ANALYSIS OF SEASONAL TIME SERIES WITH MISSING OBSERVATIONS

By

Kihoro John Mwaniki

A dissertation submitted for the partial fulfillment of the degree of

Master of Science in Mathematical Statistics, of

Kenyatta University

KENYATTA UNIVERSITY LIBRARY

June, 1998

Kihoro, John Mwaniki
Analysis of seasonal time series with

2000/258307

DECLARATION

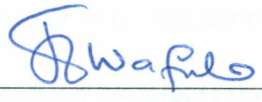
DEPARTMENT OF MATHEMAT
KENYATTA UNIVERSITY
P O. BOX 43844, NAIROB

This is my original work and has not been presented for a degree award in any other university, in part or whole.



JOHN MWANIKI KIHORO

This work has been presented with my approval as the university supervisor



DR. CHARLES WAFULA

Department of Mathematics

Kenyatta University

Abstract

In the past, Kenya's Tourist data has been analysed by among others, Mutiso (1982) who carried out Spectral Analysis to the monthly data for a series covering 10 years (1971 – 1980). Mutiso came up with general conclusions on cycles experienced by the tourism industry in Kenya. Onyango (1993) fitted non-linear model to the tourist data before testing it for linearity. In an effort to confirm or disapprove their findings we decided to analyse a longer series, that is, to collect and use data for the years 1971 to 1990. It is in this process that we encountered the problem of missing data. Monthly data for the year 1984 could not be located in the Kenya Central Bureau of Statistics (CBS) records, but quarterly totals were available. Our immediate problem was to fill in the block of missing values and as such most of the work in this project point to this direction.

In this dissertation we have used the two known segments [(1971 –1983) and (1985 – 1990)] of tourist data to estimate the block of missing values. We have considered two methods under regression (indirect) imputation methods and imputed the missing values from eventual forecasts. We have showed how two forecasts arising from the two regression methods can be combined to come up with more accurate estimates. We have also suggested a method of adjusting the estimates to incorporate the information from the already known quarterly totals.

We have proposed simpler and direct method of imputing the missing values and we have showed that the final adjusted estimates from the direct and the indirect methods are similar. Finally we modelled the two data sets arising from the direct and the indirect imputation and generated eventual forecasts which further showed that the two methods give almost similar estimates.

DEDICATION

I dedicate this work to **Mr. Takanobu KURODA** for his timely intervention when, I had unwillingly dropped out of school.

Acknowledgement

My sincere thanks first goes to the Almighty GOD (EL-SHADDAI) for His unfailing provision, protection, sustenance and the unmerited mercy He gave me during my course, certainly He is the source of all knowledge.

Secondly, I owe my gratitude to my parents for their love and proven efforts to educate me, but much gratitude goes to them that took away the fee burden from them, Mr Takanobu Kuroda of JOCV and Kenyatta University. Mr Takanobu for recommending me for a scholarship from JOCV-KESTES when he used to be my teacher and further taking the burden of paying the rest of my fees, Kenyatta University for offering me a scholarship to pursue this course.

I am greatly indebted to Dr. Wafula, who apart from being my dedicated supervisor, has been a constant source of inspiration as a great teacher and researcher. I am happy to testify that he helped me in everything except making mistakes .Any errors and omissions that remain are entirely my responsibility.

My appreciation also goes to Dr. Odongo and Dr. Njenga for their encouragement and for freely sharing their wisdom whenever I sought their assistance. I cannot forget Dr. Odondo for the way he taught me time series analysis, that aroused my interest.

I am extremely grateful to my colleagues in the struggle. Kimbu and Waweru , for their company, assistance and suggestions throughout this course. To Kimathi and Mogudo, the Laboratory Technicians, thanks for your technical assistance.

I regret that there is no way except a general expression of gratitude in which I can give thanks to friends, siblings and relatives, who encouraged me, kindly made inquiries about my progress, and voiced their hopes for successive completion of this course. To you all, may GOD bless you.

Special thanks to Mr. Wangalwa for his dedication in ensuring that all this work was complete in time. Without his computer services, typing and at times technical assistance as I worked on his computer, this work would not be what it is now. I cannot forget to appreciate sisters Rose and Beatrice for foregoing their sleep to proof read the first copy of this work. To Ann, thank you very much for proof reading part of the second draft.

Last, but not least, thanks to Kenyatta University Postgraduate Fellowship members for their constant prayers and, to Brother Githire: May God bless you for listening to my inner problems and praying with me whenever I needed you.

Table of Contents

		Page
Declaration		i
Abstract		ii
Dedication		iii
Acknowledgements		iv
Table of Contents		vi
List of Abbreviation		viii
List of Figures		x
List of Tables		xi
Overview		xii
Chapter One: Fundamental Concepts and Definitions		
1.1:	Stationary and Stochastic processes	1
1.2:	The Autocovariance and Autocorrelation Function	2
1.3:	Partial Autocorrelation	5
1.4:	Some Useful Stochastic Processes	7
1.5:	The ARIMA Process.	13
Chapter Two: Linear Versus Non – Linear Time Series Models.		
2.1:	Introduction	16
2.2:	Non – Linear and Linear Time Series Models	16
2.3:	Test for Non – Linearity/Linearity	20
2.4:	Modelling of a Time Series	22
2.5:	Model Identification	24
2.6:	Model Estimation	28
2.7:	Model Checking	35
2.8:	Forecasting	38

		Page
	Chapter Three: Estimation of a Block of Missing Observation in a Monthly Time Series given Quarterly Totals	
3.1:	Introduction	40
3.2:	Imputation Methods	40
3.3:	Regression Imputation	41
3.4:	Composite Estimates	49
3.5:	Adjusting Estimates using quarterly Information	52
3.6:	Direct Method of Imputation	54
	Chapter Four: An Empirical Study	
4.1	Introduction	56
4.2(a)	Composite forecasts for series 1	59
4.2(b)	Composite forecasts for series 2	68
4.3	Modified Hot-Deck imputation	77
4.4	Comparison between series C (Sc) and series H (Sh)	78
	Chapter Five: Conclusion and Suggested Areas for Further Research.	
5.1:	Accomplishments (Overview) and Conclusions	83
5.2:	Suggestions	84
	References	86
	Appendices: Tables of Results Based on the Empirical Study	
	Appendix A1: Tables referred to in section 4.1	91
	Appendix A2: Tables referred to in section 4.2	92
	Appendix A3: Tables referred to in section 4.3	97
	Appendix A4: Tables referred to in section 4.4	100

List of abbreviations

ACF	Autocorrelation function.
Act.-v	Actual Value.
ACVF	Autocovariance function.
Adj. Esti.	Adjusted estimate
AIC	Akaike Information Criterion
AR	Autoregressive
ARMA	Autoregressive moving average.
ARIMA	Autoregressive integrated moving average.
BIC	Bayesian Information Criterion.
BL	Bi-Linear.
comb.	Combined.
DF	Degrees of freedom
EAR	Exponential autoregressive.
F_c	Computed F statistics.
Forec./Frc	forecast.
Hd	Hot-Deck.
Holt-W	Holt-Winters
Hd-esti.	Hot deck estimates
i.i d $N(0,\sigma_e^2)$	Independent and identically distributed normal variables with mean 0 and variance σ_e^2.
MA	Moving average.
MS	Mean square.
Q-Av	Quarterly average

Q-T	Quarterly total.
Q_c	Computed Q statistic.
Q_t	Tabulated Q statistic.
RHS	Right hand side.
S1	Series 1.
S2	Series 2.
Sc	Series C.
Sh	Series H.
SACF	Sample ACF.
SARIMA	Seasonal ARIMA.
SBC	Schwartz Bayesian Criterion
SDM	State Dependent Model.
SPACF	Sample PACF.
SPSS	Statistical Package for Social Sciences.
SS	Sum of squares.
SSE	Sum of squares due to error.
TAR	Threshold autoregressive.

List of Figures

		Page
Fig.4.1:	Time Plot for Series 1	58
Fig. 4.2	Time Plot for series 2	59
Fig 4.3	SACF for S1	61
Fig 4.4	SACF for S1,one simple differencing	62
Fig 4.5	SACF for S1,one seasonal differencing	63
Fig 4.6	SACF for S1,one simple and one seasonal differencing	64
Fig.4.7:	SPACF for S1,one simple and one seasonal differencing	65
Fig 4.8	SACF for residuals S1	67
Fig 4.9	SACF for S2	71
Fig.4.10:	SACF for S2,one simple differencing	72
Fig. 4.11	SACF for S2,one seasonal differencing	73
Fig 4.12	SACF for S2,one simple and one seasonal differencing	74
Fig 4.13	SPACF for S2,one simple and one seasonal differencing	75
Fig 4.14	SACF for residuals S2	76
Fig.4.15:	Time Plot for Series C(Sc)	81
Fig. 4.16	Time Plot for series H(Sh)	82

List of Tables

		Page
Table 2.5.1	Summary of the known patterns of ACF and PACF	26
Table 4.2.1a	Final parameters for SARIMA model fitted to S1	66
Table 4.2.2a	SARIMA model variables for S1	66
Table 4.2.3a	Seasonal indices for S1	68
Table 4.2.1b	Final parameters for SARIMA model fitted to S2	76
Table 4.2.2b	SARIMA model variables for S2	77
Table 4.2.3b	Seasonal indices for S2	77
Table 4.4.1	Descriptives for Sc and Sh	83
Table 4.4.2	Test for Linearity results for Sc and Sh	83
Table 4.4.3	Model parameters for Sc and Sh	84

OVERVIEW

A time series is an ordered sequence of observations. Although the ordering is usually through time, particularly in terms of some equally spaced time intervals the ordering may be taken through other dimensions, such as space. There are various objectives of studying time series. These include the understanding and description of the generating mechanism, the forecasting of future values and optimal control of a system. In a time series the observations are dependent or correlated and the order of the observations is therefore important. Hence, statistical procedures and techniques that rely on independence assumptions are no longer applicable and different methods are needed. The body of statistical methodology available for analysing time series is referred to as Time Series Analysis.

In the past, Kenya's tourist data has been analysed by among others, Mutiso (1982) and Onyango (1993). Mutiso carried out a Spectral Analysis of monthly data for a series covering ten years (1971 – 1980) and came up with general conclusions on the cycles experienced by the tourism industry in Kenya. In so doing, Mutiso just considered one of the objectives of analysing a time series, i.e. description. Onyango later fitted non-linear models to this data but with no justification. He should have tested this particular data for linearity. For this reason, his conclusions are not reliable as far as explaining this series is concerned.

Our objective was /is to fit an appropriate model to this tourist data which could be used to forecast/predict future values. We note that forecasts are required for,

(a) Helping in making proper decisions and

(b) Planning for future.

Unlike Onyango, we are not assuming that the data follows a linear or non-linear model. We also use a longer series than that which was used in the past for more accurate results.

In an effort to obtain a longer series for this data, we failed to get all the twelve monthly observations for the year 1984. We could only get the quarterly totals for that year. We therefore had the monthly series for 1971 –1983 and 1985 to 1990. Having encountered this problem of missing data, our first task was to deal with this problem. Our major problem was thus "*Estimation of missing observation in a time series*".

Noting that the missing values partition our data into two sets, each with more than 50 observations (a necessary condition in fitting an ARIMA model), we found a basis for giving our problem an approach suggested by Abraham (1981). That is, fit a model to the known segments and extrapolate the adjacent missing values. To obtain more accurate estimates, we are going to use appropriate smoothing algorithms to generate the same and later combine these eventual forecasts from the two procedures using the regression method as suggested by Granges and Ramanathan (1984) and Triantis (1983).

We are also going to present a direct imputation method which can be used without fitting a model or using smoothing algorithms. In particular, we modify and use what is commonly known as *Hot deck* imputation procedure to come up with another complete set of the data. After the missing observations have been imputed, we then

fit an appropriate model to the whole series, which maybe used to forecast future values.

Chapter One is aimed at introducing the fundamental concepts needed to understand the rest of the work in this dissertation. Generally, it involves definition of terms to be used elsewhere in the project. In Chapter Two we introduce the concept of linear and non-linear models and present the procedures of modelling a given time series, that is, Model Identification, Estimation and Checking procedures. In Chapter Three we present the methods of imputation of the missing data, in particular, we show how Holt–Winters procedures and modelling can work together to produce better estimates through composite forecasts. Noting that we have some information about the missing values (quarterly totals), we show how this information can be utilised to produce even better estimates. That is we adjust the final estimates appropriately such that their quarterly total equals the true quarterly total. We also propose and describe a modified version of the *Hot - deck* imputation procedure and show how this can be applied in a seasonal time series.

In Chapter Four we carry out an empirical study based on Chapters two and three. It is here that we estimate the missing observations and fit an appropriate model to the complete data sets, that is, 1971 – 1990 data having been filled in two different ways. In Chapter Five we give an overview of what we have accomplished and suggest areas for further research.

CHAPTER ONE

FUNDAMENTAL CONCEPTS AND DEFINITIONS

1.1: Stationarity and stochastic processes

A time series is some sequence of observations proceeding through time where the actual order in which the values occur has importance, in that they may be related temporally. The theory and practice is very different from that for the "remainder" statistics in that:

(i) Observations elsewhere are assumed to be independent unlike in a time series

(ii) Asymptotic theory tends generally to be meaningful in practice. This is not the case with time series. In fact, the temporal or spatial dependence in time series is what is of interest and importance (Anderson, 1982).

An observed time series $(X_1, X_2 \ldots\ldots, X_n)$ can be thought of as a particular realisation of a stochastic process. Stochastic processes in general can be described by an n-dimensional probability distribution $P(x_1, x_2, \ldots\ldots x_n)$.

Let $\{X_t: t = 0, \pm 1, \pm 2, \ldots\ldots\ldots\}$ be a random sequence. Further, let the distribution of any set of n consecutive Xs say, $X_{t+1}, X_{t+2}, - - - X_{t+n}$ be $F(X_{t+1}, X_{t+2}, - - - X_{t+n})$, then if $F(.)$ is independent of t for all integral $n>0$, we say that the time series is *strictly stationary*. That is $\{X_t: t = 0, \pm 1, \pm 2, - - -\}$ is stationary in the strong sense if

$F(X_{t_1}, X_{t_2}, \ldots\ldots\ldots, X_{t_n}) = F(X_{t_1+k}, X_{t_2+k}, \ldots\ldots\ldots, X_{t_n+k})$ for all possible sets of indices $t_1, t_2 - - - t_n$ and $t_1+k, t_2+k, - - -; t_n+k$ in the set $\{1, 2, - - -\}$

The time series is said to be covariance stationary or stationary in the weak sense if

(i) $E(X_t) = \mu < \infty \quad t = 1, 2 - - - \ldots\ldots\ldots\ldots\ldots\ldots 1.1.1$

(ii) $E(X_t - \mu)(X_{t+k} - \mu) = Cov(X_t, X_{t+k}) = \sigma(k) < \infty \ldots\ldots 1.1.2$

where μ is a real number and $\sigma(k)$ is a real valued function of k.

Stationary in the weak sense implies that, the mean and the variance of the process are constant and the autocovariances

$$\sigma(k) = Cov(X_t, X_{t+k})$$

and autocorrelations

$$\rho(k) = \frac{Cov(X_t, X_{t+k})}{[V(X_t), V(X_{t+k})]^{1/2}} = \frac{\sigma(k)}{\sigma(o)}$$

depend only on the lag (or time difference) k. If joint normality is assumed, so that the distribution is entirely characterised by these first two moments, strict stationary and weak stationarity are equivalent.

1.2: The Autocovariance and Autocorrelation Function.

For a stationary process $\{X_t\}$, we have the mean $E(X_t) = \mu$ and variance $\sigma(o) = E(X_t - \mu)^2 = \sigma^2$ which are constants and the covariances $Cov(X_t, X_s)$, which are functions of the time difference $|t - s|$

Then,

$$\sigma(k) = Cov(X_t, X_{t+k}) = E(X_t - \mu)(X_{t+k} - \mu) \quad \ldots\ldots\ldots 1.2.1$$

is called the autocovariance function (ACVF) and

$$\rho(k) = \frac{Cov(X_t, X_{t+k})}{\sqrt{Var(X_t)}\sqrt{Var(X_{t+k})}} = \frac{\sigma(k)}{\sigma(o)} \quad \ldots\ldots\ldots\ldots\ldots 1.2.2$$

is called the autocorrelation funtion (ACF) in time series analysis since they represent the covariance and correlation between X_t and X_{t+k} from the same process, separated only by k time lags.

For a stationary process the ACVF and ACF have the following properties:

1. $\sigma(o) = \text{Var}(X_t)$, $\rho(o) = 1$

2. $|\sigma(k)| \leq \sigma(o)$, $\rho(k) \leq 1$.

3. $\sigma(k) = \sigma(-k)$, $\rho(k) = \rho(-k)$ and so only the positive half of ACF is usually given/plotted.

4. Both $\sigma(k)$ and $\rho(k)$ are positive semi-definite in the sense that

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\sigma|t_i - t_j| \geq 0 \quad \ldots\ldots\ldots\ldots\ldots 1.2.3$$

And

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\rho|t_i - t_j| \geq 0 \quad \ldots\ldots\ldots\ldots\ldots 1.2.4$$

for any set of time points $t_1, t_2 - - - t_n$ and any real numbers $\alpha_1, \alpha_2 - - - \alpha_n$.

Proof

Let $Y = \sum_{i=1}^{n} \alpha_i X_{t_i}$ be a random variable, then

$$0 \leq V(Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j Cov(X_{t_i}, X_{t_j})$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_j \alpha_i \sigma \mid t_i - t_j \mid$$

dividing the above inequality by $\sigma(o)$, the result for $\rho(k)$ in (1.2.4) follows immediately. As noted by Wei (1990) a necessary condition for a function to be the ACVF or ACF of some process is that it be positive semi-definite.

The ACF plays a major role in modelling the dependencies among observation since it characterises together with process mean $\mu = E(X_t)$ and variance $\sigma(o) = V(X_t)$, the stationary stochastic process describing the evolution of $\{X_t\}$.

Sample autocorrelation function (SACF)

For a given observation time series $X_1, X_2, - - - X_n$, the sample ACF is defined as

$$r(k) = \frac{\sum_{t=1}^{n-k} (X_t - \bar{X})(X_{t+k} - \bar{X})}{\sum_{t=1}^{n} (X_t - \bar{X})^2} \quad k = 0, 1, 2, - - - \qquad \ldots 1.2.5.$$

where $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$ is the sample mean of the series. A plot of $r(k)$ versus k is sometimes called a sample correlogram.

For uncorrelated observations $\rho(k) = 0$ for all $k \neq 0$, the variance of r(k) is approximately given by $1/n$ (see, for example Box – Jenkins ,1976). If n is large, r(k) in excess of $2/\sqrt{n}$ may be regarded as significantly different from zero. Bartlett (1946) has shown that for $k > q$,

$$V[r(k)] = \frac{1}{n}(1+2\rho_1^2 + 2\rho_2^2 + - - +2\rho_q^2) \ldots\ldots\ldots\ldots 1.2.6.$$

In practice ρ_i $(i = 1, 2, - - - q)$ are unknown and are replaced by their sample estimates $r(i)$.

1.3: Partial Autocorrelation

In the context of time series, a large portion of the correlation between X_t and X_{t+k} can be due to the correlation these have with the intervening lags X_{t+1}, X_{t+2}, - - - X_{t+k-1}. To adjust for this correlation, the partial autocorrelations may be calculated. The lag k partial autocorrelation is the partial regression coefficient ϕ_{kk} in the kth order autoregression.

$$X_{t+k} = \phi_{k1}X_{t+k-1} + \phi_{k2}X_{t+k-2} + - - - + \phi_{kk}X_t + e_{t+k} \ldots\ldots\ldots\ldots\ldots 1.3.1$$

where e_{t+k} is a normal error term uncorrected with X_{t+k-j}, $j \geq 1$.

Multiplying 1.3.1 by X_{t+k-j} (on both sides) and taking expectation, we get:

$$\sigma(j) = \phi_{k1}\ \sigma(j-1) + \phi_{k2}\,\sigma(j-2) + - - - + \phi_{kk}\,\sigma(j-k) \quad \ldots\ldots\ldots\ 1.3.2$$

dividing through by $\sigma(o)$ gives:

$$\rho(j) = \phi_{k1}\rho(j-1) + \phi_{k2}\rho(j-2) + - - - + \phi_{kk}\rho(j-k) \quad \ldots\ldots\ 1.3.3$$

for $j = 1, 2, - - -, k$, we have a system of equations which can be solved using Cramer's rule successively for $k = 1, 2, - - -$ to give:

$$\phi_{kk} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & - - - & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & - - - & \rho_{k-3} & \rho_2 \\ & - & - & & - & - \\ - & - & - & & - & - \\ & - & - & & - & - \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & - - - & \rho_1 & \rho_k \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & - - - & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & - - - & \rho_{k-3} & \rho_{k-2} \\ & - & - & & - & - \\ - & - & - & & - & - \\ & - & - & & - & - \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & - - - & \rho_1 & 1 \end{vmatrix}} \quad \ldots\ldots\ldots\ldots\ 1.3.4$$

For an alternative definition see Wei (1990: Section 2.3).

Sample partial autocorrelation function (SPACF)

The sample PACF $\hat{\phi}_{kk}$ is obtained by substituting ρ_i by $\hat{\rho}_i = r(i)$ in equation 1.3.4. A recursive method starting with $\hat{\phi}_{11} = r(1)$ for computing $\hat{\phi}_{kk}$ has been given by Durbin (1960) as follows:

$$\hat{\phi}_{k+1,k+1} = \frac{r(k+1) - \sum_{j=1}^{k} \hat{\phi}_{k,j} r(k+1-j)}{1 - \sum_{j=1}^{k} \hat{\phi}_{k,j} r(j)} \quad \ldots\ldots\ldots\ldots\ldots\ldots 1.3.5$$

and $\hat{\phi}_{k+1,j} = \hat{\phi}_{k,j} - \hat{\phi}_{k+1,k+1}\hat{\phi}_{k,k+1-j} \qquad j = 1, 2, - - - k.$

Under the hypothesis that the underlying process is *white noise* sequence, the variance of $\hat{\phi}_{kk}$ can be approximated by,

$$V(\hat{\phi}_{kk}) \cong 1/n$$

Hence, $\pm 2/\sqrt{n}$ can be used as critical limits on ϕ_{kk} to test the hypothesis of "white noise" process.

1.4: Some useful Stochastic Processes

(i) The white noise processes

A process $\{e_t\}$ is called a white noise process if it is a sequence of uncontrolled variables from a fixed distribution with constant mean and variance. That is,

$\mathrm{E}(e_t) = \mu$ and $\mathrm{V}(e_t) = \sigma_e^{\,2}$ and $\sigma(k) = \mathrm{cov}(e_t e_{t+k}) = 0$ for $k \neq 0$. Usually $E(e_t)$ is assumed to be 0. It follows that a white noise process is stationary with autocovariance function (ACVF),

$$\sigma(k) = \begin{cases} \sigma_e^2 & k = 0 \\ 0 & k \neq 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots 1.4.1$$

then ACF,

$$\rho(k) = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1.4.2$$

and the PACF,

$$\phi_{kk} = \begin{cases} 1 & k = 0 \\ 0 & k \neq 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1.4.3.$$

(ii) Moving Average Processes

A process $\{X_t\}$ is said to be a moving average process of order q [MA(q) process] if

$$X_t = \sum_{j=0}^{q} \theta_j e_{t-j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1.4.4$$

where $\{\theta_j\}$ are constants and $\theta_0 = 1$, $\{e_t\}$ is the white noise process. For the MA(q) process we have:

$$\sigma(k) = \begin{cases} \sigma^2 \sum_{i=1}^{q-k} \theta_i \theta_{i+k} & k = 0,1, \quad - - - q \\ \sigma(-k) & k < 0 \\ 0 & k > q \end{cases}$$

as the ACVF and the ACF is given by:

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \sum_{i\ 0}^{q-k} \theta_i \theta_{i+k} \Big/ \sum_{i\ 1}^{q} \theta_i^2 & k = 1,2, \quad - - - q \\ \rho(-k) & k < 0 \\ 0 & k > q \end{cases} \quad \ldots\ldots\ldots\ldots 1.4.5$$

Note that the ACF 'cuts off' at lag q which is a special feature of MA processes.

(iii) Autoregressive Processes

A process $\{X_t\}$ is said to be an autoregressive process of order p (AR(p) process) if:

$$X_t = \sum_{j=1}^{p} \phi_j X_{t-j} + e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1.4.6$$

where $\{\phi_j\}$ are constants.

Multiplying (1.4.6) by X_{t-k} and taking expectations on both sides, we get:

$$\sigma(k) = \phi_1\sigma(k-1) + - - - + \phi_p\sigma(k-p) \; k > 0$$

Dividing through by $\sigma(o)$ gives:

$$\rho(k) = \phi_1\rho(k-1) + - - - + \phi_p\rho(k-p) \; k > 0 \quad \ldots\ldots\ldots\ldots\ldots 1.4.7$$

the first p equations in (1.4.7) $(k = 1, 2, - - -p)$ are called Yule-walker equations which can be solved for $\rho(k)$ to obtain the ACF associated with AR(p).

Note: the solution to the difference equation in (1.4.7).

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)\rho(k) = 0$$

is

$$\rho(k) = A_1 g_1^{\,k} + - - - + A_p g_p^{\,k} \quad k = 0, 1, 2, - - -$$

where $|g_i| < 1$, B is the backward shift operator defined as:

$$B^i X_t = X_{t-i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1.4.8.$$

An autoregressive moving average process of order (p, q) ARMA (p, q) combines equations (1.4.4) and (1.4.6) into a defining equation

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + e_t + \sum_{j=1}^{q} \theta_j e_{t-j} \quad \ldots\ldots\ldots\ldots\ldots\ldots 1.4.9$$

OR

$$X_t - \sum_{i=1}^{p} \phi_i X_{t-i} = e_t + \sum_{j=1}^{q} \theta_j e_{t-j}$$

$$\Rightarrow (1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t = (1 + \theta_1 B + - - - \theta_q B^q) e_t$$

$$\Rightarrow \phi_p(B) X_t = \theta_q(B) e_t$$

where B is the backshift operator defined in (1.4.8) and $\phi_p(\bullet)$, $\theta_q(\bullet)$ are polynomials of degree p and q respectively and $\phi_p(o) = \theta_q(o) = 1$.

According to Diggle (1990), the easiest way to obtain the ACF in specific cases of ARMA is to express X_t as a general linear process,

$$X_t = \frac{\theta_q(B) e_t}{\phi_p(B)} = \psi(B) e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots 1.4.10$$

The coefficients ψ_j can be determined by a formal power series expansion of $[\phi_p(B)]^{-1}$

Example: Suppose that $X_t \sim$ ARMA(1,1). Then, the process is defined by:

$$X_t = \phi_1 X_{t-1} + e_t + \theta e_{t-1}$$

or

$$X_t - \phi_1 X_{t-1} = e_t + \theta e_{t-1}$$

$$\Rightarrow (1 - \phi B) X_t = (1 + \theta B) e_t$$

$$\Rightarrow X_t = \frac{(1 + \theta B)}{(1 - \phi B)} e_t = e_t + \psi_1 B + \psi_2 B^2 + \ldots + .$$

The weights ψ_i are obtained by equating coefficients of B^j in $(1 - \phi B)(1 + \psi_1 B + \psi_2 B^2 + \ldots)$ to those in $(1 + \theta B)$ which leads to:

$$B^1: \quad \psi_1 - \phi = \theta \quad \Rightarrow \quad \psi_1 = \phi + \theta$$

$$B^2: \quad \psi_2 - \phi\psi_1 = 0 \quad \Rightarrow \quad \psi_2 = (\phi + \theta)\phi$$

$$\vdots$$

$$B^j: \quad \psi_j - \phi\psi_{j-1} = 0 \Rightarrow \quad \psi_j = \phi\psi_{j-1} = (\phi + \theta)\phi^{j-1} \; j > 0$$

so that,

$$X_t = e + \sum_{j=1}^{\infty} (\phi + \theta)\phi^{j-1} e_{t-j} .$$

Having found the ψ weights the ACF of ARMA (1, 1) can easily be determined as follows:

$$X_t = \phi X_{t-1} + e_t + \theta e_{t-1}$$

Multiplying this equation by X_{t-k} and taking expectations, we obtain

$$\sigma(k) = \phi\sigma(k - 1) + E(e_t X_{t-k}) + \theta E(e_{t-1} X_{t-k}) \ldots\ldots\ldots\ldots\ldots\ldots \quad 1.4.11$$

If $k > 1$, $E(e_t e_{t-k}) = E(e_{t-1} X_{t-k}) = 0$ since e_t and e_{t-1} are uncorrected with X_{t-k}. therefore:

$$\sigma(k) = \phi\sigma(k - 1) \text{ for } k > 1$$

Now, $E(e_t X_t) = E[e_t(e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + - - -)] = \sigma^2$

$$E(e_{t-1} X_t) = E[e_{t-1}(e_t + \psi_1 e_{t-1} + \psi_2 e_{t-2} + - - -)] = \psi_1 \sigma^2$$

$$= (\phi + \theta)\sigma^2$$

Thus for

$k = 0$, $\sigma(o) = \phi\sigma(1) + \sigma^2 + \theta(\phi + \theta)\sigma^2$

and

$k = 1$, $\sigma(1) = \phi\sigma(o) + \theta\sigma^2$

solving these two equations leads to

$$\sigma(o) = \frac{1+\theta^2+2\phi\theta}{1-\phi^2} \text{ and } \sigma(1) = \frac{(1+\phi\theta)(\phi+\theta)\theta^2}{1-\phi^2}$$

Hence the ACF is given by

$$\rho(k) = \begin{cases} 1 & k = 0 \\ \dfrac{(1+\phi\theta)(\phi+\theta)}{1+\theta^2+2\phi\theta} & k = 1 \\ \phi\rho(k-1) & k \geq 2 \end{cases}$$

The fact that $\rho(k) = \phi\rho(k - 1)$ for $k \geq 2$ implies that the exponential decay that characterises ARMA (1, 1) process starts from $\rho(1)$ and not $\rho(o) = 1$ as in the case of AR(1). Abraham and Ledolter (1983) notes that, the decay is smooth for $\phi < 0$. They also note that the ACF and PACF of ARMA (p, q) process are both infinite in extent and tail off (die down) as the lag k increases.

Note: If $X_t \sim$ ARMA (p, q) then $X_t \sim$ MA(q) and $X_t \sim$ AR(p) when p = 0 and q = 0 respectively.

For an MA process, a restriction is placed on the parameters to make it "invertible". That is to make it possible to write the process as a converging autoregressive representation AR(∞). If $X_t \sim$ MA(q) then

$$X_t = e_t + \sum \theta_i e_{t-i} = (1 - \theta_2 B^2 - - - \theta_q B^q) e_t$$

$$= \theta_q(B) e_t \quad \ldots\ldots\ldots\

1.5: **The ARIMA Process**

The processes discussed so far deal with stationary time series. This implies that the mean, the variance and the autocovariances of the process are invariant under time translations. Many observed time series, however, are not stationary. Various methods have been suggested in dealing with non-stationary series by among others Trintner (1940), Yaglon (1955) and Box-Jenkins (1970) (1976). Box-Jenkins suggested one extension to ARMA class of processes, the ARIMA (Autoregressive Integrated Moving Average) process which deals with non-stationary time series.

A series $\{X_t\}$ is said to follow ARIMA process of order p, d, q written as $X_t \sim$ ARIMA (p, d, q) if it's d^{th} simple differencing series follows stationary, invertible ARMA (p, q) process.

Thus, if $X_t \sim$ ARIMA (p, d, q) then we have

$$\phi_p(B)(1-B)^d X_t = \theta_q(B)e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 1.5.1$$

Where e_t is white noise.

$\phi_p(\bullet)$ and $\theta_q(\bullet)$ are polynomials of degree p and q respectively with all roots of the polynomial equations $\phi(.) = 0$ and $\theta(.) = 0$ outside the unit circle.

If more than 50 observations are available, Box-Jenkins (1970) advocated the fitting of ARIMA models by differencing the observed series until it becomes stationary and then fitting an ARMA model to the differenced series.

In a time series containing seasonality, a seasonal periodic component repeats itself after every s observation (s = 12 for monthly observations). Box-Jenkings (1970) proposed that we transform $\{X_t\}$ into a stationary series by simple differencing (to remove trend) and Seasonal differencing (to remove seasonality), and then fit a Seasonal Autoregressive Integrated Moving average SARIMA model to this series The model is defined by,

$$\phi_p(B)\Phi_p(B^s)w_t = \theta_q(B)\Theta_Q(B^s)e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad 1.5.2$$

where w_t is the differenced stationary series. Since our data contains seasonality, we are going to discuss in details the SARIMA models in Chapter Two.

CHAPTER TWO

LINEAR VERSUS NON – LINEAR TIME SERIES MODELS

2.1: **Introduction**

We noted that Onyango (1993) fitted a non-linear model to tourist data before testing this data for linearity. We do not wish to do the same here and so we now present the relationship between linear and non-linear time series models. In this chapter we show how a general time series model (the state dependent model) can be used to generate both non-linear and linear models (discussed in Chapter one). We also present a general test for linearity before we discuss the problem of modelling a given time series, that is, model identification, model estimation and model checking.

2.2: **Non linear and Linear Time Series Models**

According to Priestley (1988), a linear time series model can be written as:

$$\sum_{i=-\infty}^{\infty} h_i X_{t-i} = e_t \quad \ldots\ldots\ldots 2.2.1$$

Where h is linear and $\{e_t\}$ is the white noise with the usual assumptions. In practice we usually require X_t to depend only on the past values i.e. we set $h_i = 0$ for $i = < 0$ and rewrite (2.2.1) as:

$$\sum_{i=0}^{\infty} h_i X_{t-i} = e_t \quad \ldots\ldots\ldots 2.2.2.$$

Let $B^i X_t = X_{t-i}$ be the background shift operator. Then (2.2.2) becomes:

$$H(B)X_t = e_t \quad \ldots\ldots\ldots\ldots 2.2.3$$

Inverting (2.2.3), we get:

$$X_t = H^{-1}(B)e_t = [g_0 + g_1 B + g_2 B^2 + - - -]e_t$$

$$= \sum_{i=0}^{\infty} g_i e_{t-i} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.2.4$$

Equation 2.2.4 gives the general representation of a linear model in which X_t is expressed as a linear combination of present and past values of a white noise process $\{e_t\}$. Examples of linear models, which can be represented this way, are AR, MA and ARMA models.

Priestley (1981, 1988) gives the general form of a non-linear model to be:

$$X_t = \mu + \sum_{i=0}^{\infty} g_i e_{t-i} + \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} g_{ij} e_{t-i} e_{t-j} + \sum_{i}^{\infty}\sum_{j}^{\infty}\sum_{k}^{\infty} g_{ijk} e_{t\ i} e_{t-j} e_{t-k} + - - \quad \ldots 2.2.5$$

Where $\mu = f(0), g = [\partial f / \partial e_{t-i}]_0$, which are obtained by first assuming that

$X_t = f(e_t, e_{t-i}, e_{t-2} - - -)$ is analytic and expanding it about a fixed point say about the point $\mathbf{0} = (0, 0, 0)$.

We note that non-linearity exists only if any of the higher order coefficients $\{g_{ij}\}$, $\{g_{ijk}\}$ - - - are non zero. That is (2.2.4) arises from (2.2.5) iff $\{g_{ij}\}$, $\{g_{ijk}\}$ ------- are zeros.

Tong (1990) adequately discusses the non-linear models. Three important types of non-linear models are Bi-linear models (BL-models), Threshold autoregressive (TAR) models and Exponential autoregressive (EAR) models. These three models are special cases of a general class of time series models termed by Priestley (1980) as state-dependent. Priestley shows that a general relationship between X_t and a white noise input e_t can be represented as:

$$X_t = f(X_{t-1}, - - - X_{t-p}, e_{t-1}, - - - e_{t-q}) + e_t \quad \ldots\ldots\ldots \quad 2.2.6$$

Expanding the RHS of (2.2.6) in Taylor series about some fixed point allows this relationship to be written as State Dependent Model (SDM) of order (p,q)

$$X_t + \sum_{i=1}^{p} \phi_i(Z_{t-1}) X_{t-i} = \mu(Z_{t-1}) + e_t + \sum_{j=1}^{q} \theta_j(Z_{t-1}) e_{t-j} \quad \ldots\ldots\ldots \quad 2.2.7.$$

Where $Z_t = (e_{t-q+1}, - - - e_t, X_{t-p+1}, - - - , X_t)'$ is called the state vector.

The above model has a natural interpretation as locally linear ARMA model in which the evolution of the series at time $(t-1)$ is governed by a set of AR coefficients $\{\phi_i\}$, a set of MA coefficients $\{\theta_j\}$, and a local mean μ, all of which depend on the "state" of the series at time t – 1. By choosing particular forms of these parameters we come up with a number of important time series models as special cases.

Gooijer and Kumar (1992) have taken particular forms of these parameter and have come up with both linear and non-linear models. In our case we only take one special case which gives rise to an ARMA model from which we can obtain other linear models.

Take $\mu(x)$, $\phi_i(x)$ $(i = 1, 2, - - -, p)$ and $\theta_j(x)$ $(j = 1, 2, - - -, q)$ all as constants, then (2.2.7) becomes:

$$X_t + \sum_{i=1}^{p} \phi_i X_{t-1} = \mu + e_t + \sum_{j=1}^{q} \theta_j e_{t-j} \quad \ldots\ldots\ldots\ldots 2.2.8$$

which is the ARMA (p, q) model. When $\mu = 0$ we have:

$$X_t + \sum_{i=1}^{p} \phi_i X_{t-1} = e_t + \sum_{j=1}^{q} \theta\, e_{t-j}$$

which is the well-known ARMA (p, q) model.

If $p = 0$ we get,

$X_t \sim$ MA (q)

If $q = 0$

$X_t \sim$ AR (p)

Which are examples of linear models.

We note that, the SDM links the well known linear time series models with the non-linear ones. We only need to take particular forms of its parameters and generate either a linear model or a non-linear model.

2.3: **Test for Non – Linearity or Linearity**

As noted earlier it is not logical to fit either a linear model or a non-linear one unless the data provides sufficient evidence in that direction. This requires a test to be carried out to check for this evidence. Tong (1990) suggests some informal tests for linearity before discussing in details some of the common formal test. Gooijer and Kumar (1992) also adequately discuss these formal test.

Among the formal tests suggested are two Portmanteau tests or tests with no specific alternatives, which we discuss below. Note that there are other tests with specific alternatives, which required one to be having some prior knowledge about the most likely model.

(i) **Test based on the squares of time series data**:

Granger and Newbold (1976) showed that if $\{X_t\}$ is a Gaussian distributed time series then,

$\rho_k(X_t^2) = [\rho_k(X_t)]^2$ when the series is linear, where $\rho_k = \text{Corr}\,(X_t, X_{t-k})$.

Departure from this therefore would indicate a degree of non-linearity. We note that the assumption that $\{X_t\}$ is normally distributed limits the use of this method because in most cases this is not the case.

(ii) Tukey's one – degree – of – freedom test for non – additivity

The null hypothesis is:

$$H_0 : X_t = \mu + \sum_{j=1}^{m} \phi_j X_{t-j} + e_t \quad \ldots\ldots\ldots 2.3.1$$

The initial test has been adjusted by Keenan (1985) to run as follows:

(i) Suppose we have the realization(X_1, X_2 - - - X_N) for large N,

- Regress X_t on $\{1, X_{t-1}, - - - X_{t-m}\}$, m being a large but fixed positive integer and
- Calculate the fitted values $\{\hat{X}_t\}$ and the residuals $\{\hat{e}_t\}$ for $t = m+1, m+2, - -$ N,. Calculate the residual sum of squares SSE = $\sum \hat{e}_t^{\,2}$

(ii) Regress $\hat{X}_t^{\,2}$ on $\{1, X_{t-1}, - - - X_{t-m}\}$ and calculate the residuals $\{\hat{\varepsilon}_t\}$ for

$t = (m + 1, - - - N)$

(iii) Regress $\{\hat{e}_t\}$ on $\{\hat{\varepsilon}_t\}$ and obtain η and F as follows:

$$\eta = \eta_o \left(\sum_{t=m+1}^{N} \hat{\varepsilon}_t^{\,2} \right)^{1/2} \quad \ldots\ldots\ldots\ldots\ldots 2.3.2$$

Where η_0 is the regression coefficient on regressing $\{\hat{e}_t\}$ on $\{\hat{\varepsilon}_t\}$.

$$F = \frac{\eta^2 (N - 2m - 2)}{SSE - \eta^2} \quad \ldots\ldots\ldots\ldots 2.3.3$$

Under the null hypothesis of linearity F will be distributed as $F(1, N - 2m - 2)$. Tong (1990) has noted that this test in many cases gives satisfying results.

2.4: **Modelling of a Time Series**

Once the data has been tested for linearity we may proceed to model that particular series. The modelling procedures depend on whether the series is linear or non-linear. In our case the test discussed above led to the conclusion that the tourist data is linear (Results in Chapter Four). In this section we therefore show how a linear time series may be modelled. For non-linear time series modelling see, Gooijer and Kumar (1992). We in particular show how a non-stationary linear time series may be modelled.

Consider an economic time series. In most cases they exhibit seasonal fluctuations which we informally describe as the movement of a component with a period of approximately one year. Many a time, they also show some trend. In a time series containing seasonality, a seasonal period component repeats itself after every s = 12 observations and thus we expect X_t to depend on terms such as X_{t-12} and perhaps X_{t-24} as well as terms such as X_{t-1}, X_{t-2} - - -. Box-Jenkins (1970) have generalized the ARIMA model to deal with seasonality, thus coming up with what is known as SARIMA (seasonal autoregressive integrated moving average) model given by:

$$\phi_p(B)\Phi_P(B^s)w_t=\theta_q(B)\Theta_Q(B^S)e_t \quad \ldots\ldots\ldots 2.4.1$$

ϕ_p, Φ_P, θ_q, Θ_Q are polynomials of order p, P, q, Q respectively, and $w_t = \nabla^d \nabla_s^D X_t$. where ∇^d is the simple differencing operator of order d, ∇^D is the seasonal differencing operator of order D. that is, for $d = D = 1$

$$\nabla^1 = X_t - X_{t-1} = (1-B)X_t$$

$$\nabla_s^1 = X_t - X_{t-s} = (1-B^s)X_t.$$

The variables $\{w_t\}$ are formed from the original series $\{X_t\}$ not only by simple differencing (to remove trend) but also by seasonal differencing ∇_s to remove seasonality. The above model is known as SARIMA model of order (p, d, q)(P, D, Q) commonly written as,

ARIMA (p, d, q) (P, D, Q)

Mills (1990) notes that general multiplicative model (2.4.1) is rather complicated and as such it is difficult to provide explicit expressions for its ACF and PACF. To point out some aspects of this model we consider a special case ARIMA (0, 1, 1)(0, 1, 1)s model given by;

$$\nabla\nabla_s X_t = (1 - \theta B)(1 - \Theta B^s)e_t \ldots\ldots \text{2.4.2}$$

for invertibility, we require the roots of $(1 - \theta B)(1 - \Theta B^s)$ to satisfy the conditions

$|\theta|, |\Theta| < 1$. The above model may be written as,

$$w_t = (1 - B - B^s + B^{s+1})X_t = (1 - \theta B - \Theta B^s + \theta\Theta B^{s+1})e_t$$

and since the autocovariances of $w_t = \nabla\nabla_s X_t$ are

$$\sigma(k) = E(w_t\, w_{t-k})$$

$$= E\{(e_t - \theta e_{t-1} - \Theta e_{t-s} + \theta\Theta e_{t-s-1}) \times (e_{t-k} - \theta e_{t-k-1} - \Theta e_{t-s-k} + \theta\Theta e_{t-s-1-k})\}$$

We have,

$$\sigma(0) = (1 + \theta^2)(1 + \Theta^2)\sigma_e^2$$

$$\sigma(1) = -\theta(1 + \Theta^2)\sigma_e^2$$

$$\sigma(s-1) = \theta\Theta\sigma_e^2$$

$$\left.\begin{aligned}\sigma(s) &= -\Theta(1+\theta^2)\sigma_e^2 \\ \sigma(s+1) &= \theta\Theta\sigma_e^2\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad 2.4.3$$

and $\sigma(k) = 0$ for all other k's, hence the ACF is,

$$\rho(k) = \begin{cases} -\dfrac{\theta}{1+\theta^2} & k = 1 \\ \dfrac{\theta\Theta}{(1+\theta^2)(1+\Theta^2)} & k = s-1, s+1 \\ \dfrac{-\Theta}{1+\Theta^2} & k = s \\ 0 & Otherwise \end{cases} \quad \ldots\ldots\ldots\ldots \quad 2.4.3$$

we note that the SARIMA model is just an extension of the ARIMA model and modelling in both cases follows similar procedures. That is, model identification, estimation and checking procedures are almost similar in the two cases. We now present these procedures based on a SARIMA model.

2.5: **Model Identification**

Several methods have been suggested but we confine ourselves to only two, namely:

(a) Use of ACF and PACF and

(b) Akaikes Information Criterion (AIC)

(a) **Use of ACF and PACF**

This method is widely used and has strongly been recommended by Chatfield (1975), Abraham and Ledolter (1983), Mills (1990) and Kendall (1983) among others.

Although the time plot may indicate the seasonal and trend patterns, it is not possible to decide the order p, P, q, & Q from it. According to the above authors and particularly Chatfield (1975), the seasonal time series should be differenced approximately (d times for trend removal and D times to remove seasonality) to reduce it to stationarity. Then the values of p, P, q, and Q are determined by looking at the patterns of the SACF and SPACF of the differenced series. For p and q we examine the first few values of r_k (sample ACF). While for P and Q we examine the values of r_k at k = s, 2s where s is the period.

Matching the patterns in the sample ACF and sample PACF with the theoretical patterns of the known models can help identify these orders. Table 2.5.1 gives a summary of the known patterns of ACF and PACF.

Note: For the seasonal components we also consider (P, D, Q) and observe the behaviour of SACF and SPACF at seasonal lags. The tabulated patterns hold for seasonal patterns at $k = s, 2s$ - - -.

According to Mills (1990) variance of SACF for the ARIMA (0, 1, 1) (0, 1, 1)s model is given by,

$$\text{Var (r)} = \frac{1 + 2(r_1^{\,2} + r^{2}_{s-1} + r^{2}_{s} + r^{2}_{s+1})}{n} \qquad k > s + 1 \ldots\ldots\ldots 2.5.1$$

Using this result and the known form of ACF enables ARIMA (0, 1, 1)(0, 1, 1)s model to be identified.

A rough guide to interpreting a correlogram (Sample ACF) or SPACF is using the limits $\pm 2/\sqrt{n}$ to assess individual r_k for significant departure from zero. Any value of r_k, which falls outside, $\pm 2/\sqrt{n}$ is taken to be significantly different from zero. Diggle (1990) notes that the overall pattern of r_k is more important than the significance of individual r_k.

Table 2.5.1: **Summary of the known Patterns of ACF and PACF.**

MODEL	ACF	PACF
(1, d, 0)	Exponential or oscillatory decay	$\phi_{kk} = 0 \text{ for } k > 1$
(2, d, 0)	Exponential or sine wave decay	$\phi_{kk} = 0 \text{ for } k > 2$
(p, d, 0)	Exponential and or sine wave decay	$\phi_{kk} = 0 \text{ for } k >. p$
(0, d, 1)	$\rho_k = 0 \text{ for } k > 1$	Dominated by damped exponential.
(0, d, 2)	$\rho_k = 0 \text{ for } k > 2$	Dominated by damped exponential or sine wave
(0, d, q)	$\rho_k = 0 \text{ for } k > q$	Dominated by linear combination of damped exponential and or sine waves.
(1, d, 1)	Tails off. Exponential decay after lag 1	Tails off. Dominated by exponential decay from lag 1.
(p, d, q)	Tails off after q-p lags exponential and or sine wave decay after q-p lags	Tails off after p-q lags. Dominated by damped exponentials and or sine waves after p-q lags.

(b) **Akaike Information Criterion (AIC).**

Akaike (1973, 1974a) introduced an information criterion, which can be used for statistical model identification in a wide range of situations and is not restricted to time series. Assuming that a statistical model of m parameters is fitted to data, the AIC is defined by,

$$\text{AIC}(m) = -2\,\ell n\ [\text{maximised-likelihood}] + 2m \quad \ldots\ldots\ldots\ldots\ 2.5.2$$

For ARIMA model and n effective number of observations the log-likelihood function is, Wei (1990) section 7.7.

$$\ell nL = -\frac{n}{2}\ell n 2\pi\hat{\sigma}_e^{\,2} - \frac{1}{2\sigma_e^{\,2}} S(\phi,\mu,\theta) \quad \ldots\ldots\ldots\ldots\ldots\ldots 2.5.3$$

where $S(\phi, \mu, \theta)$ is the unconditional sum of squares function given by

$$S(\phi,\mu,\theta) = \sum_{t=-\infty}^{n} E(e_t / \phi,\mu,\theta,X)^2$$

and $E(e_t / \phi,\mu,\theta,X)$ is the unconditional expectation of e_t given ϕ,μ,θ and X. Maximising ℓnL with respect to $\phi,\mu,$ and $\sigma_e^{\,2}$ we get

$$\ell n\hat{L} = -\frac{n}{2}\ell n\sigma_e^{\,2} - \frac{n}{2}(1 + \ell n 2\pi). \quad \ldots\ldots\ldots\ldots\ldots 2.5.4$$

Since the second term is a constant the AIC criterion reduces to,

$$\text{AIC}(m) = \text{n}\,\ell n\hat{\sigma}_e^{\,2} + 2m \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.5.5$$

The optimal order of the model is chosen by the value of m, which is a function of p and q, so that AIC(m) is minimum. Both Priestley (1981) and Wei (1990) note that the AIC has become a standard tool in time series model fitting. Shibata (1976) has, however, shown that the AIC tends to overestimate the order of the autoregression. As a result Akaike has developed a Bayesian extension of the AIC called BIC. Similar to BIC is Schwartz's Bayesian Criterion (SBC) suggested by Schwartz (1978) and given by

$$\text{SBC}(m) = n\,\ell n\hat{\sigma}_e^{\,2} + m\ell n\ n \qquad \text{.....2.5.6}$$

Where $\hat{\sigma}_e^2$ is the maximum likelihood estimate of σ_e^2. The computation of the AIC, BIC and SBC is available in many time series programmes which makes it easy to utilise them.

Apart from the visual inspection of the SACF, Priestley (1981) recommends the use of one of the discussed criterions because:

- The behaviour of the SACF does not always accurately reflect the behaviour of the true ACF.
- This inspection requires experience and past knowledge.

2.6: **Model Estimation**

Once the form of the model has been specified it is necessary to estimate its parameters . Let $\{X_t\}$ represent the original series and let $\boldsymbol{w} = (w_1, w_2 \ldots\ldots\ldots, w_n)$ be the vector of $n = N\text{-}d$ stationary differences. As noted earlier estimation of parameters

in SARIMA model is the same as for general ARIMA models. We now present several methods of model estimation in ARIMA models.

a) **Maximum Likelihood Estimates**

The ARIMA model can be written as:

$$e_t = \theta_1 e_{t-1} + \ldots\ldots\ldots + \theta_q e_{t-q} + w_t - \phi_1 w_{t-1} - \ldots\ldots\ldots - \phi_p w_{t-p} \quad \ldots\ldots 2.6.1$$

The joint density of $\boldsymbol{e_t} = (e_1\ e_2\ \ldots\ldots\ldots e_n)$ is given by,

$$P(e/\phi,\theta,\sigma_e^{\,2}) = (2\pi\sigma_e^{\,2})^{-n/2} \exp\left[-\frac{1}{2\sigma_e^{\,2}}\sum_{t=1}^{n} e_t^{\,2}\right] \quad \ldots\ldots 2.6.2$$

where $\phi = (\phi_1, \phi_2, \ldots\ldots, \phi_p)^1$,

$\theta = (\theta_1, \theta_2, \ldots\ldots\ldots \theta q)^1$ and $\sigma_e^{\,2}$ are parameters to be estimated.

According to Newbold(1974) the density function of w can be written in the form

$$L(\phi,\theta,\sigma_e^{\,2}/w) = g_1(\phi,\theta,\sigma_e^{\,2}) \exp\left[-\frac{1}{2\sigma_e^{\,2}} S(\phi,\theta)\right] \quad \ldots\ldots 2.6.3$$

where g_1 is a function of $(\phi,\theta,\sigma_e^{\,2})$ and $S(\phi,\theta) = \sum_{t=1-p-q}^{n} E(u_t/\varpi)^2$ .Here $\varpi = (w,\phi,\theta,\sigma_e^{\,2})$

and hence $\boldsymbol{E(u_t/\varpi)}$ is the conditional expectation of u_t given w, ϕ, θ and $\sigma_e^{\,2}$ and

$$u_t = \begin{cases} e_t & t = 1,2---n \\ g_2(e^*, w^*) & t \le 0 \end{cases}$$

g_2 being a linear function of the initial values ,$e^* = (e_{1-q}, \ldots\ldots\ldots e_{-1}, e_0)$' and $w^* = (w_{1-p}, \ldots\ldots w_{-1}, w_0)$' needed for evaluation of e in (2.6.1) above. The functions g_1 and g_2 depend on the particular model. We now illustrate the derivation of the exact

likelihood function for three important models namely AR(1), MA(1) and ARMA (1, 1)

i) Consider the AR(1) process given by,

$$e_t = (1 - \phi B)w_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 2.6.4$$

$$Or \qquad w_t = \phi w_{t-1} + e_t$$

Where $|\phi| < 1$ e_t are *i.i.d* $N(0, \sigma_e^2)$. Rewriting the process in the moving average representation we have:

$$w_t = (1 - \phi B)^{-1} e_t = \sum_{j=0}^{\infty} \phi^j e_{t-j} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad 2.6.5$$

Clearly $w_t \sim N[0, \sigma^2{}_e /(1-\phi^2)]$. However $w_t s$ are highly correlated. To derive the joint probability density function $P(w_1, w_2, - - - w_n)$ of w and hence the likelihood function for the parameters, we consider

$$a_1 = \sum_{j=0}^{\infty} \phi^j e_{1-j} = w_1$$

$$e_2 = w_2 - \phi w_1,$$

$$e_3 = w_3 - \phi w_2$$

-

-

$$e_n = w_n - \phi w_{n-1} \quad \ldots\ldots\ldots\ldots\ldots \quad 2.6.6$$

Note that $a_1 \sim N[0, \sigma^2{}_e /(1-\phi^2)]$, $e_t \sim N(0, \sigma_e^2)$ for $2 \leq t \leq n$ and they are all independent for each other. Hence the joint probability density of $(a_1, e_2, e_3 - - - e_n)$ is

$$P(a_1, e_2, ---, e_n) = \left[\frac{(1-\phi^2)}{2\pi\sigma_e^2}\right]^{\frac{1}{2}} \exp\left[\frac{-a_1^2(1-\phi^2)}{2\sigma_e^2}\right] \times \left[\frac{1}{2\pi\sigma_e^2}\right]^{\frac{n-1}{2}} \exp\left[-\frac{1}{2\sigma_e{}^2}\sum_{t=2}^{n} e_t{}^2\right] \quad ..2.6.7$$

$$\ell nL(\phi,\mu/w) = -\frac{n}{2}\ell n2\pi + \frac{1}{2}\ell n(1-\phi^2) - \frac{n}{2}\ell n\sigma_e^{\,2} - \frac{S(\phi,\mu)}{2\sigma_e^{\,2}} \quad \ldots\ldots\ldots\ldots 2.6.9$$

where $S(\phi,\mu) = w_1^{\,2}(1-\phi^2) + \sum_{t=2}^{n}(w_t - \phi w_{t-1})^2$ is the sum of squares term that is a function of ϕ and μ only.

(ii) For MA(1)

$$e_t = \theta\, e_{t-1} + w_t$$

An argument similar to the one presented above gives,

$$L(\theta,\sigma_e^{\,2}/w) = \left[\frac{1}{2\pi\sigma_e^{\,2}}\right]^{n/2}\left[\frac{1-\theta^2}{1-\theta^{2(n+1)}}\right]^{1/2}\exp\left[-\frac{1}{2\sigma_e^{\,2}}\sum_{t=0}^{n}E(e_t/w)^2\right] \quad \ldots\ldots 2.6.10$$

(iii) A detailed derivation of exact likelihood function for ARMA(1,1) process is given by Abraham and Ledolter (1983, Appendix 5.).The likelihood function is given by,

$$L(\phi,\theta,\sigma_e^{\,2}) = (2\pi\sigma_e^{\,2})^{-n/2}\left|z'z\right|^{-1/2}\exp\left[-\frac{1}{2\sigma_e^{\,2}}S(\phi,\theta)\right] \quad \ldots\ldots\ldots\ldots\ldots 2.6.11$$

where $\left|z'z\right| = \dfrac{(1-\theta^2)(1-\phi^2)+(1-\theta^{2n})(\theta-\phi^2)}{(1-\theta^2)(1-\phi^2)}$ and

$$S(\phi,\theta) = E(e_0/w)^2 + \frac{1-\phi^2}{(\theta-\phi)^2}\left[E(w_0/w) - E(e_0/w)\right]^2 + \sum_{t=1}^{n}E(e_t/w)^2 \quad \ldots\ldots 2.6.12$$

These special cases illustrate two major difficulties with maximum likelihood estimation approach. These are:

(a) The presence of the function $g_1(\phi, \theta\ \sigma^2)$ in the likelihood function makes the maximisation difficult.

(b) This approach requires the calculation of the conditional expectations $E(u^*/w)$ which in turn requires the calculation of $E(e_t/w)$ and $E(w_t/w)$ which is not easy.

Wei (1990) notes that the exact closed form of the likelihood function of a general ARMA model is complicated. Interested readers are advised to see Newbold (1974), Ansley (1979) among others for additional references. Due to this problem in exact likelihood estimation, non-linear estimation has been suggested by Wei (1990) and Abraham and Ledolter (1983) among others.

(b) Non-Linear Estimation

Non-linear least squares procedure involves an iterative search technique. This technique is called for because the MA factor in an ARMA model which is a function of e_t has non-linear parameters. We can therefore not use the method of ordinary least squares estimation which works with pure AR processes.

The least squares estimates (linear or non-linear) can always be calculated iteratively as follows:

Let $Y_t = f(X_t, \alpha) + e_t \qquad t = 1, 2, \ldots n$

Where $X_t = (X_{t1}, X_{t2}, \ldots X_{tp})^1$ is a set of independent variables corresponding to the observations. $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)^1$ is a vector of parameters .Let $Y = (Y_1, Y_2, \ldots, Y_n)'$ and $f(\alpha) = [f(X_1, \alpha), f(X_2, \alpha) \ldots f(X_n, \alpha)]'$

Step 1

Given any vector of initial guess values $\hat{\alpha}_0$. Compute the residual $\hat{e} = (Y - \hat{Y})$ and the residual sum of squares

$$S(\hat{\alpha}_0) = \hat{e}'\hat{e} = (Y - \hat{Y})'(Y - \hat{Y}), \ldots\ldots\ldots 2.6.13$$

where $\hat{Y} = f(\hat{\alpha}_0)$ is a vector of predicted values obtained by replacing the unknown parameters by initial guess values. Next we approximate $f(X_t \hat{\alpha}_0)$ with the first order Taylor series expansion about the initial value $\hat{\alpha}_0$ .That is $f(\alpha) = f(\hat{\alpha}_0) + \overline{X}_{\hat{\alpha}_0} \delta$ here $\delta = \alpha - \hat{\alpha}_0$ and $\overline{X}_{\hat{\alpha}_0} = \left((\chi_{tj})\right)$ is the n×p matrix of partial derivatives at $\hat{\alpha}_0$ in the above linear approximation. That is:

$$\chi_{tj} = \left.\frac{\partial f(X_t, \alpha)}{\partial \alpha_j}\right|_{\alpha = \hat{\alpha}_0} \qquad t = 1, 2 \ldots n, \; j = 1, 2 \ldots\ldots p \ldots\ldots\ldots\ldots .2.6.14$$

we then calculate

$$\delta = (\overline{X}'_{\hat{\alpha}_0} \overline{X}_{\hat{\alpha}_0})^{-1} \overline{X}_{\hat{\alpha}_0} \hat{e} = (\delta_1, \delta_2, \ldots \delta_p)' \ldots\ldots\ldots\ldots\ldots\ldots 2.6.15$$

Step 2

From (2.6.15) we obtain a modified estimator of $\hat{\alpha}$ as $\hat{\alpha}_1 = \hat{\alpha}_0 + \hat{\delta}$ and corresponding residual sum of squares $S(\hat{\alpha}_1)$. We note that δ_i in δ represents the change in the parameters values. For linear models step 2 gives the final least squares estimates but for non-linear model, it only leads to new initial values for further iterations.

For a given general ARMA (p, q), we can use the non-linear least squares procedure to find least squares estimates that minimise the error sum of squares $S(\phi, \theta)$ defined earlier in (2.6.11) for ARMA (11). The non-linear least squares routine starts with initial guess values of parameters. It monitors these values in the direction of the smaller sum of squares and updates the initial guess values. The iterations continue until some pre-specified convergence criteria are reached, and the relative change in the parameter values is less than a specified level or the number of iterations is greater than a certain number.

Many search algorithms have been developed and are available in computer programs e.g. Marquardt (1963) is present in SPSS – Package. Minitab programme utilised similar algorithms in estimation of parameters in Box – Jenkins ARIMA models. The availability of search algorithms has made the process of computation very easy and thus the method of exact likelihood function is rarely used. The parameters of a SARIMA model are obtained using similar search procedures.

2.7. **Model checking**

After the parameters in the model have been estimated it is necessary to check whether the model assumptions are satisfied. Model checking or diagnostic checking relies heavily on the analysis of the residuals $\{\hat{e}_t\}$ which are defined by:

Residual = Observation – fitted value.

The basic assumption in ARIMA models in that its errors are uncorrelated random variables with mean zero and constant variance. We therefore expect $\hat{e}_t$ to have similar characteristics as e_t. That is,

- The mean of $\hat{e}_t$ should be close to zero
- Variance $\hat{e}_t \cong$ constant and
- Autocorelations of $\hat{e}_t$ should be negligible.

Many methods of diagnostic checking have been advanced by Diggle (1990), who suggests, correlogram analysis , cumulative periodogram, trial overfitting and spectral checking, but the methods recommended by most time series analysts are,

- Correlogram analysis
- Box – pierce test for serially correlated residuals.

(a) **Correlogram Analysis**

This method has been recommended by Chatfield (1975), Mills (1990), Abraham and Ledolter (1983) among others. This method involves "just" looking at the first few

values of r_k (for residuals) particularly at lags 1, 2 and the first seasonal lag (if any) and see if they are significantly different from zero. That is, check whether they fall outside the limits $\pm 2/\sqrt{n}$ and if yes, modify the model. If only one (or two) values of r_k are significant at lags which have no obvious physical meaning there would be no enough evidence to reject the model (Chatfield, 1975).

(b) Box-Pierce test

Box – Pierce (1970) came up with a test for serially correlated residuals in ARMA (p,q). They showed that if the appropriate model for the differenced series is ARMA (p, q) process then the statistic

$$Q = n\sum_{k=1}^{m} \hat{r}_k^{\,2} \sim \chi^2_{(m-p-q)} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.7.1$$

Where n is the number of terms in the differenced series. Chatfield (1975) notes that m is typically chosen in the range 15 to 30. For small samples this statistic gives poor results and a better approximation has been given by Ljung and Box (1978) to be:

$$Q = n(n+2)\sum_{k=1}^{m} (n-k)^{-1} \hat{r}_k^{\,2} \sim \chi^2_{(m-p-q)} \quad \ldots\ldots\ldots\ldots\ldots 2.7.2$$

For seasonal ARMA or SARIMA (p, d, q) (P, D, Q)s the above statistic is still valid but it is now distributed as X^2 with $m - q - p - P - Q$ degrees of freedom. That is:

$$Q = n(n+2)\sum_{k=1}^{m} (n-k)^{-1} \hat{r}_k^{\,2} \sim \chi^2_{(m\ p\ q\ P\ Q)} \quad \ldots\ldots\ldots\ldots 2.7.3$$

If the calculated Q statistic exceeds the tabulated $\chi^2_{\alpha,(m-p-q-P-Q)}$,then the adequacy of the fitted model would be questioned. Mills(1990) and Chatfield (1975) have noted that the modified statistic above still gives unsatisfactory results in that it may lead to acceptance even in the presence of severe model mis-specification. It is therefore advisable to combine several methods in diagnostic checking.

2.8: **Forecasting**

Once the model has been found adequate, we may proceed to predict /forecast future values. To illustrate how forecasting is done, we consider one SARIMA model because our research is based on seasonal time series. Consider a case where

ARIMA(0 ,1 ,1)(0 ,1, 1)s has been found adequate .

This mode is given by,

$$\nabla\nabla_s X_t = (1-\theta B)(1-\Theta B^s)e_t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.8.1$$

$$\Rightarrow X_t = X_{t-1} + X_{t-s} - X_{t-s-1} + e_t - \Theta e_{t-s} - \theta e_{t-1} + \Theta\theta e_{t-s-1}$$

For forecasting origin, say $t = n.$, forecasts can be calculated directly from the difference equation,

$$X_{n+k} = X_{n+k-1} + X_{n+k-s} - X_{n+k-s-1} + e_{n+k} - \Theta e_{n+k-s} - \theta e_{n+k-1} + \Theta\theta e_{n+k-s-1} \quad \ldots\ldots 2.8.2$$

so that the forecast at $t = n$ for k–steps ahead is given by,

$$\hat{X}(n,k) = E[(X_{n+k-1} + X_{n+k-s} - X_{n+k-s-1} + e_{n+k} - \Theta e_{n+k-s} - \theta e_{n+k-1} + \Theta\theta e_{n+k-s-1}) / X_{n,1}].$$

where $X_{n,1} = X_n, X_{n-1}, \ldots\ldots\ldots, X_1$;and

$$E(X_{n+j}) = \begin{cases} X_{n+j} & j \le 0 \\ \hat{X}(n,k) & j > 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.8.5$$

$$E(e_{n+j}) = \begin{cases} e_{n+j} & j \le 0 \\ 0 & j > 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2.8.6$$

In general, for ARIMA(0 ,1 ,1)(0, 1 ,1)s ,the eventual forecast function is,

$$(1-B)(1-B^s)\hat{X}(n,k) = 0 \qquad\qquad k > s+1.$$

A similar process is used for a general SARIMA model.

CHAPTER THREE

ESTIMATION OF A BLOCK OF MISSING OBSERVATIONS IN A MONTHLY TIME SERIES GIVEN QUARTERLY TOTALS

3.1: Introduction

In the foregoing chapters we discussed the problem of modelling a given time series. In so doing we assumed that, all N observations are given in the appropriate order. This assumption may at times be violated when either a set of observations is missing or is known to have been observed erroneously. In this chapter we discuss the methods which may be used to impute (fill-in) a block of missing values in a seasonal time series in order for modelling to be carried out. In particular, we look into the case where a block of observations is missing within a given series. We first show how extrapolated values from two different models can be combined to produce better estimates and then show how the quarterly information (Already Known) may be Incorporated into the final estimates. Later we propose a direct method of imputation, which requires no pre-modelling of the known segments of the series

3.2: Imputation Methods

Imputation is a general and flexible method for handling missing data problems. In this case "Gaps" in the data are "patched up" by explicitly substituting a reasonable

replacement value whose characteristic are similar to those of observed values. This method is not without pitfalls. In the words of Dempster and Rubin(1983)

> "The idea of imputation is both seductive and dangerous, it is seductive because it can hurl the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to a real and imputed data have substantial biases."

In this chapter we describe some imputation methods which may be used in time series. We also review what has been done in the past under those imputation methods. We note that all the methods of imputation fall under either direct or indirect methods. Direct methods are straightforward and require little prior knowledge about the data's nature. These methods include mean imputation, cold deck imputation and hot deck imputation. The indirect methods involves predictions of values from a model fitted to the known data. Indirect methods include regression imputation and stochastic regression imputation among others.

Almost all the methods, which have been used in imputation of time series data, fall under regression imputation. We therefore describe this particular method before we suggest and discuss one of the direct methods which may be used to impute missing data.

3.3: **Regression Imputation**

This method involves replacing missing observations by predicted values from a model. It is basically a modelling technique where the observed series is modelled and then used to extrapolate the missing values. This method has widely been used in time

series. Before we discuss two methods under regression imputation (which we later use), we present an overview of what has been done under this method in the past.

Abraham (1991) discussed missing data problem in a non-linear time series. Abraham and Thavaneswaren (1991) did the same but looked at a linear time series model as a special case. Unfortunately, they dealt with only two missing observations. Others who have used state-dependent models and Kalman filters in estimation of missing observations are Harvey (1989), Jones (1980), Little and Rubin (1987) who also suggested the use of "EM" algorithm in pure AR models and Cartwright (1985) who, unlike others dealt with more than two observations. Although Cartwright has shown that Kalman filter recurssive techniques may be used to estimate a block of missing observations, he notes that unacceptable forecasts may be generated if appropriate restrictions are not imposed. Even with three types of restrictions, he generates forecasts with large deviations from the actual values See his paper in Anderson (1985 Ed.)

Luceño (1997) extends Ljung's (1989) method for estimating missing observations to a vector case. He avoids the use of Kalman filter algorithms, but does not solve the problem of more than two observations in a univariate case. He in particular tackles case where m observations are missing each in its own series, thus amounting to estimating a single observation in a series. We note that apart from Kalman filter technique, which has its own setbacks as noted by Cartwright all other methods mentioned above deal with at most two missing values. Our problem is therefore still unsolved.

Abraham (1981) proposed a procedure to interpolate the adjacent missing values on the basis of the known segments of an ARIMA (p, d, q). Borrowing an idea from him and considering that we have m = 12 consecutive missing observations partitioning the whole series into two segments, we may use the known segments to estimate the values. We now describe two approaches under regression imputation, which may be used to estimate the missing values.

(a) **Eventual forecasts from a model**

This involves fitting an appropriate model to each set of the data (known segments) and extrapolating the missing values from eventual forecasts. The data set following the block of missing values is dealt with in a reverse way so that the eventual "forecasts " are equivalent to "backcasts". In this case we obtain two estimates for the same observation which can be combined to give more accurate estimates

We have already discussed the methods of modelling a seasonal time series in the previous chapter and we briefly explained how eventual forecasts may be generated. In the chapter that follows we show empirically how this may be done.

(b) **Application of smoothing algorithms. (Holt - Winters forecasting procedures)**

Although modelling works well, it is sometimes hard to come up with an adequate model to generate reliable eventual forecasts. We can avoid modelling and apply appropriate smoothing algorithms, which may give reliable results. We now show how algorithms can be used. Specifically we describe Holt-Winters forecasting procedures because they deal with linear seasonal time series.

Suppose that we wish to estimate the level of a series of observations. The simplest way to do this is to use the sample mean ($\bar{x}$). However, if the purpose of estimating the level is to use this as the basis for forecasting future observations, it is more appealing to put more weight on the most recent observations. Thus the estimates of the current level of the series is taken to be

$$\mu_n = \sum_{t=0}^{n-1} \omega_t X_{n-t} \quad \ldots\ldots\ldots 3.3.1.$$

Where $\{\omega_t\}$ is a set of weights which sum to unity. This estimate is then taken to be the forecast of future observation. Following Abraham and Ledolter (1983), the *k*-step-ahead forecast made at *n* can be calculated from

$$X(n,k) = c\sum_{t=0}^{n-1} \omega^t X_{n-t} \quad \ldots\ldots\ldots 3.3.2$$

where ω is a discount coefficient such that, $|\omega| < 1$ and $c = \dfrac{1-\omega}{1-\omega^n}$ in a factor needed normalises the sum of the weights to 1, That is, since

$\sum_{t=0}^{n-1} \omega^t = \dfrac{1-\omega^n}{1-\omega}$ it follows that,

$$c\sum_{t=0}^{n-1} \omega^t = 1$$

If *n* is large, then the term.ω^n in C goes to zero and exponentially weighted 1-step ahead forecast can be written

$$\hat{X}(n,1) = (1-\omega)\sum_{j\geq 0}\omega^j X_{n-j} \quad \ldots\ldots\ldots 3.3.3$$

The coefficient $\alpha = 1 - \omega$ is called the smoothing constant and is usually chosen between 0.05 and 0.30. See Abraham and Ledolter (1983) section 3.3. Equation (3.3.3) above can be written as

$$\hat{X}(n,1) = \alpha[X_n + (1-\alpha)X_{n-1} + (1-\alpha)^2 X_{n-2} + - - -] \quad \ldots\ldots\ldots 3.3.4$$

$$= \alpha\sum_{j=0}^{\infty}(1-\alpha)^j X_{n-j}$$

$$= \alpha[1 + (1-\alpha)B + (1-\alpha)^2 B^2 + - - - (1-\alpha)^j B^j + - - -]X_n \quad \ldots 3.3.4$$

But

$1+(1-\alpha)B + (1-\alpha)^2B^2 + - - - = [1 - (1-\alpha)B]^{-1}$ so that (3.3.4) becomes:

$$\hat{X}(n,1) = \left[\frac{1}{1-(1-\alpha)B}\right]\alpha X_n$$

$$\Rightarrow \quad [1-(1-\alpha)B]\hat{X}(n,1) = \alpha X_n$$

$$\Rightarrow \quad \hat{X}(n,1) = \alpha X_n + ((1-\alpha)\hat{X}(n-1,1) \quad \ldots\ldots\ldots 3.3.5$$

$\Rightarrow$ From (3.3.5) $\hat{X}(n,1)$ is a weighted average of the current observation X_n and the previous forecast $\hat{X}(n-1,1)$. We can re-write it as

$$\hat{X}(n,1) = \hat{X}(n-1,1) + \alpha[X_n - \hat{X}(n,1,1)]$$

$$= \hat{X}(n-1,1) + \alpha e_n \quad \ldots\ldots\ldots 3.3.6$$

The recurrence equation in (3.3.5) can be used to update the smoothed statistic any time t. This algorithm due to c.c. Holt is called exponential smoothing.

The smoothing constant α smoothes the trend and seasonal components equally despite the fact that the seasonal component is normally more stable than the trend component. Having noted this, Winter (1960) modified Holts' method to come up with what is today known as Holt-Winters procedures

Holt-Winters procedures utilise three smoothing constants namely α (alpha), γ (Gamma) and δ (Delta) for updating the level, trend and seasonal components respectively . When a new observation becomes available the level μ_n the trend T_n, and the seasonal ψ_n components are updated. We now describe this procedure

Let $\{X_t\}$ be a time series with additive seasonal variation, then the observed series can be written as

$$X_{n+t} = T_{n+t} + \psi_{n+t} + e_{n+t} \quad \ldots\ldots\ldots\ldots\ldots 3.3.7$$

Where, we assume a linear trend component.

$T_{n+t} = \mu_{n+t} = \mu_n + \beta\ t$ and where the S seasonal factors

$\psi_i = \psi_{i+s} = \psi_{i+2s} = - - -$

(for i $=1,2,\text{----},s$) are restricted to add to zero. That is $\sum_{i=1}^{s} \psi_i = 0$ .The updating and error –correction equations are ,see Mills (1990).

(i) For the level,

$$\hat{\mu}_n = \alpha(x_n - \hat{\psi}_{n-s}) + (1-\alpha)(\hat{\mu}_{n-1} + \hat{\beta}_{n-1})$$

$$= \hat{\mu}_{n-1} + \hat{\beta}_{n-1} + \alpha\hat{e}_n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.3.8$$

(ii) For the trend or slope,

$$\hat{\beta}_n = \gamma(\hat{\mu}_n - \hat{\mu}_{n-1}) + (1-\gamma)\hat{\beta}_{n-1}$$

$$= \hat{\beta}_{n-1} + \alpha\gamma\hat{e}_n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.3.9$$

and

(iii) For seasonal component

$$\hat{\psi}_n = \delta(X_n - \hat{\mu}_n) + (1-\delta)\hat{\psi}_{n-s}$$

$$= \hat{\psi}_{n-s} + \delta(1-\alpha)\hat{e}_n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.3.10$$

Newbold (1988) showed that equations (3.3.7 - 3.3.10) are equivalent to seasonal ARIMA (0,1,1)(0,1,1) under some conditions.

If sufficient historical data are available we can determine the smoothing constants by minimising the sum of squares due to errors (SSE). That is we take the optimal values of α, γ and δ as the set which gives the smallest value of SSE where

$$SSE = \sum_{t=1}^{n}[X_t - \hat{X}(t-1,1)]^2$$

To obtain initial values of $\hat{\mu}_t, \hat{\beta}_t, \hat{\psi}_{j-s}(j = 1,2,---s)$. Abraham and Ledolter suggest the use of ordinary least squares in the regression model

$$X_t = \mu + \beta t + \sum_{i=1}^{s-1} Z_i(IND_{ti} - IND_{ts}) + e_t$$

where $IND_{ti} = \begin{cases} 1 & \text{if it is in seasonal period i} \\ 0 & \text{otherwise} \end{cases}$

is a seasonal indicator.

From the least squares estimates

$\hat{\mu}_o = \hat{\mu}, \hat{\beta}_o = \hat{\beta}, \psi_{j-s} = \hat{Z}_j$ for $(1 \leq j \leq s\text{-}1)$

and $\hat{\psi}_o = -\sum_{j=1}^{s-1} \hat{Z}_j$

Once these initial values are obtained we can calculate the forecasts of the future value X_{n+k} from time n and from the updated components. The forecast function is of the form

$$\left.\begin{aligned} \hat{X}(n,k) &= \hat{\mu}_n + \hat{\beta}_n k + \hat{\psi}_{n+k-s} \quad \text{for k = 1,2,---,s} \\ \hat{X}(n,k) &= \hat{\mu}_n + \hat{\beta}_n k + \hat{\psi}_{n+k-2s} \text{ for k= s+1, s+2 - - - 2s} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.3.12$$

and so on.

Note: If the seasonal variation is multiplicative, the recurrence equations change slightly as follows,

$X_n - \psi_{n-s}$ in (3.3.8) becomes X_n / ψ_{n-s}

$X_n - \mu_n$ in (3.3.10) becomes X_n / μ_n

And the forecast function in (3.3.12) becomes

$$\hat{X}(n,k) = (\hat{\mu}_n + \hat{\beta}_n k)\hat{\psi}_{n-s+k} \text{ for k =1, 2 -- -s}$$

An obvious weakness in this method as noted by Harvey (1989) is that each seasonal component is only updated after every S period and the deseasonalizing in (3.3.8) is

carried out using an estimate of the seasonal component, which is S periods outdated. Nevertheless, as Chatfield (1975) and others have argued, the method seems to work well in practice .

3.4: **Composite Estimates/Forecasts**

The two methods described in the foregoing sections give rise to the estimates from eventual forecasts for the same values. If such estimates are combined using a logical argument we come up with composite forecasts/estimates. This has been called composite imputation by Rubin and Little(1987). In this section we describe a procedure of carrying out composite imputations given estimates/forecasts from different procedures. Triants (1985), and Bates and Granger (1969) have suggested these methods. In their paper Bates and Granger (1969) showed that the combined forecasts are better than either of them on its own. The most widely used method is known as **Variance-Covariance Method** and runs as follows.

Let the variable being forecast be X, and let F_1 and F_2 be two unbiased forecasts for X. Then according to Bates and Granger(1969), F_1 and F_2 can be combined to produce new set of forecasts which is more accurate than either of its components. They assume that one wishes to minimize the variance of these forecasts errors. Now the forecasts made at a time $t - 1$ is

$$\left.\begin{aligned} X_t &= F_{1t} + e_{1t} \\ X_t &= F_{2t} + e_{2t} \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.4.1$$

Where e_1 and e_2 are forecast errors such that,

$E(e_1) = E(e_2) = 0$, $Var(e_1) = \sigma_1^2$, $Var(e_2) = \sigma_2^2$ and $Cov(e_1,e_2) = \sigma_{12}$. We need to determine the (constants) weights β_1 and β_2, which give the optimal combined forecast F_{ct}, where,

$$F_{ct} = \beta_1 F_{1t} + \beta_2 F_{2t} \quad \ldots\ldots\ldots\ldots\ldots\ldots 3.4.2$$

Now the forecast error of the combined forecast is

$$X_t - F_{ct} = X_t - \beta_1 (X_t - e_{1t}) - \beta_2 (X_t - e_{2t})$$

$$= (1 - \beta_1 - \beta_2) X_t + \beta_1 e_{1t} + \beta_2 e_{2t} \quad \ldots\ldots\ldots\ldots\ldots\ldots 3.4.3$$

For the combined forecast to be unbiased we must have, $E(X_t - F_{ct}) = 0$.

This is the true if $\beta_1 + \beta_2 = 1$ because $E(e_{1t}) = E(e_{2t}) = 0$

Now, the variance of forecast error is ,

$$\sigma_c^2 = E(X_t - F_{ct})^2$$

$$= E(1 - \beta_1 - \beta_2) X_t + \beta_1 e_{1t} + \beta_2 e_{2t})^2$$

$$= E(\beta_1 e_{1t} + \beta_2 e_{2t})^2 = E[(\beta_1 e_{1t} + (1 - \beta_1) e_{2t}]^2$$

$$= \beta^2{}_1 \sigma_1^2 + 2\beta_1 (1 - \beta_1) \sigma_{12} + (1 - \beta_1)^2 \sigma_2^2$$

$$\sigma_c^2 = \beta_1^2 (\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}) - 2\beta_1 (\sigma_2^2 - \sigma_{12}) + \sigma_2^2 \ldots\ldots\ldots\ldots 3.4.4$$

Minimizing this with respect to β_1 gives,

$$\hat{\beta}_1 = (\sigma_2^2 - \sigma_{12})/D \qquad \text{where } D = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

$$\text{and } \hat{\beta}_2 = 1 - \hat{\beta}_1 = (\sigma_1^2 - \sigma_{12})/D \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.4.5$$

The optimal linear combination has weights, which depends on the variances and covariances of the forecast errors. We note that, σ_1^2, σ_2^2 *and* σ_{12} are unknown and have to be estimated. With a sample of n observations the estimates are:

$$\hat{\sigma}_1^{\,2} = \frac{\sum e_{1t}^{\,2}}{n}, \hat{\sigma}_2^{\,2} = \frac{\sum e_{2t}^{\,2}}{n} \text{ and } \hat{\sigma}_{12} = \frac{\sum e_{1t} e_{2t}}{n} \quad \ldots\ldots\ldots\ldots 3.4.6$$

Here we have assumed that each of the forecast is unbiased and that the error variances are constant.

Granger and Ramanathan (1984) showed that the method described above can be estimated as the estimation of a regression equation,

$$X_t = \beta_1 F_{1t} + \beta_2 F_{2t} + v_t \quad \ldots\ldots\ldots\ldots 3.4.7$$

where v_t is a disturbance term with mean zero.

Since for unbiased forecasts $\beta_1 + \beta_2 = 1$ we can substitute β_1 in (3.4.7) above to get

$$X_t - F_{1t} = \beta_2 (F_{2t} - F_{1t}) + v_t \quad \ldots\ldots\ldots\ldots 3.4.8$$

Estimations using ordinary least squares produces,

$$\hat{\beta}_2 = \frac{\sum (X_t - F_{1t})(F_{2t} - F_{1t})}{\sum (F_{2t} - F_{1t})^2} \quad \ldots\ldots\ldots\ldots 3.4.9$$

But $X_t = F_{1t} + e_{1t}$, $X_t = F_{2t} + e_{2t}$ .Substituting these in (3.4.9) gives,

$$\hat{\beta}_2 = \frac{\sum e_{1t} (e_{1t} - e_{2t})}{\sum (e_{1t} - e_{2t})^2}$$

$$= \frac{\sum e_{1t}^{\ 2} - \sum e_{1t} e_{2t}}{\sum e_{1t}^{\ 2} + \sum e_{2t}^{\ 2} - 2\sum e_{1t} e_{2t}}$$

$$= \hat{\sigma}_1^{\ 2} - \hat{\sigma}_{12}^{\ 2} / \hat{D} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.4.10$$

Similarly

$$\hat{\beta}_1 = 1 - \hat{\beta}_2 = \hat{\sigma}_1^{\ 2} - \hat{\sigma}_{12}^{\ 2} / \hat{D} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3.4.11$$

Thus; To obtain the weights $\hat{\beta}_2$ and $\hat{\beta}_1$, we only need to regress X_t on F_{1t} and F_{2t}. Data on past forecast performance can be used to estimate $\hat{\beta}_2$ and $\hat{\beta}_1$ before using these weights to combine the eventual forecasts. This method works with more than two forecasts.

For biased estimates (forecasts), Grangers and Ramanathan suggests that we include an intercept term to give

$$X_t = \beta_0 + \beta_1 F_{1t} + \beta_2 F_{2t} + v_t \quad \ldots\ldots\ldots\ldots\ldots 3.4.12$$

For more details on combining forecasts see Holden and Thompson (1989), Chapter 3. Since the composite forecasts are better, we can proceed to fill in the gaps with this set of data especially if no other information may be incorporated into the final estimates of the missing observations.

3.5: **Adjusting Estimates Using Quarterly Information**

In a seasonal time series with period $S = 12$ we have four quarters each with three consecutive observations. Let X_{iT} be then i^{th} quarterly total then $X_{ij} = X_{iT} / 3$. $\{I=1,2,3,4,\ \ j=1,2,3\}$ is the j^{th} quarterly average in the i^{th} quarter. Let $\{Y_t\}$ be a series

of predicted values for $\{X_t\}$ from which we compute Y_{iT} and Y_{ij} as explained above. That is $Y_{ij} = Y_{iT}/3$ is the j^{th} forecast average corresponding to X_{ij} $\{i=1,2,3,4;\ j=1,2,3\}$

Then an appealing way of adjusting the forecast Y_j to include the information in X_{iT} is To use the quantity $X_{iT} - Y_{iT}$ as follows:

Distribute the deviation $X_{iT} - Y_{iT}$ to each Y_j $j = 1, 2, 3$ so that each receives a quantity proportional to it's size. That is,

$$\text{Adj.}\hat{Y}_j = \hat{Y}_j + \frac{\hat{Y}_j}{\sum \hat{Y}_j}(X_{iT} - Y_{iT})$$

$$\text{Or Adj.}\hat{Y}_j = \hat{Y}_j + \frac{\hat{Y}_j}{Y_{iT}}(X_{iT} - Y_{iT}) \quad \ldots\ldots\ldots\ldots 3.5.1$$

Note that $\sum_{j=1}^{3} Adj.\hat{Y}_j = X_{iT}$ as should be the case.

If the observed values are directly proportional to their predicted values the adjusted values equals the actual values. This method thus works well especially if an appropriate model had been fitted to the data.

Empirical work in the chapter that follows shows that the adjusted estimates are far much better than the unadjusted ones. We can therefore patch our gap with this data and proceed to carry out further analysis to the complete data set.

All what we have been discussing as from section 3.3 fall under regression imputation except the foregoing section which deals with quarterly information. We note that these methods are not direct and were it not for the advanced computer programs to carry out some of the necessary computations, it would be very hard to accomplish

the desired objective. We now suggest a direct method of imputing the missing values.

3.6: **Direct Method of Imputation**

Earlier we mentioned some direct methods which have been used to impute missing values in statistics. Although the methods are appealing and easy, most of them are used with the assumption that, the observations are independent. With time series, the observations are not independent and thus we can not directly use the methods. This calls for some adjustments to be made on the most appealing method as used in sample surveys. The method we wish to employ here is the Hot-deck imputation procedure but modified to take care of dependence. Instead of carrying out simple random sampling with replacement to the observed values we modify the method to run as follows

i) Form cells as the months so that all observations for December say, will be taken to be in one cell

ii) Identify the close neighbours on either side of missing value in each cell

iii) Impute the missing value from the nearest neighbours in that particular cell by taking their mean

In mathematical terms, let $X_{m,\ell}$ be the value for the m^{th} month $\{m = 1,2.....12\}$ coresponding to ℓ^{th} year $\{\ell = 1,2,.....,N/12\}$. Assuming that the twelve observations are missing in a row within the series, the Hot-deck estimate of $X_{m,\ell}$ is given by,

$$\hat{X}_{m,\ell} = \frac{\sum_{j=\ell-i} X_{m,j} + \sum_{j=\ell+i} X_{m,j}}{2I} \qquad m = 1,2,.....12,\ i = 1,2.....,I \quad ...3.6.1$$

where I is the number of observations taken on either side of $X_{m,\ell}$.

Note: Max. $I = \min(\ell - 1, N/2 - \ell)$.

This is simply the average of $2I$ values of the m^{th} month where I values are on either side of the missing value. This method gives reasonable results because

i. The values of the same months normally have the same seasonal index

ii. The trend component is weighted appropriately by taking I values above and below the missing observation. We also note that this method is very easy as it requires no modelling or complex smoothing algorithms.

To incorporate the information in the already known quarterly totals, we again adjust the imputed values as explained in section 3.5 .In the chapter that follows, we carry out an empirical study based on our suggestions.

CHAPTER FOUR

AN EMPIRICAL STUDY.

4.1: Introduction:

In this chapter we carry out an empirical study based on what we have discussed in the previous chapters. Having noted that tourism has been the second largest foreign exchange earner for Kenya, we found a need to analyze data from this industry. We in particular analyzed Kenya's world tourist data for the years 1971 to 1990 with twelve monthly values for the year 1984 missing. Our prime aim is to model this data which when well done greatly helps in the prediction of future observation after which appropriate decisions in planning may be made. Before we could fit an appropriate model to the tourist data we first had to estimate the twelve missing values for 1984 using the available information.

Table 1.1 (Appendix A1) gives the raw data in thousands, that is, the number of tourists from all over the world who visited Kenya between 1971 and 1990 with 1984 values missing, based on their arrival. Although 1984's monthly observations could not be traced, we were able to get the quarterly totals for this year from statistical abstracts. See Table 1.2 (Appendix A1).

Since the block of missing observations partition our series into two parts, we deal with each segment separately in trying to determine the salient features of the series. We will take the segment covering 1971 - 1983 to be series 1, and the segment

covering 1990 to 1985 to be series 2 but for convenience, we will discuss results for series 1 under (a) and results for series 2 under (b). We also deal with the second segment in a reverse order.

In figure 4.1 and 4.2 below, we give the time plots for the above two segment.

Fig 4.1:Time plot for series 1

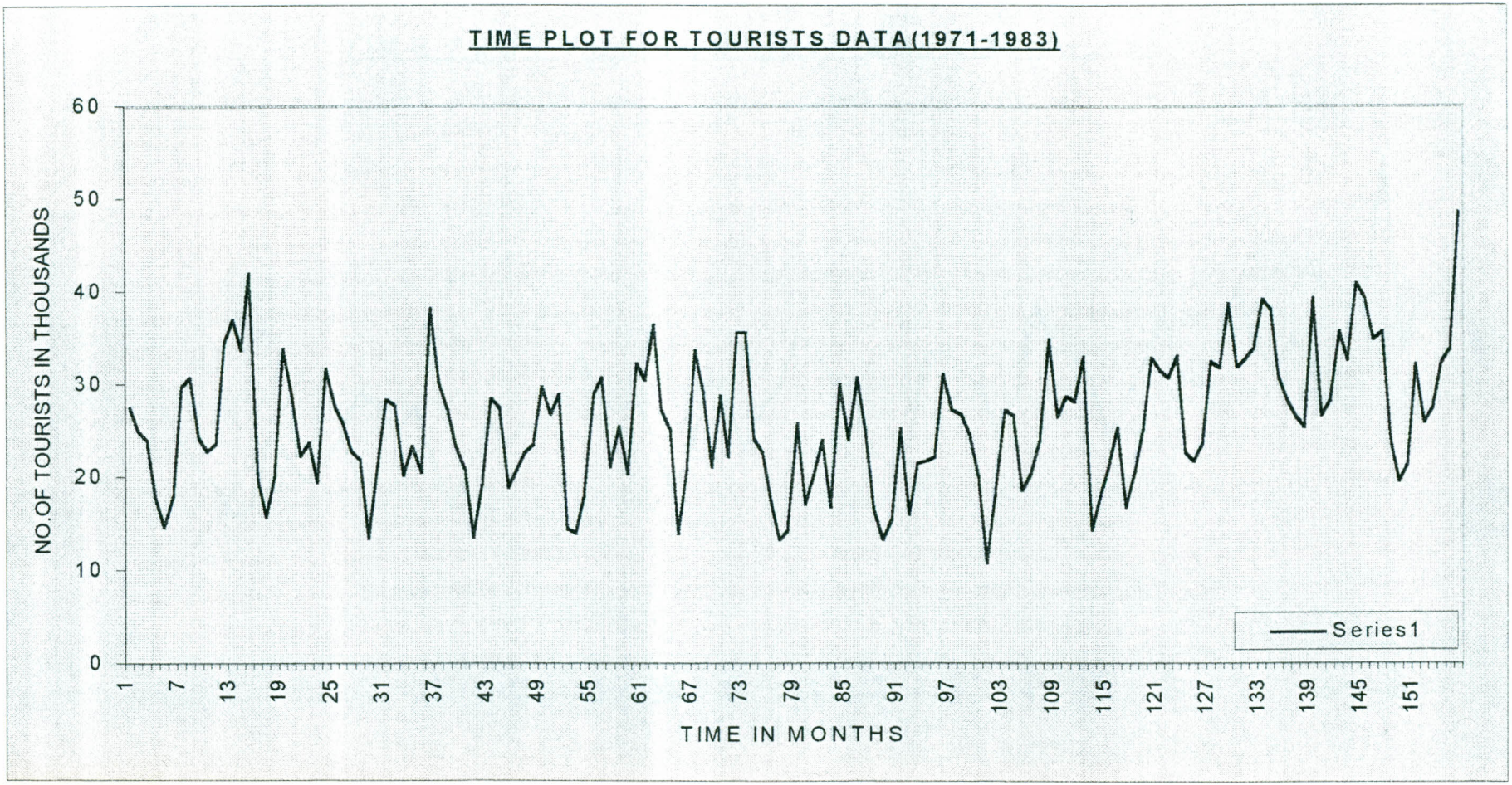
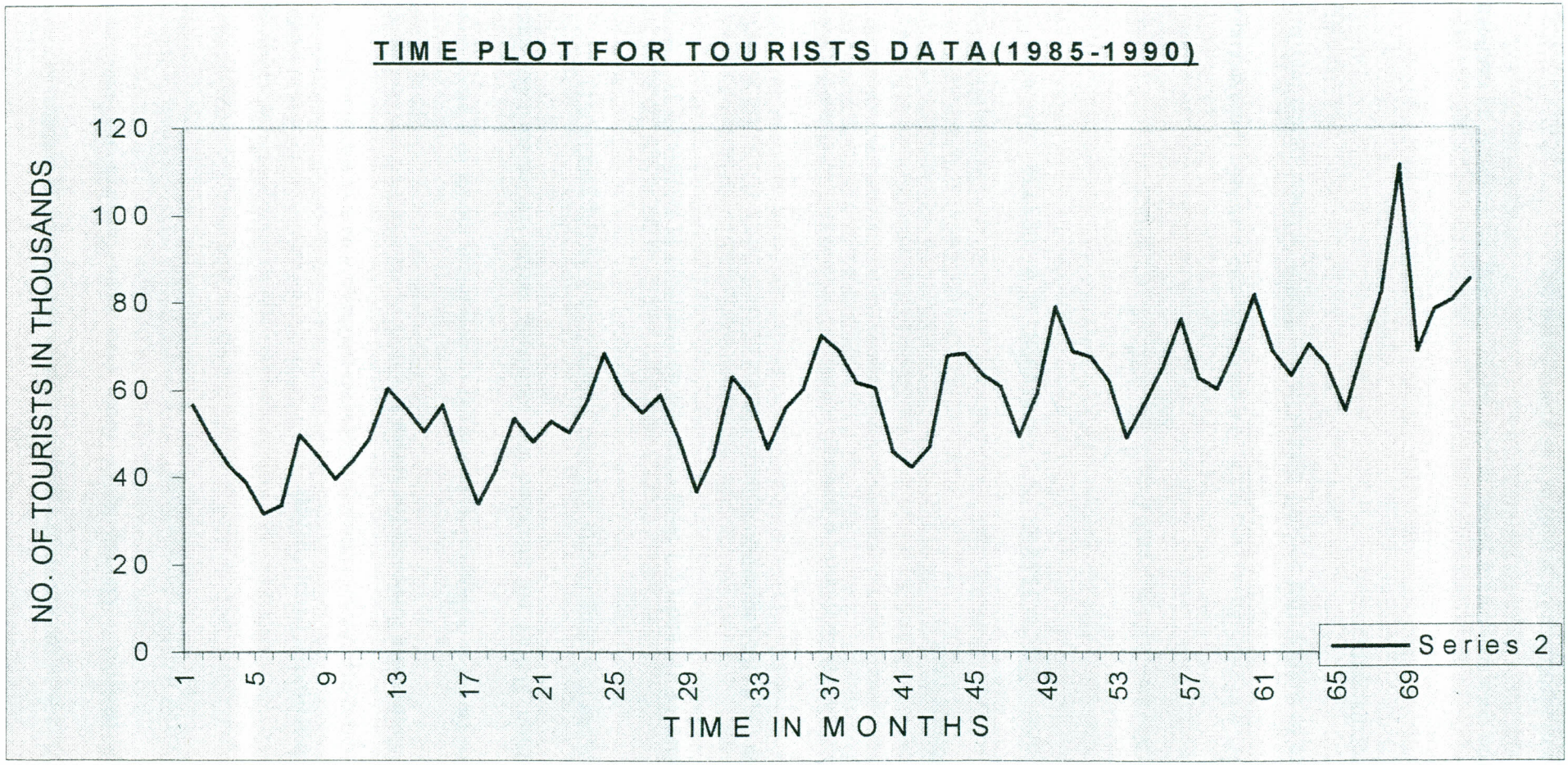


Fig 4.2: Time plot for series 2



A casual look at the above two graphs suggests that the series contains some trend and seasonal patterns . Having noted these two main features we now embark on estimation of missing observations

4.2: **Composite imputation**

4.2(a): **Composite forecasts for series 1**

As noted earlier it is illogical to fit a linear or non-linear model to a given data with no justification. Before we modelled this particular series we tested it for linearity and obtained the following results .see also Tables 2.1a-2.3a (Appendix A2).

The series $\{\hat{e}_t\}$ and $(\hat{\varepsilon}_t)$ were obtained as discussed in section (2.3). On regressing $\{\hat{e}_t\}$ on $\{\hat{\varepsilon}_t\}$, the regression equation is,

$$\hat{e}_t = 0.000 - 0.0287\,\hat{\varepsilon}_t$$

Other results were,

$$\text{SSE} = \sum \hat{e}_t^{\,2} = 3364.88\,,\ \sum \hat{\varepsilon}_t^{\,2} = 119928\,,\ \hat{\eta}_0 = -.028726\,,\ \hat{\eta} = -9.9$$

$$N - 2m - 2 = 130, \text{ and } F = \frac{\hat{\eta}^2 (N - 2m - 2}{SSE - \hat{\eta}^2} = 3.93944$$

Since $F_c = 3.93944 < F_{0.01(1,130)} \cong 6.63$ .We accept H_o and conclude that there is sufficient evidence to suggest that series 1 data is linear. That is, a linear model may be fitted to this particular data.

Since the data suggests a linear model we proceeded to model it as outlined earlier. The model identification results are as given below

Fig 4.3: Autocorrelations: S1

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.470	.079	. |**.******	35.143	.000
2	.189	.079	. |**.*	40.848	.000
3	.003	.079	. | .	40.850	.000
4	.057	.079	. |* .	41.380	.000
5	.146	.078	. |***	44.860	.000
6	.148	.078	. |***	48.449	.000
7	.087	.078	. |**.	49.694	.000
8	-.002	.077	. | .	49.695	.000
9	.006	.077	. | .	49.701	.000
10	.076	.077	. |**.	50.684	.000
11	.361	.077	. |**.****	72.860	.000
12	.586	.076	. |**.*********	131.639	.000
13	.314	.076	. |**.***	148.671	.000
14	.053	.076	. |* .	149.155	.000
15	-.073	.076	. * | .	150.080	.000
16	-.053	.075	. * | .	150.573	.000
17	.099	.075	. |**.	152.326	.000
18	.049	.075	. |* .	152.751	.000
19	.003	.075	. * | .	152.753	.000
20	-.125	.074	.** | .	155.568	.000
21	-.066	.074	. * | .	156.368	.000
22	.036	.074	. |* .	156.605	.000
23	.275	.073	. |**.***	170.668	.000
24	.438	.073	. |**.******	206.503	.000
25	.226	.073	. |**.**	216.075	.000
26	.002	.073	. * | .	216.076	.000
27	-.123	.072	.** | .	218.958	.000
28	-.133	.072	*** | .	222.366	.000
29	-.072	.072	. * | .	223.371	.000
30	-.038	.071	. * | .	223.650	.000
31	-.060	.071	. * | .	224.369	.000
32	-.155	.071	*** | .	229.116	.000
33	-.159	.071	*** | .	234.207	.000
34	-.062	.070	. * | .	234.992	.000
35	.185	.070	. |**.*	241.987	.000
36	.358	.070	. |**.****	268.365	.000

Plot Symbols: Autocorrelations * Two Standard Error Limits .
Total cases: 156 Computable first lags: 155

The correlogram above indicates that undifferenced series 1 is not stationary. We need to difference it until it becomes stationary.

Fig 4.4: Autocorrelations: S1

Transformations: difference (1)

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob
1	-.206	.080	*.** .	6.704	.01
2	-.092	.079	.** .	8.057	.01
3	-.226	.079	**.** .	16.220	.00
4	-.030	.079	. * .	16.363	.00
5	.068	.079	. * .	17.111	.00
6	.093	.078	. **.	18.525	.00
7	.034	.078	. * .	18.710	.00
8	-.110	.078	.** .	20.707	.00
9	-.097	.077	.** .	22.289	.00
10	-.214	.077	*.** .	30.007	.00
11	.045	.077	. * .	30.346	.00
12	.495	.077	. **.*******	72.090	.00
13	.015	.076	. .	72.131	.00
14	-.148	.076	*** .	75.929	.00
15	-.119	.076	.** .	78.411	.00
16	-.128	.076	*** .	81.285	.00
17	.166	.075	. ***	86.122	.00
18	.033	.075	. * .	86.319	.00
19	.081	.075	. **.	87.487	.00
20	-.192	.074	*.** .	94.135	.00
21	-.052	.074	. * .	94.627	.00
22	-.160	.074	*** .	99.288	.00
23	.071	.074	. * .	100.226	.00
24	.395	.073	. **.*****	129.206	.00
25	.015	.073	. .	129.246	.00
26	-.096	.073	.** .	131.001	.00
27	-.136	.073	*** .	134.522	.00
28	-.050	.072	. * .	134.995	.00
29	.023	.072	. .	135.096	.00
30	.082	.072	. **.	136.408	.00
31	.078	.071	. **.	137.602	.00
32	-.091	.071	.** .	139.244	.00
33	-.135	.071	*** .	142.875	.00
34	-.144	.071	*** .	147.062	.00
35	.069	.070	. * .	148.022	.00
36	.424	.070	. **.*****	184.814	.00

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 156 Computable first lags after differencing: 154

One simple differencing of series 1 data gives a correlogram above which clearly shows the presenc

of seasonal fluctuations.

Fig 4.5: Autocorrelations: S1
Transformations: seasonal difference (1 at 12)

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.290	.082	. |**.***	12.379	.000
2	.362	.082	. |**.****	31.820	.000
3	.132	.082	. |***	34.399	.000
4	.276	.082	. |**.***	45.813	.000
5	.118	.081	. |**.	47.929	.000
6	.124	.081	. |**.	50.257	.000
7	.034	.081	. |* .	50.430	.000
8	.073	.080	. |* .	51.254	.000
9	.074	.080	. |* .	52.099	.000
10	-.115	.080	.** | .	54.162	.000
11	-.053	.080	. * | .	54.602	.000
12	-.373	.079	****.** | .	76.792	.000
13	-.068	.079	. * | .	77.529	.000
14	-.152	.079	*** | .	81.268	.000
15	-.024	.078	. | .	81.360	.000
16	-.118	.078	.** | .	83.660	.000
17	.087	.078	. |**.	84.927	.000
18	-.094	.077	.** | .	86.411	.000
19	-.056	.077	. * | .	86.933	.000
20	-.161	.077	*** | .	91.350	.000
21	.021	.076	. | .	91.428	.000
22	.011	.076	. | .	91.447	.000
23	-.014	.076	. | .	91.482	.000
24	-.065	.076	. * | .	92.232	.000
25	.041	.075	. |* .	92.535	.000
26	.139	.075	. |***	95.976	.000
27	.014	.075	. | .	96.013	.000
28	-.043	.074	. * | .	96.346	.000
29	-.096	.074	.** | .	98.033	.000
30	.025	.074	. |* .	98.152	.000
31	.002	.073	. | .	98.152	.000
32	.037	.073	. |* .	98.413	.000
33	-.090	.073	.** | .	99.964	.000
34	-.012	.072	. | .	99.991	.000
35	.004	.072	. | .	99.995	.000
36	.009	.072	. | .	100.012	.000

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 156 Computable first lags after differencing: 143

One seasonal differencing of series 1 gives a correlogram above which suggests the model ARIMA(0,0,4)(0,1,1).However this model proved to be inadequate as the AIC&SBC were not the minimum. The residuals also appeared to have a pattern.

Fig 4.6: Autocorrelations: S1

Transformations: difference (1), seasonal difference (1 at 12)

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.554	.083	********.** .	44.753	.000
2	.210	.082	. **.*	51.242	.000
3	-.251	.082	**.** .	60.561	.000
4	.218	.082	. **.*	67.655	.000
5	-.112	.082	.** .	69.536	.000
6	.063	.081	. * .	70.128	.000
7	-.082	.081	.** .	71.143	.000
8	.025	.081	. .	71.238	.000
9	.122	.080	. **.	73.544	.000
10	-.164	.080	*** .	77.753	.000
11	.269	.080	. **.**	89.136	.000
12	-.448	.079	******.** .	120.929	.000
13	.286	.079	. **.***	133.948	.000
14	-.142	.079	*** .	137.188	.000
15	.147	.079	. ***	140.690	.000
16	-.216	.078	*.** .	148.299	.000
17	.261	.078	. **.**	159.523	.000
18	-.144	.078	*** .	162.946	.000
19	.094	.077	. **.	164.429	.000
20	-.206	.077	*.** .	171.615	.000
21	.145	.077	. ***	175.202	.000
22	-.002	.076	. .	175.203	.000
23	.012	.076	. .	175.227	.000
24	-.101	.076	.** .	177.007	.000
25	.001	.075	. .	177.007	.000
26	.157	.075	. ***	181.371	.000
27	-.058	.075	. * .	181.963	.000
28	.012	.074	. .	181.990	.000
29	-.127	.074	*** .	184.944	.000
30	.107	.074	. **.	187.048	.000
31	-.044	.073	. * .	187.406	.000
32	.137	.073	. ***	190.934	.000
33	-.163	.073	*** .	195.921	.000
34	.042	.073	. * .	196.252	.000
35	.027	.072	. * .	196.391	.000
36	.057	.072	. * .	197.025	.000

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 156 Computable first lags after differencing: 142

Judging from the patterns given in table (2.5.1), we roughly took ARIMA(0,1,4)(0,1,1) to be the appropriate model .Examining AIC & SBC for other competing models revealed that this is the most appropriate model for series 1.

Fig 4.7: Partial Autocorrelations: S1

Transformations: difference (1), seasonal difference (1 at 12)

Lag	Pr -Aut - Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1
1	-.554	.084	********.**
2	-.139	.084	***
3	-.287	.084	***.**
4	-.067	.084	. *
5	-.023	.084	.
6	-.031	.084	. *
7	-.053	.084	. *
8	-.090	.084	.**
9	.145	.084	. ***
10	-.050	.084	. *
11	.273	.084	. **.**
12	-.253	.084	**.**
13	-.188	.084	*.**
14	-.066	.084	. *
15	-.121	.084	.**
16	-.159	.084	***
17	.094	.084	. **.
18	.094	.084	. **.
19	.029	.084	. * .
20	-.199	.084	*.**
21	.013	.084	.
22	-.011	.084	.
23	.151	.084	. ***
24	-.206	.084	*.**
25	-.274	.084	**.**
26	.047	.084	. * .
27	.118	.084	. **.
28	-.027	.084	. *
29	.053	.084	. * .
30	.012	.084	.
31	-.005	.084	.
32	-.037	.084	. *
33	.074	.084	. * .
34	-.101	.084	.**
35	.100	.084	. **.
36	-.072	.084	. *

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 156 Computable first lags after differencing: 142

DEPARTMENT OF MATHEMATICS
KENYATTA UNIVERSITY
P. O. BOX 43844, NAIROBI

The ACFs and PACFs could not clearly give the appropriate model and as such we also examined the AIC and SBC value for different models .The model that came out to be the most appropriate for this particular series is ARIMA(0,1,4)(0,1,1) with it's parameters as given below.

Table 4.2.1a: **Final Parameters**

FINAL PARAMETERS:	
Number of residuals	143
Standard error	4.0052476
Log likelihood	-408.93702
AIC	827.87404
SBC	842.68827

Table 4.2.2a: **Variables in the Model**

	B	SEB	T-RATIO
MA1	0.693014	0.081522	8.500935
MA2	-0.08936	0.096097	-0.92986
MA3	0.229431	0.094594	2.425424
MA4	-0.14032	0.083503	-1.68046
SMA1	0.896179	0.124174	7.217134

This model was then diagnosed for adequacy using residual analysis method. The ACFs and the corresponding Box-Ljung's statistics for the residuals are as given fig.4.8. Looking at the residual ACF pattern we note that they all fall within the standard limits, they also show no definite pattern .For M = 20, the Box-Ljung statistic (Q_c) given on the RHS column is 15.414.That is $Q_c = 15.414$ and at $\alpha = 0.05$ level of significance $Q_t = 25.0$. Since $Q_c < Q_t$, we take this model to be adequate. (Recall: Q has a chi-square distribution with M-p-q-Q-P degrees of freedom.)

Fig 4.8: Autocorrelations: Error for S1 from ARIMA,

Auto -Stand.

Lag	Corr.	Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.010	.083	. .	.016	.901
2	.004	.082	. .	.018	.991
3	.030	.082	. * .	.150	.985
4	.094	.082	. ** .	1.476	.831
5	.007	.082	. .	1.483	.915
6	.026	.081	. * .	1.582	.954
7	-.046	.081	. * .	1.903	.965
8	.022	.081	. .	1.975	.982
9	.029	.080	. * .	2.106	.990
10	-.109	.080	.** .	3.950	.950
11	.011	.080	. .	3.968	.971
12	-.064	.079	. * .	4.611	.970
13	.008	.079	. .	4.621	.983
14	.045	.079	. * .	4.943	.987
15	-.059	.079	. * .	5.509	.987
16	-.097	.078	.** .	7.035	.973
17	.167	.078	. ***	11.598	.824
18	-.002	.078	. .	11.599	.867
19	-.034	.077	. * .	11.788	.895
20	-.147	.077	*** .	15.414	.752
21	.027	.077	. * .	15.538	.795
22	.009	.076	. .	15.551	.838
23	.021	.076	. .	15.628	.871
24	-.126	.076	*** .	18.376	.784
25	.065	.075	. * .	19.128	.791
26	.169	.075	. ***	24.186	.565
27	-.006	.075	. .	24.192	.620
28	-.056	.074	. * .	24.760	.641
29	-.026	.074	. * .	24.884	.684
30	.042	.074	. * .	25.214	.715
31	.064	.073	. * .	25.973	.723
32	.055	.073	. * .	26.535	.739
33	-.112	.073	.** .	28.880	.673
34	.009	.073	. .	28.896	.716
35	.112	.072	. **.	31.286	.648
36	.004	.072	. .	31.289	.692

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Having found that the model was adequate we went ahead to test it's forecasting performance by deleting 1983's observations and then generating forecasts for the same. The forecasts are reliable as can be seen in table 2.4a. We then generated forecasts for 1984(estimates of missing values)

Next we carried out Holt-Winters exponential smoothing as described in the foregoing chapter. The results were as give below. The initial values for level (μ_0), trend (β_0) and seasonal indices (ψ_i i= $0,1,2.....s\text{-}1$) found to be, μ_0= 24.13056 , β_0 = 0.04769 and

Table 4.2.3a: **Seasonal indices**.

i	ψ_i
0	5.76607
1	4.56607
2	2.96676
3	-3.74712
4	-9.98046
5	-6.26726
6	3.87392
7	1.03413
8	-2.90962
9	-0.69087
10	-2.15962
11	7.54801

The optimal smoothing constants as judged by minimizing the sum of squares due to error(SSE) are , α = 0.028, γ =0.0 and δ =0.0 which implies that only the level needed updating. At period 143(dec.1983) the updated level =33.39486 was used to generate forecasts for 1983.The forecasts were found reliable and thus the updated level at period 156 (dec.1983) = 34.42025 was used to generate forecasts for 1984. see table 2.4a.

Having obtained two estimated values at each period for 1983 and 1984, we proceeded to combine these forecasts. The optimal combining weights were obtained using the data on past forecast performance, that is we generated forecasts for periods before 1983 and used the argument in equation (3.4.8) to obtain β_2 and then β_1 from the fact that β_1+ β_2=1. These weights were found to be β_1= 0.2914012 and

β_2= 0.7085988

We then used these weights to combine the forecast from the two procedures described above as explained in chapter 3. To test the performance of the proposed combination theory, we first dealt with estimated values for 1983 and compared the combined forecasts with the actual values .Judging by the reduced SSE the combined forecasts are better than either set of the forecasts on its own. This provided a good basis of combining the estimates for 1984 .see Table 2.5a (Appendix A2) for the results.

Recalling that we had quarterly totals for 1984 and we earlier suggested a method of incorporating their information into the final estimates, we proceeded to do so, The performance of our suggested adjusting method was tested using 1983's values by computing the quarterly totals from the known actual values. Table 2.6a gives the adjusted estimates. Judging by the reduced SSE the adjusted estimates are far much better than the ordinary composite estimates. This fact motivated us to do the same to the estimates for 1984, see Table 2.7a (Appendix A2). These adjusted composite estimates could have been used to fill in the gap in the original series but we decided to model series 2 (1990-1985 observations) to find out whether it could provide additional information. We now give a summary results when the process described above in repeated on series 2.

4.2(b):Composite forecasts for series 2

This series gave results almost similar to those of series 1. The test for linearity gave the results below. See also Table 2.1b-2.3b (Appendix A2) On regressing $\{\hat{e}_t\}$ on $\{\hat{\varepsilon}_t\}$, the regression equation is

$\hat{e}_t = 0.000 - 0.0180\,\hat{\varepsilon}_t$

Other results required for the computation of the test statistic are,

$$\text{SSE} = \sum \hat{e}_t^{\,2} = 3430.75,\ \sum \hat{\varepsilon}_t^{\,2} = 322285,\quad \hat{\eta}_0 = -.0179704,\quad \hat{\eta} = -10.2018$$

$$N - 2m - 2 = 52 \text{ and } F = \frac{\hat{\eta}^2 (N - 2m - 2}{SSE - \hat{\eta}^2} = 1.62685$$

Since $F_c = 1.62685 < F_{0.01(1,52)} \cong 7.08$. We accept H_o and conclude that there is sufficient evidence to suggest that series 2 data is linear .That is ,a linear model may be fitted to this particular data.

The results given below were obtained in the model identification stage

Fig 4.9: Autocorrelations: S2

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.668	.115	. |****.********	33.474	.000
2	.418	.115	. |****.***	46.776	.000
3	.346	.114	. |****.**	56.027	.000
4	.387	.113	. |****.***	67.746	.000
5	.300	.112	. |***.**	74.906	.000
6	.258	.111	. |***.*	80.283	.000
7	.276	.110	. |***.**	86.523	.000
8	.285	.110	. |***.**	93.301	.000
9	.195	.109	. |****	96.509	.000
10	.207	.108	. |****	100.205	.000
11	.370	.107	. |***.***	112.140	.000
12	.479	.106	. |***.******	132.530	.000
13	.310	.105	. |***.**	141.181	.000
14	.146	.104	. |***.	143.142	.000
15	.086	.103	. |** .	143.841	.000
16	.091	.103	. |** .	144.627	.000
17	.089	.102	. |** .	145.388	.000
18	.079	.101	. |** .	146.011	.000
19	.085	.100	. |** .	146.741	.000
20	.049	.099	. |* .	146.988	.000
21	.025	.098	. |* .	147.055	.000
22	.072	.097	. |* .	147.613	.000
23	.154	.096	. |***.	150.198	.000
24	.216	.095	. |****	155.389	.000
25	.131	.094	. |***.	157.326	.000
26	-.045	.093	. * | .	157.559	.000
27	-.095	.092	. ** | .	158.625	.000
28	-.081	.091	. ** | .	159.420	.000
29	-.067	.090	. * | .	159.977	.000
30	-.096	.089	. ** | .	161.146	.000
31	-.060	.088	. * | .	161.618	.000
32	-.028	.087	. * | .	161.723	.000
33	-.056	.086	. * | .	162.154	.000
34	-.072	.084	. * | .	162.877	.000
35	-.025	.083	. * | .	162.965	.000
36	.061	.082	. | * .	163.511	.000

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 72 Computable first lags: 71

The correlogram above indicates that undifferenced series 2 is not stationary and thus need to be differenced until it becomes stationary.

Fig 4.10: Autocorrelations: S2

Transformations: difference (1)

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.123	.116	. ** .	1.116	.291
2	-.288	.115	*.**** .	7.335	.026
3	-.156	.115	. *** .	9.185	.027
4	.068	.114	. * .	9.538	.049
5	.017	.113	. . .	9.561	.089
6	-.054	.112	. * .	9.791	.134
7	.056	.111	. * .	10.050	.186
8	.128	.110	. *** .	11.407	.180
9	-.185	.109	**** .	14.268	.113
10	-.222	.108	**** .	18.451	.048
11	.071	.108	. * .	18.884	.063
12	.413	.107	. ***.****	33.866	.001
13	.026	.106	. * .	33.928	.001
14	-.139	.105	.*** .	35.688	.001
15	-.117	.104	. ** .	36.959	.001
16	-.036	.103	. * .	37.084	.002
17	.046	.102	. * .	37.289	.003
18	.005	.101	. . .	37.291	.005
19	.092	.100	. ** .	38.129	.006
20	-.061	.099	. * .	38.504	.008
21	-.135	.098	.*** .	40.406	.007
22	-.060	.097	. * .	40.787	.009
23	.001	.096	. . .	40.787	.013
24	.302	.095	. ***.**	50.823	.001
25	.178	.094	. ****	54.392	.001
26	-.242	.093	*.*** .	61.123	.000
27	-.116	.092	. ** .	62.706	.000
28	-.021	.091	. . .	62.758	.000
29	.072	.090	. * .	63.395	.000
30	-.036	.089	. * .	63.556	.000
31	.020	.088	. . .	63.606	.000
32	.084	.087	. **.	64.553	.001
33	-.065	.086	. * .	65.134	.001
34	-.107	.084	.** .	66.726	.001
35	-.084	.083	.** .	67.744	.001
36	.253	.082	. **.**	77.186	.000

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 72 Computable first lags after differencing: 70

From the SACF above we observe clearer seasonal patterns. This series is still not stationary.

g 4.11: Autocorrelations: S2

ansformations: seasonal difference (1 at 12)

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	.226	.126	. |*****	3.234	.072
2	-.083	.125	. ** | .	3.680	.159
3	-.045	.124	. * | .	3.810	.283
4	-.120	.123	. ** | .	4.761	.313
5	-.168	.122	. *** | .	6.664	.247
6	-.050	.120	. * | .	6.838	.336
7	.118	.119	. |** .	7.810	.350
8	.256	.118	. |*****	12.497	.130
9	.009	.117	. | .	12.503	.186
10	-.047	.116	. * | .	12.665	.243
11	-.049	.115	. * | .	12.845	.304
12	-.258	.114	***** | .	17.984	.116
13	-.064	.112	. * | .	18.313	.146
14	.271	.111	. |***.*	24.264	.043
15	.213	.110	. |****	28.013	.021
16	.066	.109	. |* .	28.386	.028
17	-.028	.108	. * | .	28.455	.040
18	-.033	.106	. * | .	28.551	.054
19	-.079	.105	. ** | .	29.120	.064
20	-.207	.104	**** | .	33.109	.033
21	-.055	.102	. * | .	33.402	.042
22	.108	.101	. |** .	34.537	.043
23	.042	.100	. |* .	34.716	.055
24	-.043	.098	. * | .	34.906	.070
25	-.028	.097	. * | .	34.989	.088
26	-.077	.096	. ** | .	35.643	.098
27	-.096	.094	. ** | .	36.678	.101
28	.030	.093	. |* .	36.784	.124
29	.026	.091	. |* .	36.868	.150
30	-.017	.090	. | .	36.904	.180
31	.002	.088	. | .	36.905	.215
32	-.083	.087	.** | .	37.821	.221
33	-.057	.085	. * | .	38.273	.242
34	.014	.084	. | .	38.300	.281
35	-.050	.082	. * | .	38.670	.307
36	.054	.080	. |* .	39.126	.331

lot Symbols: Autocorrelations * Two Standard Error Limits .

otal cases: 72 Computable first lags after differencing: 59

One seasonal differencing of the raw data (series 2) gives a correlogram above which suggests the model ARIMA (0,0,1)(0,1,1).However this model proved to be inadequate as the AIC&SBC were not the minimum. The residuals also appeared to have a pattern.

Fig 4.12: Autocorrelations: S2

Transformations: difference (1), seasonal difference (1 at 12)

Auto -Stand.

Lag	Corr.	Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.296	.127	*.***** .	5.450	.020
2	-.210	.126	.**** .	8.229	.016
3	.059	.125	* .	8.450	.038
4	-.009	.124	.	8.455	.076
5	-.109	.122	. ** .	9.244	.100
6	-.041	.121	. * .	9.356	.155
7	.016	.120	.	9.374	.227
8	.262	.119	. *****	14.226	.076
9	-.133	.118	. *** .	15.501	.078
10	-.036	.117	. * .	15.595	.112
11	.132	.115	. *** .	16.911	.111
12	-.254	.114	***** .	21.835	.039
13	-.093	.113	. ** .	22.514	.048
14	.245	.112	. ***.*	27.330	.017
15	.061	.111	. * .	27.633	.024
16	-.033	.109	. * .	27.723	.034
17	-.063	.108	. * .	28.059	.044
18	.040	.107	. * .	28.200	.059
19	.052	.105	. * .	28.444	.075
20	-.197	.104	**** .	32.041	.043
21	.003	.103	.	32.041	.058
22	.145	.101	. ***.	34.084	.048
23	.016	.100	.	34.109	.064
24	-.068	.099	. * .	34.589	.075
25	.040	.097	. * .	34.756	.093
26	-.020	.096	.	34.798	.116
27	-.095	.094	. ** .	35.814	.119
28	.092	.093	. ** .	36.805	.123
29	.019	.091	.	36.849	.150
30	-.039	.090	. * .	37.036	.176
31	.076	.088	. ** .	37.787	.187
32	-.069	.087	. * .	38.417	.202
33	-.041	.085	. * .	38.654	.229
34	.071	.083	. * .	39.389	.241
35	-.111	.082	.** .	41.237	.217
36	.133	.080	. ***	43.999	.169

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 72 Computable first lags after differencing: 58

Judging from the patterns given in table (2.5.1) we roughly took ARIMA(0,1,1)(0,1,1) to be the appropriate model .Examining AIC & SBC for other models revealed that, this is not the most appropriate model for series 2.The model with minimum AIC &SBC was found to be ARIMA(0,1,3)(0,1,1).

Fig 4.13: Partial Autocorrelations: S2

Transformations: difference (1), seasonal difference (1 at 12)

Lag	Pr -Aut - Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1
1	-.296	.130	*.**** .
2	-.326	.130	**.**** .
3	-.148	.130	. *** .
4	-.135	.130	. *** .
5	-.218	.130	.**** .
6	-.261	.130	***** .
7	-.285	.130	*.**** .
8	.073	.130	* .
9	-.066	.130	* .
10	-.025	.130	* .
11	.107	.130	** .
12	-.218	.130	.**** .
13	-.289	.130	*.**** .
14	-.057	.130	* .
15	.057	.130	* .
16	.048	.130	* .
17	-.049	.130	* .
18	-.060	.130	* .
19	.019	.130	.
20	-.004	.130	.
21	.013	.130	.
22	-.023	.130	.
23	.034	.130	* .
24	-.077	.130	** .
25	-.090	.130	** .
26	-.040	.130	* .
27	-.102	.130	** .
28	.098	.130	** .
29	.008	.130	.
30	-.103	.130	** .
31	.082	.130	** .
32	-.054	.130	* .
33	-.168	.130	*** .
34	.017	.130	.
35	-.062	.130	* .
36	.053	.130	* .

Plot Symbols: Autocorrelations * Two Standard Error Limits .

Total cases: 72 Computable first lags after differencing: 58

Diagnostic /model checking gave positive results as shown below.

Fig 4.14: Autocorrelations: Error for S2 from ARIMA,

Lag	Auto Corr.	Stand. Err.	-1 -.75 -.5 -.25 0 .25 .5 .75 1	Box-Ljung	Prob.
1	-.059	.127	. * .	.214	.644
2	.031	.126	. * .	.275	.872
3	.005	.125	. * .	.276	.964
4	.011	.124	. * .	.284	.991
5	-.156	.122	. *** .	1.897	.863
6	-.096	.121	. ** .	2.523	.866
7	-.039	.120	. * .	2.627	.917
8	.161	.119	. *** .	4.446	.815
9	-.110	.118	. ** .	5.319	.806
10	-.026	.117	. * .	5.369	.865
11	-.019	.115	. * .	5.398	.910
12	.105	.114	. ** .	6.242	.903
13	-.129	.113	. *** .	7.543	.872
14	.223	.112	. ****	11.526	.644
15	.085	.111	. ** .	12.118	.670
16	.057	.109	. * .	12.390	.717
17	-.089	.108	. ** .	13.066	.732
18	-.050	.107	. * .	13.288	.774
19	-.014	.105	. * .	13.307	.823
20	-.168	.104	.*** .	15.916	.722
21	-.091	.103	. ** .	16.706	.729
22	.092	.101	. ** .	17.533	.733
23	-.023	.100	. * .	17.586	.780
24	.017	.099	. * .	17.616	.821
25	-.062	.097	. * .	18.017	.842
26	.048	.096	. * .	18.271	.866
27	-.030	.094	. * .	18.371	.892
28	.029	.093	. * .	18.466	.914
29	-.024	.091	. * .	18.537	.933
30	-.041	.090	. * .	18.747	.945
31	-.004	.088	. * .	18.748	.959
32	-.122	.087	.** .	20.742	.937
33	-.073	.085	. * .	21.481	.939
34	.046	.083	. * .	21.792	.948
35	-.098	.082	.** .	23.222	.936
36	.102	.080	. **.	24.853	.919

Plot Symbols: Autocorrelations * Two Standard Error Limits .

The most appropriate model came out to be ARIMA(0,1,3)(0,1,1) with parameters

Table 4.2.1b: **Final Parameters**

FINAL PARAMETERS:	
Number of residuals	59
Standard error	7.5066564
Log likelihood	-202.74814
AIC	413.49627
SBC	421.80642

Table 4.2.2b: **Variable in the Model**

	B	SEB	T-RATIO
MA1	0.717513	0.135676	5.288418
MA2	0.215675	0.160263	1.345752
MA3	-0.03881	0.13406	-0.28946
SMA1	0.435416	0.145906	2.984217

The correlogram in fig 4.14 implies that the residuals have the characteristics of white noise as desired. For M =15, the Box-Ljung statistic(Q_c) given on the RHS column of this figure is 12.118 Q_c= 12.118.for $\alpha = 0.05$ level of significance.

$Q_c = 12.118 < Q_t = 19.7$, we therefore take this model to be adequate .

Forecasts for testing the performance of the model(1985's backcasts) were found reliable and thus the model was used to estimate the missing values for 1984.

Next Holt-Winters exponential smoothing was carried out and the initial values for level(μ_0), trend(β_0) and seasonal indices (ψ_i $I= 0,1,2.....s-1$) found to be,

μ_o= 78.09833 , β_0=-0.5011 and

Table 4.2.3b: **Seasonal indices**

i	ψ_I
0	11.60648
1	0.90315
2	-1.62185
3	-2.37519
4	4.14981
5	5.17481
6	-8.89796
7	-16.6352
8	-6.26852
9	3.78815
10	1.25481
11	8.92148

The optimal combining weights for the composite forecasts were obtained as, $\alpha = 0.08$, $\gamma = 0.0$ and $\delta = 0.0$ again suggesting that only the level needed updating. We again tested the forecasting performance of this procedure by predicting the values for 1985(dec.to Jan.). The forecasts appears reliable see Table 2.4b (Appendix A2). We therefore went ahead to generate forecasts (backcasts) for 1984. Note, the updated levels at period 60 and 72 backwards were found to be 47.65601 and 42.507113 respectively. To combine the forecasts we obtained the optimal combining weights to be,

$$\beta_1 = -0.979187 \text{ and } \beta_2 = 1.979184.$$

The composite estimates given in table 2.5b (Appendix A2) were adjusted using the quarterly totals. Note we first adjusted 1985's estimates to again test the performance of our suggested method. The method proved to be working as judged from the minimized SSE. See Table 2.6b (Appendix A2). Re-arranging these adjusted estimates, we obtain estimates corresponding to those obtained in section 4.2(a) (from series 1). Table 2.8 (Appendix A2) gives the adjusted estimates from the two sets of data .A look at them suggests no much difference, we therefore decided to combine them using simple averages method to come up with the final composite estimates given in the last column of the same. When these values are used to 'patch' the 'gap' in the raw data, we obtain the first complete data set, which we will henceforth refer to as series C. This process of filling in the missing values using composite estimates has been called composite imputation by Little and Rubin (1987).

4.3: **Modified Hot-Deck imputation**

Earlier we suggested this direct method of imputation, which requires no modelling. All we need to do is to arrange the data in such away that all values for month *m*, *(m=1,2...12)* are in the same cell and in the order of their occurrence. Table 1.1 (Appendix A1) is actually in the required format. We then identify the close-neighbors to the missing value and impute this particular value from them as explained in Section 3.6.

To test the performance of this method, the values for 1977 were deleted and we imputed the missing values using,

i) One value on either side of 1977 (1976's &1978's)

ii) Two values on either side of 1977

iii) Six values on either side of 1977

The estimated values proved to be reliable and so we adjusted these estimates using the information from quarterly totals (compute d from the actual values). The adjusted values again proved to be better as the SSE decreased considerably upon adjusting. See Table 3.3 (Appendix A3). We then proceeded to impute the missing values for 1984 using,

i) Two values on either side of 1984

ii) Six values on either side of 1984

The imputed values were adjusted appropriately using the quarterly totals to come up with what we refer to as Hot-Deck estimates. When the 'gap' in the raw data is patched up with this set of estimates, we get the second set of complete data, which we henceforth refer to as series H.

4.4: Comparison between series C(Sc) and series H(Sh)

In section 4.2 we obtained a complete data set out of composite imputation and called it series C and in section 4.3 above we obtained another complete data set out of modified hot-deck imputation which we have labeled series H. To compare the two imputation methods we now analyze these two sets of data. Table 4.1 (Appendix A4) gives the raw data with the missing observations having been filled in using the two different methods. A look at the two sets of estimated values reveals a very slight difference. In fact the values seem to have the same pattern as suggested by the time plots below.

Fig 4.15: **Time plot for series C**

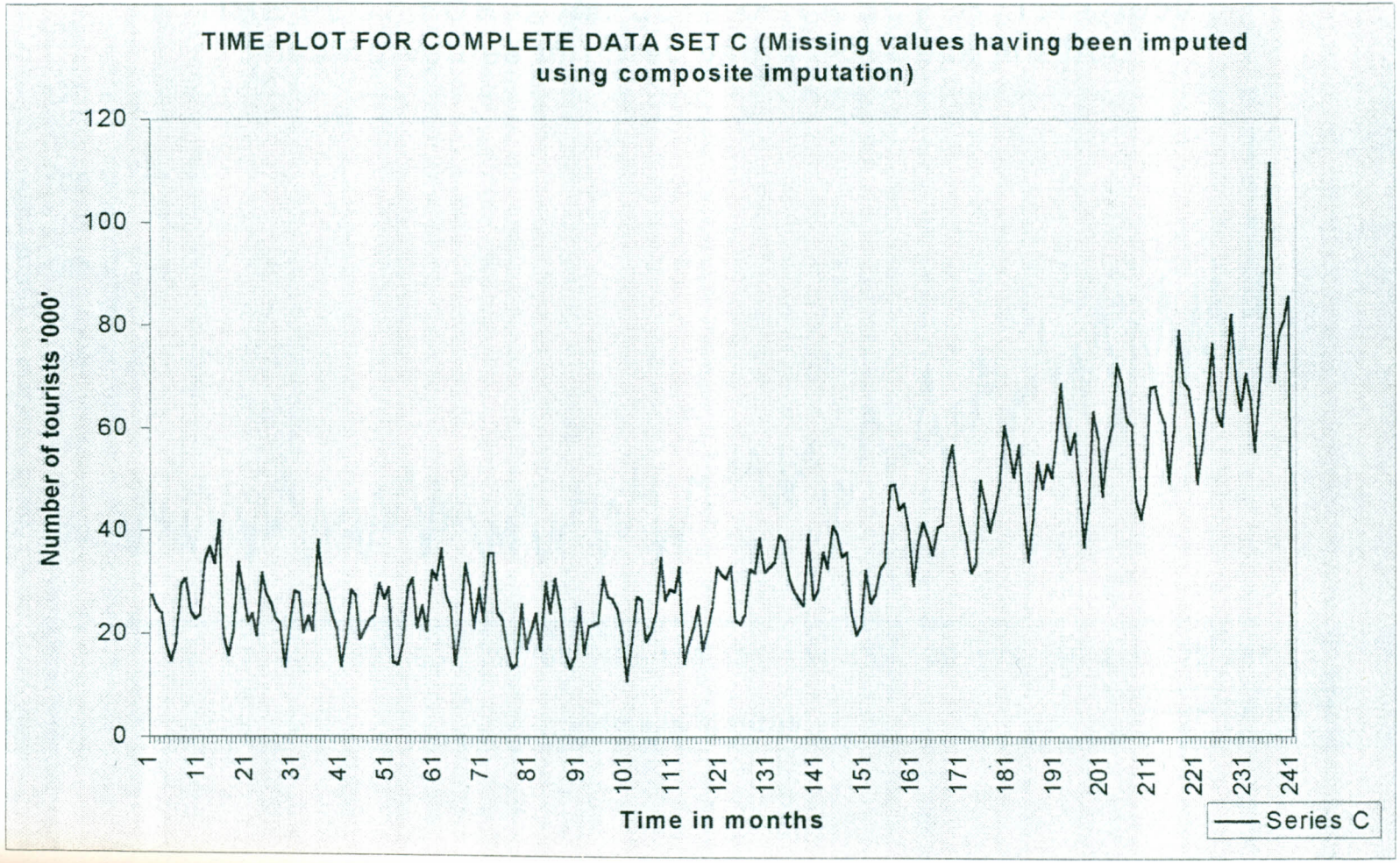
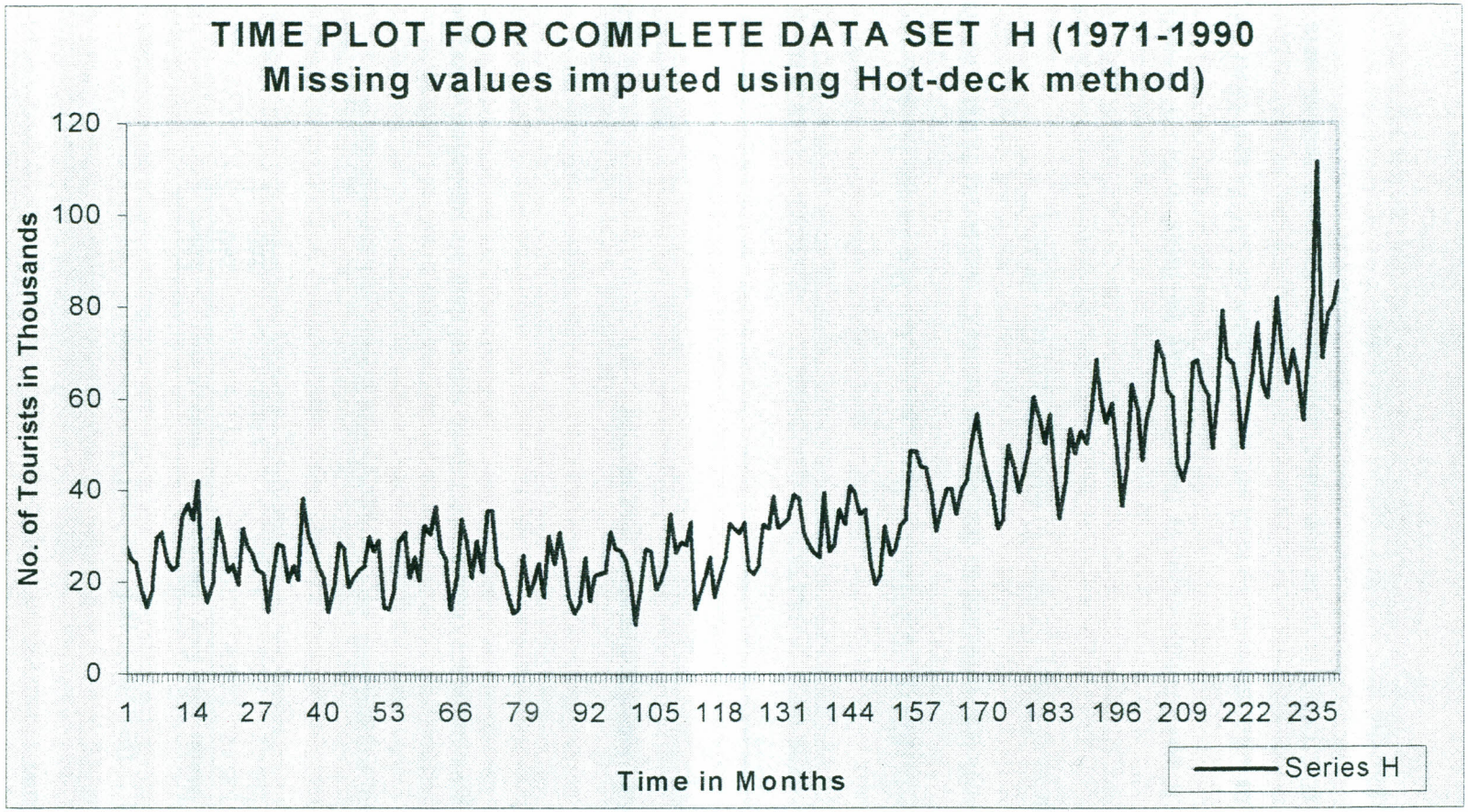


Fig 4.16: **Time plot for series H**



Since this does not bring out considerable differences we look into their descriptives given below.

Table 4.4.1: **Descriptives for sc and sh.**

Variable	Mean	Std Dev	Minimum	maximum	Label
SC	36.42	17.69	10.8	111.5	240
SH	36.42	17.69	10.8	111.5	240

They again suggest that the two-imputation methodsare equally good.

We now give summary results when these sets of data are modelled as explained earlier.

Table 4.4.2: **Test for linearity (for m=15)**

Series	$\sum \hat{e}_t^{\,2}$	$\sum \hat{\varepsilon}_t^{\,2}$	η_0	η	$\mathbf{F_c}$
Sc	6878.62	19406916	-0.00058897	-2.59460	0.201805
Sh	6858.99	19384762	-0.00056712	-2.49692	0.187418

Since F_c= 0.187418 <0.201805 < $F_{0.01(1,206)} \cong 6.63$ .We accept H_o and conclude that there is sufficient evidence to suggest that series C and series H data are linear .That is, a linear model may be fitted to each of these sets of data.

Since both sets of data suggests that they may be fitted to linear models we went ahead to fit a linear model to each set. Examination of SACFs and SPACFs suggested ARIMA (0,1,1)(0,1,1) to be the most appropriate for both sets of data .This was confirmed by the AIC and SBC values given below.

Table 4.4.3: **Model Parameters for Sc and Sh.**

Variable	AIC	SBC	MA1	SMA1
Sc	1408.5618	1415.4117	0.7419	0.8371
Sh	1408.2101	1415.06	0.7403	0.8401

This model was diagnosed to be adequate and forecasting performance found to be satisfactory. See Table 4.2 (Appendix A4). In the same table we give eventual forecasts (forecasts for the year 1991) which could have been used to plan for this particular year if this project was carried out in 1990. NOTE: We could have used a series going up to 1997 and thus forecast for 1998 but we could not get data for years after 1990.

Apart from the descriptives, we note that the two sets of data have similar statistics/features even in modelling stages. This confirms our earlier conclusion that the two imputation methods are equally good. In our last chapter we give our general conclusions/achievements and we suggest areas for further research.

CHAPTER FIVE

ACCOMPLISHMENTS AND CONCLUSION

5.1: **Introduction**

In this last chapter we briefly describe the general achievements/conclusions we have made in this project. We also suggest areas for further research, which have showed up during this project.

5.2: **Achievements/conclusions**

a) In the foregoing chapter we have showed that, apart form the indirect and quite involving imputation method, direct methods such as the one we have proposed may be used in estimation of missing observations in a seasonal time series. Since our direct method gives us reliable estimates as those given by the indirect methods, we can take this method to be better in that, it saves time, space and manpower. It also doesn't require experts in time-series analysis (modelling) to use it.

b) We have empirically showed that, two sets estimates for the same period can be combined using variance-covariance method to produce a better set of estimates. We note that this method may be extended to more than two estimates.

c) We went on to give a method of adjusting estimates with known quarterly totals and produced better estimates on adjusting.

d) We finally came up with two complete data sets, which may be analyzed. We modelled both of them and generated forecasts for periods beyond the last period (Dec 1990). In so doing we showed how an appropriate model may be used to generate future forecasts and thus help planners in the tourism industry to make better decisions based on the predicted (expected) number of tourists. We also note that, one can now proceed to carry out spectral analysis to this complete data set and establish whether Mutiso's conclusions on general cycles on Kenyan tourism industry were valid.

e) We have also established through testing, that the data on world tourists visiting Kenya is generally linear and there is no justification for anyone to fit non-linear models to this time series data. We therefore find a basis of criticising Onyango's work.

f) As noted earlier, many time series analysts have only dealt with cases were only one or two observation are missing in a given time series. Few have employed Kalman filter recursive equations but have generated estimates with large deviations from the actual values. We have in our project been able to estimate more than two observations, unlike most of the earlier work.

5.3 Suggestions

a) In our suggested direct method of imputation, we made an appealing assumption that the missing value $X_{m,\ell}$, for the m^{th} month $(m = 1,2,.....12)$ and ℓ^{th} year $(\ell = 1,2.....N/12)$ carries some information from $X_{m,\ell-1}, X_{m,\ell-2},.....$as well as $X_{m,\ell+1}, X_{m,\ell+2},....$ That is, the missing value for a given year depends on the values

for the previous and preceding years in the corresponding months, but we expect $X_{m,\ell\pm1}$ to posses more information about $X_{m,\ell}$ than $X_{m,\ell\pm2}, X_{m,\ell\pm3}, \ldots$ due to the fact that they are closest neighbors to this value. In our case we gave these values equal weights but we feel that a better estimate could be obtained if the very close neighbors receive more weight than the others. For example, weights may be given proportional to the inverse of 'distance' from the value to be imputed. How to determine those weights remains an open area for further research.

b) When for $m=1$ or $m=12$, we note that apart from the neighbors above and below $X_{m,\ell}$ we also have $X_{12,\ell-1}$ immediately to the left of $X_{1,\ell}$ and $X_{1,\ell+1}$ immediately to the right of $X_{12,\ell}$. To impute such a value it would be appealing to also consider these neighbors because they could be carrying some information about it. How to do this while still considering case (a) above remains an area for further research.

c) In imputing the deleted values directly, we used one value, two values, and then six values above and below the missing value. We compared the adjusted estimates with the actual values and found out that, the more the values the better the estimates. So far we do not know the optimal number of values which should be used and we suggest that this be established in further researches. We also note that, the missing values must be at least one year within the data so as to provide at least one neighbor above and below the missing value. Suppose this is not the case, we leave this as another open area for further research.

d) We have also suggested a method of adjusting the estimated monthly values with already known quarterly totals. Suppose we only had the sum total of this year's

observations? Although the adjustment would have been possible, we do not know whether the adjusted values would have been better. We also leave our suggested method open for criticism and if possible, an improvement or a better method be developed.

e) Our project was confined to seasonal time series. In particular we dealt with monthly data and as such we were able to define cells in the proposed direct method of imputation. In case the time series is non-seasonal we do not know how our direct method of imputation may be used. Again, this is a problem that needs to be researched on. However, we note that, since seasonal time series data are the most common especially in economics, the major problem is somehow solved.

REFERENCES

Abraham, B.(1981). Missing observations in time series. *Comm. stat..A Theory Methods* **10**, 1645-1653.

Abraham, B and Thavaneswaran, A.(1991*)*. Estimating of multivariate non-linear time series models. *Journal of stst. Plann. & infor.* **29,** 351-363.

Abraham, B and Thavaneswaran, A.(1991).A non-linear time series model and estimation of missing observations.*Ann. Inst. Stat.math* vol 43

Abraham, B and Ledolter, J (1983) *statistical methods for forecasting* .New York: John Wiley & Sons.

Anderson,O.D (1982)."Time series analysis and forecasting": Anderson,O.D and M.R Perryman(eds.),*Applied time series analysis*. New York:North-Holland Pub.c,. 522-528.

Akaike, H.(1973) Maximum likelihood identification of gaussian autoregressive moving average models, *Biometrika*,**60**, 255-265.

Akaike, H.(1974a) Markovian representation of stochastic processes and its application to the analysis of ARMA processes. *Ann. Inst. Statist.,math* **26,** 363-387.

Ansley,G.F (1979), An algorithm for the exact likelihood of a mixed autoregressive moving average process ,*Biometrika*,**66**,59-65

Bartlett,M.S (1946)."On the theoritical specification of sampling properties of autocorrelated time series" *Journal of the Royal statistical society, series* **B8**,27-41.

Bates,J.M and Granger,C.W.J (1969) "The combination of forecasts", *Journal of forecasting* Vol 3 pp.197-204.

Box,G.E.P and Pierce,D.A (1970) Distribution of the residual autocorrelations in ARIMA time series models. *J. Amer.Statist.,Biometrika*,**52**,181-192.

Box,G. E. P and Jenkins, G. M. (1970,1976). *Time analysis, Forecasting and Control.* San Francisco, Holden – Day.

Cartwright, P.A(1985) "Using state dependent models or prediction of time series with missing observations. "*Time series analysis: Theory and practice* 7; Anderson O.D(ed.) Elsevier Science Pub. B.V. North Holland.

Chatfield, C (1975). *The analysis of Time Series.* London, Chapman and Hall.

Dempster, A.P and Rubin, D.B(1983). Overview "Incomplete data in sample surveys, Vol. 2., *Theory & Annotated Bibliography* Acc. press. 3-10.

Diggle, P. J. (1990). *Time Series.* New York, Oxford University Press.

Durbin, J. (1960). "The fitting of time series models". *Review of the International Statistical Institute.* 28, 233-44

Granger, C. W. J.and Newbold, P (1976) "forecasting transformed series",*Journal of the Royal stat. Soc. Series* **B 38**,189-203.

Granger, C. W. J. and Ramanathan, R. (1984). "Improved methods of combining forecasts". *Journal of Forecasting* **vol. 3** 197 – 204.

Gooijer, J. G and Kumar, K. (1992). *Some Recent Developments in Non-Linear Time Series Modelling Testing and Forecasting.* North – Holland.

Harvey A. C. and Pierce, R.G(1984) Estimating missing observations in economic time series. *J.Amer. Statist. Assoc.***79**,125-131.

Harvey A. C. (1989). *Forecasting Structural Time Series Models and the Kalman Filter.* Cambridge University Press.

Holden, P and Thompson, J.L(1989) *Economic forecasting* . Cambridge Univ.Press.

Jones ,R.H (1980) Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics* **22**,389-395.

Keenan, D.m (1985) "A Tukey Nonadditivity Type test for Time series models" *Journal of Time series.* **9**,35-42.

Kendall, M& Stuart, A.(1966,1983). *Design and analysis and time series Vol. 3*, 4th edition ,Univ. press .Belfast Ltd. Northern Ireland

Ljung, G.M. and Box, G.E.P (1978) On a measure of lack of fit in time series models, *Biometrika,* **65,**297-303

Ljung, G.M.(1989) A note on the estimation of missing values in time series, *Commun. Stat.* **B18,** 459-465.

Luce$\tilde{n}$o,A. (1997) Estimation of missing values in possibly partially non-stationary vector time series,*Biometrika,* **84** ,495-498.

Marquartdt, P.W. (1963). An algorithm for least squares estimation of non-linear parameters, *J. Soc. Ind. Appl. Math*,**2** 431-441.

Mills T. C. (1990) *Time Series Techniques for Economists.* Cambridge University Press.

Mutiso, M. J. (1982). *Spectral Analysis of Kenya Tourist Data. (1971 – 1980).* A dissertation for MSC (University of Nairobi).

Newbold, P. (1974) "The exact likelihood function for a mixed autoregressive moving average process" *Biometrika* ,**61**,423-426.

Newbold, P. (1988). "Predictors Projecting Linear Trend Plus Seasonal Dummies" *The Statistician.* **37**, 111 – 27.

Onyango, R.K.O (1993). Non-linear Time Series Models and Kenya Tourism Data MSc project (KU).

Priestley, M. B (1980) State – dependent models: A general approach to time series analysis. *J. time series analy.* **1**,47-71.

Priestley, M. B (1981) *Spectra Analysis and Time series Vols 1 & 2.*Academic press, London.

Priestley, M. B (1988) "Currently developments in time series modelling" *Journal of Econometrics.* North – Holland.

Schwartz,G.(1978).Estimating the dimension of a model. *Ann.Statist.* **6,** 461-464.

Shibata, R (1976) Selection of the order of an autoregressive model by Akaike's Information Criterion. *Biometrika* **63**,117-126.

Tintner, G. (1940). *The Variate Difference method.* Bloomington – In. Principia Press.

Triantis, J. E. (1983). "Improving forecasts accuracy through composite forecasts". *Time Series Analysis, Theory and Practice 7.* Anderson, O. D. (ed). Amsterdam, New York. North – Holland.

Tong, H (1990). *Non – linear Time Series.* Oxford University Press.

Wei, W. W. S. (1990). *Time Series Analysis.* New York. Addison – Wesley Publishing Company.

Winters, P. R. (1960). "Forecasting sales by exponential weighted moving averages". *Management Science.* **6**, 324 – 42.

Yaglom, A. M. (1955). " The correlation theory of constitute of a stationary process" *Mat. Sb. N.S.* **37**. 141 – 196.

APPENDICES

APPENDIX A1.

Table 1.1: **Raw data of world tourists visiting Kenya between 1971-1990 in '000'.**

	JAN.	FEB.	MAR.	APR.	MAY.	JUN.	JUL.	AUG.	SEP.	OCT.	NOV.	DEC.
1971	27.6	25	24	17.8	14.7	18	29.8	30.8	24.3	22.8	23.7	34.5
1972	37	33.7	42	19.8	15.7	20.1	33.9	28.7	22.3	23.8	19.5	31.7
1973	27.7	25.8	22.8	21.9	13.5	21.5	28.4	27.8	20.2	23.4	20.6	38.2
1974	30.1	27.3	23.4	20.9	13.6	19.3	28.5	27.5	18.9	20.9	22.7	23.5
1975	29.8	26.8	29	14.5	14	17.9	29	30.8	21.1	25.5	20.4	32.3
1976	30.5	36.5	27.2	25.1	14	20.7	33.7	29.3	21.1	28.8	22.3	35.6
1977	35.6	24.1	22.7	17.4	13.3	14.2	25.8	17.1	20.3	24	16.8	30
1978	24	30.8	25.1	16.5	13.3	15.4	25.3	16	21.5	21.8	22.2	31.1
1979	27.2	26.8	24.6	20	10.8	18.1	27.2	26.7	18.6	20.4	23.9	34.9
1980	26.5	28.7	28.1	33	14.3	18	21.3	25.4	16.8	20.7	25.1	32.9
1981	31.5	30.7	33.1	22.7	21.8	23.7	32.5	31.8	38.7	31.9	32.8	33.9
1982	39.2	38.1	30.9	28.4	26.6	25.4	39.3	26.7	28.3	35.7	32.6	40.9
1983	39.3	34.9	35.7	24.1	19.6	21.4	32.2	25.9	27.5	32.4	33.8	48.6
1984												
1985	56.6	48.9	42.7	38.9	31.7	33.6	49.8	45.1	39.7	44	49.1	60.4
1986	55.8	50.4	56.7	44.2	34	41.9	53.5	48.2	52.9	50.3	57.8	68.5
1987	59.3	54.9	59	49.2	36.8	45.1	63.2	58	46.8	56.2	60.3	72.5
1988	69	61.7	60.4	45.8	42.3	47	67.8	68.2	63.4	60.8	49.3	59.3
1989	79	68.8	67.6	62	49.2	57	65.5	76.4	62.9	60.3	69.7	82
1990	68.9	63.4	70.6	65.7	55.4	69.4	82.5	111.5	68.9	78.6	80.7	85.5

Table 1.2: **Available Quarterly totals for 1984.**

Quarter	**i**	**ii**	**iii**	**iv**
Value	138.3	107.1	115.3	134.3

Source: ***Kenya Central Bureau of Statistics (CBS).***

APPENDIX A2.

COMPOSITE IMPUTATION

A2(a) :Imputation using series 1.

Test for Linearity results (m=12).

Table 2.1a: **Analysis of Variance on regressing X_t on $\{1, X_{t-1}, X_{t-2}, \ldots\ldots, X_{t-m}\}$**

SOURCE	DF	SS	MS	F
Regression	12	3904.01	325.33	12.67
Error	131	3364.88	25.69	
Total	143	7268.89		

Table 2.2a: **Analysis of Variance on regressing $\hat{X}_t^2$ on $\{1, X_{t-1}, X_{t-2}, \ldots\ldots, X_{t-m}\}$**

SOURCE	DF	SS	MS	F
Regression	12	11081510	923459	18.72
Error	131	119928	915	
Total	143	11201438		

Table 2.3a: **Analysis of Variance on regressing $\{\hat{e}_t\}$ on $\{\hat{\varepsilon}_t\}$**

SOURCE	DF	SS	MS	F
Regression	1	98.97	98.97	4.3
Error	142	3265.91	23	
Total	143	3364.88		

Table.2.4a: **Forecasts from ARIMA model and Holt-Winters procedures**

Period	ARIMA-F	HW-Forec	Actual Value		Period	ARIMA-F	HW-Forec
145	37.9046	39.17759	39.3		157	40.0467	40.23401
146	38.3838	38.02528	34.9		158	39.3008	39.0817
147	36.7546	36. 47366	35.7		159	35.8128	37.53008
148	30.6974	29.80747	24.1		160	30.925	30.86389
149	24.929	23.62182	19.6		161	25.2015	24.67824
150	28.3581	27.38271	21.4		162	28.2235	28.43913
151	38.3396	37.57158	32.2		163	38.4752	38.628
152	35.1884	34.77948	25.9		164	34.3546	35.8359
153	31.6864	30.88342	27.5		165	32.4957	31.93984
154	34.0936	33.14986	32.4		166	34.8329	34.20628
155	33.3533	31.7288	33.8		167	34.1993	32.78522
156	42.5608	41.48412	48.6		168	43.7744	42.54054

Table 2.5a: **Composite forecasts for series 1 (S1).**

Period	ACTUAL V	ARIMA FRC	HOLT-W F	COMB FRC	Period	ARIMA FRC	HOLT-W F	COMB FRC
145	39.3	37.9046	39.17759	36.21791	157	40.0467	40.23401	37.5907
146	34.9	38.3838	38.02528	35.54103	158	39.3008	39.0817	36.55682
147	35.7	36.7546	36.47366	33.9668	159	35.8128	37.53008	34.44093
148	24.1	30.6974	29.80747	27.47807	160	30.925	30.86389	28.29297
149	19.6	24.929	23.62182	21.41401	161	25.2015	24.67824	22.24199
150	21.4	28.3581	27.38271	25.07821	162	28.2235	28.43913	25.78757
151	32.2	38.3396	37.57158	35.20665	163	38.4752	38.628	35.99475
152	25.9	35.1884	34.77948	32.30991	164	34.3546	35.8359	32.81552
153	27.5	31.6864	30.88342	28.52868	165	32.4957	31.93984	29.51309
154	32.4	34.0936	33.14986	30.83614	166	34.8329	34.20628	31.80015
155	33.8	33.3533	31.7288	29.61345	167	34.1993	32.78522	30.60856
156	48.6	42.5608	41.48412	39.20914	168	43.7744	42.54054	40.31136

SS due to arima forecast errors = 316.57
SS due to holt-w forecast errors = 269.5584
SS due to combined forecast errors = 200.4917

Table 2.6a: **Adjusted composite forecasts for 1983 (using information from quarterly values)**

Period	Actual V	Forec.	Actual Q-T	Actu-q-Av	Forec. Q-T	Adj.Estim
145	39.3	36.21791	0	36.63334	0	37.64787
146	34.9	35.54103	109.9	36.63334	105.7257	36.94426
147	35.7	33.9668	0	36.63334	0	35.30788
148	24.1	27.47807	0	21.7	0	24.18298
149	19.6	21.41401	65.1	21.7	73.97029	18.84611
150	21.4	25.07821	0	21.7	0	22.07091
151	32.2	35.20665	0	28.53333	0	31.37781
152	25.9	32.30991	85.6	28.53333	96.04524	28.7961
153	27.5	28.52868	0	28.53333	0	25.42609
154	32.4	30.83614	0	38.26667	0	35.52111
155	33.8	29.61345	114.8	38.26667	99.65873	34.11266
156	48.6	39.20914	0	38.26667	0	45.16623

Sum of squares due to forecasts errors= 200.4915
Sum of squares due to adjusted estimates errors= 43.08199

Note: "0" in actual Q-T or forecast Q-T (Forec Q-T) column have no physical meaning. They come up as a result of centering the quarterly total.

Table 2.7a: **Adjusted composite forecasts for 1984. (Using of information from quarterly values)**

Period	Actual Q-T	Actual Q-Av	Forec. Q-T	Forec. Q-Av	Forecast	Adj.Estimate
157	0	46.1	0	36.19615	37.5907	47.87612
158	138.3	46.1	108.5884	36.19615	36.55682	46.55936
159	0	46.1	0	36.19615	34.44093	43.86452
160	0	35.7	0	25.44084	28.29297	39.70226
161	107.1	35.7	76.32253	25.44084	22.24199	31.21119
162	0	35.7	0	25.44084	25.78757	36.18655
163	0	38.43333	0	32.77446	35.99475	42.20965
164	115.3	38.43333	98.32336	32.77446	32.81552	38.48149
165	0	38.43333	0	32.77446	29.51309	34.60886
166	0	44.76667	0	34.24002	31.80015	41.57669
167	134.3	44.76667	102.7201	34.24002	30.60856	40.01876
168	0	44.76667	0	34.24002	40.31136	52.70456

A2(b): Imputation using series 2.

Test for Linearity(m=9).

Table 2.1b: **Analysis of Variance on regressing X_t on $\{1, X_{t-1}, X_{t-2}, \ldots\ldots\ldots, X_{t-m}\}$**

SOURCE	DF	SS	MS	F
Regression	9	4731.73	525.75	8.12
Error	53	3430.75	64.73	
Total	62	8162.48		

Table 2.2b: **Analysis of Variance on regressing $\hat{X}_t^2$ on $\{1, X_{t-1}, X_{t-2}, \ldots\ldots\ldots, X_{t-m}\}$**

SOURCE	DF	SS	MS	F
Regression	9	61248800	6805422	19.16
Error	53	322285	6081	
Total	62	61571084		

Table 2.3b: **Analysis of Variance on regressing $\{\hat{e}_t\}$ on $\{\hat{\varepsilon}_t\}$**

SOURCE	DF	SS	MS	F
Regression	1	104.08	104.08	1.91
Error	61	3326.68	54.54	
Total	62	3430.75		

Table.2.4b: Forecasts from ARIMA model and Holt-Winters procedures

Period	ARIMA-Fore.	HW-Fore.T	Actual-V.Ue	Period	ARIMA-Forc.	HW- Fore.
61	58.9193	59.07518	60.4	73	54.8645	53.6125
62	48.6464	47.87074	49.1	74	43.5099	42.405
63	45.7846	44.84463	44	75	39.8202	39.38195
64	43.5153	43.59018	39.7	76	37.7335	38.12749
65	53.5219	49.61407	45.1	77	46.6414	44.15139
66	51.0588	50.13796	49.8	78	44.9745	44.67527
67	36.3385	35.56408	35.6	79	30.4064	30.1014
68	28.1102	27.32574	31.7	80	23.3695	21.86306
69	37.8634	37.1913	38.9	81	32.5795	31.72862
70	48.564	46.74686	42.7	82	41.4296	41.28418
71	45.5274	43.71241	48.7	83	40.0506	38.24973
72	51.8424	50.87797	56.6	84	46.4898	45.41529

Table 2.5b: Composite forecasts for series 2 (S2).

Period	ARIMA Frc	Holt-W	Comb Frc	Actual V	Period	ARIMA Frc	Holt-W Frc	Comb Frc
61	58.9193	59.07518	60.77918	60.4	73	54.8645	53.6125	53.93793
62	48.6464	47.87074	48.66259	49.1	74	43.5099	42.40805	42.8805
63	45.7846	44.84463	45.4756	44	75	39.8202	39.38195	40.50419
64	43.5153	43.59018	45.21487	39.7	76	37.7335	38.12749	40.06465
65	53.5219	49.61407	47.33895	45.1	77	46.6414	44.15139	43.26458
66	51.0588	50.13796	50.78765	49.8	78	44.9745	44.67527	45.93364
67	36.3385	35.56408	36.35715	35.6	79	30.4064	30.1014	31.35412
68	28.1102	27.32574	28.10898	31.7	80	23.3695	21.86306	21.93935
69	37.8634	37.1913	38.08456	38.9	81	32.5795	31.72862	32.44682
70	48.564	46.74686	46.51892	42.7	82	41.4296	41.28418	42.69315
71	45.5274	43.71241	43.48657	48.7	83	40.0506	38.24973	38.03772
72	51.8424	50.87797	51.48499	56.6	84	46.4898	45.41529	45.91452

SS due to arima forecast errors = 174.2458

SS due to holt-w forecast errors = 135.654

SS due to combined forecast errors = 120.9757

Table 2.6b: **Adjusted composite estimates for 1985 (using information from quarterly totals).**

Period	Actual Q-T	Actual Q-Av	Forec. Q-T	Forec. Q-Av	Actual Val	Forecast-V	Adj.Estima
61	0	51.16667	0	51.63913	60.4	60.77918	60.22309
62	153.5	51.16667	154.9174	51.63913	49.1	48.66259	48.21737
63	0	51.1666	0	51.63913	44	45.4756	45.05954
64	0	44.86667	0	47.78049	39.7	45.21487	42.45751
65	134.6	44.86667	143.3415	47.78049	45.1	47.33895	44.45205
66	0	44.86667	0	47.78049	49.8	50.78765	47.69044
67	0	35.4	0	34.18356	35.6	36.35715	37.65094
68	106.2	35.4	102.5507	34.18356	31.7	28.10898	29.10925
69	0	35.4	0	34.18356	38.9	38.08456	39.43982
70	0	49.33333	0	47.16349	42.7	46.51892	48.65911
71	148	49.33333	141.4905	47.16349	48.7	43.48657	45.48724
72	0	49.33333	0	47.16349	56.6	51.48499	53.85365

Sum of squares due to forecasts errors= 120.9757
Sum of squares due to adjusted estimates errors= 78.9918

Table 2.7b: **Adjust the composite forecasts for 1984 (using of information from quarterly values)**

Period	Actual Q-T	Actual Q-Av	Forecast	Forec. Q-T	Forec. Q-V	Adj.Estim
73	0	44.76667	53.93793	0	45.7742	52.7507
74	134.3	44.76667	42.8805	137.3226	45.7742	41.93666
75	0	44.76667	40.50419	0	45.7742	39.61265
76	0	38.43333	40.06465	0	43.08763	35.7369
77	115.3	38.43333	43.26458	129.2629	43.08763	38.59118
78	0	38.43333	45.93364	0	43.08763	40.97192
79	0	35.7	31.35412	0	28.5801	39.16509
80	107.1	35.7	21.93935	85.74029	28.5801	27.4049
81	0	35.7	32.44682	0	28.5801	40.53001
82	0	46.1	42.69315	0	42.21513	46.62201
83	138.3	46.1	38.03772	126.6454	42.21513	41.53816
84	0	46.1	45.91452	0	42.21513	50.13983

Table 2.8: **Combined estimates of missing values for 1984 (series 1's and series 2's).**

Period	S1-Esti	S2-Esti	Final Esti
157	47.87612	50.13983	49.00798
158	46.55936	41.53816	44.04876
159	43.86452	46.62201	45.24326
160	39.70226	40.53001	40.11613
161	31.21119	27.4049	29.30804
162	36.18655	39.16509	37.67582
163	42.20965	40.97192	41.59078
164	38.48149	38.59118	38.53633
165	34.60886	35.7369	35.17288
166	41.57669	39.61265	40.59467
167	40.01876	41.93666	40.97771
168	52.70456	52.7507	52.72763

APPENDIX A3.

MODIFIED HOT-DECK IMPUTATION

Imputed (deleted) values for 1977.

Table 3.1: Using 1 value on either side of the value to be imputed.

Period	Actual-V	Hd - Esti.	Actual Q-T	Actu..Q- Av	Esti. Q-T	Est-. Q-Av	Adj.Esti.
73	35.6	27.25	0	27.46667	0	29.01667	25.79437
74	24.1	33.65	82.4	27.46667	87.05	29.01667	31.8525
75	22.7	26.15	0	27.46667	0	29.01667	24.75313
76	17.4	20.8	0	14.96667	0	17.5	17.78895
77	13.3	13.65	44.9	14.96667	52.5	17.5	11.674
78	14.2	18.05	0	14.96667	0	17.5	15.43705
79	25.8	29.5	0	21.06667	0	24.48333	25.38325
80	17.1	22.65	63.2	21.06667	73.45	24.48333	19.48918
81	20.3	21.3	0	21.06667	0	24.48333	18.32757
82	24	25.3	0	23.6	0	26.96667	22.14141
83	16.8	22.25	70.8	23.6	80.89999	26.96667	19.47219
84	30	33.35	0	23.6	0	26.96667	29.18641

Sum of squares due to hd-estimates errors = 287.44
Sum of squares due to adjusted hd-esti. errors = 185.8216

Table.3.2: **Using 2 values on either side of the value to be imputed.**

Period	Actual-V	Hd - Esti.	Actual Q-T	Actual Q-Av	Hd Esti.Q-T	Hd Est.. Q-Av	Adj.Estimate
73	35.6	27.875	0	27.46667	0	28.19167	27.15815
74	24.1	30.225	82.4	27.46667	84.575	28.19167	29.44771
75	22.7	26.475	0	27.46667	0	28.19167	25.79415
76	17.4	19.025	0	14.96667	0	16.69167	17.05886
77	13.3	13.025	44.9	14.96667	50.075	16.69167	11.67893
78	14.2	18.025	0	14.96667	0	16.69167	16.16221
79	25.8	28.8	0	21.06667	0	25.025	24.24455
80	17.1	25.7	63.2	21.06667	75.075	25.025	21.6349
81	20.3	20.575	0	21.06667	0	25.025	17.32055
82	24	24.125	0	23.6	0	26.6	21.40414
83	16.8	22.2	70.8	23.6	79.8	26.6	19.69624
84	30	33.475	0	23.6	0	26.6	29.69963

Sum of squares due to hd-estimates errors = 253.0755
Sum of squares due to adjusted hd-esti. errors = 163.1099

Table 3.3: **Using 6 values on either side of the value to be imputed.**

period	Actual-v	hd-esti.	actu. q-t	act.q-av	est.q-t	est.. q-av	adj.esti.
73	35.6	30.86667	0	27.46667	0	30.03889	28.22356
74	24.1	30.425	82.4	27.46667	90.11667	30.03889	27.81972
75	22.7	28.825	0	27.46667	0	30.03889	26.35673
76	17.4	22.05833	0	14.96667	0	19.33056	17.07865
77	13.3	15.975	44.9	14.96667	57.99167	19.33056	12.36863
78	14.2	19.95833	0	14.96667	0	19.33056	15.45272
79	25.8	30.09167	0	21.06667	0	26.88333	23.58082
80	17.1	27.28333	63.2	21.06667	80.65	26.88333	21.38012
81	20.3	23.275	0	21.06667	0	26.88333	18.23906
82	24	25.675	0	23.6	0	28.49444	21.26485
83	16.8	24.96667	70.8	23.6	85.48332	28.49444	20.67818
84	30	34.84166	0	23.6	0	28.49444	28.85697

Sum of squares due to hd-estimates errors = 385.8509
Sum of squares due to adjusted hd-esti. errors = 135.4793

Imputed (missing) values for 1984.

Table.3.4: **Using 2 values on either side of the value to be imputed.**

Period	Hd - Est.	Actual Q-T	Actual Q-Av	Est. Q-T	Est.. Q-Av	Adj.Est.
157	47.725	0	46.1	0	44.08333	49.90826
158	43.025	138.3	46.1	132.25	44.08333	44.99326
159	41.5	0	46.1	0	44.08333	43.39849
160	33.9	0	35.7	0	30.81667	39.27193
161	27.975	107.1	35.7	92.45	30.81667	32.40803
162	30.575	0	35.7	0	30.81667	35.42004
163	43.7	0	38.43333	0	39.09166	42.96406
164	36.475	115.3	38.43333	117.275	39.09166	35.86073
165	37.1	0	38.43333	0	39.09166	36.47521
166	40.6	0	44.76667	0	46.175	39.36171
167	43.325	134.3	44.76667	138.525	46.175	42.00359
168	54.6	0	44.76667	0	46.175	52.93471

Table.3.5: **Using 6 values on either side of the value to be imputed.**

Period	Hd - Esti.	Actual Q-T	Actual Q-Av	Est. Q-T	Est.. Q-Av	Adj.Esti.
157	48.025	0	46.1	0	45.79722	48.34251
158	44.825	138.3	46.1	137.3917	45.79722	45.12135
159	44.54167	0	46.1	0	45.79722	44.83615
160	37.54167	0	35.7	0	33.95278	39.47357
161	29.65	107.1	35.7	101.8583	33.95278	31.1758
162	34.66667	0	35.7	0	33.95278	36.45063
163	46.675	0	38.43333	0	44.60834	40.21392
164	46.65834	115.3	38.43333	133.825	44.60834	40.19956
165	40.49166	0	38.43333	0	44.60834	34.88652
166	42.75833	0	44.76667	0	47.24722	40.51345
167	44.775	134.3	44.76667	141.7417	47.24722	42.42424
168	54.20833	0	44.76667	0	47.24722	51.36231

APPENDIX A4.

Table 4.1: **Data for world tourists visiting Kenya (1971-1990. Missing values for 1984 having been imputed)**

	JAN.	FEB.	MAR.	APRI.	MAY.	JUN.	JUL.	AUG.	SEP.	OCT.	NOV.	DEC.
1971	27.6	25	24	17.8	14.7	18	29.8	30.8	24.3	22.8	23.7	34.5
1972	37	33.7	42	19.8	15.7	20.1	33.9	28.7	22.3	23.8	19.5	31.7
1973	27.7	25.8	22.8	21.9	13.5	21.5	28.4	27.8	20.2	23.4	20.6	38.2
1974	30.1	27.3	23.4	20.9	13.6	19.3	28.5	27.5	18.9	20.9	22.7	23.5
1975	29.8	26.8	29	14.5	14	17.9	29	30.8	21.1	25.5	20.4	32.3
1976	30.5	36.5	27.2	25.1	14	20.7	33.7	29.3	21.1	28.8	22.3	35.6
1977	35.6	24.1	22.7	17.4	13.3	14.2	25.8	17.1	20.3	24	16.8	30
1978	24	30.8	25.1	16.5	13.3	15.4	25.3	16	21.5	21.8	22.2	31.1
1979	27.2	26.8	24.6	20	10.8	18.1	27.2	26.7	18.6	20.4	23.9	34.9
1980	26.5	28.7	28.1	33	14.3	18	21.3	25.4	16.8	20.7	25.1	32.9
1981	31.5	30.7	33.1	22.7	21.8	23.7	32.5	31.8	38.7	31.9	32.8	33.9
1982	39.2	38.1	30.9	28.4	26.6	25.4	39.3	26.7	28.3	35.7	32.6	40.9
1983	39.3	34.9	35.7	24.1	19.6	21.4	32.2	25.9	27.5	32.4	33.8	48.6
1984	49.00798	44.04876	45.24326	40.11613	29.30804	37.67582	41.59078	38.53633	35.17288	40.59467	40.9771	52.72763
1984	48.34251	45.12135	44.83615	39.47358	31.17579	36.45063	40.21392	40.19957	34.88652	40.51345	42.42424	51.3623
1985	56.6	48.9	42.7	38.9	31.7	33.6	49.8	45.1	39.7	44	49.1	60.4
1986	55.8	50.4	56.7	44.2	34	41.9	53.5	48.2	52.9	50.3	57.8	68.5
1987	59.3	54.9	59	49.2	36.8	45.1	63.2	58	46.8	56.2	60.3	72.5
1988	69	61.7	60.4	45.8	42.3	47	67.8	68.2	63.4	60.8	49.3	59.3
1989	79	68.8	67.6	62	49.2	57	65.5	76.4	62.9	60.3	69.7	82
1990	68.9	63.4	70.6	65.7	55.4	69.4	82.5	111.5	68.9	78.6	80.7	85.5

-Composite estimates.

-Hot-deck estimates.

Table 4.2: **Forecasts for 1990 and eventual forecasts after 1990.**

Period	Sc-forec.	Sh-forec.	Actual		Period	Sc-forec.	Sh-forec.
229	78.4826	78.4379	68.9		241	88.41	88.367
230	73.9292	74.006	63.4		242	83.27	83.363
231	73.4554	73.4366	70.6		243	84.174	84.154
232	65.5854	65.5561	65.7		244	76.638	76.615
233	57.6517	57.814	55.4		245	68.052	68.214
234	62.7864	62.7195	69.4		246	74.813	74.741
235	74.9711	74.8866	82.5		247	87.367	87.257
236	73.2823	73.3968	111.5		248	91.046	90.989
237	68.6808	68.6674	68.9		249	79.975	79.942
238	70.5254	70.5277	78.6		250	82.951	82.921
239	71.5476	71.642	80.7		251	84.48	84.512
240	82.2969	82.2108	85.5		252	94.334	94.223

KENYATTA UNIVERSITY LIBRARY