 0.0496 & -0.4151 \end{bmatrix} \cdot \begin{bmatrix} \mathrm{D(THALASSEMIA)_{t-2}} \\ \mathrm{D(IRON)_{t-2}} \end{bmatrix} .$$

5.4.6 The residual tests:

In order to validate the estimated model we must make sure that it undergoes the residuum normal distribution and that it link to itself.

1. **The probability distribution of the residual**: use the test Jarque –Bera[88]

Table 5-10 Normal distribution test of residuals

VAR Residual Normality Tests

Orthogonalization : cholesky(lutkepohl)

Null hypothesis: residuals are multivariate normal

Sample: 2005-2015 , Included observations:8

Component	Jarqur- Bera	Df	Prob.
1	0.431051	2	0.8061
2	0.834086	2	0.6590
Joint	1.265137	4	0.8673

The test indicates that the null hypothesis should not be rejected for each of the residuals for the first and second equation and that. Not to reject hypothesis of normal distribution of residuals at the 5% significance level should not be rejected.

2. Autocorrelation of residuals test: use the Ljung-Box test[89]

Table 5-11 Autocorrelation of residuals

VAR Residual portmanteau Test for Autocorrelations

Null Hypothesis: no residual autocorrelations up to lag h

Sample: 2005-2015 Included observations :8

Lag	Q-stat	Prob	Adj Q-stat	Prob	Df
1	6.237209	NA*	7.128239	NA*	NA*
2	11.25064	NA*	13.81281	NA*	NA*

3	14.76834	0.0052	19.44113	0.0006	4
4	16.24727	0.0390	22.39899	0.0042	8
5	17.89003	0.1191	26.77968	0.0083	12
6	18.78529	0.2800	30.36072	0.0162	16

The autocorrelation test indicates that the null hypothesis should be rejected, not which means a lack of self-link when the level of significance is 5%.

5.5 Forecasting:

Table5-12: Forecasting thalassemia mortality and iron, for the years 2016-2020

Year	[1]L Mortality thalassemia	[2]U mortality thalassemia	L iron	U iron
2016	11.84799	12.79583	5.157853	5.261010
2017	14.84125	16.02855	7.502795	7.652851
2018	16.02047	17.30211	8.758657	8.933830
2019	16.87972	18.23010	9.367777	9.555132
2020	17.85826	19.28692	10.37876	10.58634

(1) Minimum of forecasting

(2) Upper limit of forecasting

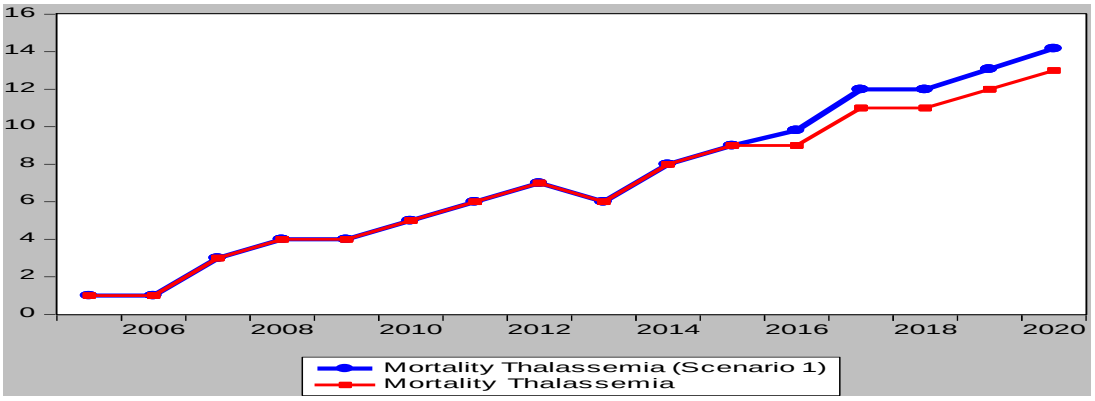


Figure 5-3 Graph of the values of forecasting of thalassemia mortality

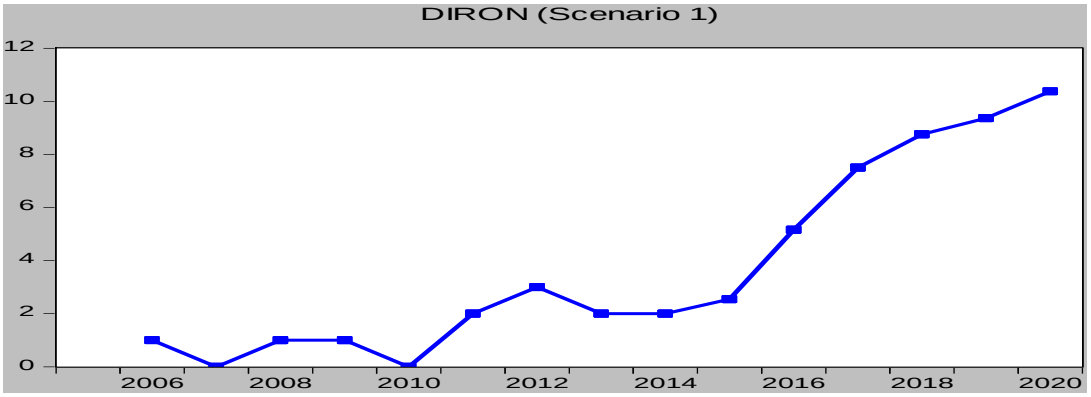


Figure 5-4Graph of the values of forecasting of iron

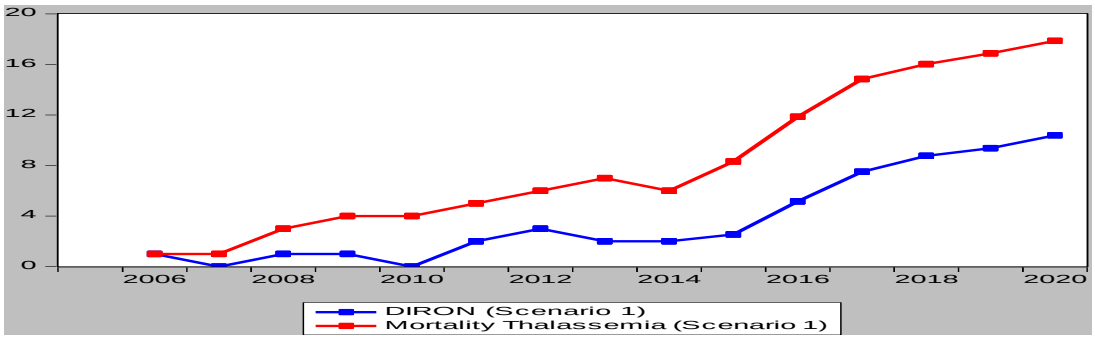


Figure 5-5 Graph of the values of forecasting of thalassemia mortality and iron

Result:

Through a graphs and tables, we can forecast the numbers of mortalities for the coming years where there is a very clear increase in the number of mortalities as well as a very clear correlation between the deaths and the increase of iron in the blood. The highest recorded cases of death are for 2020 (19.28692) and it is clear that the number of deaths will increase in the coming years, and whenever iron is increased the proportion of patient will increase.

5.6 Cause of mortality for patients with thalassemia:

The accumulation of iron levels in the body resulting from hundreds of blood transfusions,

Along with the increase in the absorption of iron from food, cause vital organs and glands in patients with thalassemia to experience induced early mortality. In this study,

examined which diseases represent the greatest that cause of mortality. We took the cases of death of thalassemia patients in the province of Maysan for the period 2005-2015 and the number of mortalities during this period, which was 73 as well as information from the thalassemia center in the province and based on the data collection we found that the primary causes of these deaths were

- Cardiac disease
- Infections
- The liver
- The spleen

Table 5-13 the mortality cause for thalassemia patients for the years 2005-2015

The mortality cause Years	Cardiac disease	Infections	The liver	The spleen
2005	2	1	1	1
2006	3	1	2	0
2007	2	2	1	0
2008	1	1	0	1
2009	5	1	0	1
2010	4	2	2	0
2011	5	3	1	2
2012	3	1	1	0
2013	3	0	3	1
2014	4	1	2	1
2015	3	1	4	0
Sum	35	14	17	7
Percentage	48%	19%	23%	10%

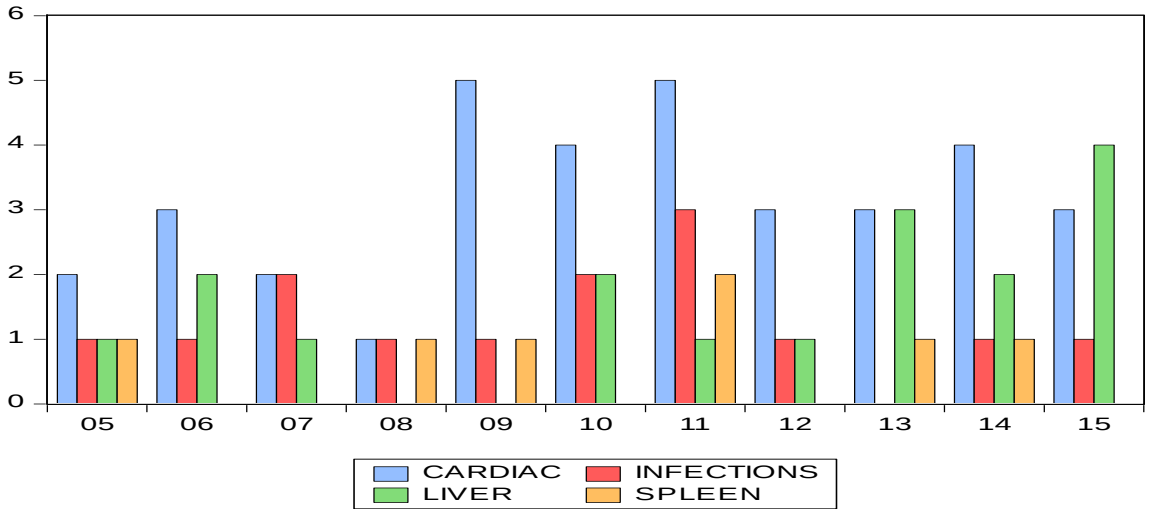


Figure 5-6 the graph for the diseases causing the death, distributed according to the years 2005-2015

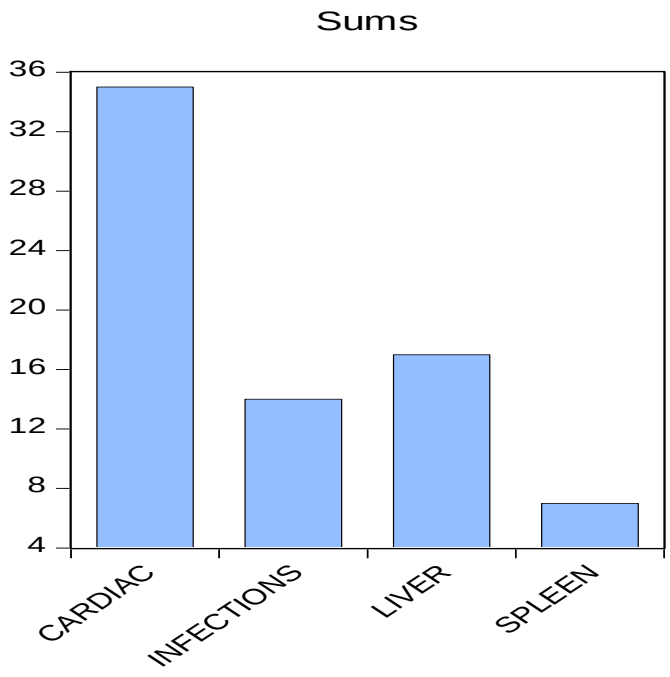


Figure 5-7 the percentage of diseases that cause the death

Result:

Through a table and graphs we found that the main cause of death was heart disease There were 35 deaths due to heart disease (48%), followed by disease of the liver (17 cases, 23%), infections(14 cases, 19%), and disease of the spleen (7 cases, 10%)

Conclusions:

The VAR (Vector autoregressive) model has the ability to detect a causal relationship between thalassemia and iron in the long term, high level of iron in the blood causes several diseases because of deposition of iron in the body's organs, which the leading cause of death among people with thalassemia is related cardiac issues.

Recommendations:

1. The model could use in the prediction of other diseases, and the adoption of forecasts given by the study put of many of the problems solutions of thalassemia patients to reduce the burden of this disease.
2. Thalassemia is a reality in our country. It must accept as a public health problem in the long term and awareness campaigns should be initiated that make people more familiar with the seriousness of the disease in an attempt to reduce its prevalence by conducting tests before marriage.
3. Cardiac-related complications are the leading cause of death among people with thalassemia, and it is therefore very important for patients who do not have cardiac issues to take steps to prevent these complications from developing, and for patients who already do have cardiac issues to work to reverse them or prevent them from getting worse.

6 SUMMARY AND CONCLUSIONS

6.1 Introduction

This is section gives the summary of the study the results in consonance with the objectives set out in the introduction to the study the methods reviewed finally conclusions commensurate with the findings were drawn

6.2 Summary and Conclusions

(1) ARIMA model was suitable for application to Thalassemia data and analysis of other similar medical data.

(2) By application of ARIMA model，we were able to forecast future cases easily and accurately.

(3) Cases of Thalassemia will increase within coming years, which means that, currently, no serious efforts offered to solve or treat this disease in Iraq.

(4) There is a relationship between the number of blood that takes thalassemia patients and the high proportion of iron in the blood doses.

(5) Showed this study that thalassemia beta major was the most common type in the patient’s thalassemia.

(6) Our results demonstrated that the blood type O+ was the most common type affected.

(7) Ages varied among people infected with Thalassemia and ranged between (1- 44) and they were subdivided in to categories (1-4), (5-9), (10-14), (15-24), (25-34) and (35-44) year’s. Results showed that the most infected age groups is the category (1-4). Males more affected than females. We also noticed that data were absent for patients from category (35-44).

(8) The VAR (vector autoregressive) model has the ability to detect a causal relationship between thalassemia and iron in the long term.

(9) VAR models are flexible and are relatively easy to work with both in theory and practice.

(10) In general，thalassemia patients live with more risks than people without thalassemia do.

(11) Despite the use of iron chelating therapy in recent years in Iraq, serious complications can result in deaths who suffer from high iron levels.

(12) High levels of iron in the blood cause several diseases because of deposition of iron in the body's organs.

(13) The leading cause of death among people with thalassemia is related cardiac issues.

6.3 Answers to Key Questions

We now reconsideration our original key questions from Section 1 to see if we succeeded in answer all of them.

Table 6-1 answer to key questions

No	Question	Answer
Q 1	Is there an increase in the number infected with Thalassemia in the coming years 2016-2018?	The increasing number of cases of thalassemia diseases in the coming years.
Q2	Is there a relationship between the blood doses which given for thalassemia patients and increase the proportion of iron?	There is a very strong correlation between the number of doses that given to patients with thalassemia and high iron and the greater the number of potions whenever we noticed that there is an increase in the accumulation of iron.

Q 3	What are the physiological factors that affect patients with thalassemia?	There are several factors that influence the thalassemia 1. The type of thalassemia 2. Blood type 3. the age and gender
Q 4	Is there an increase in the number of deaths in patients with thalassemia in the coming years 2016-2020?	There is a large increase in the number of deaths for the coming years due to complications Thalassemia.
Q 5	Which diseases caused by Thalassemia?	There are several diseases, including 1. Cardiac disease 2. Infections 3. The liver swell 4. The spleen swell
Q 6	What are the most diseases that cause mortality for patients with thalassemia?	Cardiac disease, The proportion was 48%.

References

[1] Enders, W., *Applied Econometric Time Series, by Walter.* Technometrics, 2004. 46(2): p. 264.

[2] Lütkepohl, H., *New introduction to multiple time series analysis*. 2005: Springer Science & Business Media.

[3] Tsay, R.S., *Analysis of financial time series*. Vol. 543. 2005: John Wiley & Sons.

[4] Anderson, T.W., *The statistical analysis of time series*. Vol. 19. 2011: John Wiley & Sons.

[5] Grimmett, G. and D. Stirzaker, *Probability and random processes*. 2001: Oxford university press.

[6] Salas, J.D., *Applied modeling of hydrologic time series*. 1980: Water Resources Publication.

[7] Hipel, K.W. and A.I. McLeod, *Time series modelling of water resources and environmental systems*. Vol. 45. 1994: Elsevier.

[8] Lauritzen, S.L. and N. Wermuth, *Graphical models for associations between variables, some of which are qualitative and some quantitative.* The annals of Statistics, 1989: p. 31-57.

[9] Wei, W.W.-S., *Time series analysis*. 1994: Addison-Wesley publ Reading.

[10] Cryer, J.D. and N. Kellet, *Time series analysis*. Vol. 101. 1986: Springer.

[11] Chatfield, C., *Time-series forecasting*. 2000: CRC Press.

[12] Box, G.E., et al., *Time series analysis: forecasting and control*. 2015: John Wiley & Sons.

[13] Engle, R.F. and C.W. Granger, *Co-integration and error correction: representation, estimation, and testing.* Econometrica: journal of the Econometric Society, 1987: p. 251-276.

[14] Amisano, G. and C. Giannini, *From VAR models to structural VAR models*, in *Topics in Structural VAR Econometrics*. 1997, Springer. p. 1-28.

[15] Cooley, T.B., E. Witwer, and P. Lee, *Anemia in children: With splenomegaly and peculiar changes in the bones report of cases.* American Journal of Diseases of Children, 1927. 34(3): p. 347-363.

[16] Powell, W., J. Rodarte, and J. Neel, *The occurrence in a family of Sicilian ancestry of the traits for both sickling and thalassemia.* Blood, 1950. 5(10): p. 887-897.

[17] Weatherall, D. and J. Clegg, *Inherited haemoglobin disorders: an increasing global health problem.* Bulletin of the World Health Organization, 2001. 79(8): p. 704-712.

[18] Rund, D. and E. Rachmilewitz, *β-Thalassemia.* New England Journal of Medicine, 2005. 353(11): p. 1135-1146.

[19] Duster, T., *Backdoor to eugenics*. 2003: Psychology Press.

[20] Raiola, G., et al., *Growth and puberty in thalassemia major.* Journal of pediatric endocrinology & metabolism: JPEM, 2003. 16: p. 259-266.
[21] Win, N., et al., *Use of intravenous immunoglobulin and intravenous methylprednisolone in hyperhaemolysis syndrome in sickle cell disease.* Hematology, 2004. 9(5-6): p. 433-436.
[22] Hasani, V., et al., *The Prevalence of [Beta] Thalassemia in the Department of Pediatric, Regional Hospital of Vlore. 2007-2011.* Journal of Chemical, Biological and Physical Sciences (JCBPS), 2013. 3(2): p. 1169.
[23] Davis, B.A., et al., *Value of sequential monitoring of left ventricular ejection fraction in the management of thalassemia major.* Blood, 2004. 104(1): p. 263-269.
[24] Cunningham, M.J., et al., *Complications of β-thalassemia major in North America.* Blood, 2004. 104(1): p. 34-39.
[25] Pennell, D.J., et al., *Randomized controlled trial of deferiprone or deferoxamine in beta-thalassemia major patients with asymptomatic myocardial siderosis.* Blood, 2006. 107(9): p. 3738-3744.
[26] Nielsen, P., et al., *Using SQUID biomagnetic liver susceptometry in the treatment of thalassemia and other iron loading diseases.* Transfusion science, 2000. 23(3): p. 257-258.
[27] Tanner, M., et al., *Myocardial iron loading in patients with thalassemia major on deferoxamine chelation.* Journal of Cardiovascular Magnetic Resonance, 2006. 8(3): p. 543-547.
[28] Guyatt, G.H., et al., *Laboratory diagnosis of iron-deficiency anemia.* Journal of general internal medicine, 1992. 7(2): p. 145-153.
[29] Fischer, R., et al., *Assessment of iron stores in children with transfusion siderosis by biomagnetic liver susceptometry.* American journal of hematology, 1999. 60(4): p. 289-299.
[30] Eldor, A. and E.A. Rachmilewitz, *The hypercoagulable state in thalassemia.* Blood, 2002. 99(1): p. 36-43.
[31] Piga, A., et al., *Randomized phase II trial of deferasirox (Exjade, ICL670), a once-daily, orally-administered iron chelator, in comparison to deferoxamine in thalassemia patients with transfusional iron overload.* haematologica, 2006. 91(7): p. 873-880.
[32] Voskaridou, E., et al., *Treatment with deferasirox (Exjade®) effectively decreases iron burden in patients with thalassaemia intermedia: results of a pilot study.* British journal of haematology, 2010. 148(2): p. 332-334.
[33] Wahidiyat, I. and P. Wahidiyat, *Genetic problems at present and their challenges in the future: Thalassemia as a model.* Paediatrica Indonesiana, 2016. 46(5): p. 189-94.
[34] Perrotta, S., P.G. Gallagher, and N. Mohandas, *Hereditary spherocytosis.* The Lancet, 2008. 372(9647): p. 1411-1426.

[35] Kantz, H. and T. Schreiber, *Nonlinear time series analysis*. Vol. 7. 2004: Cambridge university press.
[36] Hamilton, J.D., *Time series analysis*. Vol. 2. 1994: Princeton university press.
[37]Madsen, H., *Time series analysis*. 2007: CRC Press.
[38]Chatfield, C., *The analysis of time series: an introduction*. 2016: CRC press.
[39] Shumway, R.H. and D.S. Stoffer, *Time series analysis and its applications: with R examples*. 2010: Springer Science & Business Media.
[40] Tay, F.E. and L. Cao, *Application of support vector machines in financial time series forecasting.* Omega, 2001. 29(4): p. 309-317.
[41] Cogley, T. and J.M. Nason, *Effects of the Hodrick-Prescott filter on trend and difference stationary time series Implications for business cycle research.* Journal of Economic Dynamics and control, 1995. 19(1): p. 253-278.
[42]Hylleberg, S., *Modelling seasonality*. 1992: Oxford University Press.
[43] Gomez, V., *The use of Butterworth filters for trend and cycle estimation in economic time series.* Journal of Business & Economic Statistics, 2001. 19(3): p. 365-373.
[44] Wagner, A.K., et al., *Segmented regression analysis of interrupted time series studies in medication use research.* Journal of clinical pharmacy and therapeutics, 2002. 27(4): p. 299-309.
[45] Kristoufek, L., *Measuring correlations between non-stationary series with DCCA coefficient.* Physica A: Statistical Mechanics and its Applications, 2014. 402: p. 291-298.
[46] Williams, B.M. and L.A. Hoel, *Modeling and forecasting vehicular traffic flow as a seasonal stochastic time series process*. 1999.
[47] Sfetsos, A., *A comparison of various forecasting techniques applied to mean hourly wind speed time series.* Renewable energy, 2000. 21(1): p. 23-35.
[48] De Gooijer, J.G. and R.J. Hyndman, *25 years of time series forecasting.* International journal of forecasting, 2006. 22(3): p. 443-473.
[49]Robinson, P.M., *Time series with long memory*. 2003: Oxford University Press, USA.
[50] Fuller, W.A., *Introduction to statistical time series*. Vol. 428. 2009: John Wiley & Sons.
[51] Huang, N.E., et al. *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis*. in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. 1998. The Royal Society.
[52] Dickey, D.A., D.P. Hasza, and W.A. Fuller, *Testing for unit roots in seasonal time series.* Journal of the American Statistical Association, 1984. 79(386): p. 355-367.
[53] Dickey, D.A. and W.A. Fuller, *Likelihood ratio statistics for autoregressive time series with a unit root.* Econometrica: Journal of the Econometric Society, 1981: p. 1057-1072.

[54] Zhang, G.P. and M. Qi, *Neural network forecasting for seasonal and trend time series.* European journal of operational research, 2005. 160(2): p. 501-514.
[55] Hannan, E.J., *Multiple time series*. Vol. 38. 2009: John Wiley & Sons.
[56] Tong, H., *Threshold models in non-linear time series analysis*. Vol. 21. 2012: Springer Science & Business Media.
[57] Ma, Y. and M.G. Genton, *Highly robust estimation of the autocovariance function.* Journal of time series analysis, 2000. 21(6): p. 663-684.
[58] Harvey, A.C., *The econometric analysis of time series*. 1990: Mit Press.
[59] Zhang, G.P., *Time series forecasting using a hybrid ARIMA and neural network model.* Neurocomputing, 2003. 50: p. 159-175.
[60] Kadilar, C. and C. Erdemir, *Modification of the akaike information criterion to account for seasonal effects.* Journal of Statistical Computation and Simulation, 2003. 73(2): p. 135-143.
[61] McQuarrie, A.D. and C.-L. Tsai, *Regression and time series model selection*. 1998: World Scientific.
[62] Sen, L.K. and M. Shitana, *The Performance of AlCC as an Order Selection Criterion in ARMA Time Series Models.* Pertanika Journal of Science & Technology, 2002. 10(1): p. 25-33.
[63] Cochrane, J.H., *Time series for macroeconomics and finance.* Unpublished book manuscript, 1997.
[64] Brown, R.G., *Smoothing, forecasting and prediction of discrete time series*. 2004: Courier Corporation.
[65] Faruk, D.Ö., *A hybrid neural network and ARIMA model for water quality time series prediction.* Engineering Applications of Artificial Intelligence, 2010. 23(4): p. 586-594.
[66] Musallam, K.M., et al., *Non-transfusion-dependent thalassemias.* haematologica, 2013. 98(6): p. 833-844.
[67] Britton, R.S., K.L. Leicester, and B.R. Bacon, *Iron toxicity and chelation therapy.* International journal of hematology, 2002. 76(3): p. 219-228.
[68] Guralnik, J.M., et al., *Anemia in the elderly: a public health crisis in hematology.* ASH Education Program Book, 2005. 2005(1): p. 528-532.
[69] Porter, J.B., *Practical management of iron overload.* British journal of haematology, 2001. 115(2): p. 239-252.
[70] Harvey, A.C., *Forecasting, structural time series models and the Kalman filter*. 1990: Cambridge university press.
[71] Phillips, P.C. and P. Perron, *Testing for a unit root in time series regression.* Biometrika, 1988: p. 335-346.
[72] Kwiatkowski, D., et al., *Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit*

root? Journal of econometrics, 1992. 54(1-3): p. 159-178.
[73] Bourbonnais, R. and M.M. Maftei, *THE CONTRIBUTION OF ECONOMETRICS.* Romanian Journal of Economic Forecasting, 2012: p. 144.
[74] Klein, A. and G. Melard, *FISHER'S INFORMATION MATRIX FOR SEASONAL AUTOREGRESSIVE-MOVING AVERAGE MODELS.* Journal of time series analysis, 1990. 11(3): p. 231-237.
[75] Moon, H.R., B. Perron, and P.C. Phillips, *Incidental trends and the power of panel unit root tests.* Journal of Econometrics, 2007. 141(2): p. 416-459.
[76] George, E., et al., *Types of thalassemia among patients attending a large university clinic in Kuala Lumpur, Malaysia.* Hemoglobin, 1992. 16(1-2): p. 51-66.
[77] Lal, A., et al., *Combined chelation therapy with deferasirox and deferoxamine in thalassemia.* Blood Cells, Molecules, and Diseases, 2013. 50(2): p. 99-104.
[78] Wood, J.C., et al., *Onset of cardiac iron loading in pediatric patients with thalassemia major.* Haematologica, 2008. 93(6): p. 917-920.
[79] Galanello, R. and R. Origa, *Beta-thalassemia.* Orphanet journal of rare diseases, 2010. 5(1): p. 11.
[80]Quenouille, M., *The analysis of multiple time-series.* 1957.
[81] Debelak, K.A. and C.A. Sims, *Stochastic modeling of an industrial activated sludge process.* Water Research, 1981. 15(10): p. 1173-1183.
[82] Shumway, R.H. and D.S. Stoffer, *Time series regression and exploratory data analysis.* Time Series Analysis and Its Applications: With R Examples, 2006: p. 48-83.
[83] Cromwell, J., et al., *Multivariate tests for time series models. Quantitative Applications in the Social Sciences Series, No. 100*. 1994, Thousand Oaks, CA: Sage Publications, Inc.
[84] Tamura, Y., et al., *A procedure for tidal analysis with a Bayesian information criterion.* Geophysical Journal International, 1991. 104(3): p. 507-516.
[85] Hatemi-J, A., *A new method to choose optimal lag order in stable and unstable VAR models.* Applied Economics Letters, 2003. 10(3): p. 135-137.
[86] Brockwell, P.J. and R.A. Davis, *Time series: theory and methods*. 2013: Springer Science & Business Media.
[87] Kirchgässner, G. and J. Wolters, *Introduction to modern time series analysis*. 2007: Springer Science & Business Media.
[88] Kilian, L. and U. Demiroglu, *Residual-based tests for normality in autoregressions: Asymptotic theory and simulation evidence.* Journal of Business & Economic Statistics, 2000. 18(1): p. 40-50.
[89] Engle, R., *Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models.* Journal of Business & Economic Statistics, 2002. 20(3): p. 339-350.

Acknowledgement

I would like to express my gratitude to my advisor Prof. Liu Jicheng for his supported to me in my study and research, for the patience, motivation. His supervision helped me in all the time of research this thesis. I want also to thank all Professors in my defense committee. I am indebted to my mother and to my family for their patience and efforts that helped me to be what I am today. also I would like to give a special thanks to my husband DHEYAA he always supports me and help me to cross all the problem thanking my nephews FATEMA,REDA and MOHAMED for being a constant reminder to worry about the important things in life.

In addition, I must thank all my all friends who helped keep me sane during the process including classmate here in my second home china.

Appendix Publications

[1] Rana Sabeeh Abood Alsudani and Jicheng Liu. “Forecasting mortality patterns of thalassemia major patients in Iraq by using VAR model and reasons for this mortality” journal of Advances in mathematics, Vol 12, 2016,no.11,6785-6798.

[2] Rana Sabeeh AlSudani, Jicheng Liu and Juan Wu. “Estimation of Co-Integration of the Relationship between Blood Transfusion and Iron Deposits in Thalassemia Patients and Study of the Effects of Some Physiological Factors” International Mathematical Forum, Vol. 11, 2016, no. 22, 1089 – 1102.

[3] Rana Sabeeh Abbood Alsudani and Jicheng Liu. “The Use of Some of the Information Criterion in Determining the Best Model for Forecasting of Thalassemia Cases Depending on Iraqi Patient Data Using ARIMA Model” Journal of Applied Mathematics and Physics, Published during the this month.