Forecasting the water consumption using the ARIMA-model

Ploemen, E.M.M.

Award date:
1986

Link to publication

DEPARTMENT OF ELECTRICAL ENGINEERING
EINDHOVEN UNIVERSITY OF TECHNOLOGY
Group Measurement and Control

FORECASTING THE WATER CONSUMPTION USING THE ARIMA-MODEL

by E.M.M. Ploemen

This report is submitted in partial fulfillment of the requirements for the degree of electrical engineer (M.Sc.) at the Eindhoven University of Technology.
The work was carried out from nov. 1984 until dec. 1985
in charge of Prof.Dr.Ir. P. Eykhoff
under supervision of Dr.Ir. A.A.H. Damen
Dr.Ir. A.J.W. van den Boom

De afdeling der elektrotechniek van de Technische Hogeschool aanvaardt geen verantwoordelijkheid voor de inhoud van stage- en afstudeerverslagen.

Summary.

FORECASTING THE WATER CONSUMPTION USING THE ARIMA-MODEL.

This report gives a comprehensive survey of the methods, developed by Box and Jenkins, to build, identify, fit and diagnose ARIMA-models for stochastic time series.
Those methods are implemented on a VAX 11/750 computer. The software is written in standard FORTRAN 77.
Experiments with simulated data, in order to test the programs and to gain more insight in practical aspects using the methods, are discussed.
Finally the methods are applied to practical data. Model building and forecasting time series obtained from the "NV Tilburgsche" water-company will be described. The forecasting results are compared with results obtained from a so-called 'naive' method.

Samenvatting.

HET VOORSPELLEN VAN HET WATERVERBRUIK MET BEHULP VAN HET ARIMA-MODEL.

Dit verslag geeft een uitvoerig overzicht over de door Box en Jenkins ontwikkelde methoden om voor stochastische tijdreeksen ARIMA-modellen te bouwen, te identificeren, te schatten en te testen.
Deze methoden zijn geimplementeerd op een VAX 11/750 computer. De software is geschreven in standaard FORTRAN 77.
Experimenten met gesimuleerde data, uitgevoerd om de programma's te testen en om meer inzicht te verkrijgen in de praktische aspecten van het gebruik van deze methoden, zullen worden beschreven.
Tenslotte worden de methoden toegepast op praktijk data. Het modelbouwen en het voorspellen van tijdreeksen, beschikbaar gesteld door de NV Tilburgsche waterleiding-maatschappij, zullen uitgebreid worden behandeld. De voorspelresultaten worden vergeleken met de resultaten van een zogenaamde 'naieve' methode.

CONTENTS

INTRODUCTION

Many data in business, economic and engineering occur in the form of time series where the observations are dependent. Future values of such stochastic time series can only be described in terms of a probability distribution.
Forecasting of future values of a time series from current and past values can provide a basis for e.g. economic planning, production planning, inventory and production control, etc.
Planners and decision makers have a wide choice of ways to forecast, ranging from purely intuitive or judgemental approaches to highly structured and complex quantitative methods.

Yule proposed the idea that a stochastic time series usefully can be regarded as the output from a linear filter whose input is a series of independent observations, mostly called white noise. Based on this linear filter model, Box and Jenkins developed methods to build, identify, fit and diagnose the so-called ARIMA-models for stochastic time series. Once an adequate model has been found, it can be used to compute forecasts. Because those methods are rather close related to the work on system identification and parameter estimation done within the group Measurement and Control of the department of Electrical Engineering, they were chosen to be studied and to examine their use to forecast the water consumption. That is, in order to be able to plan and control the production of water in an optimal way, the "Tilburgsche" water-company needs to have forecasts of the consumption. A water-company has to satisfy the demand for water at each moment. Especially due the limited store-capacity for purified water the quality demands on the forecasts are rather severe.

So the goal of the work, described in this report, was twofold.
First : to study the methods developed by Box and Jenkins and to implement them on a VAX-computer.
Second to examine the usefulness and feasability of using the ARIMA-model to forecast the water consumption as a solution for the water-company problem.

In chapter 1 a survey of the methods to develop the ARIMA-model will be given.
In chapter 2 the model building using the implemented interactive programs will be described. Experiments with simulated data in order to test the programs will be discussed. Finally attention is paid to some practical aspects.
The results obtained by using the ARIMA-model in forecasting the water-consumption will be described in chapter 3. The data of 1983, obtained from the "Tilburgsche" water-company, is used for model building and forecasting. The forecasting results will be discussed in detail and will be compared with those obtained from a 'naive' method.
Some general conclusions are presented in chapter 4.

Chapter 1 FORECASTING BY USING THE ARIMA-MODEL.

1.1 Introduction.

The aim of this chapter is to introduce briefly the techniques described by Box and Jenkins [1]. They developed methods to build, identify, fit and diagnose models for time series and dynamic systems. Once an adequate model has been found, forecasts can be computed of future values of the series from current and past values.
Stochastic models and their forecasting will be discussed first; after that the stochastic model building section describes an iterative model building methodology whereby the stochastic models are related to actual time series. Finally seasonal models will be discussed.

1.2 Stochastic models and their forecasting.

1.2.1 Time series and stochastic models.

A time series is a set observations,generated sequentally in time. If the set is continuous, the time series is said to be continuous. If the set is discrete, the time series is said to be discrete. A discrete time series may arise by sampling a continuous time series or by accumulating a variable over a period of time, e.g. half-hourly demands of water.
We consider only discrete time series where observations are equispaced.
If future values of a time series can be exactly determind by some mathematical function, the time series is deterministic.
If future values can be described only in terms of a probability distribution, the series is non-deterministic or stochastic.
A stochastic process or probability model describes the probability structure of a statistical series. Such a series $z_1, z_2, \ldots\ldots, z_N$ of N observations is regarded as a sample realisation from an infinite population of such time series that could have been generated by the process.
An important class of stochastic models for describing time series is the so called class of stationary models, which assume that the process remains in equilibrium about a constant mean level. Usually a stationary time series can be usefully described by its mean, variance and auto-

correlation function.

A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin. Hence it appears that a stationary stochastic process has a constant mean

$$\mu = E[z_t] = \int_{-\infty}^{\infty} z.p(z)\ dz$$

which defines the level about which it fluctuates, and a constant variance

$$\sigma_z^2 = E[\ (z_t-\mu)^2\ ] = \int_{-\infty}^{\infty} (z-\mu)^2.p(z)\ dz$$

which measures its spread about this level.

Given N observations, the mean of the stochastic process can be estimated by :

$$\bar{z} = \frac{1}{N} \sum_{t=1}^{N} z_t \tag{1.1}$$

and the variance by :

$$\sigma_z^2 = \frac{1}{N} \sum_{t=1}^{N} (z_t - \bar{z})^2 \tag{1.2}$$

The covariance between z_t and z_{t+k}, separated by k intervals of time, is defined by :

$$\gamma_k = \text{cov}\ [z_t, z_{t+k}] = E\ [\ (z_t - \mu)(z_{t+k} - \mu)] \tag{1.3}$$

similarly the autocorrelation at lag k is

$$\rho_k = \frac{E[(z_t - \mu)(z_{t+k} - \mu)]}{\sqrt{E[(z_t-\mu)^2]\ E[(z_{t+k}-\mu)^2]}} = \frac{\gamma_k}{\sigma_z^2} \tag{1.4}$$

since for a stationary process, the variance $\sigma_z^2 = \gamma_0$ is the same at time t+k as at time t.

Thus, the autocorrelation at lag k is $\rho_k = \frac{\gamma_k}{\gamma_0}$.

The autocovariance matrix Γ_n and the autocorrelation matrix P_n associated with a stationary process for N observations are

$$\Gamma_n = \begin{vmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 \cdot & \cdot\ \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 \cdot & \cdot\ \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & & \vdots \\ \vdots & & & & \vdots \\ \gamma_{n-1} & \cdot & & & \gamma_0 \end{vmatrix} = \sigma_z^2 \begin{vmatrix} 1 & \rho_1 & \rho_2 & \cdot & \cdot & \cdot & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdot & \cdot & \cdot \\ \rho_2 & & & & & & \\ \vdots & & & & & & \\ \rho_{n-1} & \cdot & & & & & \cdot \end{vmatrix} = \sigma_z^2\ P_n$$

For a statonary process both matrices are positive-definite.

- Gaussian processes. If the probability distribution associated with any set of times is a multivariate Normal distribution, the process is called a Normal or Gaussian process. Since the multivariate Normal distribution is fully characterized by its moments of first and second orders, the existence of a fixed mean and an autocovariance matrix for all n, would be sufficient to ensure the stationarity of a Gaussian process.

Estimation of autocovariance and autocorrelation.

From N observations we can estimate the k´th lag autocorrelation by

$$r_k = \frac{c_k}{c_0} \qquad (1.5)$$

where c_k , the estimate of the autocovariance γ_k, is estimated by

$$c_k = \frac{1}{N}\sum_{t=1}^{N-k} (z_t - \bar{z})(z_{t+k} - \bar{z}) \qquad k= 0,1,2,\dots,K \qquad (1.6)$$

For a stationary Normal process the variance of the estimated autocorrelation coefficient is given by Bartlett [2] :

$$\mathrm{var}[r_k] \approx \frac{1}{N}\sum_{-\infty}^{\infty} \{ \rho_\nu^2 + \rho_{\nu+k}\rho_{\nu-k} - 4\rho_k\rho_\nu\rho_{\nu-k} + 2\rho_\nu^2\rho_k^2\} \qquad (1.7)$$

For any process for which ρ_ν are zero for $\nu > q$ Bartlett´s approximation gives for lags k greater than q :

$$\mathrm{var}[r_k] \approx \frac{1}{N}\{1+2\sum_{\nu=1}^{q} \rho_\nu^2 \} \qquad k > q \qquad (1.8)$$

In practice, in equation (1.8) the estimated autocorrelation r_k are substituted for the theoretical ρ_k.

1.2.2 Linear stationary models.

The stochastic models, which will be employed, are based on the idea (Yule [3]) that they can be usefully regarded as the output from a linear filter, whose input is white noise.

The models will be described in the time domain using the following operators :

- the backward shift operator B : $Bz_t = z_{t-1}$ and $B^m z_t = z_{t-m}$.

The inverse operation is performed by

- the forward shift operator F : $F = B^{-1}$ and $F^m z_t = z_{t+m}$.

Later on we will use

- the backward difference operator ∇ : $\nabla z_t = z_t - z_{t-1} = (1-B)z_t$.

and its inverse

- the summation operator S :

$$\begin{aligned} Sz_t = \nabla^{-1} z_t &= \sum_{j=0}^{\infty} z_{t-j} \\ &= z_t + z_{t-1} + z_{t-2} + \ldots \\ &= (1 + B + B^2 + \ldots) z_t \\ &= (1 - B)^{-1} z_t . \end{aligned}$$

1.2.2.1 The general linear process:

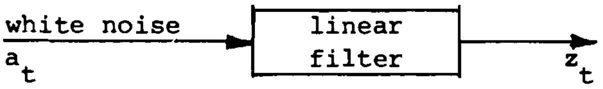


fig.1.1 representation of a time series as the output of a linear filter.

The white noise as input of the linear filter, as shown in fig.1.1, consists of a sequence of uncorrelated random variables $a_t, a_{t-1}, \ldots.$ (also called "shocks") with mean zero and constant variance. Usually the white noise is assumed to have a Normal distribution.

The linear filter tranforms the white noise process a_t to the process z_t. The linear filtering operation takes a weighted sum of present and previous values of the white noise process, that is

$$z_t = \mu + a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \ldots\ldots$$

$$\tilde{z}_t = z_t - \mu = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} = \psi(B)\, a_t \qquad (1.9)$$

where in general, μ is a parameter that determines the level of the process, and $\psi(B) = 1 + \psi_1 B + \psi_2 B^2 + \ldots\ldots\ldots$

$$= 1 + \sum_{j=1}^{\infty} \psi_j B^j = \sum_{j=0}^{\infty} \psi_j B^j \qquad \text{with } \psi_0 = 1$$

If the sequence $\psi_1, \psi_2, \psi_3, \ldots$ is finite or infinite, and such that the series $\psi(B)$ converges for all $\| B \| < 1$, that is on or within the unit circle, the filter is said to be stable and the process z_t to be stationary. Otherwise the process is non stationary.

The parameter μ is the mean about which the stationary process varies.

The model (1.9) implies that, under suitable conditions, $\tilde{z}_t$ is a weighted sum of past values of the $\tilde{z}_t$´s plus an added shock a_t, that is

$$\tilde{z}_t = \pi_1 \tilde{z}_{t-1} + \pi_2 \tilde{z}_{t-2} + \ldots\ldots\ldots + a_t$$

$$\tilde{z}_t = \sum_{j=1}^{\infty} \pi_j \tilde{z}_{t-j} + a_t \qquad (1.10)$$

Also $$(1 - \sum_{j=1}^{\infty} \pi_j B^j) \tilde{z}_t = a_t \quad ; \qquad \pi(B) \tilde{z}_t = a_t \qquad (1.11)$$

where $$\pi(B) = (1 - \sum_{j=1}^{\infty} \pi_j B^j).$$

From (1.9) and (1.11) we obtain $\psi(B).\pi(B) = 1$ (1.12)

The relationship (1.12) may be used to derive the π-weights, knowing the ψ-weights and vice versa.

If the π-weights are such that the series $\pi(B)$ converges for all $\|B\| \leqslant 1$, that is on or within the unit circle, the process is said to be invertible.

To illustrate the basis idea of invertibility, consider the model

$$z_t = (1 - \theta B) a_t \qquad (1.13)$$

note : From now on we will write z_t instead of $\tilde{z}_t$, except if it is necessary for understanding.

Whatever the the value of θ, (1.13) defines a stationary process.

Expressing the a_t´s in terms of the z_t´s, (1.13) becomes

$$a_t = (1 - \theta B)^{-1} z_t = \frac{1 - (\theta B)^{k+1}}{(1 - \theta B)} \cdot \frac{1}{1 - (\theta B)^{k+1}} z_t$$

$$a_t = (1 + \theta B + \theta^2 B^2 + \ldots. + \theta^k B^k)(1 - \theta^{k+1} B^{k+1})^{-1} z_t \qquad (1.14)$$

that is $$z_t = - \theta z_{t-1} - \theta^2 z_{t-2} - \ldots - \theta^k z_{t-k} + a_t - \theta^{k+1} a_{t-k-1}$$

and if $|\theta| < 1$, on letting k tend to infinity, we obtain the infinite series $$z_t = - \theta z_{t-1} - \theta^2 z_{t-2} - \ldots\ldots. + a_t \qquad (1.15)$$

and the π-weights of the model in the form of (1.10), are $\pi_j = - \theta^j$.

However, if $|\theta| \geqslant 1$, z_t in (1.14) depends on $z_{t-1}, z_{t-2}, \ldots, z_{t-k}$ which weights increase if k increases. We avoid this situation by requiring that $|\theta| < 1$. The series is then invertible. In general the series is invertible if $\pi(B)$ converges for all $\|B\| \leqslant 1$.

The representations (1.9) and (1.10) of the general linear process are not very useful in practice, because they contain an infinite number of parameters. We will now consider some special cases of the general linear process, which later turn out on to be very useful in practice.

1.2.2.2 The autoregressive process.

Consider the special case of (1.10) in which only the first p of the weights are nonzero, that is

$$z_t = \phi_1 z_{t-1} + \phi_2 z_{t-2} + \ldots\ldots\ldots + \phi_p z_{t-p} + a_t \qquad (1.16)$$

The process defined by (1.16) is called an autoregressive process of order p , or more succintly an AR(p)-process.
We can write (1.16) in the equivalent form

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots\ldots - \phi_p B^p) z_t = a_t$$

or $\phi(B)\ z_t = a_t$ (1.17)

(1.17) implies $z_t = \phi^{-1}(B)\ a_t$ (1.18)

$\phi(B)$ can be written as : $\phi(B) = (1-G_1 B)(1-G_2 B)\ldots\ldots(1-G_p B)$ (1.19)

where G_i^{-1} are the roots of $\phi(B) = 0$, which may be referred as the zero's of the polynomial $\phi(B)$.
Using (1.19) and expanding in partial fractions , (1.18) becomes

$$z_t = \phi^{-1}(B).a_t = \sum_{i=1}^{p} \frac{K_i}{(1-G_i B)} \ .\ a_t$$

The AR(p) process is stationary if $\psi(B) = \phi^{-1}(B)$ is a convergent series for $|B| < 1$. Hence it appears that we must have $|G_i| < 1$, $i=1,2,..,p$
So the stationary condition may be expressed by saying that the zero's of $\phi(B)$ must lie outside the unit circle.
Since $\pi(B) = \phi(B)$ is finite,no restrictions are required on the parameters of an AR(p) process to ensure invertibility.

The autocorrelation function (a.c.f) of an AR(p) process.

By taking the expected values of

$$z_{t-k}z_t = \phi_1 z_{t-k}z_{t-1} + \phi_2 z_{t-k}z_{t-2} + \dots + \phi_p z_{t-k}z_{t-p} + z_{t-k}a_t \qquad (1.20)$$

we obtain

$$\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \dots\dots + \phi_p\gamma_{k-p} \qquad k > 0$$

dividing by γ_0 :

$$\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \dots\dots + \phi_p\rho_{k-p} \qquad k > 0 \qquad (1.21)$$

We see that the a.c.f satisfies the equation $\phi(B).\rho_k = 0$.

Writing $\phi(B) = \prod_{i=1}^{p} (1 - G_i B)$, the general solution of (1.21) is

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \dots\dots\dots + A_p G_p^k \qquad (1.22)$$

where the G_i^{-1} are the roots of $\phi(B) = 0$.

Stationarity requires $|G_i| < 1$. If G_i is real, a term $A_i G_i^k$ decays geometrically to zero as k increases. If a pair of roots is complex they contribute a term $d^k \sin(2\pi fk+F)$ to the a.c.f. In general the a.c.f of a stationary AR(p) process will consist of a mixture of damped exponentials and damped sine waves.

If we substitute k = 1,2,....,p in (1.21), we obtain a set of linear equations for $\phi_1, \phi_2, \dots, \phi_p$ in terms of $\rho_1, \rho_2, \dots, \rho_p$, that is

$$\begin{array}{llll} \rho_1 = \phi_1 & + \phi_2\rho_1 & + \dots + & \phi_p\rho_{p-1} \\ \rho_2 = \phi_1\rho_1 & + \phi_2 & + \dots + & \phi_p\rho_{p-2} \\ \vdots & & & \vdots \\ \rho_p = \phi_1\rho_{p-1} & + \phi_2\rho_{p-2} & + \dots + & \phi_p \end{array} \qquad \text{or} \quad \underline{\rho}_p = P_p \underline{\phi}$$

These are usually called the Yule-Walker equations. We obtain Yule-Walker estimates of the parameters by replacing ρ_k by the estimated autocorrelations r_k , that is $\underline{\phi} = P_p^{-1} \underline{r}_p$.

When k = 0 in (1.20), by taking the expected values we obtain

$$\gamma_0 = \phi_1\gamma_{-1} + \phi_2\gamma_{-2} + \dots + \phi_p\gamma_{-p} + \sigma_a^2$$

On dividing throughout by $\gamma_0 = \sigma_z^2$, and substituting $\gamma_{-j} = \gamma_j$,the variance may be written :

$$\sigma_z^2 = \frac{\sigma_a^2}{1 - \rho_1\phi_1 - \dots\dots - \rho_p\phi_p}$$

The partial autocorrelation function (p.a.c.f).

The p.a.c.f is a device which exploits the fact that whereas an AR(p) process has an a.c.f which is infinite in extent, it can by its very nature be described in terms of p nonzero functions of the autocorrelations.

Denote by ϕ_{kj}, the j´th coefficient in an AR(k) model, so that ϕ_{kk} is the last coefficient. From (1.21) the ϕ_{kj} satisfy the set of equations :

$$\rho_j = \phi_{k1}\rho_{j-1} + \phi_{k2}\rho_{j-2} + \dots\dots + \phi_{kk}\rho_{j-k} \quad j=1,2,..,k \qquad (1.23)$$

Solving these Yule-Walker equations for $k = 1,2,3,...$

we obtain $\phi_{11}, \phi_{22}, \phi_{33}, \dots\dots\dots$

The quantity ϕ_{kk}, regarded as a function of lag k, is called the partial autocorrelation function. For an AR(p) process ϕ_{kk} will be nonzero for k less or equal to p and zero for k greater then p.

Estimation of the p.a.c.f.

By substituting estimates r_j for the theoretical autocorrelations in (1.23), and solving the equations by a recursive method (Durbin). It can be shown that for an AR(p) process, the estimates for $k > p+1$ are approximately independently distributed and, if n is the number of observations used in fitting, then : $\text{var}[\hat{\phi}_{kk}] \simeq \frac{1}{n} \quad k > p+1 \quad (1.24)$

1.2.2.3 The moving average process.

Consider the special case of (1.9) in which only the first q of the weights are nonzero, that is

$$z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots\dots - \theta_q a_{t-q} \qquad (1.25)$$

This process is called moving average process of order q, or more succintly a MA(q) process.

We can write (1.25) in the equivalent form

$$z_t = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q)a_t = \theta(B)\, a_t \qquad (1.26)$$

Since the series $\psi(B) = \theta(B)$ is finite, no restrictions are needed to ensure stationarity.

It can be shown that the invertibility condition is satisfied if the roots of the characteristic equation $\theta(B) = 0$ lie outside the unit circle.

The autocorrelation of an MA(q) process.

Using (1.26) the a.c.f of a MA(q) process is

$$\gamma_k = E[(a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q})(a_{t-k} - \theta_1 a_{t-k-1} - \ldots - \theta_q a_{t-k-q})]$$

Hence, the variance of the process is :

$$\gamma_0 = (1+ \theta_1^2 + \theta_2^2 + \ldots + \theta_q^2) \sigma_a^2 \qquad (1.27)$$

and

$$\gamma_k = \begin{cases} (-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \ldots + \theta_{q-k}\theta_q) \sigma_a^2 & k = 1,2,\ldots,q \\ 0 & k > q \end{cases}$$

thus

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \ldots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \ldots + \theta_q^2} & k = 1,2,\ldots,q \\ 0 & k > q \end{cases} \qquad (1.28)$$

We see that the a.c.f of a MA(q) process is zero, beyond the order q.

- If $\rho_1, \rho_2, \ldots, \rho_q$ are known, the q equations (1.28) may be solved for the parameters $\theta_1, \theta_2, \ldots, \theta_q$. However, these equations are non linear and have to be solved iteratively.

The partial autocorrelation of a MA(q) process.

Since a finite MA(q) process is equivalent to an infinite AR process its p.a.c.f is infinite in extent and is dominated by damped exponentials and/or damped sine waves.

1.2.2.4 The autoregressive-moving average process.

In practice, to obtain a parsimonius parameterization, it will sometimes be necessary to include both autoregressive and moving average terms in the model. Thus

$$z_t = \phi_1 z_{t-1} + \ldots + \phi_p z_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q} \qquad (1.29)$$

that is $(1- \phi_1 B - \ldots - \phi_p B^p) z_t = (1- \theta_1 B - \ldots - \theta_q B^q) a_t$

or $\phi(B) z_t = \Theta(B) a_t$ (1.30)

where $\phi(B)$ and $\theta(B)$ are polynomials of degree p and q in B.

We refer to this process as an ARMA(p,q) process.

(1.30) will define a stationary process if the roots of $\phi(B)=0$ lie outside the unit circle, and the process is invertible if the roots of $\theta(B)=0$ lie outside the unit circle.

The autocorrelation function of an ARMA(p,q) process.

On multiplying throughout in (1.29) by z_{t-k} and taking expectations, we see that the a.c.f satisfies the equation:

$$\gamma_k = \phi_1\gamma_{k-1} + \cdot\cdot + \phi_p\gamma_{k-p} + \gamma_{za}(k) - \theta_1\gamma_{za}(k-1) - \ldots - \theta_q\gamma_{za}(k-q) \qquad (1.31)$$

where $\gamma_{za}(k) = E[\, z_{t-k}.a_t]$.

Since z_{t-k} depends only on shocks which have occured up to time t-k, it follows that $\gamma_{za}(k) = 0 \quad k > q$ and $\gamma_{za}(k) = 0 \quad k < 0$

(1.31) implies : $\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + \ldots + \phi_p\gamma_{k-p} \qquad k > q+1$

and hence $\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + \ldots + \phi_p\rho_{k-p} \qquad k > q+1$

or $\phi(B)\,\rho_k = 0 \qquad k > q+1$

Thus, for an ARMA(p,q) process, the values $\rho_q, \rho_{q-1}, \ldots, \rho_1$ depend directly through (1.31) on the q MA-, as well on the p AR-parameters. Also the p values $\rho_q, \rho_{q-1}, \ldots, \rho_{q-p+1}$ provide the necessary starting values for the equation $\phi(B)\,\rho_k = 0 \quad k > q+1$ which then entirely determines the autocorrelations at higher lags. If $q-p < 0$, the whole a.c.f ρ_j, for j = 0,1,2,.... will consist of a mixture of damped exponentials and/or damped sine waves, whose nature is dictated by the polynomial $\phi(B)$ and the starting values.

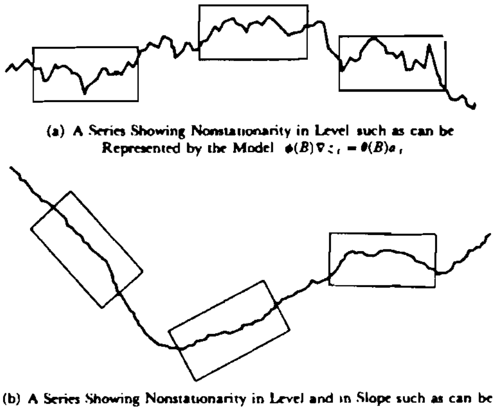
If, however, $q-p > 0$ there will be q-p+1 initial values which do not follow this general pattern. These facts are useful in identifying mixed series, as we will see later on.

The p.a.c.f of an ARMA(p,q) process.

The process (1.30) may be written : $a_t = \theta^{-1}(B)\,\phi(B)\,z_t$ and $\theta^{-1}(B)$ is an infinite series in B. Hence, the p.a.c.f of a mixed process is infinite in extent. After the first p-q lags it will be dominated by damped exponentials and/or sine waves.

1.2.3 Linear nonstationary models.

In the previous section we have seen that an ARMA process is stationary, if the roots of $\phi(B) = 0$ lie outside the unit circle. If one or more roots lie inside the unit circle the process exhibits explosive nonstationary behaviour. If, however, one of more roots lie on the unit circle, the process turns out to be of great value in representing so called homogeneous nonstationary time series. Such series behave as though they have no fixed mean; however they exhibit homogeneity in the sense that, apart from local level, or perhaps local level and trend, one part of the series behaves much like any other. Fig.1.2 shows two kinds of homogeneous nonstationary behaviour.



(a) A Series Showing Nonstationarity in Level such as can be Represented by the Model $\phi(B)\nabla z_t = \theta(B)a_t$

(b) A Series Showing Nonstationarity in Level and in Slope such as can be Represented by the Model $\phi(B)\nabla^2 z_t = \theta(B)a_t$

fig.1.2 Two kinds of homogeneous nonstationary behaviour.

Consider the model $$\varphi(B)\,\tilde{z}_t = \theta(B)\,a_t \qquad (1.32)$$

where $\varphi(B)$ is a nonstationary AR operator of order p+d with d roots equal to unity and the remainder outside the unit circle. Then we can express the model (1.32) in the form $\varphi(B)\,\tilde{z}_t = \phi(B)(\,1-B\,)^d\,\tilde{z}_t = \theta(B)\,a_t$ where $\phi(B)$ is a stationary AR operator of order p.

By using the backward difference operator, we can write

$$\phi(B)\,\nabla^d\,\tilde{z}_t = \theta(B)\,a_t \qquad (1.33)$$

where $$\nabla^d\,\tilde{z}_t = \nabla^d\,z_t \quad \text{for } d > 1\,.$$

Equivalently the process is defined by the two equations :

$$\phi(B)\,w_t = \theta(B)\,a_t \qquad (1.34)$$

$$w_t = \nabla^d\,z_t \qquad (1.35)$$

By using the summation operator (1.35) can be written as $z_t = S^d\,w_t$. This implies that the process (1.33) can be obtained by summing (or "integrating") the stationary process (1.34) d times. Therefore the process (1.33) is called, an autoregressive integrated moving average (ARIMA) process.

- General form of the ARIMA(p,d,q) model :

$$\varphi\,(B)\,z_t = \phi(B)\,\nabla^d\,z_t = \theta_0 + \theta(B)\,a_t \qquad (1.36)$$

where order of $\varphi(B)$ is p+d, order of $\phi(B)$ is p and order of $\theta(B)$ is q. By permitting θ_0 to be nonzero, a deterministic polynomial trend of degree d is included in the model.

- A widening of the range of useful applications of the model (1.36), is achieved by allowing the possiblity of some nonlinear transformation of z_t.

1.2.4 Forecasting with the ARIMA model.

We will now consider how the ARIMA model may be used to forecast future values of an observed time series.

With $\theta_0 = 0$, the general model (1.36) may be written in difference equation form :

$$z_t = \varphi_1 z_{t-1} + \, .. \, + \varphi_{p+d} z_{t-p-d} + a_t - \theta_1 a_{t-1} - \, .. \, - \theta_q a_{t-q} \qquad (1.37)$$

and in terms of current and previous shocks

$$z_t = \psi(B)\, a_t = \sum_{j=0}^{\infty} \psi_j a_{t-j} \qquad (1.38)$$

where $\psi_0 = 1$ and the ψ-weights may be obtained by equating

$$\varphi(B) . \psi(B) = \theta(B) \qquad (1.39)$$

We are concerned with forecasting a value $z_{t+\ell}$, $\ell > 1$, when we are currently standing at time t. This forecast is said to be made at origin t for leadtime ℓ.

Such an observation $z_{t+\ell}$, generated by the process, may be expressed by using (1.37):

$$z_{t+\ell} = \varphi_1 z_{t+\ell-1} + \, .. \, + \varphi_{p+d} z_{t+\ell-p-d} + a_{t+\ell} - \theta_1 a_{t+\ell-1} - \, .. \, - \theta_q a_{t+\ell-q} \qquad (1.40)$$

or by using (1.38) :

$$z_{t+\ell} = \sum_{j=0}^{\infty} \psi_j a_{t+\ell-j} \qquad \psi_0 = 1 \qquad (1.41)$$

Now suppose, we are to make a forecast $\hat{z}_t(\ell)$ of $z_{t+\ell}$ which is to be a linear function of current and previous observations $z_t, z_{t-1}, z_{t-2}, \ldots$. Then it will also be a linear function of current and previous shocks $a_t, a_{t-1}, a_{t-2}, \ldots$. Suppose then, that the best forecast is :

$$\hat{z}_t(\ell) = \psi_\ell^* a_t + \psi_{\ell+1}^* a_{t-1} + \psi_{\ell+2}^* a_{t-2} + \, \ldots..$$

Then, using (1.41), the mean square error of the forecast is

$$E[z_{t+\ell} - \hat{z}_t(\ell)]^2 = (1 + \psi_1^2 + \, .. \, + \psi_{\ell-1}^2)\, \sigma_a^2 + \sum_{j=0}^{\infty} (\psi_{\ell+j} - \psi_{\ell+j}^*)^2 \, \sigma_a^2 \qquad (1.42)$$

which is minimized by setting : $\psi_{\ell+j} = \psi_{\ell+j}^*$

We have then :

$$z_{t+\ell} = (a_{t+\ell} + \psi_1 a_{t+\ell-1} + \ldots + \psi_{\ell-1} a_{t+1}) + (\psi_\ell a_t + \psi_{\ell+1} a_{t-1} + \ldots) \qquad (1.43)$$

$$z_{t+\ell} = e_t(\ell) + \hat{z}_t(\ell) \qquad (1.44)$$

where $e_t(\ell)$ is the error of the forecast $z_t(\ell)$ at leadtime ℓ.

Denote by $\underset{t}{E}[\ z_{t+\ell}]$ the conditional expection of $z_{t+\ell}$ at time t,

that is $\underset{t}{E}[\ z_{t+\ell}] = E[\ z_{t+\ell} \mid z_t, z_{t-1}, z_{t-2}, \ldots\ldots]$

then certain important facts emerge :

- $$\hat{z}_t(\ell) = \psi_\ell a_t + \psi_{\ell-1} a_{t-1} + \ldots\ldots = \underset{t}{E}[\ z_{t+\ell}] \qquad (1.45)$$

  Thus, the minimum mean square error forecast at origin t, for leadtime ℓ is the conditional expection of $z_{t+\ell}$, at time t.

- The forecast error for lead time ℓ is :

  $$e_t(\ell) = a_{t+\ell} + \psi_1 a_{t+\ell+1} + \ldots + \psi_{\ell-1} a_{t+1} \qquad (1.46)$$

  Since $\underset{t}{E}[\ e_t(\ell)] = 0$, the forecast is unbaised.

  The variance of the forecast error is :

  $$V(\ell) = \mathrm{var}[\ e_t(\ell)\ ] = (\ 1 + \psi_1^2 + \psi_2^2 + \ldots . + \psi_{\ell-1}^2)\ \sigma_a^2 \qquad (1.47)$$

- The one step ahead forecast error is :

  $$e_t(1) = z_{t+1} - \hat{z}_t(1) = a_t$$

  Hence, the residuals a_t which generate the process, turn out to be the one step ahead errors.

  So, for a minimum mean square error forecast, the one step ahead forecast errors must be uncorrelated. As shown in [1] forecast errors for longer leadtimes in general will be correlated.

Calculation of the forecasts.

It is usually simplest in practice to compute the forecast directly from the difference equation (1.40), by taking the conditional expections, that is

$$\hat{z}_t(\ell) = \varphi_1 \underset{t}{E}[z_{t+\ell-1}] + .. + \varphi_{p+d} \underset{t}{E}[z_{t+\ell-p-d}] + \underset{t}{E}[a_{t+\ell}] - \theta_1 \underset{t}{E}[a_{t+\ell-1}] - .. - \theta_q \underset{t}{E}[a_{t+\ell-q}] \qquad (1.48)$$

where $\underset{t}{E}[z_{t-j}] = z_{t-j} \qquad j = 0,1,2,....$

$\underset{t}{E}[z_{t+j}] = \hat{z}_t(j) \qquad j = 0,1,2,....$

$\underset{t}{E}[a_{t-j}] = a_{t-j} = z_{t-j} - \hat{z}_{t-j-1}(1) \qquad j = 0,1,2,...$

$\underset{t}{E}[a_{t+j}] = 0 \qquad j = 0,1,2,...$

Probability limits of the forecasts.

The variance of the ℓ steps ahead forecast error is given by (1.47). Assuming that the a´s are Normal, it follows that, given information up to time t, the conditional probability distribution $p(z_{t+\ell} | z_t, z_{t-1}, ..)$ of a future value $z_{t+\ell}$ of the process will be Normal with mean $\hat{z}_t(\ell)$ and standard deviation $V(\ell)^{\frac{1}{2}} = \{ 1 + \sum_{j=1}^{\ell-1} \psi_j^2 \}^{\frac{1}{2}} \sigma_a$

Thus the limits for each desired level of probability ε , are

$$\hat{z}_{t+\ell}(\pm) = \hat{z}_t(\ell) \pm \mu_{\varepsilon/2}(1 + \sum_{j=1}^{\ell-1} \psi_j^2)^{\frac{1}{2}} \sigma_a \qquad (1.49)$$

where $\mu_{\varepsilon/2}$ is the deviate exceeded by a proportion $\varepsilon/2$ of the unit Normal distribution.

In practice σ_a is replaced by an estimate s_a , of the standard deviation of the white noise process.

Updating.

When a new deviation z_{t+1} comes at hand, the forecast may be updated to origin t+1 , by calculating the new forecast error $a_{t+1} = z_{t+1} - \hat{z}_t(1)$ and using the difference equation (1.48) with t+1 replacing t.

To provide more theoretical insight in the nature of the forecasts we may express them in another way.

Taking conditional expections at time t in (1.48) , we have, for $\ell > q$:

$$\hat{z}_t(\ell) - \varphi_1 \hat{z}_t(\ell-1) - \dots - \varphi_{p+d}\hat{z}_t(\ell-p-d) = 0 \qquad (1.50)$$

or $$\varphi(B)\ \hat{z}_t(\ell) = 0 \qquad (1.51)$$

where B operates on ℓ and $\hat{z}_t(-j) = z_{t-j}$ $j \geq 0$.

(1.50) has the solution :

$$\hat{z}_t(\ell) = b_0^{(t)} f_0(\ell) + b_1^{(t)} f_1(\ell) + \ldots + b_{p+d-1}^{(t)} f_{p+d-1}(\ell) \qquad \ell > q-p-d \qquad (1.52)$$

which is called the eventual forecast function ; "eventual" because when $q > p+d$, the function supplies the forecasts only for lead time $\ell > q-p-d$.

The function will pass through p+d "pivotal" values, which can consist of forecasts or actual values of the series depending on the value of q.

We see from (1.51) that it is the general AR operator $\varphi(B)$ which determines the mathematical form of the forecast function.

The MA operator is influential in determining how the function is to be "fitted" to the data and hence how the coefficients $b_0^{(t)},\ldots,b_{p+d-1}^{(t)}$ (constant for a given origin t) are to be calculated.

1.3 Stochastic model building.

We have seen that the general ARIMA(p,d,q) model provides a class of models for representing time series which, although not necessarily stationary, are homogeneous and in statistical equilibrium. Now we have to relate a model of this kind to the data. Usually this is best achieved by a three stage procedure based on identification, estimation and diagnostic checking.

Basic idea in model building is the principle of parsimony, that is we want to employ the smallest possible number of parameters for adequate representation.

1.3.1 Identification.

The aim is to identify an appropriate subclass of models from the general ARIMA model, which is worthy of further investigation. In other words to obtain some idea of the values of p,d and q needed to represent a given time series.

Principal tools for identification will be the autocorrelation and partial autocorrelation function.

Degree of differencing.

We have seen that for a stationary ARMA(p,q) process the a.c.f satisfies

$$\phi(B)\,\rho_k = 0 \qquad k > q$$

If $\phi(B) = \prod_{i=1}^{p} (1 - G_i B)$, the solution is of the form :

$$\rho_k = A_1 G_1^k + A_2 G_2^k + \ldots + A_p G_p^k \qquad k > q-p \qquad (1.53)$$

where $|G_i| < 1$ because of the stationarity requirement.

Inspection of (1.53) shows, that the a.c.f will quickly "die out" for moderate and large k.

Now suppose that a real root, say G_1 , approaches unity, so that $G_i = 1-\delta$, where δ is some small positive quantity. Then, since for k large $\rho_k \approx A_1(1 - k\delta)$, the a.c.f will not die out quickly and will fall off slowly and very nearly linearly. This tendency not to die out quickly, is taken as indication that a root close to unity may exist. It is assumed that the degree of differencing has been reached when the a.c.f of $w_t = \nabla^d z_t$ dies out fairly quickly.

The estimated a.c.f tends to follow the behaviour of the theoretical a.c.f however; it need not happen that the estimated correlations are extremely high even at low lags.

Identification of the resulting ARMA process.

The characteristic behaviour of the theoretical a.c.f and p.a.c.f for AR, MA and ARMA processes, as described in section 1.2 are summarized in table 1.3.

process	autocorrelation function	partial autocor. function
AR(p)	infinite, damped exponentials and/or damped sine waves.	$\phi_{kk} = 0 \qquad k > p$
MA(q)	$\rho_k = 0 \qquad k > q$	infinite, damped exponentials and/or damped sine waves.
ARMA(p,q)	infinite, dominated by damped exponentials and/or damped sine waves after first q-p lags.	infinite, dominated by damped exponentials and/or damped sine waves after first p-q lags.

table 1.3 Behaviour of theoretical a.c.f and p.a.c.f

Using this knowledge, we now study the general appearence of the estimated a.c and p.a.c functions of the (differenced) series to provide clues about the choice of the orders for the AR- and MA-operators.

- It should be noted here, that the estimated autocorrelations can have rather large variances and can be highly (auto)correlated with each other. In particular, moderately large estimated autocorrelations can occur after the theoretical function has damped out, and apparent ripples and trends can occur in the estimated function which have no basis in the theoretical function.

Initial estimates of the parameters.

After the identification of the model orders, initial estimates of the parameters will be made. As already denoted in section 1.2.2 , for a pure AR(p) process this can be done by solving the Yule-Walker equations. Also for a pure MA(q) process this can be done by solving the non linear equations (1.28) e.g. using a Newton-Raphson algorithm.

A general method for an ARMA(p,q) process is based on the first p+q+1 autocorrelations of $w_t = \nabla^d z_t$, and proceeds as follows :

i) Use $\phi(B)\, r_k = 0 \quad k > q+1$, where r_k is the estimated correlation, to obtain initial estimates $\hat{\underline{\phi}}$ for the ϕ parameters.

ii) By writing $$w_t' = \phi(B)\, w_t \qquad (1.54)$$ the process can be treated as an MA-process : $$w_t' = \theta(B)\, a_t \qquad (1.55)$$ Then the autocorrelations for the process (1.55) can be calculated from (1.54) and used in an iterative calculation to obtain initial estimates $\hat{\underline{\theta}}$ for the $\underline{\theta}$ parameters.

1.3.2 Estimation.

In the identification stage we obtained a tentative formulation for the model; now we need to obtain efficient estimates of the parameters. This is done by the maximum likelihood method.

The joint probability density function of N observations $\underline{z} = (z_1, z_2, \ldots\ldots, z_N)$, given by $p(\underline{z} | \underline{\beta})$, depends on the unknown parameters $\underline{\beta}$.

After we have obtained a set observations $\underline{z}$, the maximum likelihood estimation $\hat{\underline{\beta}}$ for $\underline{\beta}$ is that value of $\underline{\beta}$, which maximizes the likelihood function $L(\underline{\beta} | \underline{z})$.
$L(\underline{\beta} | \underline{z})$ is of the same form as $p(\underline{z} | \underline{\beta})$, in which now $\underline{z}$ is fixed and $\underline{\beta}$ is a variable.

Now, from N observations $\underline{z}$ of an ARIMA(p,d,q) model we can generate a series $\underline{w}$ of $n = N - d$ differences, where $w_t = \nabla^d z_t$.
Thus the general problem of fitting the parameters $\underline{\phi}$ and $\underline{\theta}$ of the ARIMA model is equivalent to fitting the w´s to the stationary, invertible ARMA(p,q) model which may be written :

$$a_t = \tilde{w}_t - \phi_1 \tilde{w}_{t-1} - \dots - \phi_p \tilde{w}_{t-p} + \theta_1 a_{t-1} + \dots + \theta_q a_{t-q} \qquad (1.56)$$

$$\text{where } \tilde{w}_t = w_t - \mu \qquad , \mu = E[w_t]$$

If $\mu \neq 0$, μ is included as an additional parameter θ_0 to be estimated.
The a´s cannot be calculated immediately from (1.56) because of the difficulty of starting up the difference equation. However, suppose that the p values $\underline{w}_*$ and the p values $\underline{a}_*$ prior to the commencement of the w series were given. Then a set values $a_t(\underline{\phi},\underline{\theta} | \underline{w}_*,\underline{a}_*,\underline{w})$ $t = 1,2,\dots,n$ could be calculated.
Assuming that the a´s are Normally distributed,

$$p(a_1,a_2,\dots,a_n) \propto \sigma_a^{-n} \exp\{ -(\sum_{t=1}^{n} \frac{a_t^2}{2\sigma_a^2})\}$$

Given a particular set of data $\underline{w}$, the log-likelihood associated with the parameter values $(\underline{\phi},\underline{\theta},\sigma_a)$, conditional on the choice of $(\underline{w}_*,\underline{a}_*)$ would then be

$$l_*(\underline{\phi},\underline{\theta},\sigma_a) = - n \ln \sigma_a - \frac{S_*(\underline{\phi},\underline{\theta})}{2 \sigma_a^2}$$

where :

$$S_*(\underline{\phi},\underline{\theta}) = \sum_{t=1}^{n} a_t^2(\underline{\phi},\underline{\theta} | \underline{w}_*,\underline{a}_*,\underline{w}) .$$

and the starscripts emphasize that the results are conditional to the choice of the starting values.
We notice that on the Normal assumption the maximum likelihood estimates are the same as the least squares estimates.

For some purposes a sufficient approximation to the unconditional likelihood which, strictly, is what we need for parameter estimation, is obtained by using the conditional likelihood with suitable values substituted for $\underline{w}_*$ and $\underline{a}_*$. For example by setting the $\underline{w}_*$´s and $\underline{a}_*$´s equal to their unconditional expections, that is $\underline{a}_* = 0$ and $\underline{w}_* = \mu$. Or to use (1.56) to calculate the a´s from a_{p+1} onwards, setting previous a´s equal to zero.
However, certainly for seasonal series, discussed in section 1.4 , the conditional approximation is not very satisfactory and the unconditional becomes more necessary.

It is shown by Box and Jenkins that, corresponding to the N = n+d observations, assumed to be generated by an ARIMA(p,d,q) model, the unconditional likelihood is given by :

$$L(\underline{\phi},\underline{\theta},\sigma_a | \underline{z}) = (2\pi\sigma_a^2)^{-n/2} . \left| M_n^{(p,q)} \right|^{\frac{1}{2}} . \exp\left\{ - \frac{S(\underline{\phi},\underline{\theta})}{2\,\sigma_a^2} \right\} \qquad (1.57)$$

where $\left| M_n^{(p,q)} \right|^{-1} . \sigma_a^2$ is the n×n covariance matrix of the w´s of the resulting ARMA(p,q) process,

and $$S(\underline{\phi},\underline{\theta}) = \sum_{t=-\infty}^{n} E[a_t | \underline{w}_n,\underline{\phi},\underline{\theta}]^2 \qquad (1.58)$$

So the log-likelihood is given by :

$$l(\underline{\phi},\underline{\theta},\sigma_a^2) = f(\underline{\phi},\underline{\theta}) - n \ln \sigma_a - \frac{S(\underline{\phi},\underline{\theta})}{2\,\sigma_a^2} \qquad (1.59)$$

Usually, $f(\underline{\phi},\underline{\theta})$ is of importance only for small n. For moderate and large values of n, (1.59) is dominated by $\frac{S(\underline{\phi},\underline{\theta})}{2\,\sigma_a^2}$. So it follows that minimizing the sum of squares (1.58) usually provide very close approximations to the likelihood estimates.

A procedure which can supply the unconditional sum of squares to any desired degree of approximation, for any ARIMA model is as follows.

The stationary forward model generating the w´s

$$\phi(B)\ \tilde{w}_t = \theta(B)\ a_t \qquad (1.60)$$

can be written as $\tilde{w}_t = \phi^{-1}(B)\ \theta(B)\ a_t = \sum_{j=0}^{\infty} a_{t-j}\psi_j$

that is $$\tilde{w}_t \approx \sum_{j=0}^{Q} \psi_j a_{t-j} \qquad (1.61)$$

because of the stationary character of the AR operator $\phi(B)$, the ψ-weights die out rather quickly, so we can approximate (1.60) to any desired accuracy by a MA(Q) process.

This implies that the $E[\ a_t]$ ´s are negligible beyond the point $t = -Q$, and the infinite summation (1.58) can be replaced by a finite sum from the point $t = 1 - Q$.

In the calculation of the sum of squares in practice, the $E[\ a_t]$ ´s are computed recursively by taking conditional expections in (1.60).

The values $E[\ -w_j]$, $j = 0,1,\ldots,Q-1$ are obtained by so called "backforecasting". Using the fact that the probability structure of $w_1,\ldots\ldots,w_n$ is equally explained by the forward model (1.60) , as by the backward model : $\phi(F)\ \tilde{w}_t = \theta(F)\ e_t$ (1.62)

The value w_{-i} thus bears exactly the same probability relationship to the sequence $w_1,w_2,\ldots\ldots,w_n$, as does the value w_{i+1} to the sequence $w_n,w_{n-1},\ldots\ldots,w_1$. So the values w_{-j} , $j = 0,1,\ldots,Q-1$ can be obtained by forecasting the reversed series.

Due to the MA parameters the $E[\ a_t]$ ´s are nonlinear functions of the parameters. Minimizing of the sum of squares wil take place using an iterative nonlinear least squares method, which is an extension of the Marquardt-algorithm. For a more comprehensive discussion of the estimation procedure, one is referred to Box and Jenkins.

1.3.3 Diagnostic checking.

Diagnostic checks are applied with the estimated model of uncovering possible lack of fit and diagnosing the cause. Diagnostic checks must be sensitive to discrepancies which are likely to happen. No system of diagnostic checks can ever be comprehensive. However, if checks, which have been thoughtfully devised, are applied to a model fitted to a reasonable large body of data and fail to show serious discrepancies, then we shall feel more comfortable about using that model.

Overfitting.

One technique which can be used is overfitting. That is, to estimate a more elaborate model containing additional parameters covering feared directions of discrepancy and to analyse whether the additions are needed. If the analysis fails to show that the additions are needed, e.g if their values are less or equal to one or two times their standard deviations, then it is only proved that we can not improve the model in that direction. However it does not prove that the model is correct.

Analysis of the residuals.

Procedures less dependent upon the knowledge of feared discrepancies are based on the analysis of residuals :

$$\hat{a}_t = \hat{\theta}^{-1}(B)\ \hat{\phi}(B)\ \nabla^d z_t$$

where ($\hat{\underline{\theta}},\hat{\underline{\phi}}$) are the obtained M.L estimates.

It is possible to show that, if the model is adequate :

$$\hat{a}_t = a_t + O\left(\frac{1}{\sqrt{n}}\right)$$

As the series length increases, the $\hat{a}_t$'s become close to the white noise a_t's. If the form of the model were correct and we knew the true parameters $\underline{\phi}$ and $\underline{\theta}$, then the estimated autocorrelations $r_k(a)$ of the a's would be uncorrelated and distributed approximately Normally about zero with a standard error of $n^{-½}$.

In practice, we can calculate the autocorrelations $r_k(\hat{a})$ of the $\hat{a}$'s. It can be shown that for moderate and high lags the use of $n^{-½}$ as standard error for $r_k(\hat{a})$ can be employed; for low lags, however, we may seriously under-estimate the significance of apparent discrepancies, because for low lags a reduction in variance can occur and at low lags the $r_k(\hat{a})$ can be highly correlated.

Rather than considering the $r_k(\hat{a})$'s individually, we can use the sum of the first K correlations, taken as a whole, to indicate inadequacy of the model. Box and Pierce showed that, if the fitted model is appropriate,

$$Q = n \sum_{k=1}^{K} r_k^2(\hat{a}) \qquad (1.63)$$

is approximately distributed as $\chi^2(K-p-q)$, where $n = N-d$ is the number of w´s used to fit the model. If the model is inappropriate, the average value of Q will be inflated. Therefore, an approximate or general test of the hypothesis of model adequacy may be made by referring an observed value of Q to a table of the percentage points of χ^2.

Cumulative periodogram check.

The power spectrum p(f) for white noise has a constant value $2\sigma_a^2$ over the frequency domain 0 - ½ cycles. Consequently, the cumulative spectrum for white noise :

$$P(f) = \int_0^f p(g)\, dg \tag{1.64}$$

plotted against f is a straight line running from (0,0) to ($\frac{1}{2},\sigma_a^2$)

For a time series a_t , t = 1,2,.... the periodogram $I(f_i)$ defined as

$$I(f_i) = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t \cos 2\pi f_i t \right)^2 + \left(\sum_{t=1}^{n} a_t \sin 2\pi f_i t \right)^2 \right] \tag{1.65}$$

where $f_i = \frac{i}{n}$ is the frequency,

provides an estimate of the power spectrum at frequency f.

Also $$P(f_j) = \frac{1}{n}\sum_{i=1}^{j} I(f_i) \tag{1.66}$$

provides an unbiased estimate of the integrated spectrum.

We shall refer to : $$C(f_j) = \frac{P(f_j)}{s^2} \tag{1.67}$$

where s is an estimate of σ_a as the normalized cumulative periodogram.

Now, if the model were adequate and the parameters known exactly, then the a´s could be computed from the data and would yield a white noise series. Model inadequacy would produce nonrandom a´s , whose cumulative peridogram could show systematic deviations from the straight line for white noise.

For large samples the periodogram for the estimated residual $\hat{a}$´s , which we have in practice, will have similar properties to that for the a´s. Thus inspection of the periodogram of the $\hat{a}$´s can provide a useful check, particularly for indicating periodicities inadequately taken account of.

Use of the residuals to modify the model.

Suppose that the residuals b_t from the model

$$\phi_0(B)\ \nabla^{d_0} z_t = \theta_0(B)\ b_t \qquad (1.68)$$

appear to be nonrandom.

We may now identify a model

$$\phi_1(B)\ \nabla^{d_1} b_t = \theta_1(B)\ a_t \qquad (1.69)$$

for the b_t series.

On eliminating b_t between (1.68) and (1.69) , we get a new model

$$\phi_0(B)\ \phi_1(B)\ \nabla^{d_0}\ \nabla^{d_1}\ z_t = \theta_0(B)\ \theta_1(B)\ a_t$$

which can now be fitted and diagnostically checked.

1.4 Seasonal models.

In general, we say that a series exhibits periodic behaviour with period s , when similarities in the series occur after s basic time intervals. A model, which can provide a useful representation of such series with remarkably few parameters is the general multiplicative seasonal model :

$$\phi_p(B)\ \Phi_P(B^s)\ \nabla^d\ \nabla_s^D\ z_t = \theta_q(B)\ \Theta_Q(B^s)\ a_t \qquad (1.70)$$

where : $\phi_p(B)$ and $\theta_q(B)$ are polynomials in B of order p and q.

$\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are polynomials in B^s of order P and Q.

and $\nabla = \nabla_1 = 1 - B$

$\nabla_s = 1 - B^s$, s is period in basic time intervals.

This model, called an $\underline{\text{ARIMA}(p,d,q)\times(P,D,Q)_s}$ model, will now be elucidated.

day:	Sun.	Mon.	Tues.	Wedn.	Thurs.	Frid.	Satur.
week : 1	z_1	z_2	z_3	z_4	z_5	z_6	z_7
2	z_8	z_9	z_{10}	z_{11}	z_{12}	z_{13}	z_{14}
3	z_{15}	z_{16}	z_{17}	z_{18}	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	z_t	.	.

table 1.4 Seasonal data.

The arrangement of the data in table 1.4 emphasizes that in periodic data there are two time intervals of importance.

For this example we expect relationships to occur

i) between observations for successive days in a particular week

ii) between the observations for the same day in successive weeks.

Suppose, as denoted, the t'th observation z_t is for a Thursday.

The relationship between z_t and previous observations for Thursdays t_{t-ks} , k = 1,2,..... can be described by an

ARIMA model : $\Phi_P(B^s)\ \nabla_s^D\ z_t = \Theta_Q(B^s)\ A_t$ (1.71)

Similary, a model : $\Phi_P(B^s)\ \nabla_s^D\ z_{t-1} = \Theta_Q(B^s)\ A_{t-1}$

might be used to link current behaviour for Wednesday with previous Wednesday observations, and so on, for each of the seven days. Moreover, it would usually be reasonable to assume that the parameters $\underline{\Phi}$ and $\underline{\Theta}$ contained in these seven models would be approximately the same for each day.

Now the error components $A_t, A_{t-1}, \ldots\ldots$ in these models would, in general, not be uncorrelated.

To take care of such relationships, a second model is introduced :

$$\phi_p(B)\ \nabla^d\ A_t = \theta_q(B)\ a_t \qquad (1.72)$$

where a_t is a white noise process.

Substituting (1.72) in (1.71) , we obtain the general multiplicative model (1.70).

One modification which is sometimes useful allows the mixed AR or MA operator to be non-multiplicative.

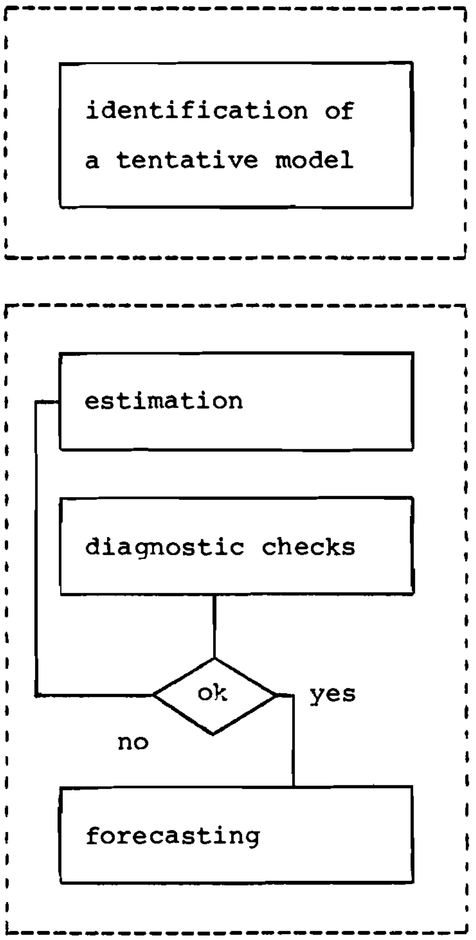
The extension to the seasonal model does not fundamentally alter the methods for identification, estimation and diagnostic checking as discussed in the previous section. This will also become more clear in the next chapters, where the described methods will be used with mainly seasonal data.

Chapter 2 MODEL BUILDING : INTERACTIVE COMPUTING AND EXAMPLES WITH SIMULATED DATA.

2.1 Introduction.

Based on the methods developed by Box and Jenkins, as described in the previous chapter, a number of interactive programs to perform the various steps in model building and forecasting were implemented on a VAX 11/750 computer. The programs will be discussed in detail in appendix V.
In this chapter these interactive model building procedures will be discussed and demonstrated with examples on simulated data. The aim of performing simulations is twofold. First, the examples can be seen as test of the programs. Second, one gains experience with the methods. Finally attention is paid to some practical aspects, which may be important, if using these methods on real data.

2.2 The model building procedure and examples.



program: DIFAC - differencing
- a.c.function
- p.a.c.function

program: MEP - preliminary est.
- estimation
- χ^2 test residuals
- forecasting

fig.2.1 Basic steps in model building.

The main steps in model building for a stochastic process are shown in the flow diagram in fig.2.1. Also the two main programs to perform these steps are denoted. As mentioned already, the programs will be discussed in detail in appendix V; here only their use in the model building procedure will be explained.

First step in identification is to plot the data and to examine on possible nonstationary behaviour.
As described in section 1.3.1 a tentative model has to be identified by studying the autocorrelation function (a.c.f) and the partial autocorrelation function (p.a.c.f).
The program DIFAC is used to perform the differencing and to compute the esimated a.c.f and the estimated p.a.c.f. The a.c.f and p.a.c.f are stored in files, so that plots can be generated. As standard error for the estimated a.c.f and p.a.c.f is used $n^{-\frac{1}{2}}$, where n = N-d-D and N is the number of observations of the series and d resp. D are the order of non-seasonal and seasonal differencing.
After deciding on the degree of differencing and deciding on a seasonal or a non-seasonal model, the orders of the remaining ARMA-model have to be determined. A useful aid in model identification non-seasonal models can be to compare the estimated a.c.f and p.a.c.f with plots of theoretical a.c.f and p.a.c.f, as given in appendix VI. For the seasonal case also in appendix VI the covariance structure for a number of seasonal models is given. However these lists are not comprehensive and it should be emphasized that in doubtful cases it is useful to determine several possible tentative models to use for estimation.
The next steps in model building are performed with the program MEP. Preliminary estimates of the parameters are obtained using the a.c. function of the (differenced) data. Initial estimates for the AR-part are generated by solving the Yule-Walker equations and initial estimates for the MA-part are obtained by a Newton-Raphson quadratic convergence procedure.
In the estimation routine an iterative modified Marquardt-algorithm is used, to fit simultaneously all parameters. This is a non linear least-squares fit. Parameter estimates are obtained together with appropriate standard errors. If the estimated parameters violate the stationary or

invertibility conditions the estimation program is stopped and the program types out a warning.

The residual series, which are also returned by the estimation routine, are used to perform diagnostic checking. The autocorrelation function of the residuals will be calculated and the Box-Pierce value Q together with the corresponding tail probability of the chi-square distribution is computed. The a.c.f of the residuals has to be examined i.e. by plotting on significant values. A value is significant if $r_k(\hat{a}) > 2n^{-\frac{1}{2}}$. These significant values indicate the remaining model of the residuals and so they suggest in what way the model should be modified.

Overfitting as diagnostic checking can be performed by running the estimation program with a more elaborate model.

Finally, if there is no doubt on the adequacy of the model left, it is ready to be used in forecasting. The forecasting routine uses the 'state' set as returned by the estimation routine. This state set contains the minimum amount of time series information needed to construct forecasts. Depending on the model-order that are samples of the differenced series, data to reconstitute the original series and samples of the residual series. If new observations come to hand this state set can be updated.

The program offers three possible ways of forecasting :

- the so called modelfitting. The model is used to predict the original time series used for estimation. This prediction will be performed as one step ahead with updating.
- one step ahead forecasting with updating.
- ℓ-steps ahead forecasting.

The model building procedure will now be demonstrated with some simulation examples used to test the programs.

example 1.

Two time series were generated by a non-seasonal stationary ARMA(2,0) model:

$$(1+ 0.4B -0.32B^2) z_t = a_t \qquad (2.1)$$

or written in difference equation form :

$$z_t = -0.4 z_{t-1} + 0.32 z_{t-2} + a_t$$

where : a_t is normally distributed with $\sigma_a = 1$ and $\mu_z = 0$.

Series 1a has N_a = 100 observations,

series 1b has N_b = 400 observations.

- note : In future one is referred to appandix II for the figures of chapter 2.

The a.c.f and p.a.c.f are shown in fig. 2.2. Also the ± 2 σ lines are drawn in the plots, where σ is approximated by the large lag standard-error : $n^{-\frac{1}{2}}$.

For time series 1b it is rather clear to identify the right model order from the a.c.f and p.a.c.f in fig. 2.2 , however this is not true for series 1a.

Table 2.1 shows the results of fitting an AR(2) model to the series.

series	N	ϕ_1 (σ)	ϕ_2 (σ)	res. var.	Q	df	ε
1a	100	- 0.436 (0.098)	0.264 (0,098)	1.14	11.8	13	0,544
1b	400	- 0.397 (0.048)	0.318 (0.048	1.01	28.4	35	0.697

table 2.1 Example 1 results of fitting an AR(2) model.

- note : Explanation of abbreviations used in the tables.

σ : estimated standard error of the estimates

res. var. : est. residual variance S_{min}/ N-p-q(-1 if c estimated) *)
where S_{min} is the final sum of squares (see page 25).

Q : Box-Pierce value $Q = n \sum_{k=1}^{K} r_k^2(\hat{a})$ (see page 27)

df : degrees of freedom df = K-p-q *)

ε : tail probability $\varepsilon = \Pr[\ \chi^2(df) > Q\ ]$

*) for seasonal models the degrees of freedom have to be reduced for the number of seasonal parameters.

A small value of the tail probability ε means that there is ground for questioning the randomness of the residuals.

From table 2.1 we see that for both series there is no doubt on the adequacy of the model. Also the plots of the a.c.f of the residuals in fig.2.3 show that there are no significant values.

If we compare the estimates of the parameters for both series, we notice that the results for time series 1b closely approximate the true parameter values. However this is not true for the series 1a; yet the estimates of the parameters are within in the 95 percent confidence interval, that is ± 1.96 σ.

Both from the estimated a.c.f and from the estimates of the parameters it is clearly shown that the series 1b hold more information about the generating model.

example 2.

A seasonal series was generated by an ARIMA (0,0,2)(1,0,0)7 model, that is a seasonal multiplicative model of order (0,0,2)(1,0,0)7 ; that is with a non-seasonal MA-part of order 2 and a seasonal AR-part of order 1. The period of seasonality is 7.

The model : $(1 - \Phi B^7)(z_t - \mu) = (1 - \theta_1 B - \theta_2 B^2) a_t$ (2.2)

or $(z_t - \mu) = \Phi(z_{t-7} - \mu) + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$

where : $\Phi = 0.92$

$\theta_1 = 0.60$, $\theta_2 = - 0.80$, $\mu = 40.0$, $\sigma_a^2 = 6.25$ and $N = 400$.

Fig. 2.4 shows the first 50 observations of the original series and of the first order seasonal difference at period 7 of the series.

Fig. 2.5 shows the estimated a.c.f for the original and for the differenced series. From those functions we see that is possible to identify either the orignal model of order (0,0,2)(1,0,0)7 as given by equation (2.2) or an alternative model of order (0,0,2)(0,1,0)7 ; that is with a seasonal differencing and no seasonal parameters as given by equation (2.2a).

$w_t = (1 - \theta_1 B - \theta_2 B^2) a_t$ where $w_t = \nabla_7 z_t$ (2.2a)

modelorder	θ_1 (σ)	θ_2 (σ)	c (σ)	Φ (σ)	res. var.	Q df	ε
(002)(100)7	0.596 (0.028)	- 0.834 (0.028)	40.2 2.1	0.936 (0.020)	6.76	27.0 30	0.62
(002)(010)7	0.597 (0.028)	- 0.832 (0.028)	—	—	7.03	29.5 30	0.49

table 2.2 Example 2 : estimation results.

The estimation results are shown in table 2.2. Fig. 2.6 shows the estimated a.c.f of the residuals for both models. We see that there is no doubt on the adequacy of both models. Looking at the residual variance, which is equal to the residual sum of squares divided by the degrees of freedom, the original model given by equation (2.2) should be preferred. However on real data there may be an advantage in employing the nonstationary model as given by equation (2.2a), because if Φ is near to unity, it is not really known whether the mean of the series has a meaning or not.

Many more examples were performed, but will not be described here. On the one hand to avoid confusion, on the other hand especially because they all showed, like the two examples given, that for a stationary and invertible generating model the estimated parameters closely approximate the true parameter values. So there is no doubt on the programs.

2.3 Practical aspects.

In this section attention is paid to some practical aspects in order to be alert to the difficulties that may occur and to avoid these difficulties. First near-redundant factors will be discussed, then attention is paid to the estimation of the mean of the series and to aspects of differencing.

Finally simulation examples will be discussed of which the generated data is disturbed by a deterministic trend, so that the model is outside the model-set.

Although these last experiments were done on account of an examination of the water-company data and on difficulties encountered during working with this data, they will be discussed here as it concerns general aspects.

2.3.1 Near redundant factors.

Instability in the parameter estimates can be expected in the estimation procedure if a model is fitted which contains a near redundant factor. In such a situation, combinations of parameter values yielding similar residuals and consequently similar likelihoods can be found. A change of an AR-parameter value can be nearly compensated by a suitable change of a MA-parameter. This will be illustrated in example 3.

Example 3.

Five time series were generated by the ARMA(2,1) model :

$$(1 + 0.40 B - 0.32 B^2) z_t = (1 - 0.50 B) a_t \qquad (2.3a)$$

with $\mu_z = 0$ and a_t normally distributed with $\sigma_a = 1$.

By factorizing the AR-part, this model can also be written as :

$$(1 - 0.40 B)(1 + 0.80 B) z_t = (1 - 0.50 B) a_t \qquad (2.3b)$$

So that the near cancellation factors (1 - 0.40 B) and (1 - 0.50 B) become clear.

The results of fitting an ARMA(2,1) model to the generated series is given in table 2.3.

series	N	ϕ_1 (σ)	ϕ_2 (σ)	θ (σ)	res. var.	Q	degree freed.	ε	numb.of iterat.
3a	200	- 0.32 (0.39)	0.30 (0.32)	0.56 (0.36)	1.09	16.0	22	0.82	20
3b	200	- 0.34 (0.17)	0.65 (0.16)	0.87 (0.14)	1.13	12.3	22	0.95	11
3c	200	- 0.66 (0.42)	0.13 (0.36)	0.34 (0.40)	0.99	20.8	22	0.53	3
3d	400	- 0.16 (0.25)	0.48 (0.21)	0.71 (0.23)	1.10	22.6	37	0.97	33
3e	400	- 0.05 (0.17)	0.61 (0.15)	0.83 (0.15)	1.01	38.7	37	0.39	14
true par.		- 0.40	0.32	0.50					

table 2.3 Example 3 : results of fitting an ARMA(2,1) model.

From table 2.3 we notice the big variety of the estimated values of the parameters, also the big values of the standard errors are remarkable. However, the chi-square tests gives no reason to doubt on the adequacy of the fitted models and none of the a.c.f of the residuals showed significant values.

If we rewrite the model of equation (2.3a) as an infinite AR-model, that is :

$$(1 - 0.40 B)(1 + 0.8 B) z_t = (1 - 0.50 B) a_t$$

$$(1 + 0.10 B + 0.05 B^2 + 0.025 B^3 + \dots)(1 + 0.8 B) z_t = a_t$$

$$(1 + 0.90 B + 0.09 B^2 + 0.045 B^3 + \dots) z_t = a_t$$

We see that this model can be very nearly approximated by an AR(1) model:

$$(1 - \phi B) z_t = a_t \quad \text{where } \phi \approx - 0.9 \qquad (2.4)$$

The results of fitting an AR(1) model to the series, as given in table 2.4 , confirm this statement.

series	θ (σ)	res.var.	Q	d.f.	prob.	num.iter.
3a	- 0.78 (0.045)	1.10	19.6	24	0.72	2
3b	- 0.84 (0.039)	1.13	15.6	24	0.90	2
3c	- 0.85 (0.037)	1.02	30.6	24	0.17	2
3d	- 0.80 (0.030)	1.11	27.0	39	0.93	2
3e	- 0.82 (0.029)	1.02	44.5	39	0.25	2

table 2.4 Example 3 : fitting an AR(1) model.

From table 2.4 we notice much more stability of the estimated parameters and much smaller values of the standard errors. All these estimations were performed in two iteration steps and except maybe series 3c there is no reason to doubt on this model.

From this example we see that the instability of the estimates by fitting an ARMA(2,1) model with 2 near redundant factors will occur, because we are trying to fit three parameters in a situation that can be represented by one parameter.

So we should avoid to fit ARMA-processes containing near common factors and we should be alert to the difficulties that can result. In general this can be done by using the identification and estimation procedures with care.

For the series in this example, as they have a small number of observations, their estimated a.c.f will hardly be distinguishable from that from data generated by an AR(1)-model. As we can see from fig. 2.7 , which shows the a.c.f for series 3d and for a series generated with an AR(1)-model with $\phi = - 0.8$.

So in the identification procedure corresponding to the parsimony principle an AR(1) model will be identificated for the ARMA(2,1) model, which will turn out to be adequate.

2.3.2 Estimation of the mean and differencing.

The estimation procedure fits a model of desired order to the series w_t where

$$w_t = \nabla^d \nabla_s^D z_t - c \qquad (2.5)$$

where the scalar c is the expected value of the differenced series and w_t is assumed to follow a zero-mean stationary ARMA model.

The scalar c can be estimated as a additional parameter or can be given a constant value.

One should take care of using the scalar in the right way, as now will be illustrated.

i)If the orders of differencing are zero, then $c = E[z_t] = \mu_z$.

Now, difficulties will appear if c is given a wrong value and so a bias remains in the series w_t. Due to this bias the fitted AR part of the estimated model will contain a zero for $B = 1$, so the fitted model is nonstationary and the program will stop, this is illustrated in example 4. Although this result suggests a differencing of the data, which will eliminate the bias, we should take care to do so, because the data need not to be nonstationary at all.

Example 4.

A series z_t was generated with the ARMA(1,1) model :

$$(1 - 0.9 B)(z_t - 55.0) = (1 + 0.6 B) a_t \qquad (2.6)$$

The results of fitting an ARIMA(1,0,1) model to the series $w_t = z_t - c$ with c as an additional parameter to be estimated were:

$\phi = 0.88 \pm 0.03$

$\theta = -0.60 \pm 0.04$

$c = 55.1 \pm 0.7$

and if c was taken a constant value equal to $\mu_z = 55.0$ then the results were identical, that is $\phi = 0.88 \pm 0.03$

$\theta = -0.60 \pm 0.04$

Both results were obtained after 5 iterations.

However, during fitting an ARMA(1,1) model with c = 0 the program stopped after 34 iterations and typed out a warning that the estimated parameters : $\phi = 1.0$ and $\theta = -0.578$ violate the stationary conditions.

A similar result was obtained after 33 iterations fitting an ARMA(2,1) model with c = 0. The final estimated parameters were : $\phi_1 = 0.4983$

$\phi_2 = 0.5117$

$\theta = -0.9418$

If we factorize the AR part, that is :

$$1 - 0.4883 B - 0.5117 B^2 = (1 - B)(1 + 0.5117 B)$$

the zero for B = 1 is clearly seen.

ii) If d and/or D are nonzero, then by permitting $c = \mu_w$ to be nonzero a deterministic trend is included in the model. However, in many applications, where no physical reason for a deterministic component exists, the mean of w can be assumed to be zero unless such an assumption is contrary to facts presented by the data.

Differencing or not in doubtful cases.

As we saw in example 2, if a root of the AR operator is rather near to unity, it is almost equally well to use a stationary model with a root near to unity or to use a nonstationary model, that is to perform a difference on the series. On real data, however, it may be an advantage to employ the nonstationary model, which does not include a mean μ because it is not really known wether the mean of the series has a meaning or not. Using the nonstationary model to produce forecasts, these forecasts will not in any way depend on an estimated mean, calculated from a previous period, which may have no relevance to the future level of the series.

Wrongly differencing.

It may be helpful to realize the effects of wrongly differencing, for instance if it is not clear whether a seasonal or nonseasonal differencing should be used. To illustrate consider the simple model :

$$(1 - \phi B) z_t = a_t$$

Nonseasonal differencing yields :

$$w_t = \frac{1 - B}{1 - \phi B} a_t = \left(1 - (1-\phi)B - (1-\phi)\phi B^2 - (1-\phi)\phi^2 B^3 - \ldots . \right) a_t$$

We now notice that if ϕ is near to unity and thus $(1-\phi)$ near to zero, the result of the justly differencing is : $w_t \approx a_t$

However if ϕ is near to zero the result is : $w_t \approx (1-B) a_t$. So the process w_t is non invertible.

For values of ϕ between 0 and 1 , the resulting process w_t can be approximated by a suitably chosen q-order MA process, because for higher order $(1- \phi) \phi^k$ $k = q+1, q+2, \ldots .$ will be negligible. However because of the superfluous differencing we violate the parsimony principle.

2.3.3 Data disturbed by a deterministic trend.

The watercompany time series of the daily consumption of water exhibits not only an evident weekly periodicity but also an additional, however relative small, yearly season, as we can see from fig. 2.8. That is the consumption in winter is lower than in summer; of course there is a gradual change-over.

Theoretically it is possible, using similar arguments as in section 1.4, to obtain multiplicative models with three or more periodic components to take care of multiple seasonalities.

However such an extension of the program was abandoned because of the availibility of only one year data and also lack of time.

From an other point of view, this yearly trend can be seen as an additional deterministic trend.

In order to gain insight of the possible aspects of such an deterministic trend, several experiments were done, which will be discussed here.

example 5.

Time series $z5_t$ was generated by the stationary seasonal model of order (2,0,0)(1,0,0)7 with : $\phi_1 = 0.6$

$\phi_2 = -0.4$

$\Phi = 0.8$, $\sigma_a = 2.5$, $N = 364$ and

the mean of the series $\mu_{z5} = 32.0$

As deterministic trend was a sine wave added, in two slightly different ways, that is :

trend 1 : each data sample was increased as follows :

$$z5_t := z5_t + A.\sin\left(\pi \frac{t-1}{364} \right) \qquad t = 1,2,\ldots.,364$$

yielding the series: T1a with A = 4.0
T1b with A = 8.0

trend 2 : each sample of the same week was increased with the same value :

$$z5_t := z5_t + A.\sin\left(\pi \frac{k}{52} \right)$$ where k equals the number of integers of (t-1)/7 $t = 1,2,\ldots.,364$.

yielding the series : T2a with A = 4.0
T2b with A = 8.0

The a.c.f of the series hardly show a difference, as is illustrated in fig. 2.9 which shows the a.c.f of the series $z5_t$, T1a and T2b.

A (2,0,0)(1,0,0)7 model was fitted to the series and in accordance with the practically similar a.c.f also the estimated parameters do not differ much, as is shown in table 2.5.

series	ϕ_1 (0.6)	ϕ_2 (-0.4)	Φ (0.8)	c	(σ)
z5	0.671	- 0.446	0.781	32.2	0.8
T1a	0.673	- 0.436	0.793	34.4	0.8
T1b	0.679	- 0.414	0.824	36.5	1.0
T2a	0.673	- 0.436	0.793	34.4	0.8
T2b	0.679	- 0.414	0.824	36.5	1.0
	for all estimates (σ) $<$ 0.05				

Table 2.5 Example 5,results fitting an (2,0,0)(1,0,0)7 model.

Also the a.c.f of the residual series showed no significant values for all series.

We note that the estimated mean values c are roughly equal to $32 + 2A/\pi$, where $2A/\pi$ is the mean of the trend.

Example 6.

Time series z6a and z6b were generated by the the stationary seasonal model of order (2,0,0)(1,0,0)7 with : $\phi_1 = 0.7$
$\phi_2 = -0.3$
$\Phi = 0.6$ and mean $\mu = 20.0$
series z6a $\sigma_a = 10.0$
series z6b $\sigma_a = 1.0$

To both series a trend, similar to trend 1 of example 5 now with A = 10.0 was added.

Fig. 2.10 shows the first 182 samples of the series. We see that for series z6b with $\sigma_a = 1.0$ and A = 10.0 the trend clearly dominates.

Fig. 2.11 shows the a.c.f for both series. The dominating trend of series z6b finds expression in the high correlation coefficients for the series.

Table 2.6 shows the results of fitting a (2,0,0)(1,0,0)7 model to the series.

series	ϕ_1 (0.7)	ϕ_2 (-0.3)	Φ (0.6)	c	(σ)
z6a	0.749	- 0.331	0.550	27.3	1.9
z6b	0.715	- 0.284	0.905	25.5	1.1
	for all estimates (σ) $<$ 0.05				

Table 2.6 Example 6, results fitting an (2,0,0)(1,0,0)7 model.

Where this model was found to be adequate for series z6a : Q = 23.9 with 31 degrees of freedom so that the tail probablity ε = 0.82 , the a.c.f of the residuals of series z6b showed several significant values and the chi-square test Q = 83.8 with d.f = 31 , ε = 0.0 confirms the doubt on the adequacy.

For series z6b the estimated value of Φ is near to unity, therefore a model (2,0,0)(0,1,1)7 , thus with seasonal differencing of order one, was fitted. This resulted to : $\phi_1 = 0.720$

$$\phi_2 = -0.302$$

$$\Phi = 0.213$$

However, the a.c.f of the residuals still showed significant values and also Q = 93.1 with d.f = 31 , ε = 0.0 gives reason to doubt on the adequacy of this model.

From both fitted models for series z6b, we remark that the nonseasonal parameters seem not to be affected by the relative big trend, however the estimation of the seasonal part fails.

Several other multiplicative models with different orders of seasonal and/or nonseasonal differencing, turned out to be inadequate to fit the series. Finally a non-multiplicative AR model of order 9 was fit :

$\phi_1 = 0.771$	$\phi_4 = -0.051$	$\phi_7 = 0.531$	for all estimates
$\phi_2 = -0.389$	$\phi_5 = 0.109$	$\phi_8 = -0.335$	(σ) $<$ 0.05 ;
$\phi_3 = 0.010$	$\phi_6 = 0.096$	$\phi_9 = 0.166$	c = 21.4 ± 7.9

The a.c.f of the residuals showed no significant values and also Q = 31.5 with d.f = 25 , ε = 0.18 gives no doubt on the adequacy of the fitted model.

Example 7.

The supposed yearly trend in the water company time series was approximated by fitting an second order polynomial with (least squares) coefficients to the series minus the mean. See fig 2.8 and also fig. 2.12 which shows the approximated trend.

This trend was added to two time series generated with the same model as used in example 5, however now with μ = 3200 and different noise,

that is : σ_a= 250 yielding series : z7a without and T7a with trend.
σ_a= 50 series : z7b without and T7b with trend.

Fig 2.13 shows the a.c.f for series z7a and T7b, where the a.c.f for series T7a and z7b are similar to that for series z7a. We notice for series T7b that the samples are high correlated.

The results of fitting an (2,0,0)(1,0,0)7 model to the series, are shown in table 2.7.

series	ϕ_1 (0.6)	ϕ_2 (-0.4)	Φ (0.8)	c	(σ)	Q	d.f	ε
z7a	0.645	- 0.420	0.752	3224	65	23.4	31	0.83
T7a	0.649	- 0.414	0.760	3201	69	23.4	31	0.83
z7b	0.645	- 0.420	0.752	3205	13	23.4	31	0.83
T7b	0.692	- 0.352	0.892	3150	35	46.2	31	0.04
	all estimates (σ) $<$ 0.050							

Table 2.7 Example 7, results fitting an (2,0,0)(1,0,0)7 model.

From table 2.7 we see that for series T7a the trend does not affect the estimation of the parameters, however this is not true for series T7b. Also the chi-square check gives reason to doubt on this model for series T7b. Just like in example 6, it was difficult to fit an adequate model to series T7b. For a model with a second order seasonal differencing and a nonseasonal AR(2), that is (200)(P2Q)7 we noticed, that the estimated nonseasonal parametervalues 0.63 and -0.41 approximate closely to the true values, however no adequate seasonal model could be fitted.

Based on the chi-square test, the best however still doubtful model fitted was a (200)(210) model : ϕ_1= 0.68 , ϕ_2= -0.38

Φ_1= -0.07 , Φ_2= -0.09

where for estimates $\sigma < 0.06$, we can doubt on the seasonal parameters. Chi-square value Q(31) = 40.1 so that ε = 0.13.

From example 5 and 6 we saw, that adding a relatively small sine wave trend to a series does not affect either the a.c.f or the estimation of the model, used to generate the series. By a relatively small trend is meant the magnitude of the sine wave in proportion to the standard deviation of the white noise. A relatively small trend is almost buried by the randomness of the series and will be hard to distinguish.

However, from the series 6b we saw, that a relatively big and dominating trend affects the a.c.f as well as the estimation. The dominating trend finds its expression in high correlation coefficients and so the identification of the model-order is obstructed. Also the finally adequate model fitted to the series differs from the generating model.

Example 6 also shows that differencing, in contrary to data showing a stochastic trend, need not to be an improvement in identification and estimation.

Example 7 confirm these statements, here a second order trend fitted to practically data was used to disturb the time series.

Although the experiment is too small to prove the observed behaviour, it can be very useful to know the effect a deterministic trend might have on the a.c.f, the identification and finally the estimation of an adequate model. So, since for practical data, the variance of the noise is not known, it might be possible to get an indication from the a.c.f whether it is affected by a supposed trend or not.

Chapter 3 THE WATER-COMPANY DATA : MODEL BUILDING AND FORECASTING.

3.1 Introduction.

After implementing the methods and using them with simulated data, as described in the previous chapters, we will now apply them to practical data. As mentioned already in the introduction of this report the "Tilburgsche" water-company made available the half-hourly readings of the water consumption for one year, that is 1983.
In this chapter we will first briefly elucidate the background of the need for forecasts by the water-company, and also discuss the data more comprehensively.
Then the trials and errors leading to the final models, fitted to the data, will be described.
Next forecasting results with these models will be evaluated and compared with forecasts produced by a so called 'naive-method'. Also a cross validation will be done with data of 1984, which meanwhile had become available.
Finally conclusions will be given on the results and on the usefulness of these methods as a solution for the water-company problem.
- note : One is refered to appendix III for the figures of chapter 3.

3.2 The water-company.

3.2.1 The need for forecasts.

The water-company has to satisfy the demand for water at each moment. The water has to meet certain requirements like sufficiently high pressure and quality restrictions.
To meet the quality restrictions the spring-water has to be purified by, among other things, filtering, etc. The production process should be equable, to ensure a good and constant quality of the water.
A water-tower and high-pressure pumps are used to provide the sufficient high pressure.
In order to avoid technical trouble, one should take care of limiting the on and off switching of the spring-water-, filtering- and high-pressure pumps. Also from the point of view of efficiency a peak-consumption of electricity, by switching on too many pumps simultaneous, should be avoided. As a high peak-consumption increases the costs of electricity.

For an average day the total consumption of water exceeds the capacity of the water-tower and the store-basin for purified water. For the data of 1983 fig. 3.1 shows the graphs of the average daily consumption of water for the various days of the week.

During the daytime the consumption will be greater than the production of water, and the stock of water will decrease. This decrease is limited due the quality requirements, especially due the requirement to satisfy the demand for water at each moment. At night the stock will be filled up again.

So, the production of water has to be controlled in such a way that at all times a deficiency of water should be avoided. Also a too high, short time, rapid increase of the production on account of a threating deficiency should be avoided.

Now, it will be clear that it is very important to have a good forecast of the water consumption, in order to control the production process in an optimal way, that is to provide an equable production in the most efficient way.

Due to the half-hourly readings of the water-consumption it is possible to interfere with the production process at each half hour. This, however, might lead to unnecessary switching of the pumps. Considering the average daily pattern, a good and reliable forecast of the total daily consumption might be sufficient for efficient control. Therefore we have chosen to examine the possiblity to get a good forecast of the total daily consumption by the Box-Jenkins methods, as will be described in the next sections.

For the sake of completeness we will mention the way how the water-company forecasted the water-consumption at the time the data became available. Then they used the daily pattern of the corresponding day of the previous week as starting point. During the day they adjusted their prognosis to the actual consumption. Also weather changes were taken into account immediatly. In spite of this adjusting the total daily consumption differed considerably from the total final prognosis.

3.2.2 The water-company data.

The daily totals of the water-company data are listed in appendix IV. It is recommendable to examine the data carefully, before starting the identification and estimation procedures, in order to gain as much insight and information about the data as possible. As it concerns the water-company data, we can ask ourselves by what circumstances the demand for water will be influenced.

First of all we recognize the kinds of and the ratio between the various consumers, as there are for example industries, (pig)-farmers and private persons. However, such data can not be measured directly and we will leave them out of consideration unless there is a strong indication of a change from the data.

In the second place we notice the weather, that is the temperature and the quantity of rainfall, which can both be measured directly. Especially in spring and summer this might influence the use by private persons. The question whether it is necessary to use these data, will be dealt with later on.

It is also known that on feast- and other holidays like Carnival, Easter, Ascension day, Whitsuntide, Christmas and perhaps other days the consumption is considerably lower and roughly equal to the consumption on Sundays.

Fig. 3.2 shows the graph of the total daily consumption for 1983, starting at Sunday 2 January, and fig. 3.3 gives a more clear plot of the first 8 weeks of the same data.

From these two plots we notice :

- A rather regular pattern with a periodicity of 7; of course this seasonal character comes up to ones expections. The pattern looks so regular because the consumption on Sun- and Saturdays are considerably lower than the other days and,on the contrary, the consumption on Monday, the traditional washing-day, is higher.
- Except a period of about 10 weeks at the time of the summer-holidays the data looks very stationary. During that summer-period the series appears to be nonstationary in the level.
- As mentioned already in section 2.3 fig. 3.2 suggests an additional, small, yearly trend; that is in springtime the consumption gradually increases and it decreases in autumn.
- In fig. 3.3 an example of the lower consumption during holidays can be seen. Samples 44 and 45, that is Carnival Monday and Tuesday, the consumption is considerably lower in comparison to other Mondays and Tuesdays.

From fig. 3.1 we see that the average daily pattern for the various days are quite similar, except the levels. Also we notice a time shift as it concerns Saturday and Sunday-night.

These average daily patterns were calculated without making allowance for the deviations on holidays.

A graph of all Wednesdays, as shown in fig. 3.4 gives insight into the spread about the average Wednes-pattern. Fig. 3.5 shows the graph of the total daily consumption of these 52 Wednesdays and the mean of the daily-totals for Wednesday and the whole year. From these two plots we notice :

- Most Wednesdays follow the average pattern within a rather small deviation and also the daily-total of these days do not differ much from the mean value.
- As could be expected, the days which show the greatest deviations from the average pattern are the same whose daily-total differs relatively much from the mean total. These effects concern mainly holidays or days within the summer-period.
- The peak-days show a rather big deviation in the evening.

The weather influence.

The maximum temperature and the quantity of rainfall for each day of 1983 are also available. Using these we might get more insight about the weather influence on the consumption, especially for the summer-period, because the stationary pattern for the other parts of the year give no reason to suppose a great influence.

For the summer-period, fig. 3.6 shows the graphs of the temperature, of the daily totals, and of the deviations from the specific daily-means, that is Sundays, etc.

During the summer-period there was very little rainfall during only a few days.

From fig. 3.6 some influence from the temperature on the consumption can be noticed. For instance if we consider the first two weeks then it is very likely that the rather big difference is due the big rise of temperature. However, a much more important fact which emerges from this figure, is the influence of the summer-holidays. During this period, if many factories etc. are for some time and many people are away on holiday, we notice that the level of the consumption is considerably lower in spite of high temperatures. In this period also the regularly pattern fades away.

Summing up we can conclude that certainly during the summer-period the temperature influences the water-consumption. During a great part of this period, the summer-holidays, the consumption will be more influenced by nonmeasurable causes like changes of kind and number of users.

3.3 The way leading to the final models fit to the data.

In this section the way leading to the adequate models, fitted to the data, will be described.

Although these models, and especially their forecasts-results are our main goal, as illustration of the use of the methods on practical data, it is worthwhile and interesting to know what problems had to solved to attain this results. To avoid confusion and to avoid that the reader gets buried under a torrent of tables, figures and results this road will be described only in main lines.

From the examination of the daily-totals we noticed the following deviations from the rather regular pattern with period 7 :

- the summer-period looks rather nonstationary.
- there might be an additional year-trend.
- a lower consumption occurs on holidays.

Because it was not clear whether and in what measure these deviations would influence the fitting of an adequate model to all 52x7=364 samples (the whole year New Yearsday excluded), it was first tried to fit a model to the complete series, without making any adjustment on the data. It turned out to be impossible to fit an adequate model. The estimation failed in a rather peculiar way, which presently will be described. In order to examine if this failure was due the nonstationary summer-period, the next attempt was to fit a model to the first 182 samples. Also these estimations failed in the same way, which will be illustrated now.

Example 3.1

Fig. 3.7a shows the estimated a.c.f for the undifferenced and for the first order seasonal differenced series of the first 182 samples.

The a.c.f for the undifferenced series clearly indicate a first order seasonal AR-model with a rather big parameter value and a first, maybe second, order nonseasonal MA-model that is (0,0,1)(1,0,0)7 or (0,0,2)(1,0,0)7. For the differenced series the a.c.f indicates a first order seasonal MA-model. With respect to the nonseasonal model one can doubt between a first order AR- or a second order MA-model. So the model-order is (1,0,0)(0,1,1)7 or (0,0,2)(0,1,1)7.

> note - With regard to the identification it should be noted here, that it turned out to very useful to compare the estimated a.c.f

with a.c.functions calculated from simulated data. This holds particularly for seasonal models containing autoregressive parameters. Appendix VI shows some examples of a.c functions of simulated data.

Fitting a (2,0,0)(1,0,0)7 model turned out to be inadequate; that is, the a.c.f of the residual series showed one rather big, significant value at lag 7 : $\hat{r}_7 = 0.38 \pm 0.07$
A model (0,0,0)(0,0,1)7 fittted to the residual series proved to be adequate. According to the theory (section 1.3) a modified model (2,0,0)(1,0,1)7 was estimated. However this estimation failed, because the seasonal AR-para-meter became equal to unity independent from the fact whether the constant c was estimated as an additional parameter or not. Also it should be noticed that the preliminary estimates $\Phi = 0.91$ and $\Theta = 0.38$ give no reason to suppose near-redundant factors.
Next, based on the a.c.f of the first order seasonal differenced series a model (0,0,2)(0,1,1)7 was estimated. Now the estimation failed because the seasonal MA-parameter became equal to unity and so this result suggests that the differencing is unnecessary.
This result is especially remarkable because the a.c.f of the residual series of fitting an (0,0,2)(0,1,0)7 model, similar as above, showed only one rather big significant value at lag 7. So it suggests a modified model of order (0,0,2)(0,1,1)7.

So for the undifferenced series the estimation results for the seasonal AR polynomial show a pole for $B = 1.0$, suggesting a first order seasonal differencing and, on the contrary, estimation results for the first order seasonal differenced series suggest that the differencing is unnecessary.

Similar contradictory results as described in example 3.1 were obtained with other orders of differencing. Also model-extension, as overfitting and fitting related model orders, in order to improve these results, turned out to be useless.
To find out what causes this failure of fitting a multiplicative model to the data we might start at the beginning of the procedure, that is to doubt the identification.

- First possiblity is to doubt the model orders identified on the basis of the a.c. functions. This looks quite unlikely and therefore in the first instance this possibility will not be taken in consideration.
- Second possibility is to suppose that the estimated a.c. funtion is disturbed by, for example, an additional year-trend. However, referring to the experiments described in section 2.3 and considering the a.c.f in fig. 3.7a one can doubt this possiblity too.

Nevertheless, to find out if the supposed yearly trend disturbed the a.c.f and/or the estimation, the data was adjusted by subtracting a second order trend (see section 2.3). This turned out to be no improvement. The a.c.f for the adjusted series hardly differs from the a.c.f of the original series as can be seen from fig. 3.7b. In accordance with this also the estimation results were quite similar.

- Considering the data it looked rather unlikely that the few deviations at holidays would be the cause of the failure.

Summarizing it is quite unlikely that the failure of fitting a model to the unadjusted, original data is caused by wrong identification.

Next (as it turned out later) a fruitful idea about the possible cause did arise considering the data, the results as described in example 3.1 together with example 4 section 2.3 about differencing, and estimation of the mean.

The idea is that the failure is caused by the big differences between the mean values of the various days. This will be explained by the following interpretation.

As we saw in section 1.4 the fundamental idea of the multiplicative seasonal model is the assumption that the parameters of the models which describe the seasonal behaviour are equal.

For example the model which describes the relationship between the

Sundays : $\Phi(B)\, \nabla_s^D\, z_{t-1} = \Theta(B)\, a_{t-1}$

and the model which describes the relationship between the Mondays :

$$\Phi(B)\, \nabla_s^D\, z_t = \Theta(B)\, a_t$$

contain the same parameters $\underline{\Phi}$ and $\underline{\Theta}$. As the seasonal models only concern Sundays, Mondays, etc. , we can write $z_{t-1} = \mu_{su} + \tilde{z}_{t-1}$ and $z_t = \mu_m + \tilde{z}_t$

where the mean μ_{su} of the Sunday-samples and similar μ_m of the Mondays are the levels about which the data fluctuate and $\tilde{z}_t$ in fact is the process we want to model.

Now if D is greater than zero, the mean values will be eliminated. However if the processes $\tilde{z}_t$ are stationary no differencing is needed and the different mean values remain in the data. That is, estimation of the general multiplicative model fits a model to $w_t = \nabla^d \nabla_s^D z_t - c$, as already seen in example 4 section 2.3, where c is the estimated mean of w_t. So c will be the mean of the total data, that is of μ_{su}, μ_m, etc. and the series w_t shows the similar regularly pattern as the series z_t.

In other words the extension to the general multiplicative model, that is supposing the same seasonal model for each day, is only true for the seasonal differenced series or in, case D is equal to zero, for the deviations from the levels about which the data fluctuates.

From a somewhat different viewpoint one could say that we do not want to model the rather deterministic regularly pattern but the stochastic fluctuations about the mean pattern.

Table 3.1 gives a summary of the mean values for the various days and the deviation from the mean value of the total data. All these means are calculated from the original data whithout any adjustment for the holidays.

i	day	mean μ_i	Δ_i
1	Sunday	2364	+ 807
2	Monday	3518	- 347
3	Tuesday	3387	- 216
4	Wednesd.	3390	- 219
5	Thurstd.	3358	- 187
6	Friday	3376	- 205
7	Saturd.	2806	+ 365
total data		3171	

$\Delta_i = \mu_w - \mu_i$, where

μ_w = mean total data

Table 3.1 Mean values of the days and deviations from total mean.

From the estimation procedure it is clear that it will be quite similar if the pattern is subtracted from the data so that the mean of the remaining series equals zero or if the data is adjusted in a way so that the mean of the adjusted series equals the mean of the total data.

We have chosen for the latter by adding Δ_i as given in table 3.1. Fig. 3.8 shows the adjusted series. From this plot it is obvious that now the deviations on holidays can not be neglected any longer. This was also confirmed by failing attempts to fit an adequate model; however these attempts will not be described here. The adjusting of the holiday-samples based on a priori knowledge will presently be discussed.
Also the nonstationary behaviour of the summer-period can be seen much more clearly from fig. 3.8. As already discussed in section 3.2 this nonstationarity is caused by the temperature and especially the summer-holidays; it is questionable if it would be possible to adjust these data even if data of more years were available. It is more likely to suppose that we have to fit a separate model to the data of this period.

Adjusting the deviations at holidays.

Of course the best way to adjust the deviations at holidays would be to use as much information as possible about the consumption on these days, especially for those holidays which do not fall on a fixed day of the week. To do so the data of more years should be available. However, having only one year's data the adjustments will be rather arbitrary. This will specially be true for doubtful cases as, for example, Saturdays preceding feastdays, where it is not always clear if the deviation, mostly not as big as on the holidays, is due the next holiday or is just coincidence. However on deciding whether a sample value should be adjusted or not, we should recall that from another viewpoint it is desirable to limit the number of adjustments in order to avoid too much manipulation of the data. We will meet this problem once more later on. Also the question how to adjust can not always clearly be answered. One possibility is to substitute the mean value for the particular day of the week. More likely the best way to adjust is by interpolation, considering the neighbouring sample values.
In accordance with these considerations only the very clearly deviations, as they emerge from fig. 3.8 were adjusted to the mean values. Fig. 3.9 shows the first 168 samples of the obtained series.

Based on the assumption that we have to fit a separate model to the summer-period, a model was fit to these 168 samples which turned out to be adequate and which will be described in the next section.
So we decided to use the series, adjusted for the average week-pattern and additionally adjusted for the holidays, as a starting point to obtain and to examine forecasts results. Also it was decided to divide the year into three parts in order to check if a separate model has to be fitted to the summer-period or to examine the changes in the parameter-values. If this is true it is rather likely that the series might be explained more exactly by the different models for each part. A disadvantage of dividing the series is that the smaller series might contain less information and that the estimated a.c functions have bigger standard deviations because $\mathrm{var}\left[\hat{r}_k\right] = N^{-1}$.

3.4 Final models and forecasts results.

In this section we will first describe results for part one of the series only adjusted for the clear deviations on holidays. Based on this result the results for a series with additional adjustments will be described and used for a cross-validation with the data of 1984.
As a criterion to evaluate the forecast results we will use two acuracy measures which are defined as follows :

- the mean absolute percentage error $\mathrm{m.a.p.e} = \frac{1}{N} \sum_{t=1}^{N} \frac{|x_t - \hat{x}_t|}{x_t} \cdot 100$

- the mean square error $\mathrm{m.s.e} = \frac{1}{N} \sum_{t=1}^{N} |x_t - \hat{x}_t|^2$

where x_t is the actual value and $\hat{x}_t$ is the one-step ahead forecast-value.

Also the forecast results will be compared with a so-called 'naive' or 'no-change' method. On account of the seasonality for model-fitting this can be described as $\hat{x}_{i-1}(1) = \hat{x}_i = x_{i-7} \qquad i = 8,9,\ldots,N.$

Series zca_t.

We shall refer to series zc_t as the series adjusted for the average week-pattern, as described in the last section and to zca_t as the series additionally adjusted for only the clear deviations at holidays.

- note : See appendix IV for a complete list of the series.

These series zca_t will be divided into three parts, that is :

$zca1_t$	samples : 1- 168	adjusted samples : 44,45,93,131,132,142
$zca2_t$	samples : 169 - 245	
$zca3_t$	samples : 246 - 364	adjusted samples : 359-363.

Most samples are adjusted to the daily-mean or the general mean of all days.

Looking at the residual variance and the Box-Pierce chi-square test the best model fit to series $zca1_t$ was a (0,1,2)(2,0,0)7 model. Table 3.2 shows the estimation results for both the scalar c estimated and not estimated.

$\theta_1 \pm \sigma$	0.72 ± 0.08	0.71 ± 0.08
$\theta_2 \pm \sigma$	0.10 ± 0.08	0.08 ± 0.08
$\Phi_1 \pm \sigma$	0.22 ± 0.08	0.22 ± 0.08
$\Phi_2 \pm \sigma$	0.22 ± 0.09	0.23 ± 0.09
$c \pm \sigma$	2.1 ± 1.9	———
Q(39) / ε	37.1 / 0.56	37.0 / 0.56
res.var.	0.66E+04	0.66E+04

table 3.2 Fitting a (012)(200)7 model to series $zca1_t$

From table 3.2 we see that both results are quite similar. The non-seasonal second order parameter θ_2 differs hardly from the standard error; however a (0,1,1)(2,0,0)7 model :

$$(1 - 0.21 B^7 - 0.25 B^{14}) \nabla zca1_t = a_t(1 - 0.75 B)$$

turned out to be less adequate. Also a (2,0,0)(2,0,0)7 model :

$$(1 - 0.33 B - 0.29 B^2)(1 - 0.29 B^7 - 0.36 B^{14}) zca1_t = 3151 + a_t$$

turned out to be adequate, having a bigger residual variance.

These 4 models were used for the so called 'model-fitting', that is to predict the series used for estimation. Table 3.3 shows the accuracy measures for the various models, and also for the naive method. The naive method was performed on the original series z_t similarly adjusted for the deviations on holidays as series zc_t, thus not adjusted for the average week-pattern. In order to compare the naive method correctly with the results of the other models, m.a.p.e_2 was calculated from the series resulting from the modelfitting re-adjusted for the average week-pattern.

model	scalar c	m.a.p.e_1	m.a.p.e_2	m.s.e
A: (012)(200)7	estimated	1.89	1.95	0.648E+04
B: (012)(200)7	not est.	1.92	1.98	0.675E+04
C: (200)(200)7	estimated	1.94	1.99	0.683E+04
D: (011)(200)7	not est.	1.93	2.00	0.684E+04
naive method			2.51	1.040E+04

Table 3.3 Model-fitting: accuracy measures for the series $zca1_t$.

note: In case of model-fitting it will take some time lags before the forecasts 'adapt' to the series, due to the fact that the state set contains a number (depending on the modelorder) of samples and residuals to construct the forecasts. Because there is no real link between the end and the beginning of the series the state set has to be updated several times before the forecasts fit to the series. Therefore calculating the accuracy measures of model-fitting mostly will be done excluding the first 7 or 14 values.

From table 3.3 we notice that the accuracy measures differ hardly for the various models; model A, the most adequate, has the smallest values.
The results obtained with models A-D are clearly better than those obtained with the naive method.
Fig. 3.10a shows the series $zca1_t$ together with the predicted series and fig. 3.10b shows both series re-adjusted for the average week-pattern.
From fig. 3.10a we notice that the predicted series follows the general trend of the series however not the biggest peaks, which leads to rather big errors as can be seen from fig. 3.11 which shows the absolute percentage errors (a.p.e_1) and the square errors (s.e) for this series.
From fig. 3.10 also the time needed to adapt to the series can clearly be seen.
For the 'naive' method the original and the predicted series are shown in fig. 2.12. Fig. 3.13 shows the plots of the errors for the naive method as well for the re-adjusted series as shown in fig. 3.10b, thus a.p.e_2 and s.e which of course is not affected by the re-adjusting.

Using these plots for comparison of both methods, it is remarkable that both methods show the biggest errors for the same samples, that is where the series deviates from the general trend. For the more complex or sophisticated method the errors are smaller than those for the naive method as already appeared from the mean error values.

From a further examination of those samples that lead to the biggest errors it followed that an evident cause of the deviation can not be indicated for all errors. The major part of the errors, however, are on feast-days or days next to feast-days, which in the first instance do not emerge as a clear deviation of the data.
As already discussed earlier from using only one year data it can not be judged whether this is a coincidence or not. Based on the supposition that the data of more years justifies the additional adjustment of these deviations, later on the results for a series with additional adjustments will be described.

So far we only considered the results of fitting models to the first part of series zca_t. Now the possible changes in parameters and/or models will be examined.
Table 3.4 gives the results of fitting a (0,1,2)(2,0,0)7 model to the various series.

series	$zca1_t$	$zca2_t$	$zca3_t$	zca_t
n=N-d-sxD	167	76	118	364
$\theta_1 \pm \sigma$	0.72 ± 0.08	0.01 ± 0.11	0.69 ± 0.07	0.17 ± 0.05
$\theta_2 \pm \sigma$	0.10 ± 0.08	0.45 ± 0.11	—	0.36 ± 0.05
$\Phi_1 \pm \sigma$	0.22 ± 0.08	0.19 ± 0.13	0.39 ± 0.10	0.19 ± 0.05
$\Phi_1 \pm \sigma$	0.22 ± 0.09	0.15 ± 0.12	0.12 ± 0.11	0.01 ± 0.05
Q/df/ε	39/37.1/0.56	15/9.6/0.84	29/21.3/0.85	39/71.7/0.001
res. var.	0.66E+04	6.5E+04	0.41E+04	2.1E+04

Table 3.4 Fitting a (0,1,2)(2,0,0)7 model to the parts of series zca_t

From table 3.4 we see that the model of order (0,1,2)(2,0,0)7 is adequate for the three separate parts of the year but not for the whole series. The estimated models show rather big changes in the nonseasonal

parameters. This is especially true for the parameters fitted to series $zca2_t$ compared with those for the other series.
To judge if a parameter-change is significant we can compare the difference with the standard error of this difference. For example :

For series $zca2_t$ $\Phi_1 = 0.19 \pm 0.13$ and for $zca3_t$ $\Phi_1 = 0.39 \pm 0.10$

So the difference is 0.20 and the standard error: $\sqrt{0.13^2 + 0.10^2} = 0.164$
The difference is 1.22 times its standard error and it is rather doubtful to consider this as a real change.
However we have to remark that due the smaller number of observations in the series the standard errors of the estimated parameters increase. So for smaller series a difference sooner will be considered as not significant due the bigger standard errors of the estimates.
Also remarkable from table 3.4 is the rather big residual variance for the model fit to the series $zca2_t$. Although these model turned out to be adequate we should remember that it need not to be the best to series $zca2_t$ because the model order was determined from series $zca1_t$.

The fitted models were used for model-fitting in order to compare the mean errors with those obtained with the naive method. These results are given in table 3.5, together with, in the third column, the mean errors calculated from forecasting (one step ahead with updating) the series $zca2_t$ and $zca3_t$ using the model-parameters fitted to series $zca1_t$.

series	modelfitting		naive method		forecasting	
	mape	mse	mape	mse	mape	mse
$zca1_t$	1.89	0.668E+04	2.51	1.04E+04		
$zca2_t$	5.84	6.190E+04	11.5	22.70E+04	7.87	11.20E+04
$zca3_t$	1.63	0.419E+04	1.91	0.60E+04	2.57	1.97E+04
zca_t	2.87	1.980E+04	4.40	5.93E+04		

Table 3.5 Accuracy measures for the various parts of series zca_t.

From table 3.5 we notice that for all series the results obtained with the naive method are worse. Even for series zca_t this is true, although the model fit to this series was inadequate. As already mentioned, according to the big residual variance the results for series $zca2_t$ are rather bad. From fig 3.14 which shows the series $zca2_t$ and the series resulting from the model-fitting, we notice that the predicted series seems to be shifted in time. This is likely due the differencing and the small values of the parameters of the model-fit to the series.
The comparison of the model-fitting and forecasting results shows clearly the influence of the changes in the parameters. Especially for series $zca3_t$ the mean errors are considerably bigger, as might be expected from the comparison of the parameters.

Series zcb_t.

Based on the assumption that the additional adjustments of those sample-values which caused the biggest errors by modelfitting series zca_t will be justified by the data of more years, we will now consider series zcb_t and examine the possible improvements obtained with these series.
Just like series zca_t, series zcb_t is obtained by adjusting series zc_t; where zc_t is the original series adjusted for the average week-pattern. Series zcb_t will also be divided into three parts, different in comparison to zca_t, that is :

$zcb1_t$	samples: 1 - 147	adjusted samples : 42,44,45,92,101,130,132, 141,142.
$zcb2_t$	samples: 148 - 245	
$zcb3_t$	samples: 246 - 364	adjusted samples : 358- 363.

Most of the samples are adjusted by interpolation.

In contrast with the procedure followed with series zca_t we now fitted the 'best' models to the various parts of zcb_t . 'Best' means having the smallest residual variance. The results are given in table 3.6.

series	$zcb1_t$	$zcb2_t$	$zcb3_t$
n=N-d-sxD	139	97	111
model	83-1: (011)(011)7	83-2: (013)(001)7	83-3: (012)(011)7
$\theta_1 \pm \sigma$	0.88 ± 0.04	- 0.02 ± 0.10	0.59 ± 0.09
$\theta_2 \pm \sigma$	—	0.38 ± 0.10	0.20 ± 0.10
$\theta_3 \pm \sigma$	—	0.26 ± 0.10	—
$\Theta_1 \pm \sigma$	0.80 ± 0.06	- 0.21 ± 0.11	0.81 ± 0.07
$Q/df/\varepsilon$	23.6/31/0.83	13.6/20/0.85	16.8/20/0.67
res.var	0.332E+04	5.327E+04	0.401E+04
model-fiting:	1.52 0.362E+04	5.32 5.39E+04	1.52 0.401E+04
(mape / mse)			
naive meth.:	1.78 0.569E+04	10.1 18.50E+04	1.91 0.603E+04

Table 3.6 : Series zcb_t. Models fit to the various parts, and accuracy measures for model-fitting, using these models and a naive method.

As emerges from table 3.6 now models of different order are fitted to the various series and it is meaningless to consider the changes in the parameters. In contrast with the other series, series $zcb2_t$ does not need to be seasonal-differenced.

In comparison to series $zca1_t$, series $zcb1_t$, which is additionally adjusted most, shows a considerable reduction of the residual variance. This may also be influenced by the different division of the samples by which the data close to the summer-period now belong to series $zcb2_t$.

Fig. 3.15 shows the accuracy measures for series $zcb1_t$. Now the magnitude of the peaks are considerably smaller.

Likely due to the greater number of samples and the fact that now the best model is fitted to the series also the residual variance of series $zcb2_t$ is smaller than those of series $zca2_t$, however the improved accuracy measures are still bad compared to the other series.

For series $zcb3_t$, with only one additional adjustment, the best model-fit turns out to give only a small improvement of the results obtained from series $zca3_t$.

For the naive method quite the same conclusions can be drawn. The results for series 1 are considerably improved; remarkable is the fact that the difference between the two methods has become smaller.
Also for series 2 the mean errors are considerably smaller; this, however, will mainly be due the greater number of samples.
For series 3 the only one additional adjustment does not improve the results of the naive method.

Sofar the estimated models were mainly used to predict the series used for estimation, the so called model-fitting.
Now we will describe the results of real forecasting that is using the models fitted to the various parts of zcb_t to forecast the corresponding parts of data of 1984. Because meanwhile the data of 1984 had become available, it is possible to produce forecasts by one step ahead forecasting with updating. To do so, first the data of 1984 has to be adjusted using of course the knowledge gained from the data of 1983, that is :

i)- Adjusting for the average week-pattern.
The data of 1984 was adjusted in the similar way as the data of 1983 by using the average week-pattern of 1983, for the average week-pattern of 1984 is, strictly, not yet known.

ii)- Adjusting the deviations on holidays.
Here we have to distinguish the holidays on a fixed day of the week and the others. The fixed-holidays were adjusted in the similar way as the data of 1983 by adding the similar difference-values. The others were adjusted to the corresponding daily mean of 1983.
Two exceptions had to be made. First: New-years day, which was not included in the 1983-data, had to be adjusted and second: Queen-day. In 1983 Queen-day was on a Saturday and no deviation emerged from the data however in 1984 Queen-day was on a Monday and the consumption was considerably lower. This example emphasis the fact that for optimum adjustment of the deviations on holidays, data of several years should be available.

note : Some daily totals had to be determined by interpolation because they were not measured due to technical troubles. That is : the samples 62, 63, 84, 90, 278, 288, 289, 296-299.

To produce these forecasts the state set was initialized on the data next to the part of the year to be forecasted. In this way the adapting-phenomenon as occured performing model-fitting can be avoided.
Table 3.7 shows the mean errors of the forecasting results and also those obtained with the naive method.

data 1984		model*	forecasting		naive method	
	samples		m.a.p.e	m.s.e	m.a.p.e	m.s.e
series 84-1:	1 - 147	83-1	2.42	1.26 E+04	3.16	1.97 E+04
series 84-2:	148 - 245	83-2	4.35	3.37 E+04	8.93	13.0 E+04
series 84-3:	246 - 364	83-3	1.78	0.651E+04	2.27	0.971E+04

Table 3.7 Mean errors forecasting time series 1984.
* see table 3.6.

At first sight the forecasting resuls look rather good, certainly compared to the results of the naive method. However to gain more insight we have to consider the a.p.e.'s and the s.e.'s instead of their mean values.
Fig. 3.16 shows the plots for the a.p.e of the three series and fig. 3.17 the plots of the s.e. From these plots considerable peaks emerge. For example for series 84-1 sample 71 shows an a.p.e. of 16.8 %, due to a very big peak consumption on a 'normal' Sunday; 'normal' that is there is no clearly demonstrable cause. Also the a.p.e plots for series 84-2 shows lots of values greater than 10 %. About the maximum error, which can be accept we will discuss later on.
Fig. 3.18 a,b,c shows the adjusted series. From these plots also the outliers leading to the greatest errors can be seen. Remarkable from series 84-2 is the similar lower consumption during the summer-holidays, as we also noticed for the data of 1983.
Examination of the errors of the adjusted samples shows : 5 that is 25 % is bigger than 5 % and even 1 is greater than 10 %. Some of the other big errors are on days before, or after feastdays, which showed no deviation in 1983, as for example 5 May. These results confirm the doubt to adjust these deviations based on only one year data.
In order to check the adjustment of the average week-pattern, this was also calculated for 1984. Table 3.8 shows both for 1983 and 1984. We notice a clear difference for the Saturday, which might have influenced

the results. So one can doubt too about this adjustment based on year data. We will discuss the adjustment more comprehensively in the next section.

i	day	1983 mean μ_i	∇_i	1984 mean μ_i	∇_i	$\nabla_i = \mu_w - \mu_i$
1	Sunday	2364	+ 807	2342	+ 807	
2	Monday	3518	- 347	3490	- 341	
3	Tuesday	3387	- 216	3354	- 205	
4	Wednesd.	3390	- 219	3388	- 239	
5	Thurstd.	3358	- 187	3360	- 211	
6	Friday	3376	- 205	3370	- 221	
7	Saturd.	2806	+ 365	2734	+ 415	
total data μ_w		3171		3149		

Table 3.8 Mean values of the days and deviations from the mean.

Finally we checked the changes in the parameters and/or model. From table 3.9 we notice for series 84-1 a significant change in the parameter θ_1 (difference is 2.1 times its standard deviation). From the diagnostic check we find this model of the same order adequate, however having a considerable greater residual variance. This of course is due the outliers as already described earlier.

model	83-1 : (011)(011)7	83-2 : (013)(011)7
series	84-1	84-2
θ_1	0.73 ± 0.06 (0.88 ± 0.04)	- 0.08 ± 0.10 (- 0.02 ± 0.10)
θ_1	—	0.13 ± 0.10 (0.38 ± 0.10)
θ_1	—	- 0.03 ± 0.11 (0.26 ± 0.11)
θ_1	0.80 ± 0.06 (0.80 ± 0.06)	- 0.53 ± 0.09 (- 0.21 ± 0.09)
Q/df/ε	9.8/20/0.98	10.3/20/0.96
res.var.	1.22E+04	2.96E+04
m.a.p.e	2.28 (2.42)	4.35 (4.45)
m.s.e	1.25E+04 (1.26E+04)	3.1E+04 (3.4E+04)

Table 3.9 Estimation results and mean errors of modelfitting series 84-1 and 84-2. Within brackets the parameters of the models used for forecasting and the obtained errors.

The esimated model was used for model-fitting and these results turn out to be a rather small improvement.
For series 84-2 all parameters changed considerably; however the model of this order is adequate. Similar to series 84-1 the model-fitting results are only a small improvement.
For series 84-3 estimation of model 83-3 failed, that is the estimated parameters turned out to be non-stationary. So for this series there is a change of model order.

3.5 Evaluation of the results, conclusions and recommendations.

Summary of the results.

In order to estimate adequate models to the data of 1983 we had to adjust the data for the average week-pattern. After this adjustment it turned out to be necessary to adjust additionally for the deviations on holidays.
Due the nonstationary summer-period the data had to be divided into (at least) three parts.
Using the estimated ARIMA-models for model-fitting the mean errors for the part containing the summer-period were considerably worse than the other parts.
Additional adjusting of more deviations improved the results.
Compared to a pure naive method the results with the ARIMA-models all along the line were better, but the improvements were relatively small especially for part 1 and 3 (thus not the summer-period) after the additional adjustments.
The estimated models to the various parts showed considerable changes in the parameters.
Changes in parameters and model orders occurred due additional adjustments and/or changing the number of samples a little.

Using the finally best models fit to the data of 1983 to forecast the coresponding parts of the 1984 data, we noticed :

- Relative acceptable mean errors compared to the model-fitting results of 1983.
- Considerable outliers. So the summer-period shows errors up to 22.9 %, even part 1 and 2 showed some big errors.

- Also important is the fact, that 25 % of the adjusted deviations on holidays showed errors greater than 5 % and that still errors emerged on days before or after holidays, for example Easter-Tuesday.
- Estimation of a model with the same model order as the model used for forecasting on the data of 1984 showed considerable changes in the parameters; for part 3 even the estimation failed.

Comparison of the average week-pattern of 1983 to that of 1984 showed some small differences for Tuesday-Friday and a big difference for the Saturday.

Remarks.

1 - Adjusting the data.

- As already mentioned before, it is questionable to use only one year's data to adjust the data. This doubt is confirmed by the results obtained. To gain more insight in the deviations at holidays, data of more years should be used. However using more year data one should pay attention to a possible change in the kind and number of the users, which will influence these deviations. If there are such changes it is questionable if more insight will be obtained.

- The means calculated of all data, so inclusive the deviations on holidays, were used to adjust for the average week-pattern. As a matter of course we can ask ourselves if the results can be improved by calculating the means exclusive these holidays or by calculating them of only those samples used for model building.

- Another more or less arbitrary chosen starting-point is the division into the various parts. The, in comparision to the other parts, different models fit to the summer-period confirm the dividing of the data in at least three parts. At least suggests the question, whether dividing in more parts will improve the results. However, smaller series will contain less information and will have considerably bigger standard deviations for the estimates. So it is very likely that the results will not be improved by dividing the data into more parts.

2 - Forecasting quality : requirements on the accuracy.

- Thus far we judged the obtained results by comparing them with those obtained with a naive method or by considering the improvements in comparision to other models.
  However the judgement on the forecasting quality can be different depending on the goal the forecast will be used for. So it is quite possible that for one goal the mean errors will be sufficient and some big errors do not matter. In contrast, for another goal the magnitude of the biggest errors is the most important.
  The latter is true in case of the water-company. The water-company has to avoid a deficiency of water at all times and so the deviation will be restricted by the capacity of the buffer-store. So the limit of the absolute percentage error depends on the total daily consumption, and this limit especially will be small (≈ 5%) for peak-days in the summer-period, etc. Thus errors as obtained by forecasting 1984 up to 22.9 % can not be accepted.
  It is not possible to compare correctly the obtained results with those obtained by the water-company due to the different leadtimes; as already mentioned they dated up their prognosis to the actual consumption during the day.

3 - Practical aspects about the use of the Box-Jenkins method.

- It turned out that especially for the seasonal data the identification caused some troubles, certainly for small series, due to the rather big standard deviations of the estimated a.c.f this is true.

- The fact the identification of the right model order has to be done by examination the a.c.f and p.a.c.f; this can be seen as a disadvantage. That is the method can not be used automatically if it is likely that changes of modelorders will occur, unless one fits all possible models up to a specific order and one accepts the model having the smallest residual variance as the best.

Conclusions.

1 - In view of the fact that the ARIMA-models fit to the series turned out to be adequate, that is the residuals can be supposed to be white noise and no information is left in the residual series, the obtained results have to be considered as the optimum results in model-building and forecasting the daily totals of the water-company data.

2 - Forecasting results obtained with the ARIMA-model are considerably better than those obtained with a 'naive' method.

3 - Although the results obtained with the ARIMA-model in comparison with the results obtained with a 'naive' method can provide a better starting-point to control the production of water; it is considering : - the requirements of the water-company to limit the the magnitude of the errors,

- the obtained results, that is the rather big errors in forecasting incidental deviations of the data,
- the problems met to adjust the data, in order to be able to fit adequate models,
- the practical aspects(see remark 3),
- the time, costs and trouble using this method in comparison to a 'naive' method,

questionable that these method, practised in the way used here, will be a solution to the water-company problem.

Recommendations.

Finally some remarks about other ways to obtain eventually sufficient results will be discussed.

- It is worthwhile to examine if better results will be obtained using the ARIMA-model to forecast smaller periods than one day, for example periods of 2,4 or 6 hours. In that way there will be some updating to the actual consumption during the day. However in order to be able to fit the models, we will meet again the problem of when and how the data has to be adjusted.

- Although the weather clearly has a influence it is questionable if using a multivariate or transfer-model with the temperature as input-variable, will sufficiently improve the results; the weather-influence is 'disturbed' by unmeasurable influences, as can be seen from the summer-holiday data.

- In view of the unmeasurable influences (especially : changes in kind or number of consumers, the deviations on holidays and the summer-holidays) it is recommended to gain as much a priori knowledge as possible about these deviations. Maybe this knowledge, together with a more or less sophisticated 'naive' method and actual information about the weather, will be sufficient to produce short-term forecasts. This might be improved by updating during the day.

CHAPTER 4. GENERAL CONCLUSIONS.

- In chapter 1 the ARIMA-model is described as a class of stochastic models capable of representing equispaced, discrete, stochastical time series which, although not necessarily stationary, are homogeneous and in statistical equilibrium. The use of this ARIMA-model to produce minimum square error forecasts is elucidated. Also described is the relation of a model of this kind to a given series by a three stage interactive procedure based on identification, estimation and diagnostic checking.

- The interactive programs DIFAC and MEP are implemented on a VAX 11/750 computer to perform the several stages in building the ARIMA-model, given a time series and to use this model to produce forecast of future values of the series. The programs are written in FORTRAN 77.

The program DIFAC can be used :

- to perform differencing of the time series
- to compute the estimated autocorrelation function
- to compute the estimated partial autocorrelation function.

The program MEP can be used :

- to compute preliminary estimates of the parameters using the estimated autocorrelation function
- to estimate the parameters
- for diagnostic checking by:- overfitting
  - eximination of the residuals
- for forecasting:- modelfitting
  - one step ahead with updating
  - ℓ- step ahead.

The program PRER can be used to calculate forecast errors that is :

- the mean absolute percentage error
- the mean square error.

The structure of all data-files satisfies the requirements of the plotprogram GRA.

- Experiments with simulated data showed very good results.

- Identification of seasonal practical data turned out to be rather difficult.

- Forecasting results obtained with the ARIMA-model were considerably better than those obtained with a 'naive' method.

- Using the ARIMA-model to forecast the daily totals of the water-consumption we noticed :
  - quite acceptable mean errors
  - considerable outliers, that is, the fitted adequate ARIMA-model turned out not to be able to produce forecasts of incidental deviations of the data.

APPENDIX I Symbols.

a_t	input, white noise
z_t	output, time series
w_t	differenced time series
$\tilde{z}_t = z_t - \bar{z}_t$	deviation of the estimated mean of the series
B ; $B\ z_t = z_{t-1}$	backward shift operator
F ; $F\ z_t = z_{t+1}$	foreward shift operator
∇ ; $\nabla\ z_t = z_t - z_{t-1}$	difference operator
d ; $\nabla^d = \nabla^{d-1}.\nabla$	order differencing
S ; $S\ z_t = \nabla^{-1}\ z_t$	summation operator
∇_s ; $\nabla_s z_t = z_t - z_{t-s}$	seasonal difference operator
D ; $\nabla_s{}^D = \nabla_s{}^{D-1}.\nabla_s$	order of seasonal differencing
$\phi(B) = 1 - \phi_1 B - \ldots. - \phi_p B^p$	Autoregressive operator of order p
$\varphi(B) = \phi(B)(\ 1 - B\ )^d$	general AR-operator of order p+d
$\theta(B) = 1 - \theta_1 B - \ldots. - \theta_q B^q$	moving average operator of order q
$\Phi(B^s) = 1 - \Phi_1 B^s - \ldots. - \Phi_P B^{sxP}$	seasonal AR-operator of order P
$\Theta(B^s) = 1 - \Theta_1 B^s - \ldots. - \Theta_Q B^{sxQ}$	seasonal MA-operator of order Q
$\hat{z}_t(\ell)$	prediction for leadtime ℓ at origin t
$e_t(\ell)$	forecast error for leadtime ℓ
$\hat{a}_t$	residuals of estimation
$\hat{}$ e.g $\hat{\phi}$	estimated paramater
$\bar{z}$	estimation of μ, mean of stationary process
s^2	estimation of σ^2, variance of stationary process
c_k	estimation of γ_k, autocovariance at lag k
r_k	estimation of ρ_k, autocorrelation at lag k
$\hat{\phi}_k$	estimation of ϕ_k, partial autocorrelation at lag k

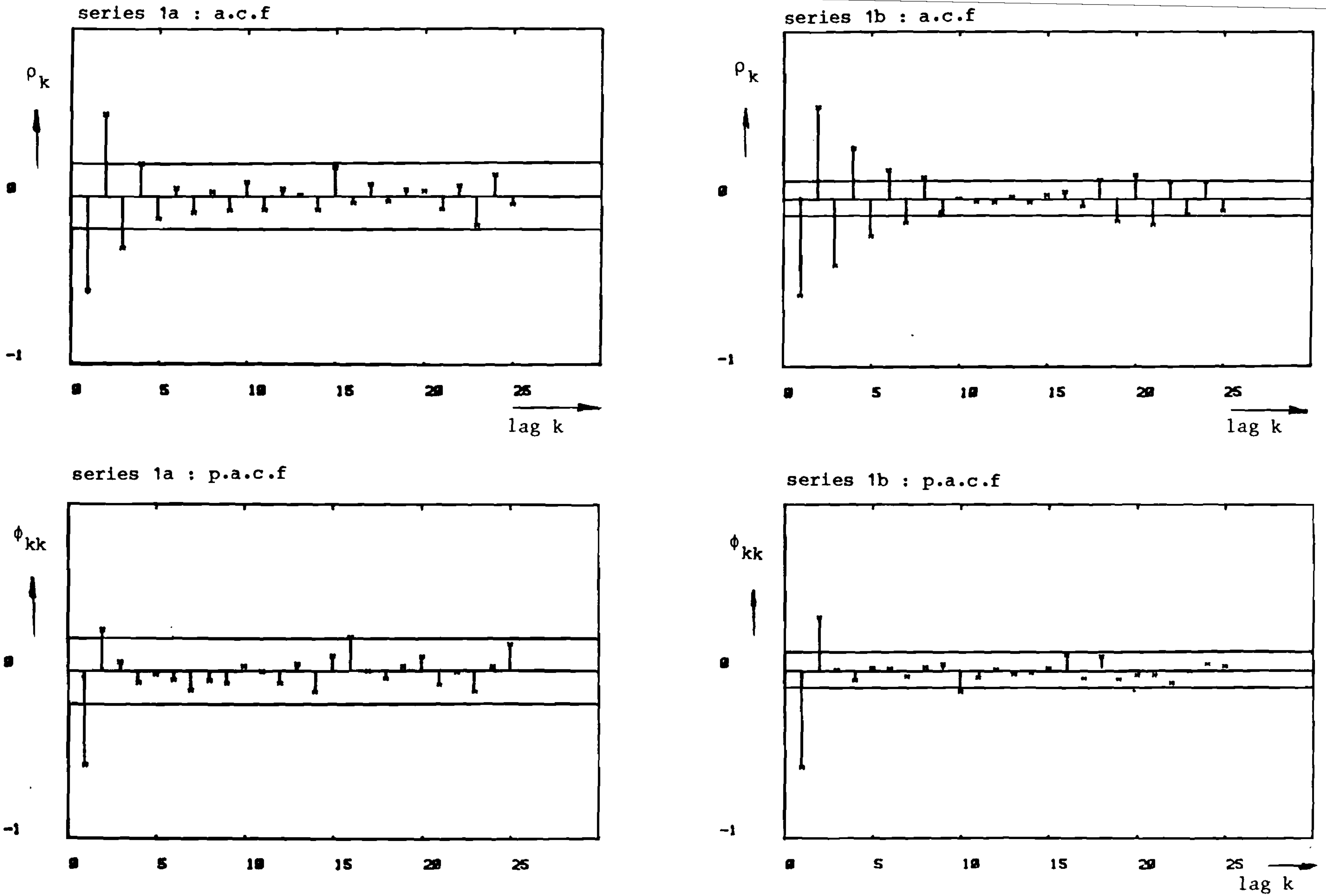
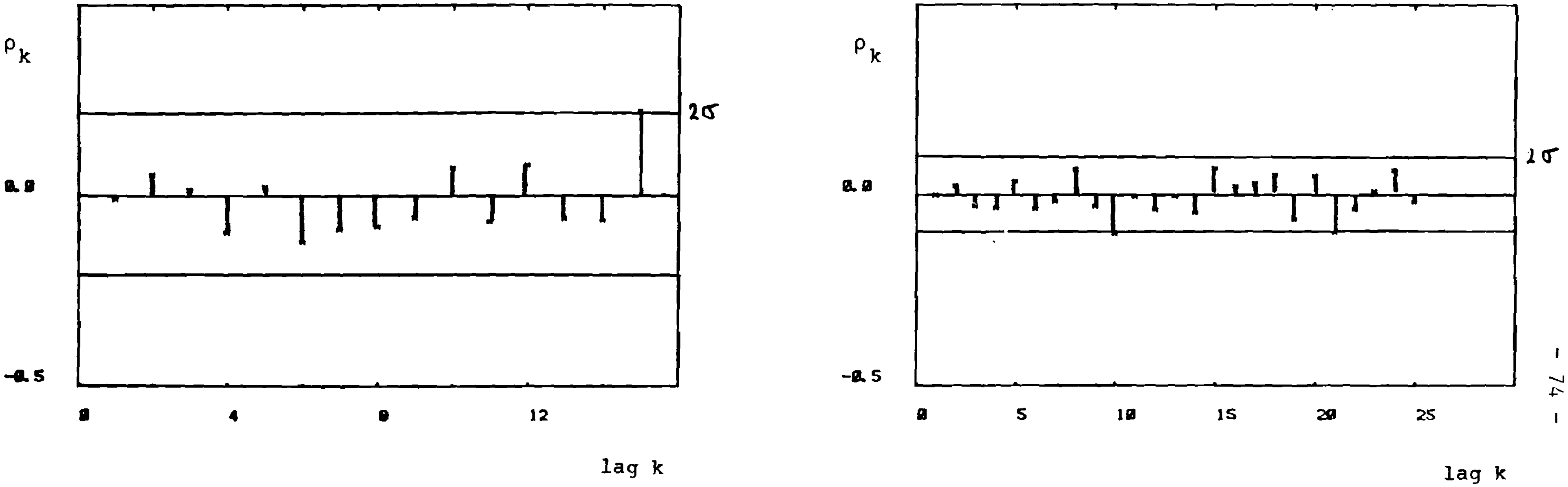


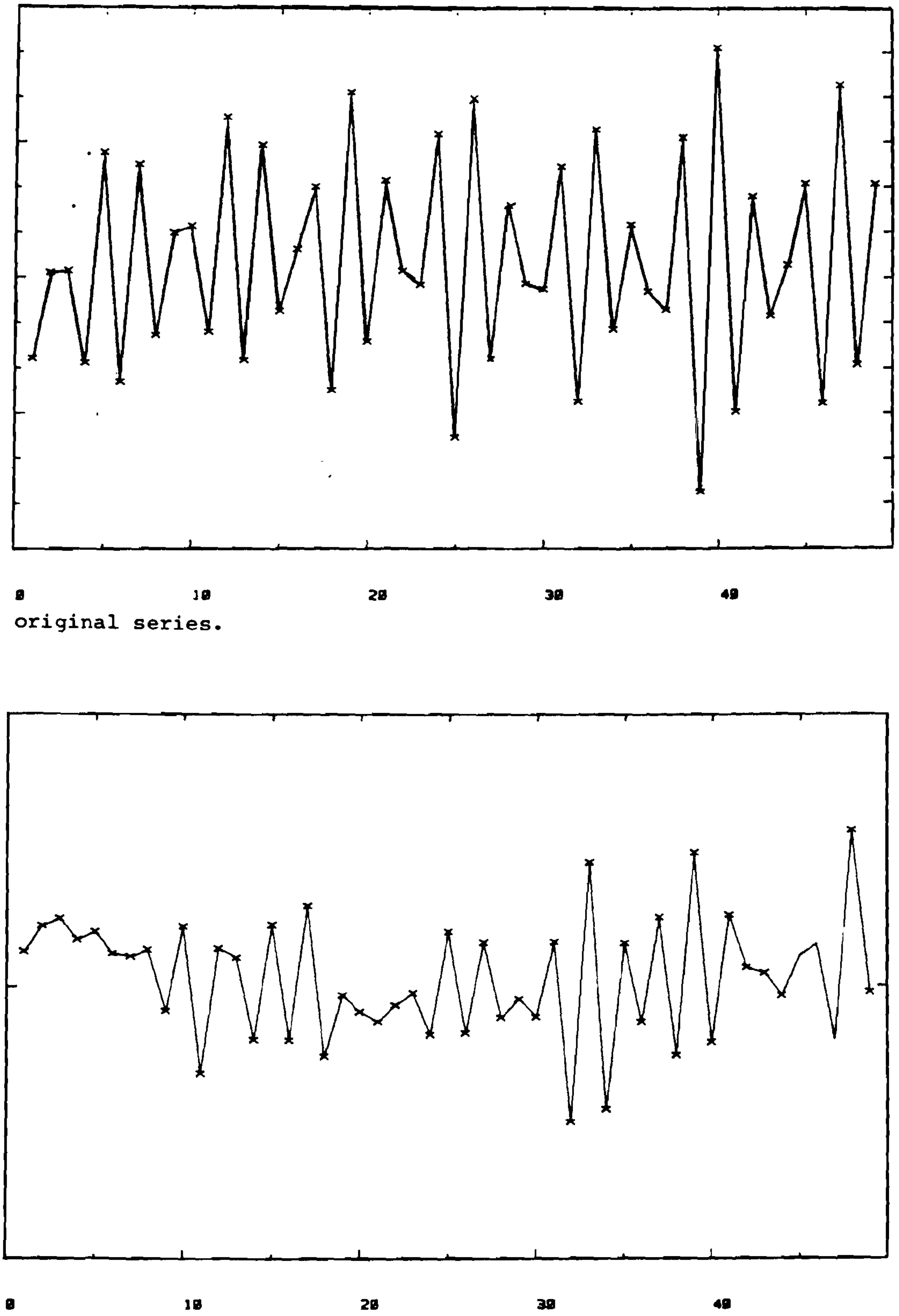
fig. 2.2 Example 1 : a.c.f and p.a.c.f of series 1a and 1b.



series 1a

series 1b

fig. 2.3 Example 1 : a.c.f of the residual series.



original series.

first order seasonal differenced series.

fig. 2.4 Example 2 : first 50 observations of original and first order seasonal differenced series.

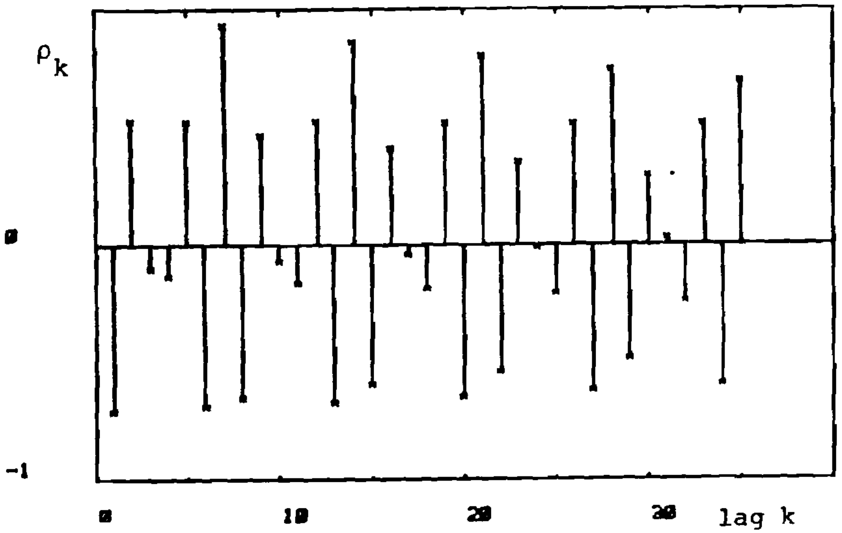


fig. 2.5 Ex. 2 : a.c.f original series.

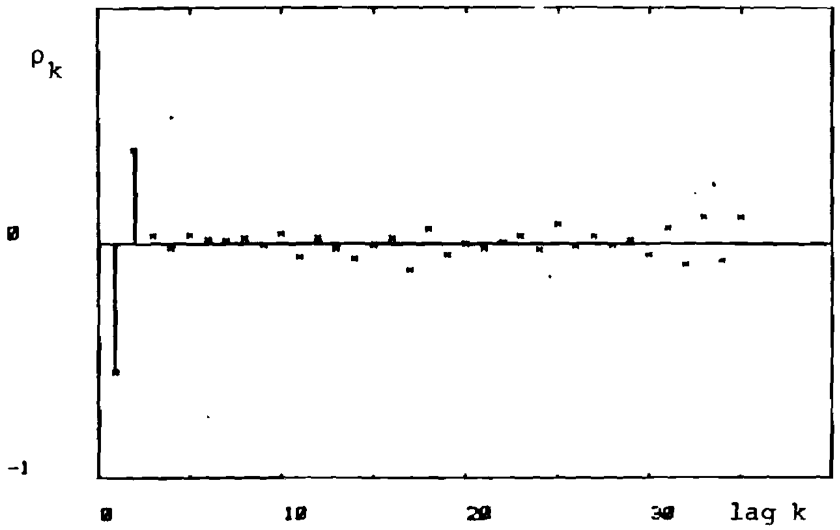


fig. 2.5 Ex. 2 : a.c.f differenced series.

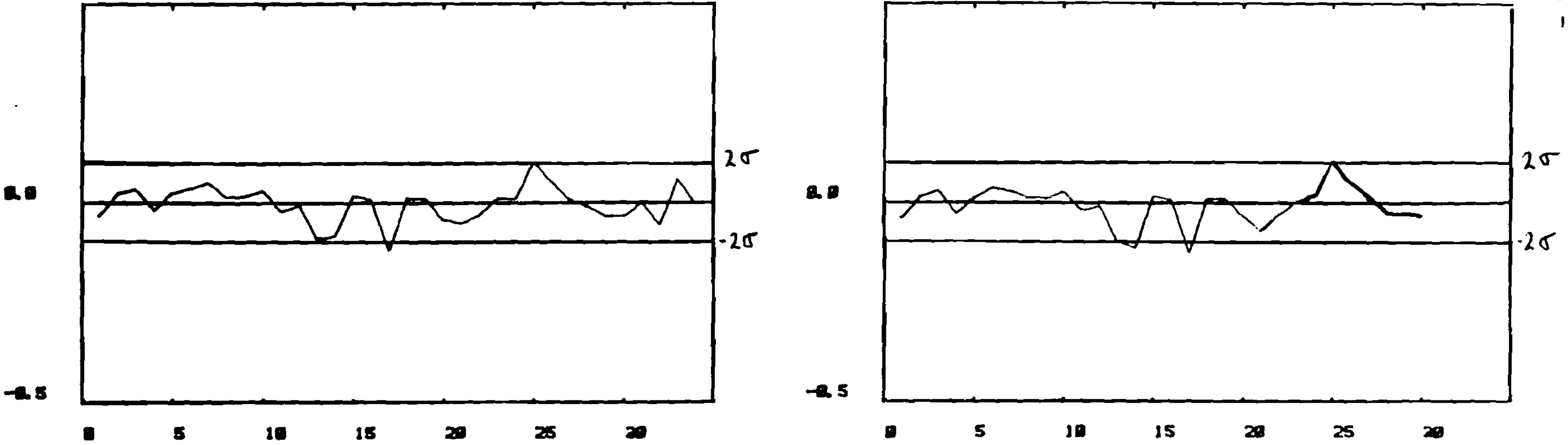
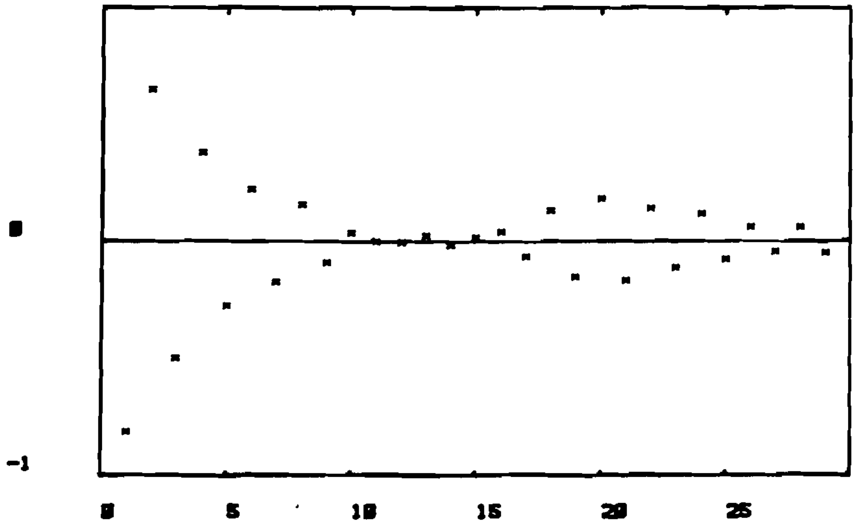
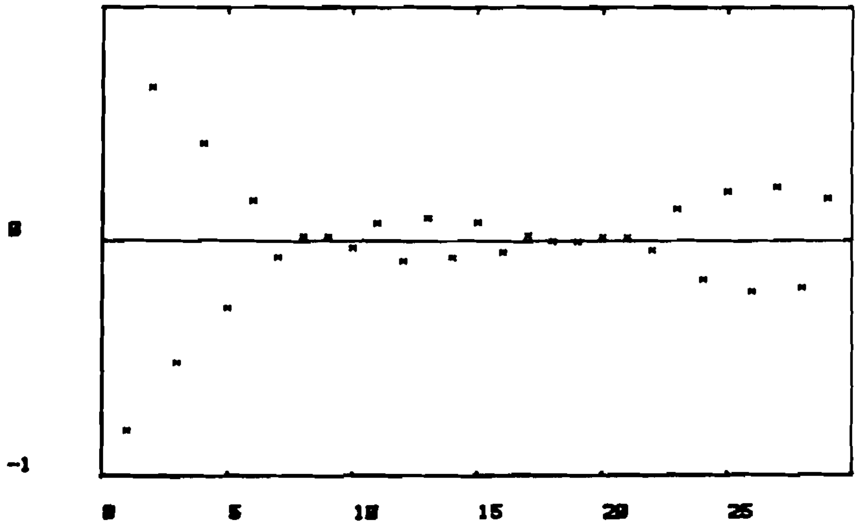


fig. 2.6 Example 2 : a.c.f residual series.





a.c.f of an AR(1) model.

fig. 2.7 Example 3 : a.c.f of series 3d and of an AR(1) model.

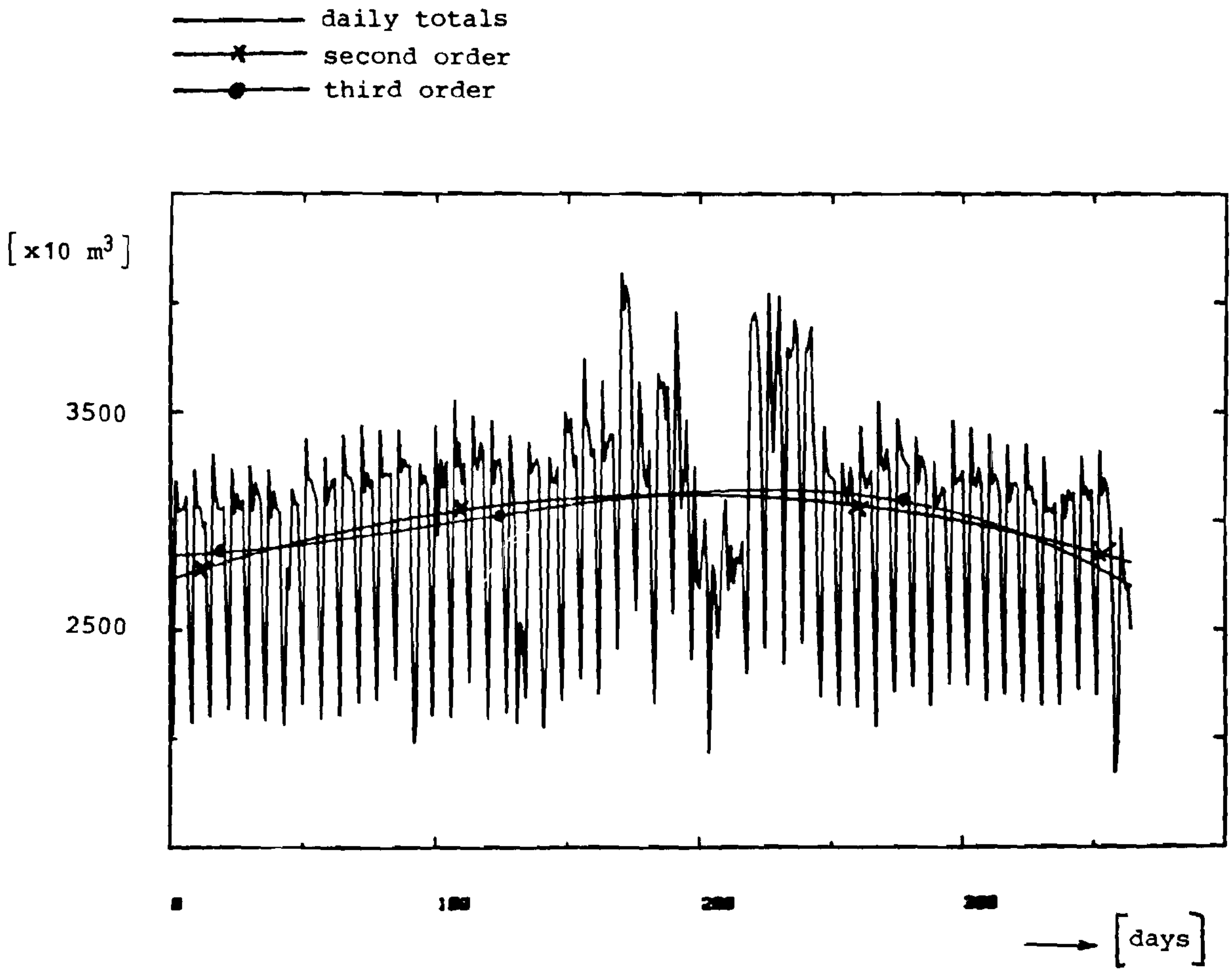
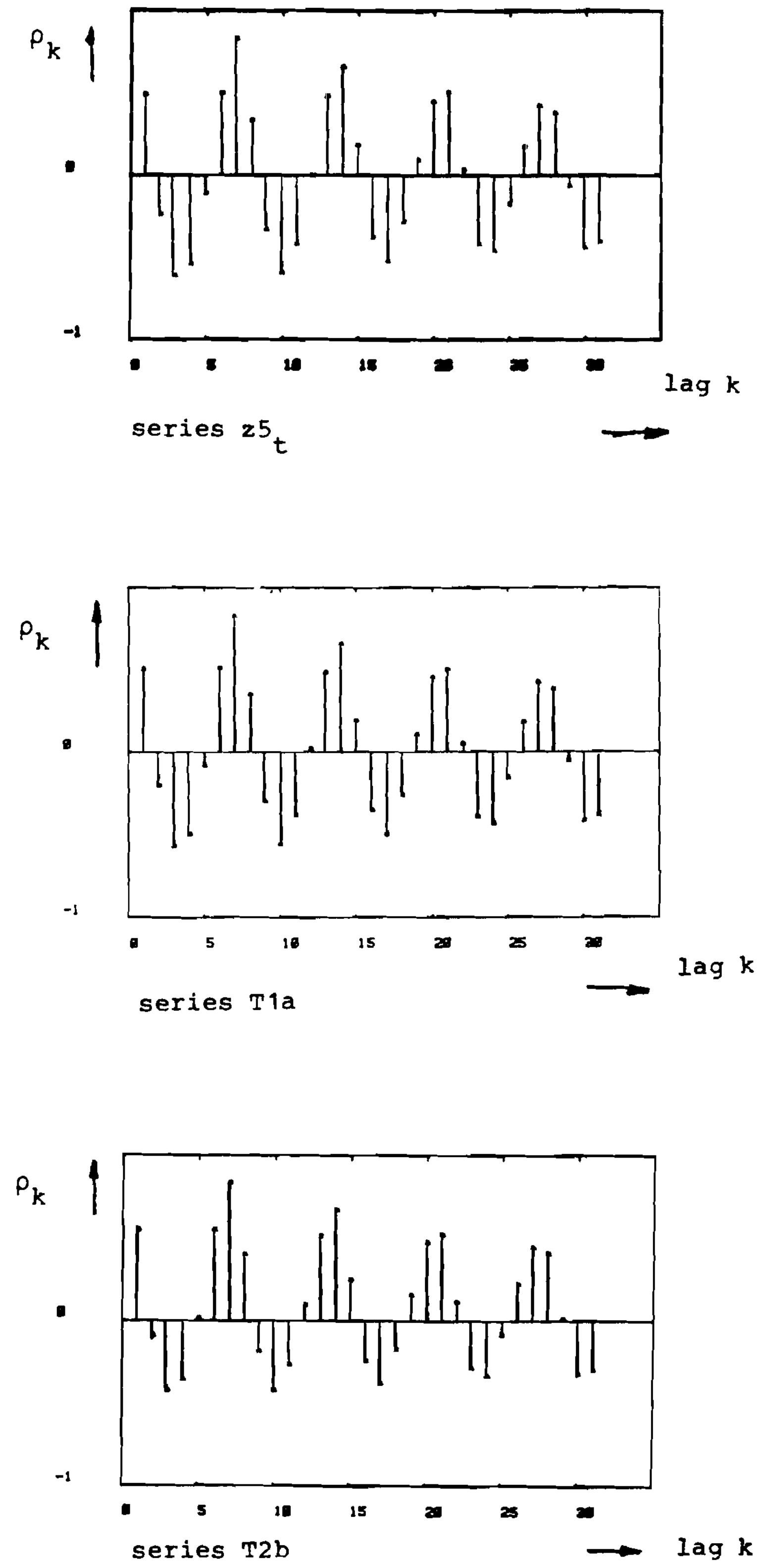


fig. 2.8 Daily totals of the water-company data and fitted year-trend.



series $z5_t$

series T1a

series T2b

fig. 2.9 Example 5 : a.c.f of series $z5_t$, T1a, and T2b.

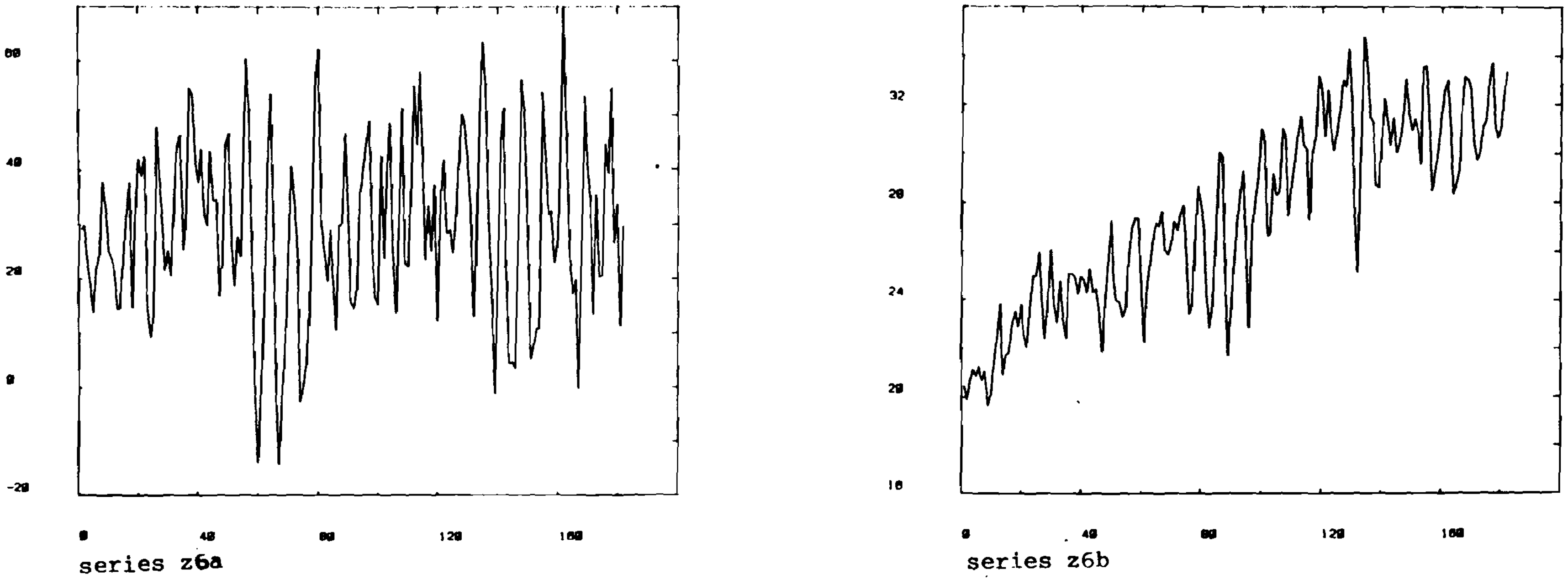


fig. 2 10 Example 6 : first 182 samples.

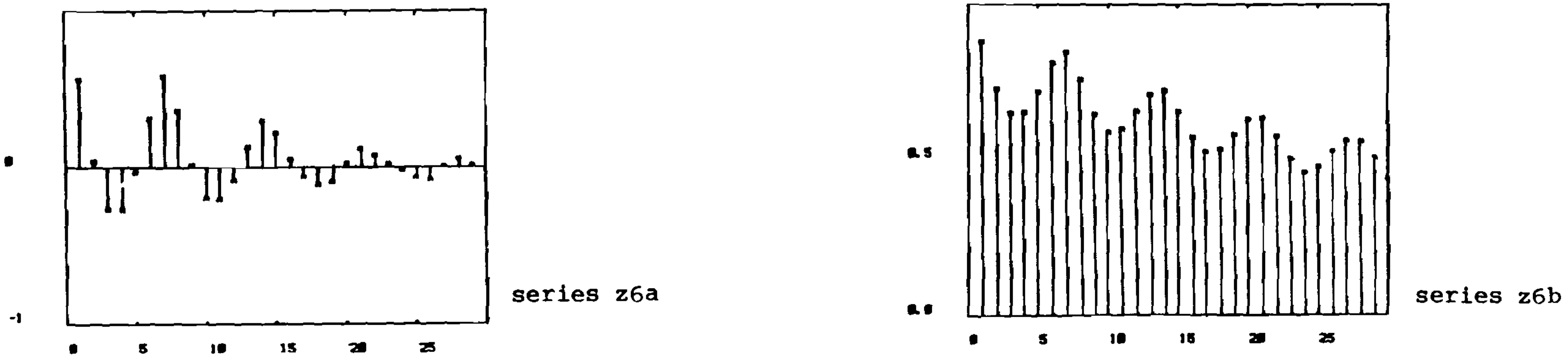


fig. 2.11 Example 6 : a.c.f of series z6a and z6b.

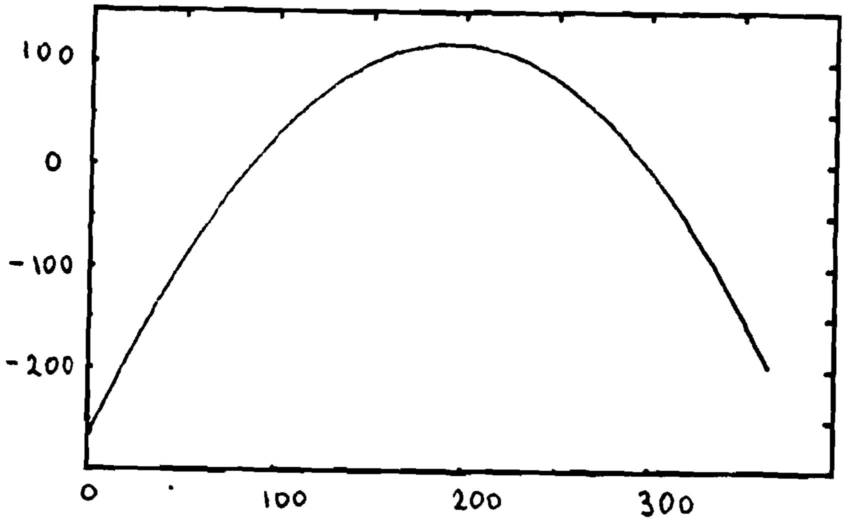


fig. 2.12 Example 7 : second order year-trend fitted to the data.

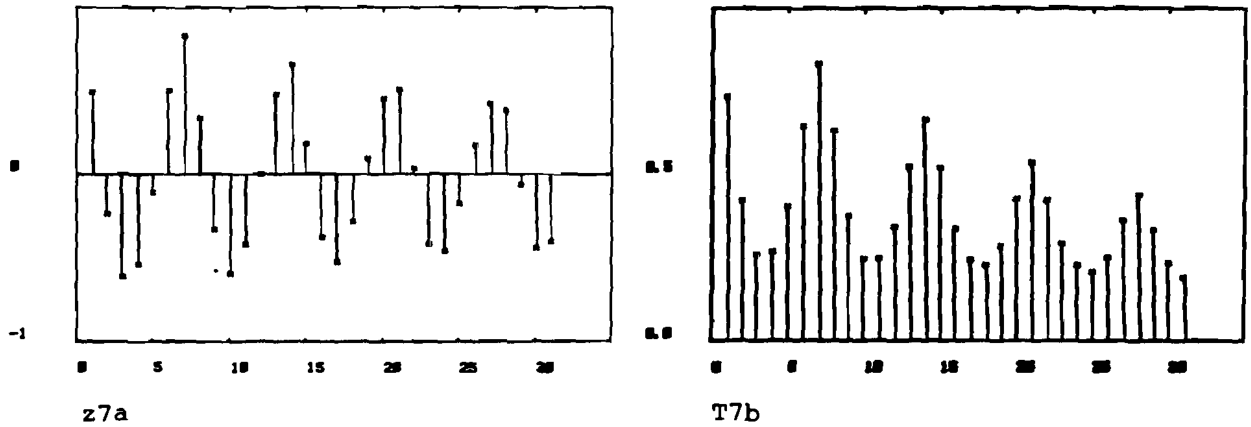


fig. 2.13 Example 7 : a.c.f of series z7a and T7b.

APPENDIX III Figures of chapter 3.

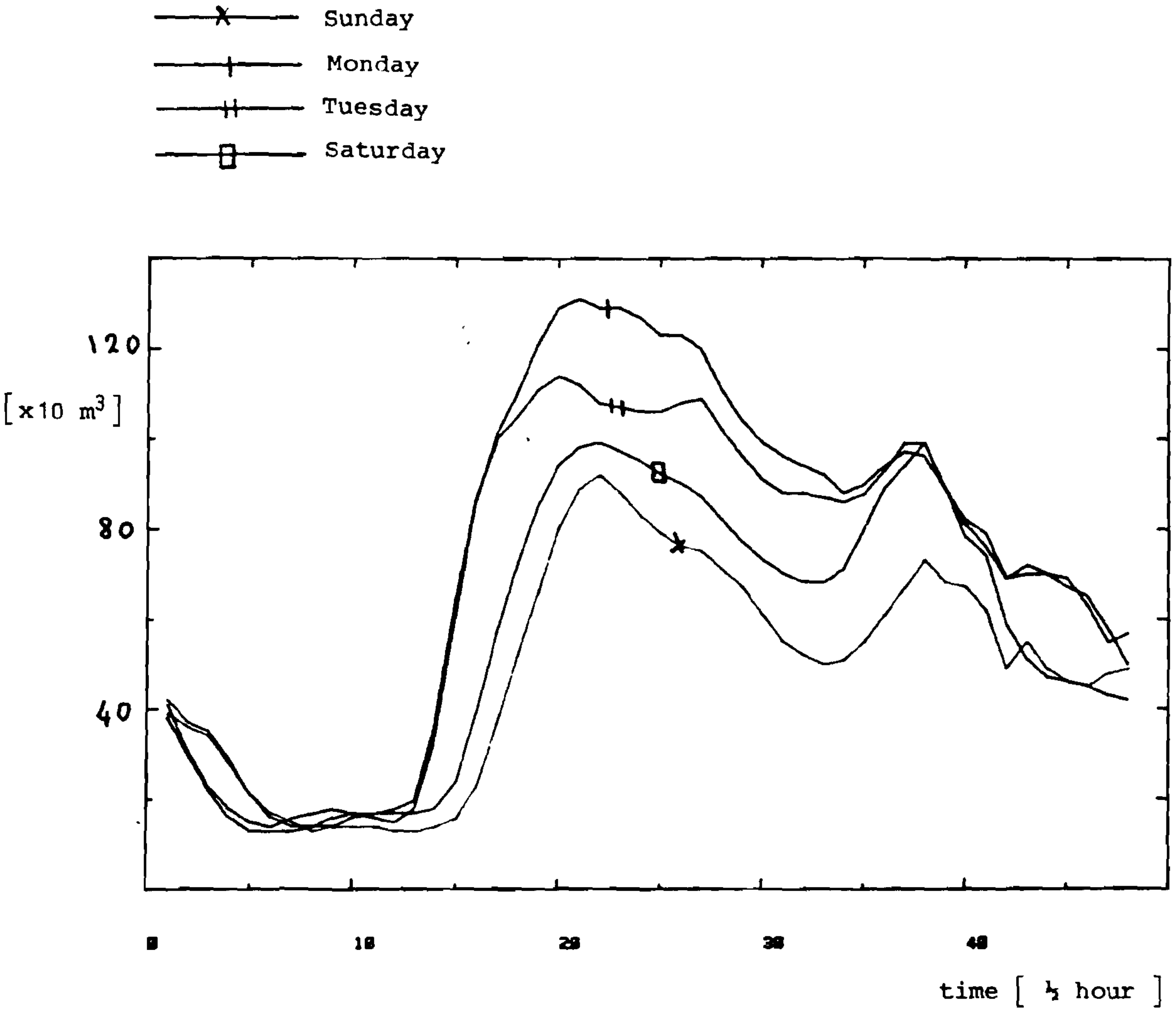


fig. 3.1 : Average daily-pattern for the various days of the week.

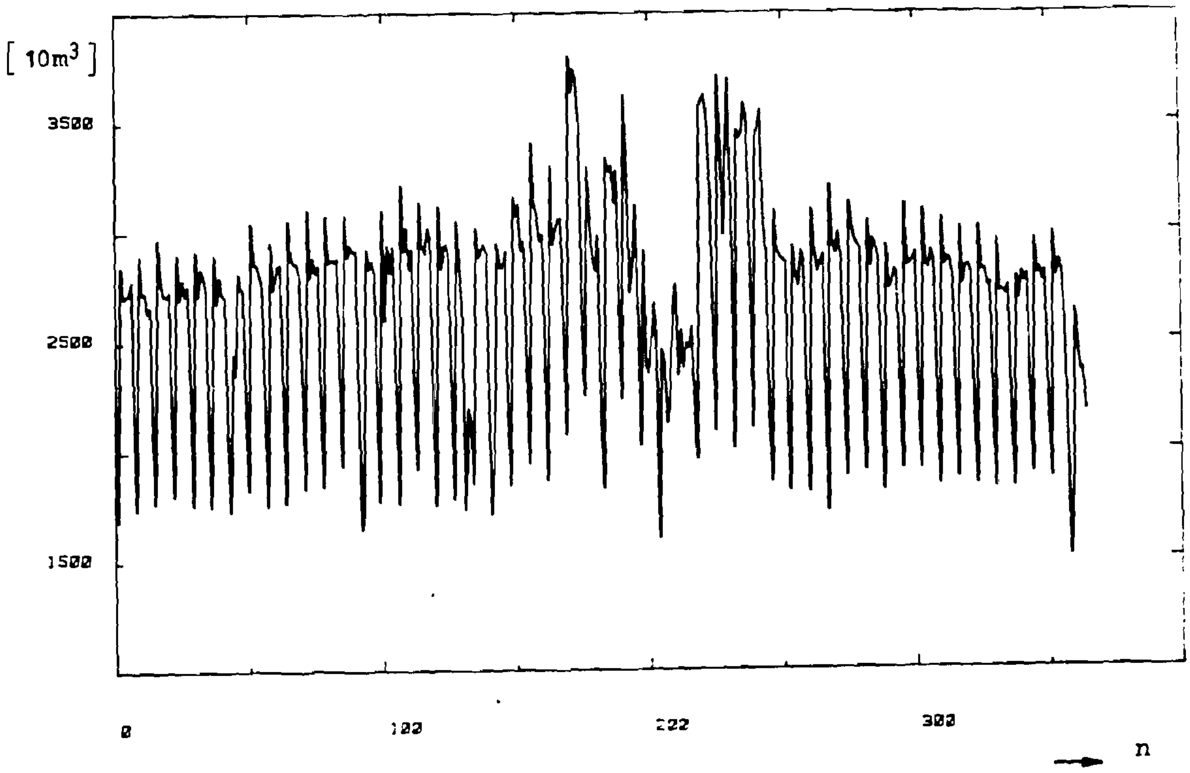


fig. 3.2 : series of daily totals of 1983.

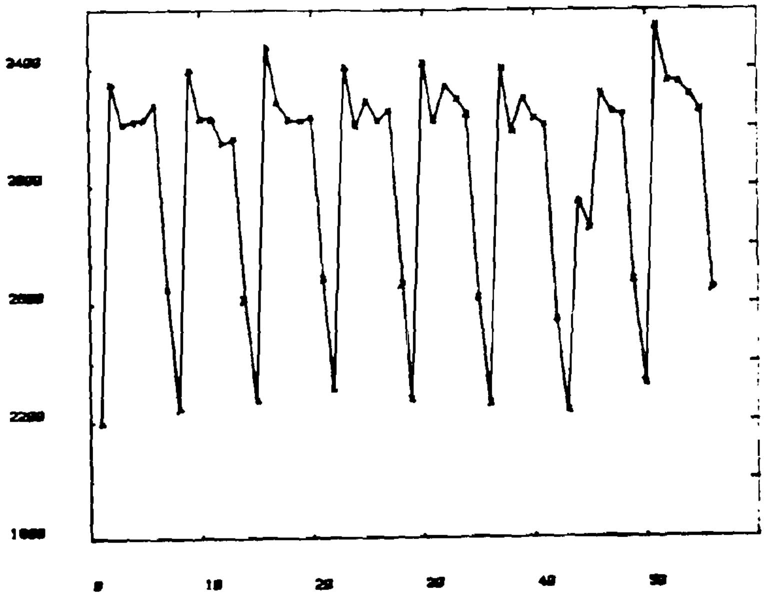


fig. 3.3 : first 8 weeks of daily totals series 1983.

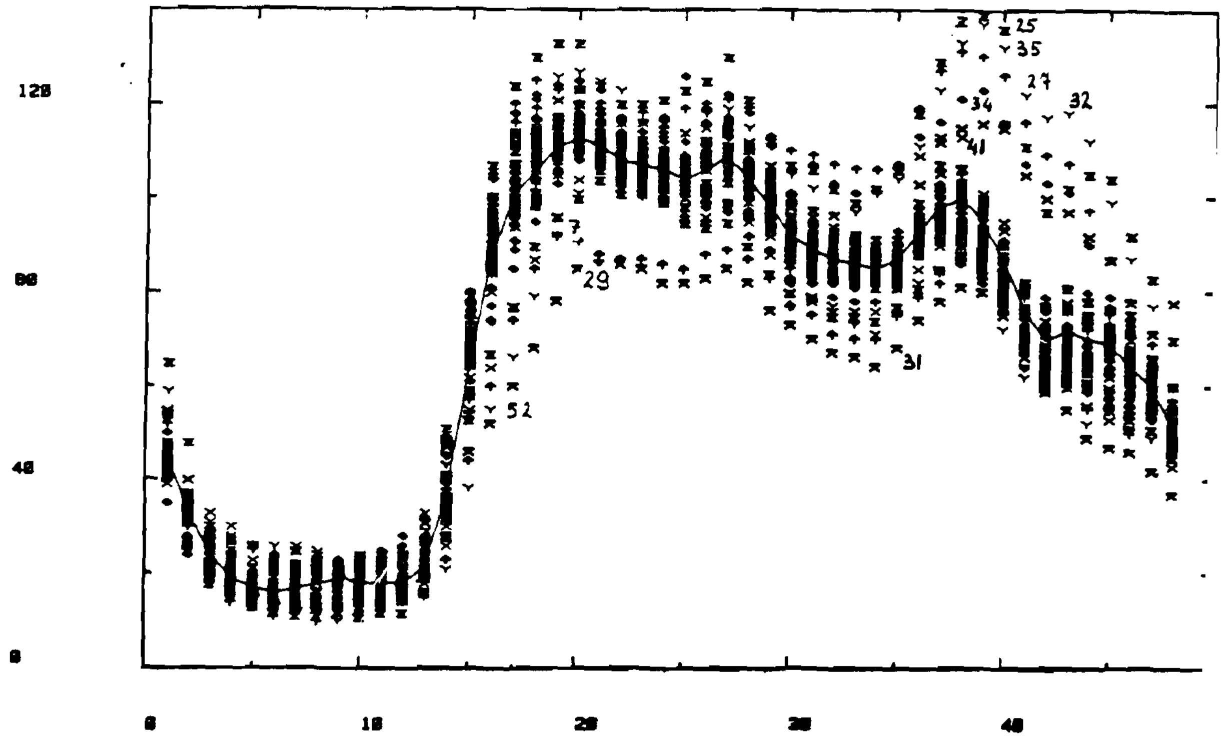


fig. 3.4 : Spread of the Wednesdays about the average Wednesday-pattern

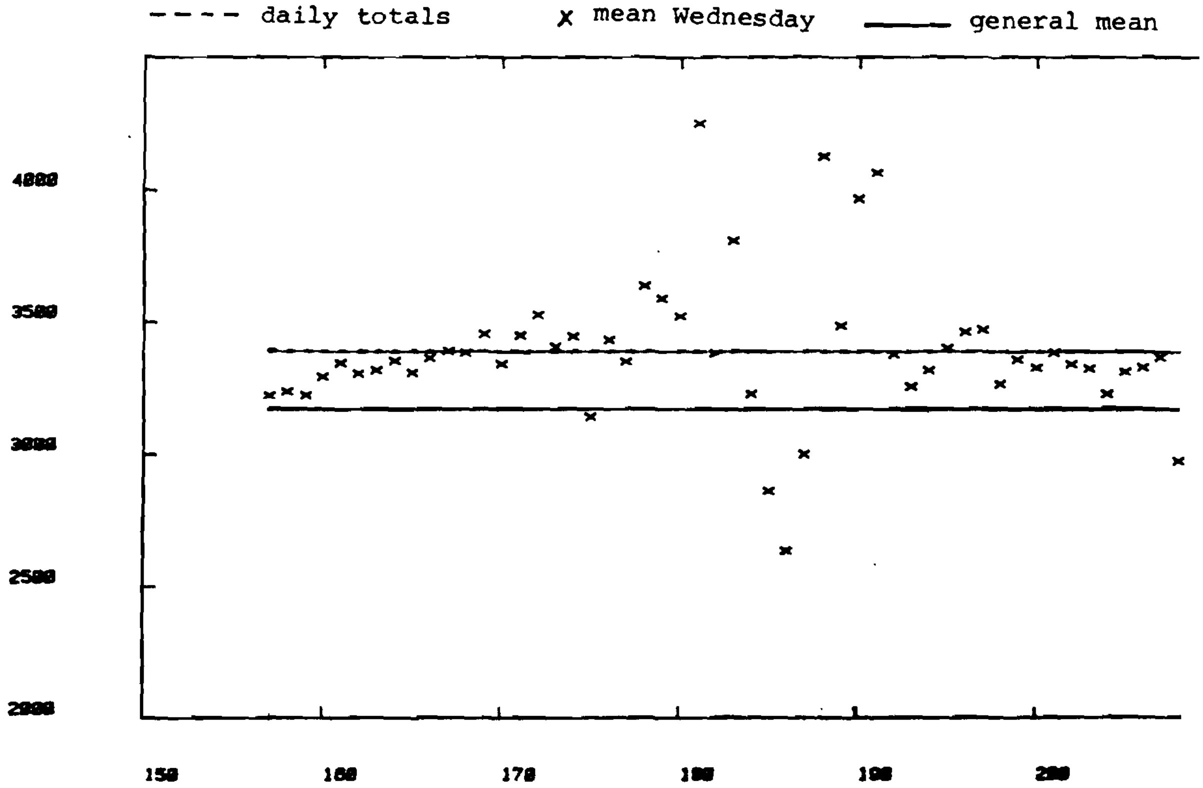


fig. 3.5 : Daily totals Wednesday, mean Wednesday, and general me

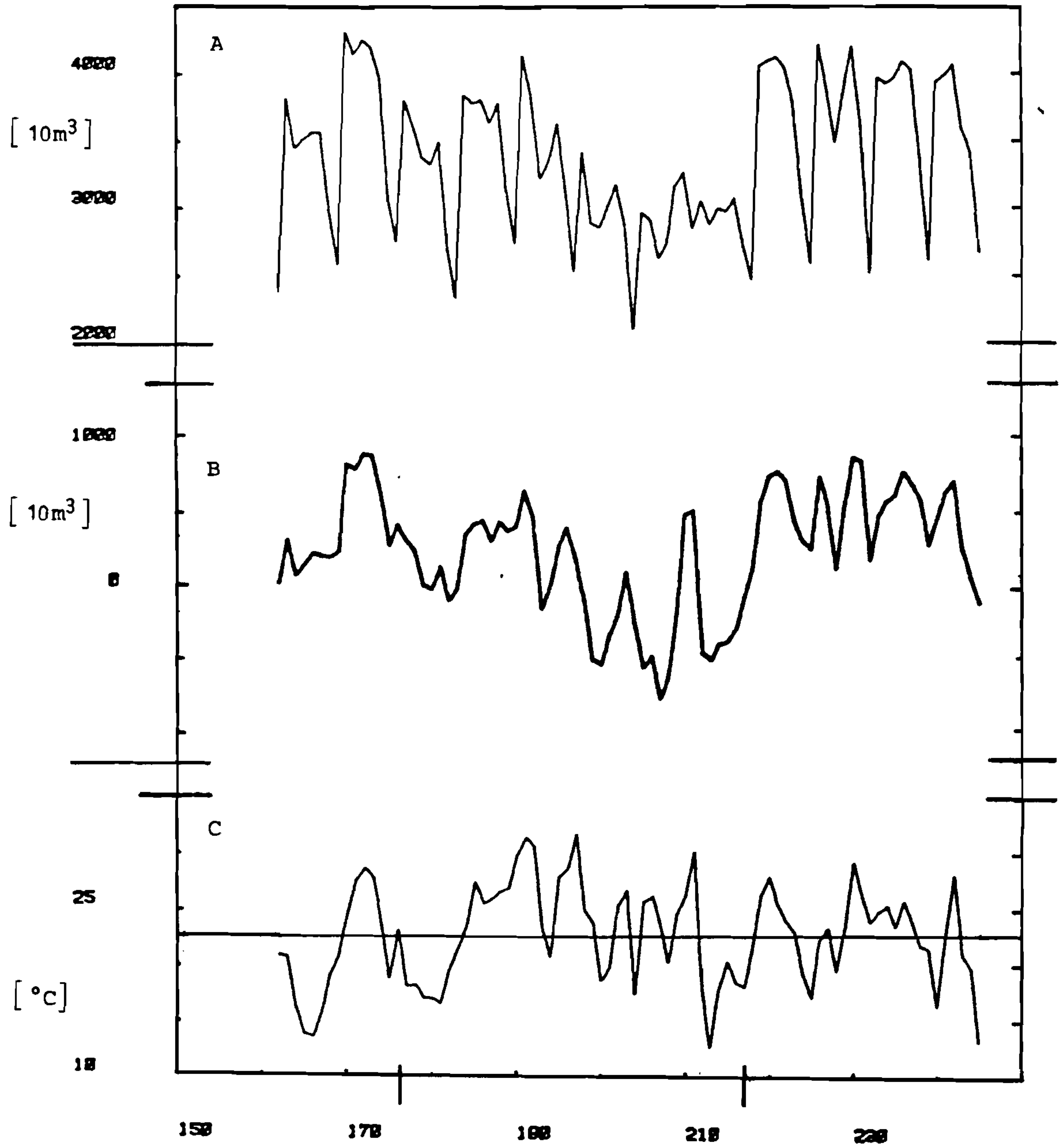


fig. 3.6 Weather influence ,

A : daily totals

B : deviation of the mean values

C : temperature

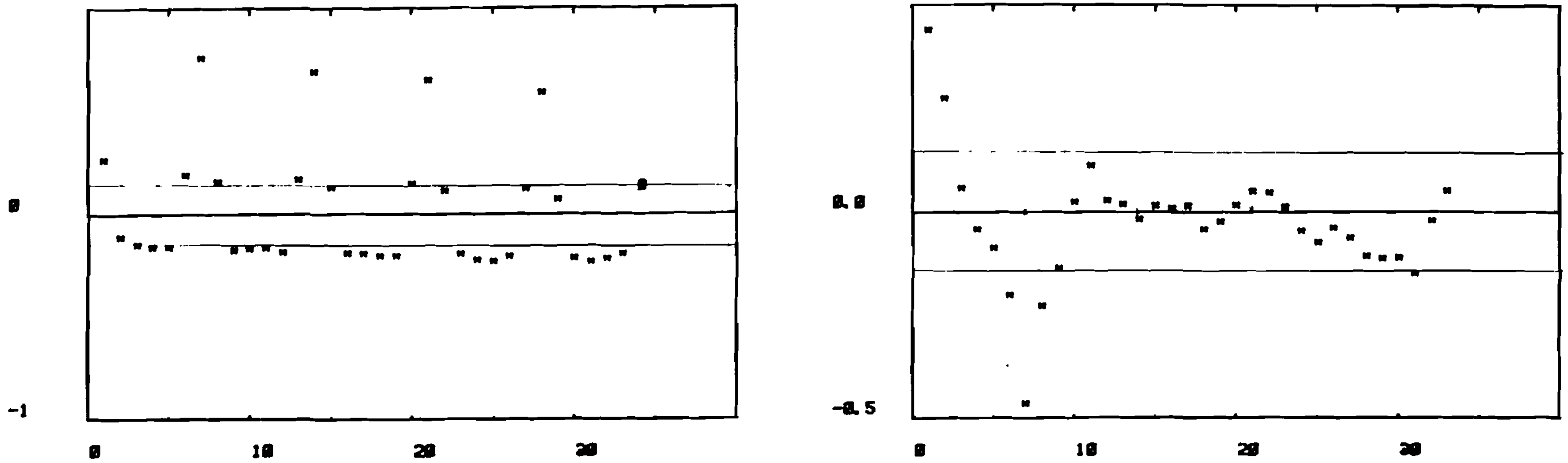
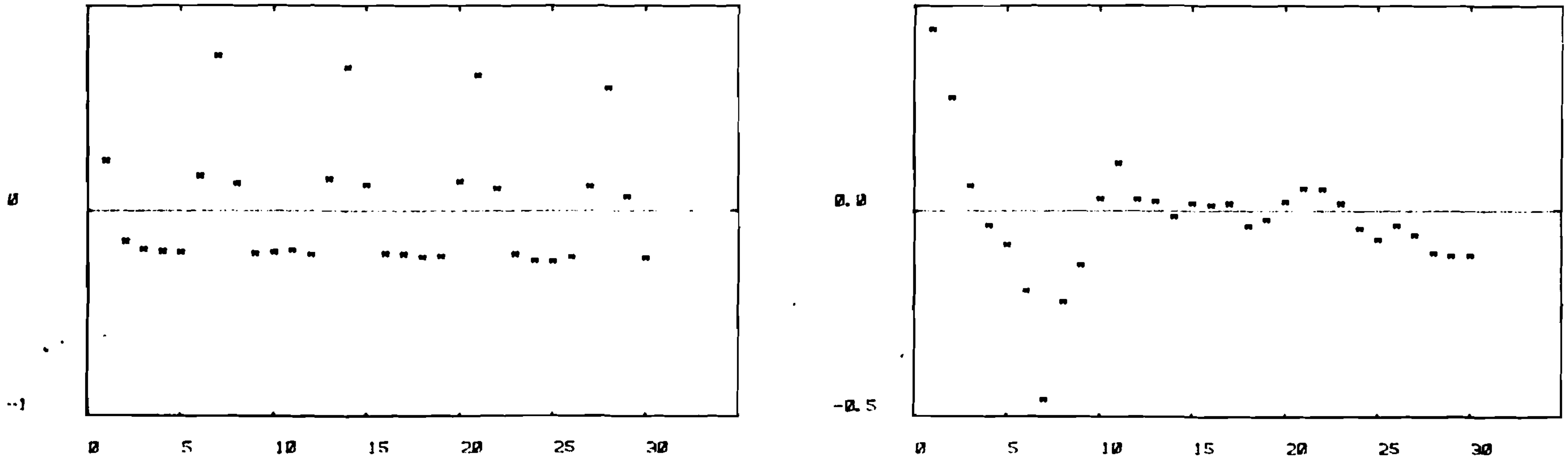


fig. 3.7a : a.c.f. first 182 samples original series and of' first order seasonal differenced series.



fia. 3.7b : a.c.f after subtracting the 'year-trend'.

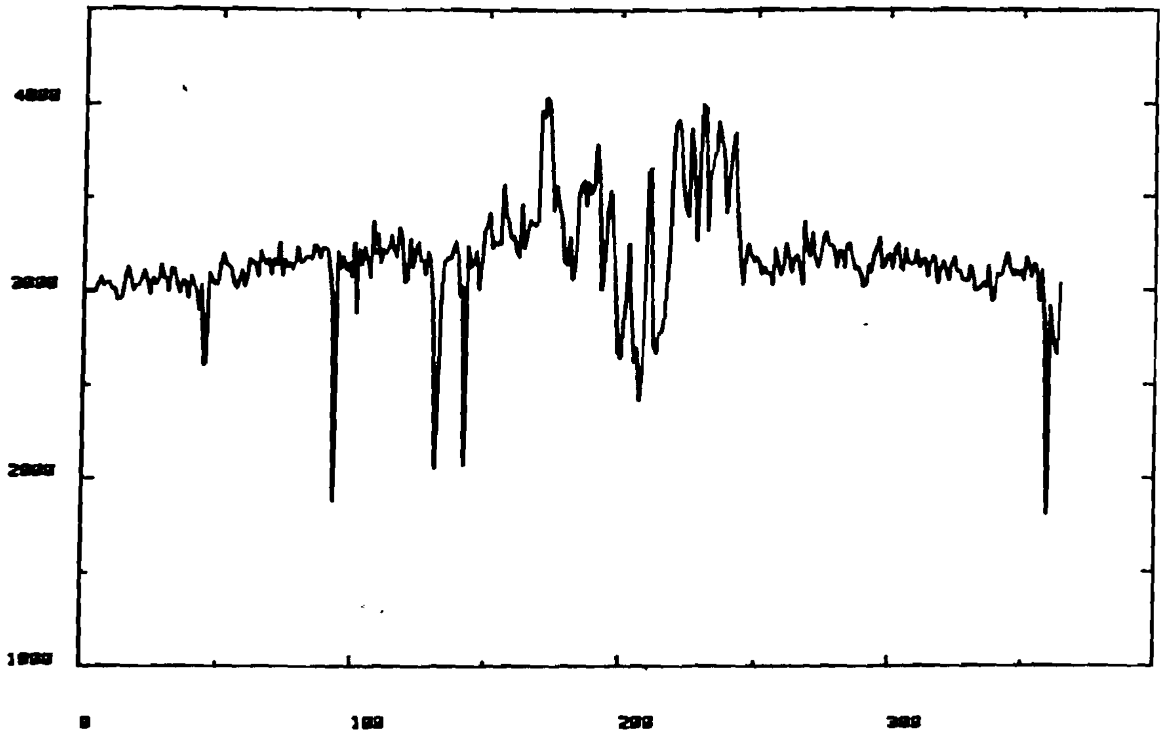


fig. 3.8 : series of daily totals adjusted for the average week-pattern.(series zc_t)

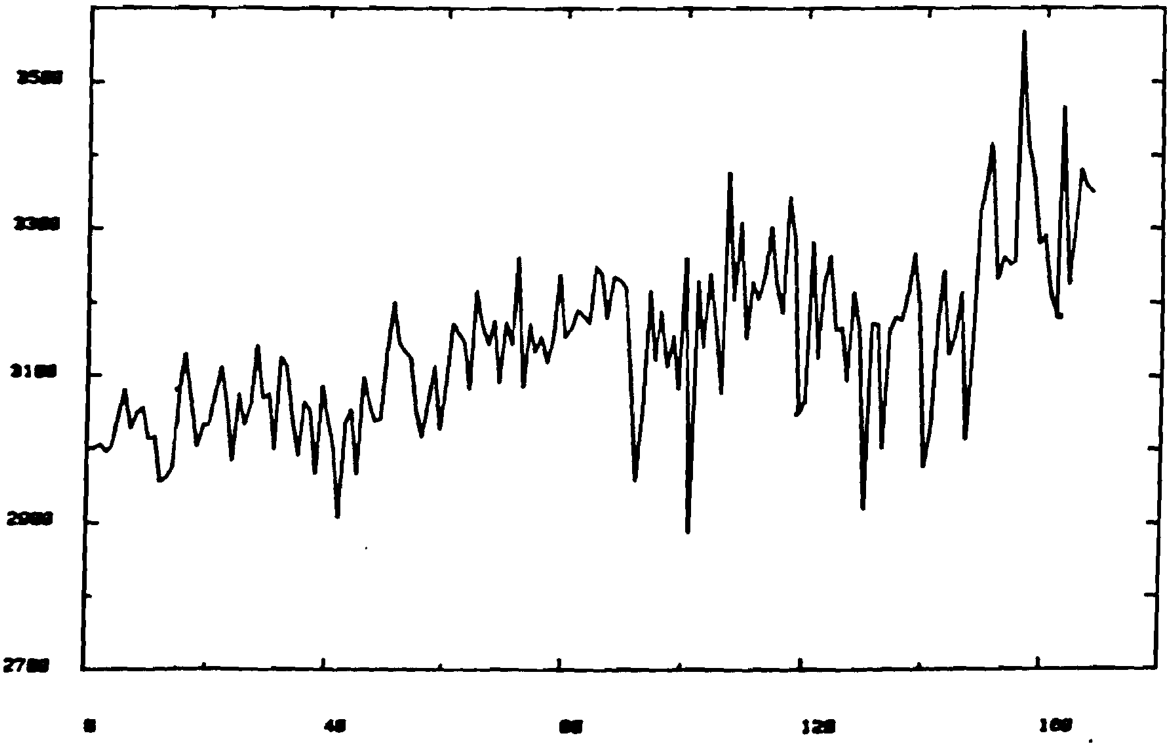


fig. 3.9 : first 168 samples of series zc_t adjusted for the clear deviations on holidays.(series $zca1_t$)

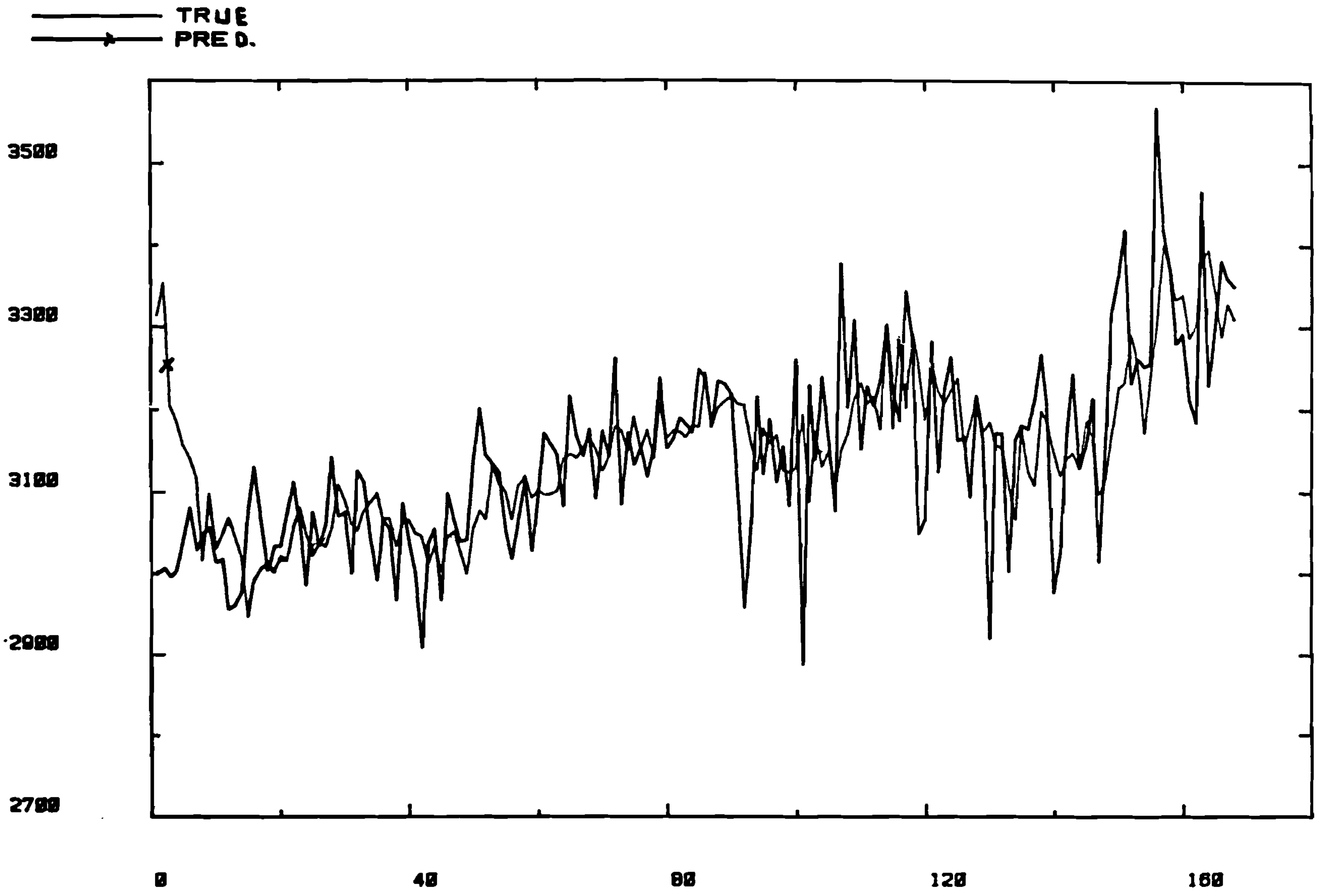


fig. 3.10a : series zca1_t, original and forecasted series.

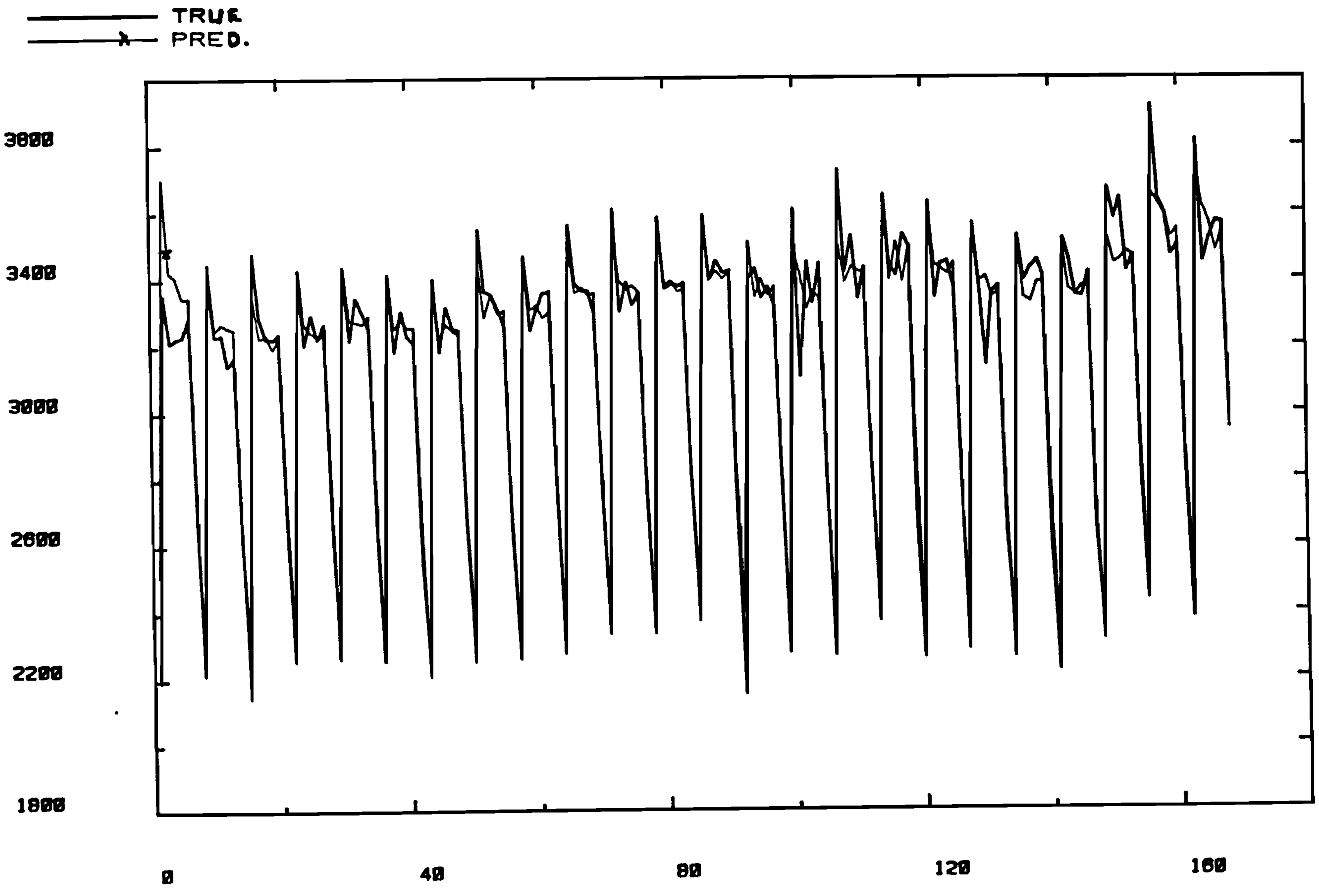


fig. 3 10b : series $zca1_t$, after re-adjusting for the average weekpattern.

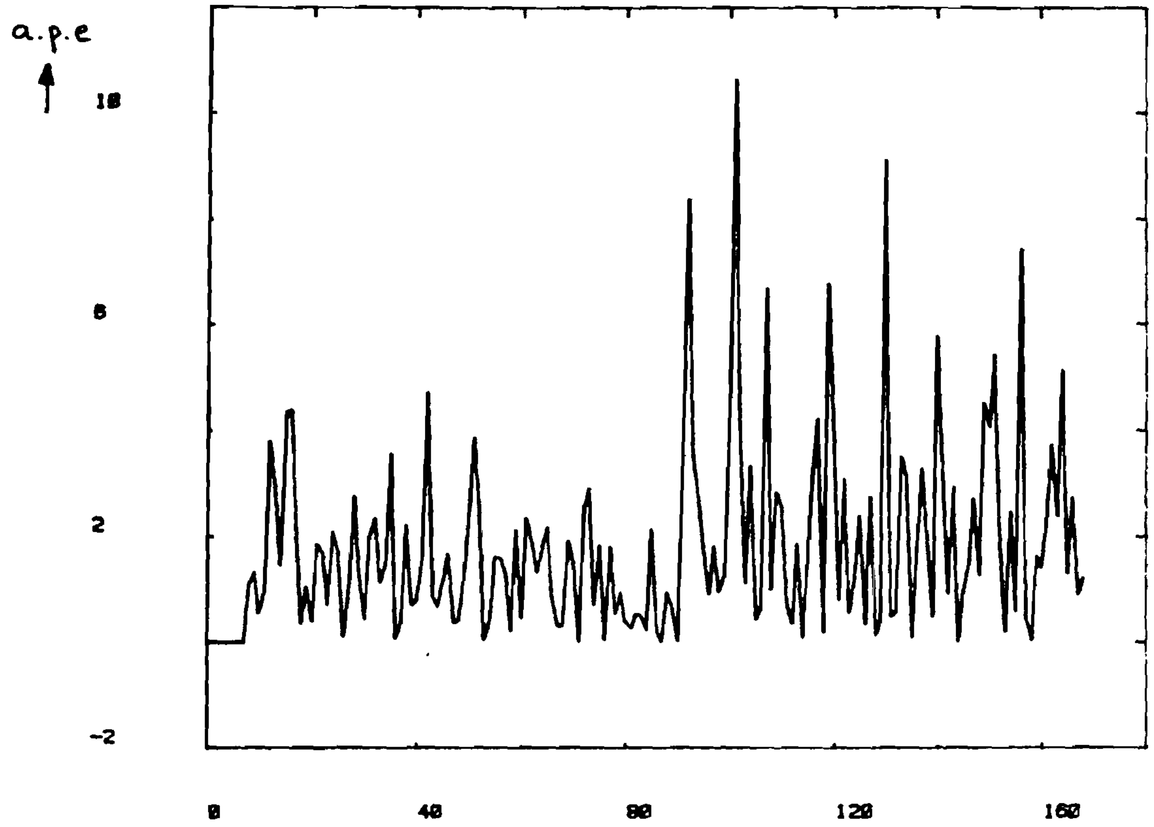
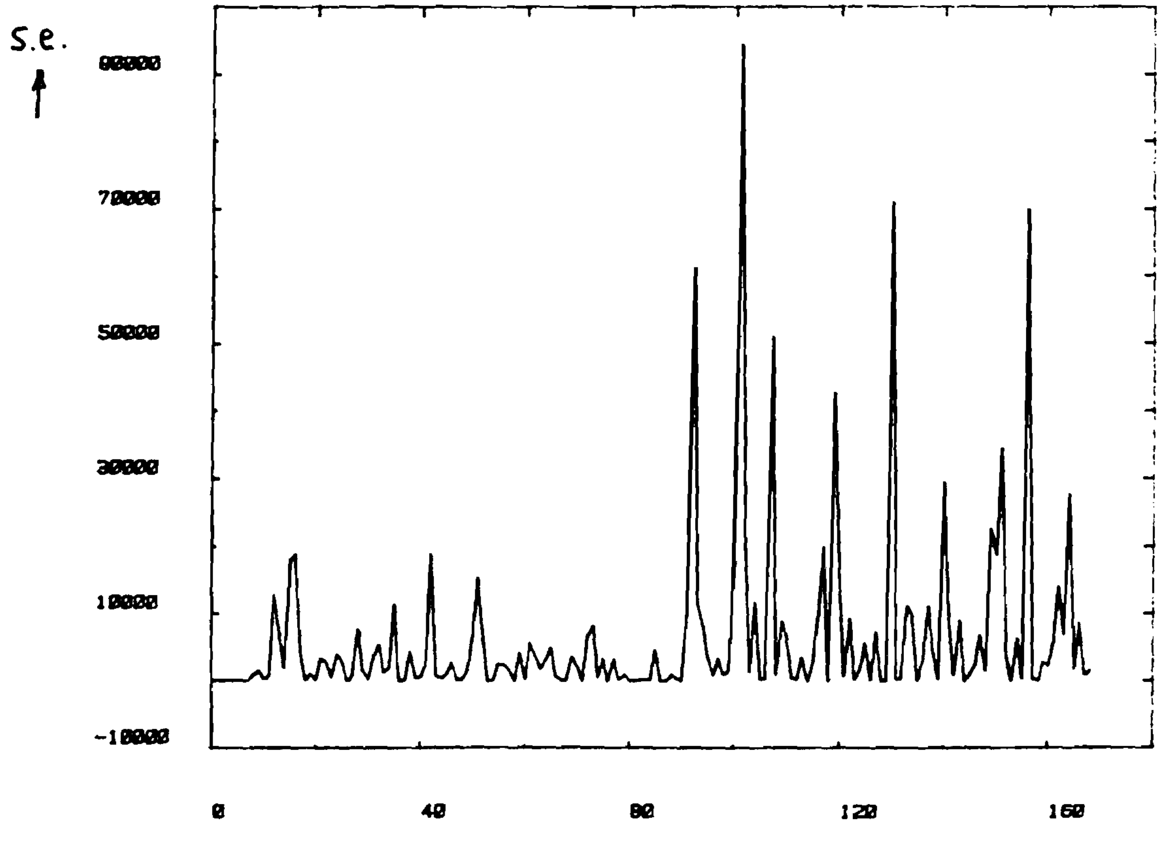


fig. 3.11 : forecast errors of series zca1_t.



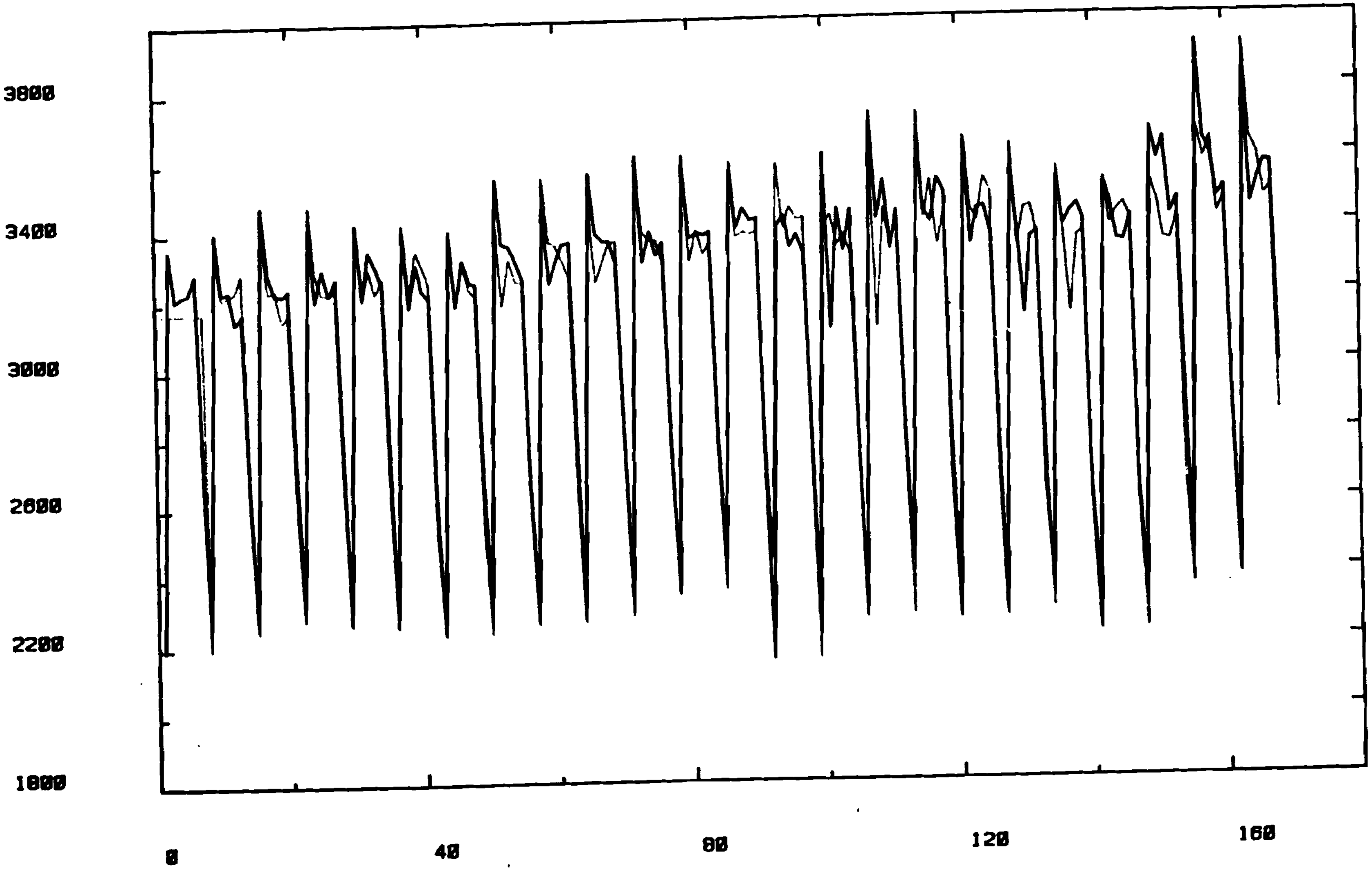
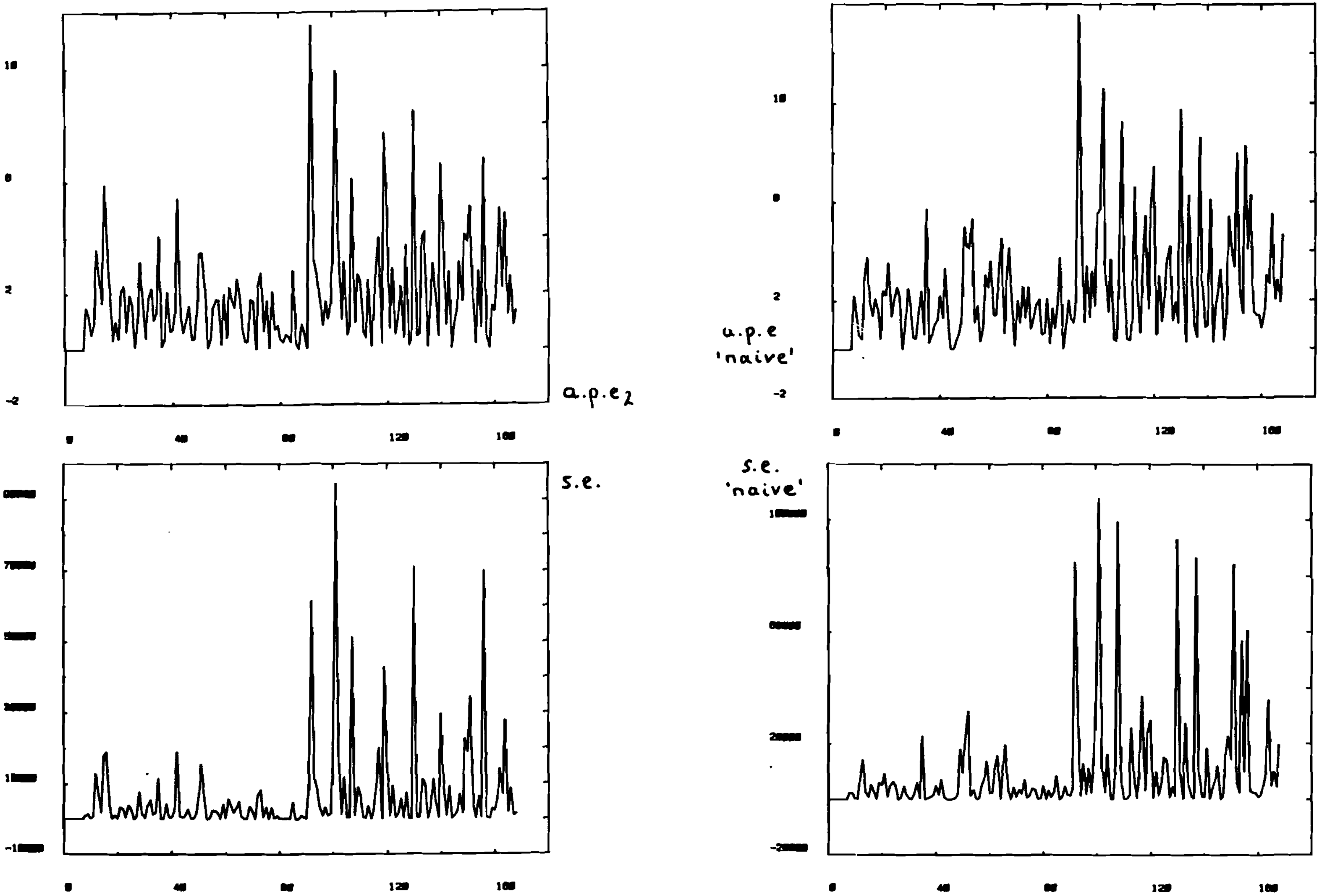


fig. 3.12 : series zca1$_t$, original series and forecasted ('naive' method) series.



fiq. 3.13 : series zca1 , forecast errors. (ARIMA-model and 'naive' method)

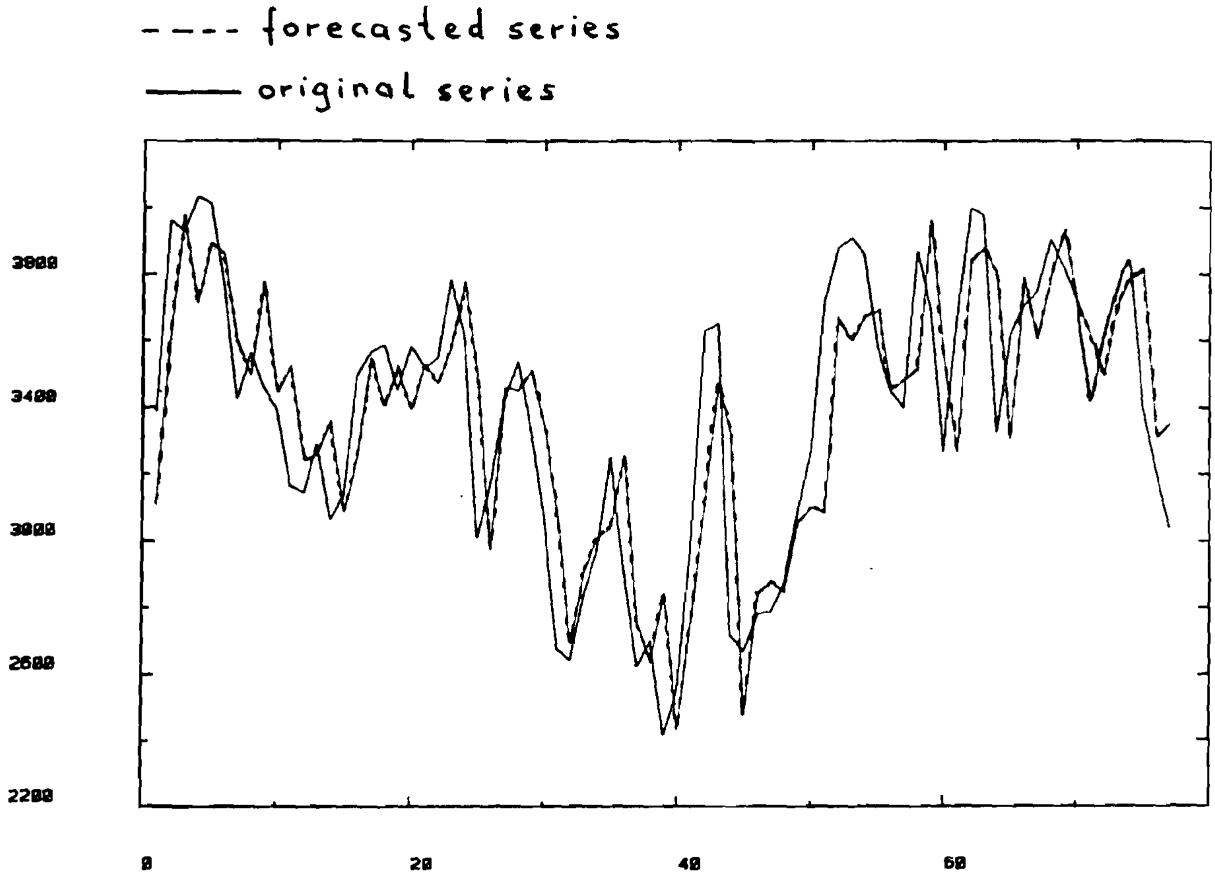


fig. 3.14 : series $zca2_t$, original and forecasted series.

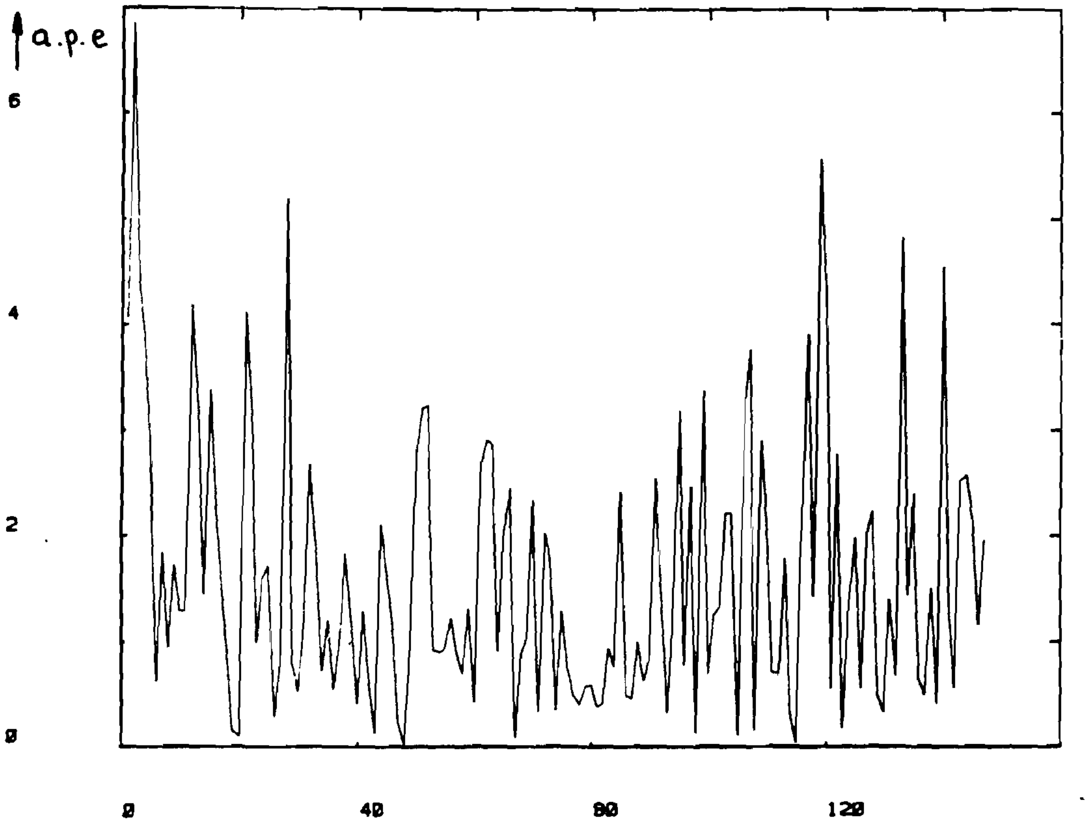
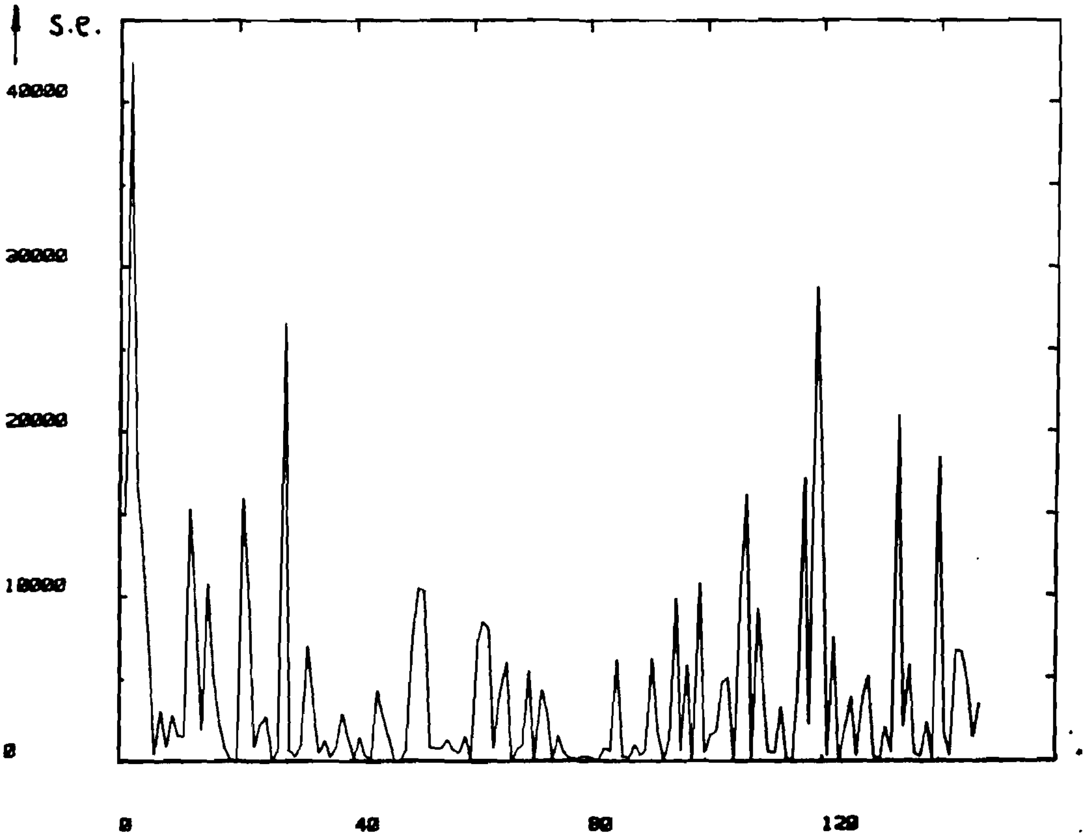
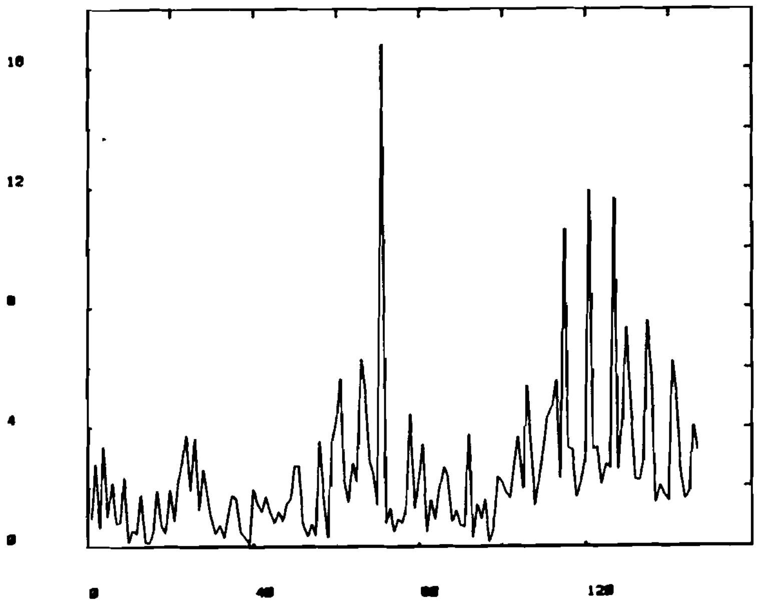
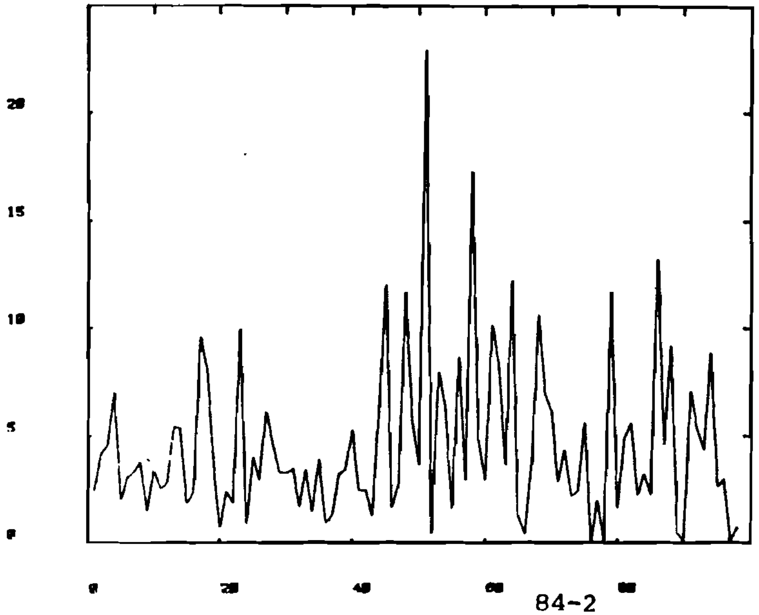


fig. 3.15 : series $zcb1_t$, forecast errors.

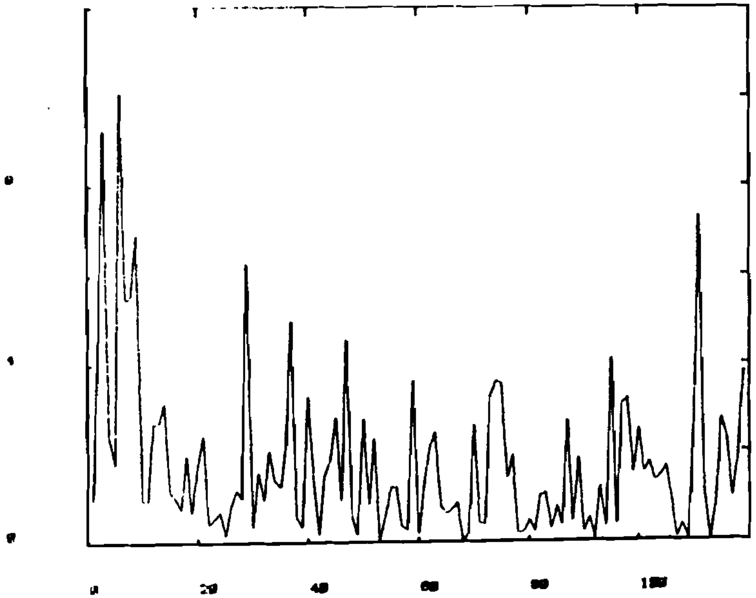




84-1

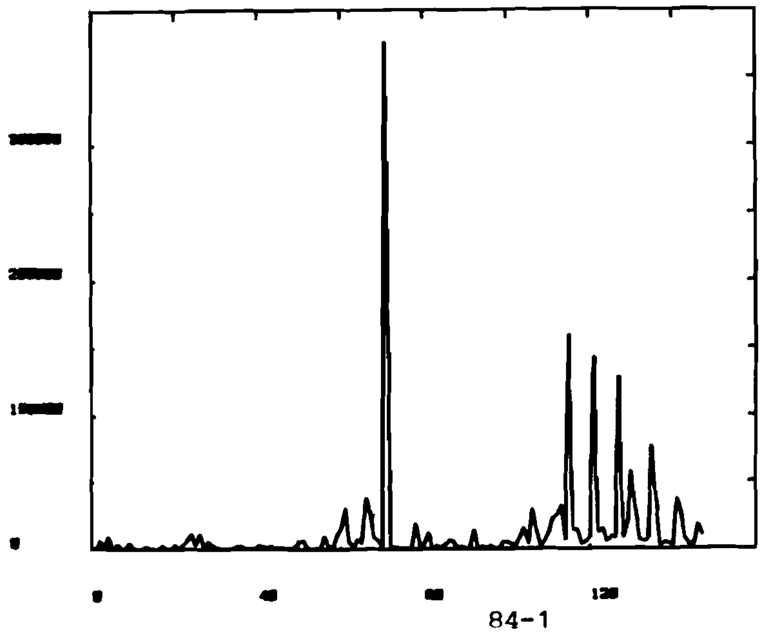


84-2

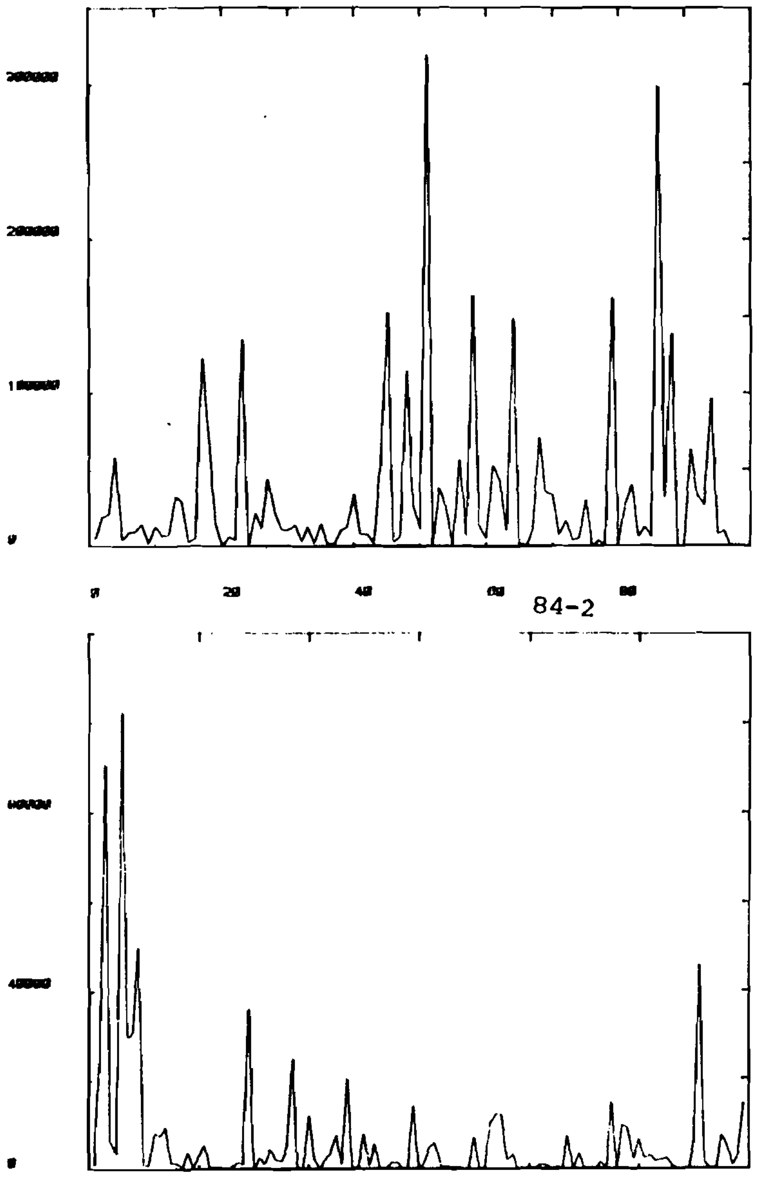


84-3

fig. 3.16 : series 84, absolute percentage errors.



84-1



84-2

fig. 3.17 : series 84, square errors.

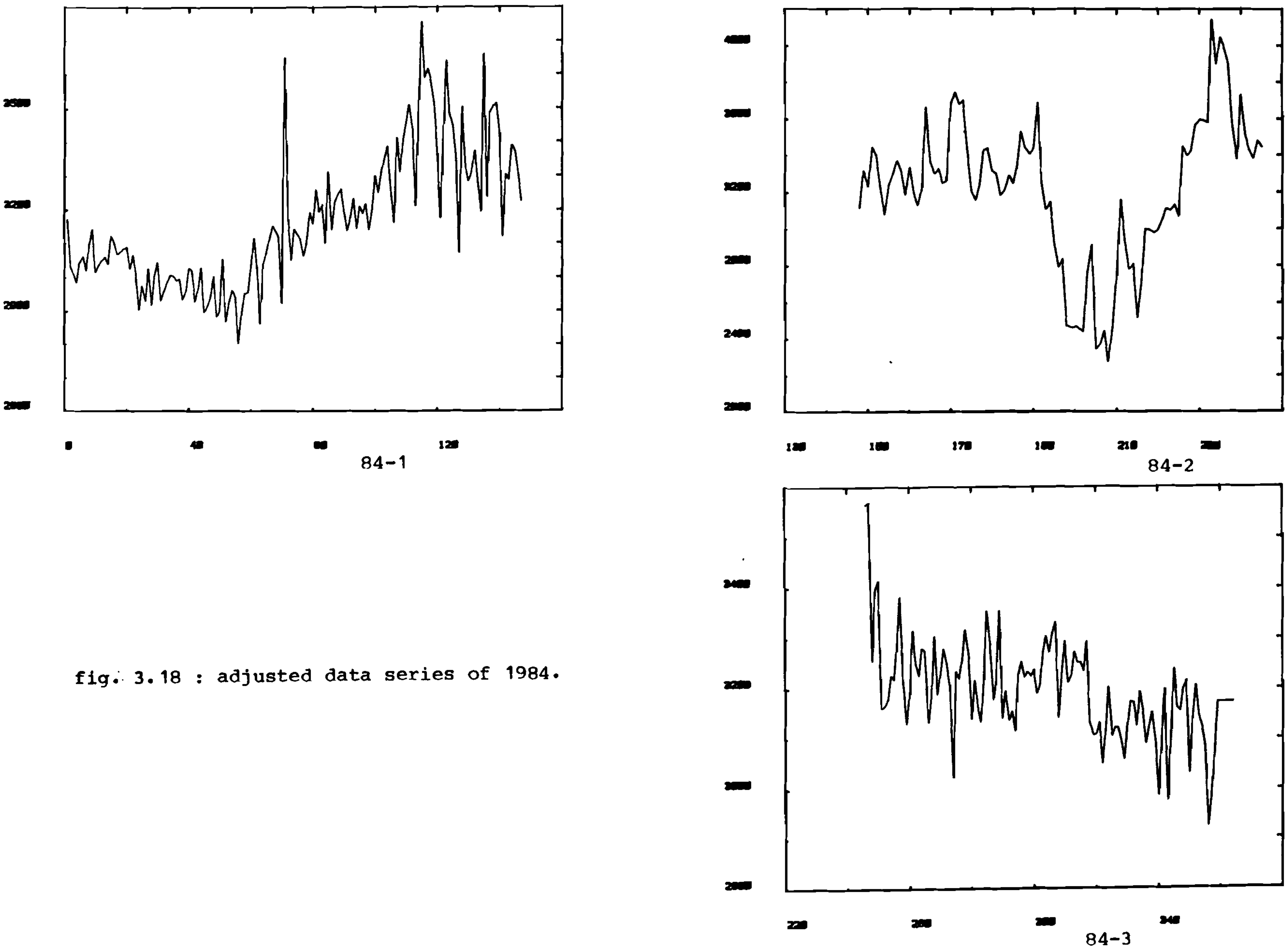


fig. 3.18 : adjusted data series of 1984.

APPENDIX 4 The water-company data.

The original series of the daily totals of 1983 and 1984 are listed on the next pages; that is page 99 series 1983,
page 100 series 1984.

To the adjusted series is refered as follows:

series zc_t : the original series of 1983 adjusted for the average week-pattern of 1983. (see page 55).

series zca_t : series zc_t adjusted for only the clear deviations, (see page 55), divided into :

series $zca1_t$	samples :	1 - 168
series $zca2_t$	samples :	169 - 245
series $zca3_t$	samples :	246 - 364.

series zcb_t : series zc_t adjusted more comprehensively in comparison to series zca_t, (see page 60), divided into :

series $zcb1_t$	samples :	1 - 147
series $zcb2_t$	samples :	148 - 245
series $zcb3_t$	samples :	246 - 364.

The adjusted series of 1984 (page 62) is divided into :

series 84-1	samples :	1 - 147
series 84-2	samples :	148 - 245
series 84-3	samples :	246 - 364.

1	0.21940E+04	53	0.33530E+04	105	0.28000E+04	157	0.36340E+04	209	0.31270E+04	261	0.36030E+04	313	0.23580E+04
2	0.33540E+04	54	0.33100E+04	106	0.22690E+04	158	0.35890E+04	210	0.32680E+04	262	0.34050E+04	314	0.32840E+04
3	0.32120E+04	55	0.32550E+04	107	0.37240E+04	159	0.34670E+04	211	0.28490E+04	263	0.33180E+04	315	0.27790E+04
4	0.32240E+04	56	0.26520E+04	108	0.34180E+04	160	0.34970E+04	212	0.30590E+04	264	0.33650E+04	316	0.23770E+04
5	0.32310E+04	57	0.22610E+04	109	0.35270E+04	161	0.28460E+04	213	0.28830E+04	265	0.23990E+04	317	0.35180E+04
6	0.32860E+04	58	0.34610E+04	110	0.33380E+04	162	0.23770E+04	214	0.30000E+04	266	0.27550E+04	318	0.32560E+04
7	0.26640E+04	59	0.32440E+04	111	0.34330E+04	163	0.28130E+04	215	0.29780E+04	267	0.21170E+04	319	0.33400E+04
8	0.22430E+04	60	0.33070E+04	112	0.28390E+04	164	0.34450E+04	216	0.30760E+04	268	0.37160E+04	320	0.33450E+04
9	0.34040E+04	61	0.33580E+04	113	0.24290E+04	165	0.35190E+04	217	0.27200E+04	269	0.34460E+04	321	0.33160E+04
10	0.32400E+04	62	0.33630E+04	114	0.36500E+04	166	0.35680E+04	218	0.24680E+04	270	0.34000E+04	322	0.26930E+04
11	0.32370E+04	63	0.27790E+04	115	0.34400E+04	167	0.35640E+04	219	0.40660E+04	271	0.34590E+04	323	0.23420E+04
12	0.31430E+04	64	0.22760E+04	116	0.34050E+04	168	0.29850E+04	220	0.41000E+04	272	0.33750E+04	324	0.35230E+04
13	0.31670E+04	65	0.35630E+04	117	0.35300E+04	169	0.25820E+04	221	0.41280E+04	273	0.27640E+04	325	0.33190E+04
14	0.26120E+04	66	0.33840E+04	118	0.34900E+04	170	0.43090E+04	222	0.40430E+04	274	0.23840E+04	326	0.33240E+04
15	0.22740E+04	67	0.33630E+04	119	0.26840E+04	171	0.41440E+04	223	0.37830E+04	275	0.35340E+04	327	0.32490E+04
16	0.34770E+04	68	0.33630E+04	120	0.22610E+04	172	0.42520E+04	224	0.30790E+04	276	0.35440E+04	328	0.32600E+04
17	0.37830E+04	69	0.32970E+04	121	0.36290E+04	173	0.41990E+04	225	0.25910E+04	277	0.34610E+04	329	0.26830E+04
18	0.32230E+04	70	0.28090E+04	122	0.33400E+04	174	0.39710E+04	226	0.42160E+04	278	0.34760E+04	330	0.23260E+04
19	0.32210E+04	71	0.23360E+04	123	0.34430E+04	175	0.30620E+04	227	0.39110E+04	279	0.34200E+04	331	0.34600E+04
20	0.32390E+04	72	0.36090E+04	124	0.34510E+04	176	0.27580E+04	228	0 34850E+04	280	0.27610E+04	332	0.32160E+04
21	0.27080E+04	73	0.33000E+04	125	0.33670E+04	177	0.38060E+04	229	0.38610E+04	281	0.24100E+04	333	0.32260E+04
22	0.23050E+04	74	0.33910E+04	126	0.28020E+04	178	0.36050E+04	230	0.42050E+04	282	0.35540E+04	334	0.32080E+04
23	0.34070E+04	75	0.33200E+04	127	0.22870E+04	179	0.33800E+04	231	0.36110E+04	283	0.33400E+04	335	0.32510E+04
24	0.32020E+04	76	0.33590E+04	128	0.35600E+04	180	0.33280E+04	232	0.25150E+04	284	0.34680E+04	336	0.26520E+04
25	0.32940E+04	77	0.27530E+04	129	0.33760E+04	181	0.34950E+04	233	0.39680E+04	285	0.34430E+04	337	0.23290E+04
26	0.37210E+04	78	0.23500E+04	130	0.31380E+04	182	0.26980E+04	234	0.39250E+04	286	0.33750E+04	338	0.22900E+04
27	0.32670E+04	79	0.35850E+04	131	0.22430E+04	183	0.23350E+04	235	0.39680E+04	287	0.27520E+04	339	0.31840E+04
28	0.27770E+04	80	0.33690E+04	132	0.27070E+04	184	0.38440E+04	236	0.40910E+04	288	0.27200E+04	340	0.33110E+04
29	0.22640E+04	81	0.33840E+04	133	0.26370E+04	185	0.37850E+04	237	0.40250E+04	289	0.34390E+04	341	0.32650E+04
30	0.34230E+04	82	0.33770E+04	134	0.23570E+04	186	0.38080E+04	238	0.33580E+04	290	0.32390E+04	342	0.32990E+04
31	0.32160E+04	83	0.33870E+04	135	0.35270E+04	187	0.36390E+04	239	0.26100E+04	291	0.32630E+04	343	0.27940E+04
32	0.33450E+04	84	0.28070E+04	136	0.33920E+04	188	0.37870E+04	240	0.39410E+04	292	0.33270E+04	344	0.23950E+04
33	0.33000E+04	85	0.24410E+04	137	0.34320E+04	189	0.31570E+04	241	0.39890E+04	293	0.33070E+04	345	0.34630E+04
34	0.32470E+04	86	0.35850E+04	138	0.34530E+04	190	0.27440E+04	242	0.40630E+04	294	0.28260E+04	346	0.32950E+04
35	0.26270E+04	87	0.33950E+04	139	0.34060E+04	191	0.41310E+04	243	0.35780E+04	295	0.24200E+04	347	0.33280E+04
36	0.22580E+04	88	0.34530E+04	140	0.26110E+04	192	0.38260E+04	244	0.34080E+04	296	0.36300E+04	348	0.32670E+04
37	0.34020E+04	89	0.34180E+04	141	0.22220E+04	193	0.32260E+04	245	0.26700E+04	297	0.33380E+04	349	0.33120E+04
38	0.31830E+04	90	0.34230E+04	142	0.24180E+04	194	0.33570E+04	246	0.23610E+04	298	0.33570E+04	350	0.26910E+04
39	0.33050E+04	91	0.27420E+04	143	0.34590E+04	195	0.36310E+04	247	0.36010E+04	299	0.33700E+04	351	0.23700E+04
40	0.32290E+04	92	0.21500E+04	144	0.33480E+04	196	0.31720E+04	248	0.34070E+04	300	0.34050E+04	352	0.34920E+04
41	0.32060E+04	93	0.22320E+04	145	0.33440E+04	197	0.25310E+04	249	0.33750E+04	301	0.27550E+04	353	0.32870E+04
42	0.25430E+04	94	0.34310E+04	146	0.34180E+04	198	0.34200E+04	250	0.33670E+04	302	0.24160E+04	354	0.33650E+04
43	0.22320E+04	95	0.33400E+04	147	0.26490E+04	199	0.28870E+04	251	0.33520E+04	303	0.36000E+04	355	0.33150E+04
44	0.29520E+04	96	0.33750E+04	148	0.23490E+04	200	0.28590E+04	252	0.27240E+04	304	0.33400E+04	356	0.31500E+04
45	0.28580E+04	97	0.33170E+04	149	0.36690E+04	201	0.30270E+04	253	0.23250E+04	305	0.33280E+04	357	0.27480E+04
46	0.33170E+04	98	0.27900E+04	150	0.35790E+04	202	0.31810E+04	254	0.34380E+04	306	0.34150E+04	358	0.20120E+04
47	0.32510E+04	99	0.22760E+04	151	0.36380E+04	203	0.28860E+04	255	0.33120E+04	307	0.33560E+04	359	0.21540E+04
48	0.32450E+04	100	0.36070E+04	152	0.34190E+04	204	0.21040E+04	256	0.32520E+04	308	0.27730E+04	360	0.41370E+04
49	0.26770E+04	101	0.31030E+04	153	0.34670E+04	205	0.29690E+04	257	0.34150E+04	309	0.23460E+04	361	0.29750E+04
50	0.23300E+04	102	0.34480E+04	154	0.28870E+04	206	0.29160E+04	258	0.33610E+04	310	0.35640E+04	362	0.28540E+04
51	0.35480E+04	103	0.33250E+04	155	0.24490E+04	207	0.26330E+04	259	0.27200E+04	311	0.33410E+04	363	0.28660E+04
52	0.33610E+04	104	0.34440E+04	156	0.39150E+04	208	0.27420E+04	260	0.23140E+04	312	0.33830E+04	364	0.26740E+04

series : daily totals of 1983.

1	0.21420E+04	53	0.31450E+04	105	0.28820E+04	157	0.35860E+04	209	0.26790E+04	261	0.36620E+04	313	0.34590E+04
2	0.33770E+04	54	0.31480E+04	106	0.23520E+04	158	0.35360E+04	210	0.23910E+04	262	0.34630E+04	314	0.34560E+04
3	0.32230E+04	55	0.31440E+04	107	0.37580E+04	159	0.33710E+04	211	0.23480E+04	263	0.34430E+04	315	0.28860E+04
4	0.32010E+04	56	0.24330E+04	108	0.35240E+04	160	0.35390E+04	212	0.32790E+04	264	0.34660E+04	316	0.24250E+04
5	0.32290E+04	57	0.20780E+04	109	0.36250E+04	161	0.28280E+04	213	0.29900E+04	265	0.34780E+04	317	0.36390E+04
6	0.32660E+04	58	0.32970E+04	110	0.36420E+04	162	0.21860E+04	214	0.30240E+04	266	0.27650E+04	318	0.33450E+04
7	0.26510E+04	59	0.31690E+04	111	0.37140E+04	163	0.23870E+04	215	0.26970E+04	267	0.23740E+04	319	0.33230E+04
8	0.22890E+04	60	0.32450E+04	112	0.30710E+04	164	0.38740E+04	216	0.28970E+04	268	0.36500E+04	320	0.32950E+04
9	0.34870E+04	61	0.32990E+04	113	0.21920E+04	165	0.35790E+04	217	0.26300E+04	269	0.34020E+04	321	0.33360E+04
10	0.32280E+04	62	0.32320E+04	114	0.24670E+04	166	0.34840E+04	218	0.21820E+04	270	0.34500E+04	322	0.26820E+04
11	0.32520E+04	63	0.24930E+04	115	0.39700E+04	167	0.35310E+04	219	0.3[illegible]20E+04	271	0.34650E+04	323	0.23090E+04
12	0.32360E+04	64	0.22300E+04	116	0.38050E+04	168	0.28830E+04	220	0.32100E+04	272	0.34540E+04	324	0.35480E+04
13	0.32640E+04	65	0.28890E+04	117	0.38020E+04	169	0.24550E+04	221	0.4[illegible]660E+04	273	0.28360E+04	325	0.33180E+04
14	0.26720E+04	66	0.29170E+04	118	0.37870E+04	170	0.40320E+04	222	0.32980E+04	274	0.22130E+04	326	0.33400E+04
15	0.23140E+04	67	0.33690E+04	119	0.31480E+04	171	0.39590E+04	223	0.33040E+04	275	0.35800E+04	327	0.33060E+04
16	0.34480E+04	68	0.33200E+04	120	0.25680E+04	172	0.38970E+04	224	0.27460E+04	276	0.34340E+04	328	0.32990E+04
17	0.32810E+04	69	0.33200E+04	121	0.27780E+04	173	0.38860E+04	225	0.22580E+04	277	0.34760E+04	329	0.26900E+04
18	0.32920E+04	70	0.25540E+04	122	0.36160E+04	174	0.36460E+04	226	0.37930E+04	278	0.35030E+04	330	0.23130E+04
19	0.32690E+04	71	0.28410E+04	123	0.38570E+04	175	0.28330E+04	227	0.36100E+04	279	0.34720E+04	331	0.35180E+04
20	0.32920E+04	72	0.35150E+04	124	0.36650E+04	176	0.23440E+04	228	0.36400E+04	280	0.27720E+04	332	0.33850E+04
21	0.26560E+04	73	0.32620E+04	125	0.36520E+04	177	0.35850E+04	229	0.37570E+04	281	0.24090E+04	333	0.33410E+04
22	0.22550E+04	74	0.33590E+04	126	0.29890E+04	178	0.36440E+04	230	0.37950E+04	282	0.35080E+04	334	0.33790E+04
23	0.33510E+04	75	0.33100E+04	127	0.22620E+04	179	0.36570E+04	231	0.32190E+04	283	0.33470E+04	335	0.33580E+04
24	0.31160E+04	76	0.34140E+04	128	0.38490E+04	180	0.34980E+04	232	0.37480E+04	284	0.34380E+04	336	0.27220E+04
25	0.31910E+04	77	0.26950E+04	129	0.35480E+04	181	0.34990E+04	233	0.44850E+04	285	0.35400E+04	337	0.23140E+04
26	0.31140E+04	78	0.22890E+04	130	0.34990E+04	182	0.28150E+04	234	0.41100E+04	286	0.34900E+04	338	0.34980E+04
27	0.32300E+04	79	0.35370E+04	131	0.34920E+04	183	0.24010E+04	235	0.42590E+04	287	0.28100E+04	339	0.33060E+04
28	0.25480E+04	80	0.33710E+04	132	0.35790E+04	184	0.36760E+04	236	0.41680E+04	288	0.24140E+04	340	0.32030E+04
29	0.21970E+04	81	0.34760E+04	133	0.29130E+04	185	0.34610E+04	237	0.41020E+04	289	0.37000E+04	341	0.32850E+04
30	0.33890E+04	82	0.33770E+04	134	0.23830E+04	186	0.35560E+04	238	0.31890E+04	290	0.33550E+04	342	0.34020E+04
31	0.31430E+04	83	0.34180E+04	135	0.40030E+04	187	0.37130E+04	239	0.25650E+04	291	0.34140E+04	343	0.26080E+04
32	0.31750E+04	84	0.27420E+04	136	0.34490E+04	188	0.36410E+04	240	0.40740E+04	292	0.33220E+04	344	0.22920E+04
33	0.31720E+04	85	0.25030E+04	137	0.37070E+04	189	0.30400E+04	241	0.37200E+04	293	0.33590E+04	345	0.35840E+04
34	0.32090E+04	86	0.44840E+04	138	0.36930E+04	190	0.26340E+04	242	0.36430E+04	294	0.27480E+04	346	0.33740E+04
35	0.26370E+04	87	0.34350E+04	139	0.37180E+04	191	0.40320E+04	243	0.35650E+04	295	0.24190E+04	347	0.33710E+04
36	0.21810E+04	88	0.44630E+04	140	0.30550E+04	192	0.34590E+04	244	0.36820E+04	296	0.36000E+04	348	0.33850E+04
37	0.33410E+04	89	0.34420E+04	141	0.23090E+04	193	0.33720E+04	245	0.30760E+04	297	0.34370E+04	349	0.34190E+04
38	0.31460E+04	90	0.34010E+04	142	0.36480E+04	194	0.33450E+04	246	0.37570E+04	298	0.34520E+04	350	0.26630E+04
39	0.31770E+04	91	0.27690E+04	143	0.35010E+04	195	0.31130E+04	247	0.39160E+04	299	0.34100E+04	351	0.23280E+04
40	0.32120E+04	92	0.23710E+04	144	0.36070E+04	196	0.34200E+04	248	0.34690E+04	300	0.34420E+04	352	0.35520E+04
41	0.32200E+04	93	0.35780E+04	145	0.35560E+04	197	0.30280E+04	249	0.36160E+04	301	0.28240E+04	353	0.33570E+04
42	0.25580E+04	94	0.33560E+04	146	0.35040E+04	198	0.28110E+04	250	0.36000E+04	302	0.23980E+04	354	0.33390E+04
43	0.21560E+04	95	0.34270E+04	147	0.28570E+04	199	0.26720E+04	251	0.33640E+04	303	0.36190E+04	355	0.32700E+04
44	0.33730E+04	96	0.33230E+04	148	0.23070E+04	200	0.26850E+04	252	0.27990E+04	304	0.35190E+04	356	0.31270E+04
45	0.31080E+04	97	0.34220E+04	149	0.36640E+04	201	0.26170E+04	253	0.23710E+04	305	0.34890E+04	357	0.26170E+04
46	0.31320E+04	98	0.27710E+04	150	0.34440E+04	202	0.26390E+04	254	0.35720E+04	306	0.34980E+04	358	0.22660E+04
47	0.31250E+04	99	0.23940E+04	151	0.33670E+04	203	0.23920E+04	255	0.34320E+04	307	0.35360E+04	359	0.26420E+04
48	0.32050E+04	100	0.36460E+04	152	0.29720E+04	204	0.21030E+04	256	0.35080E+04	308	0.27740E+04	360	0.18270E+04
49	0.25150E+04	101	0.34640E+04	153	0.21520E+04	205	0.25860E+04	257	0.35690E+04	309	0.24210E+04	361	0.19320E+04
50	0.20910E+04	102	0.35770E+04	154	0.27090E+04	206	0.25920E+04	258	0.34150E+04	310	0.36400E+04	362	0.27200E+04
51	0.33980E+04	103	0.35270E+04	155	0.24340E+04	207	0.26600E+04	259	0.27630E+04	311	0.34250E+04	363	0.25790E+04
52	0.30810E+04	104	0.35910E+04	156	0.36440E+04	208	0.24550E+04	260	0.23840E+04	312	0.34440E+04	364	0.24070E+04

series : daily totals of 1984.

APPENDIX V Computer programs.

The programs DIFAC and MEP, which can be used to perform the model building procedure in an interactive way, are based on the routines of the section G13 - Time series Analysis of the NAG - library. One is refered to the NAG FORTRAN Library Routine Documents for further information about these routines.
In this apppendix the structure of DIFAC and MEP will be discussed; due to the comments added to the program-listings, this knowledge of the structure also satisfies to read the listings of the programs.
As already mentioned the programs can be used to perform the model building procedure in an interactive way; that is the user has to answer the questions displayed on the screen. Due the straightforward character of these questions they will not be dealt with in detail. Again the knowledge of the structure of the programs will be sufficient to understand the programs and to be able to run these programs.

The flow-diagram in fig. V-1 shows the main structure of the program DIFAC. It is started by the call : r[un] DIFAC.
First the name of the data file containing the time series has to be entered. The requirements on the structure of files will be dealt with later on. It is possible to select each serried part of the series to be used.
If differencing of the (selected part of the) series is desired, this is carried out by the routine G13AAF. Before one has to enter the orders of the differencing. The differenced series can be written into a file.
Next the routine G13ABF can be used to compute the autocorrelation function of the (differenced) series. G13ABF also computes the sample mean, the sample variance and the Box-Pierce statistic Q. These results, together with information about the used time series (see structure of the files later on), can be written into a so called AC-file.
The routine G13ACF calculates the partial correlation coefficients given a set of autocorrelation coefficients. Also the calculated partial autocorrelation function can be written into a so called PAC-file.
Finally, by selecting one of the options the user can either stop the program or repeat any part of the program.

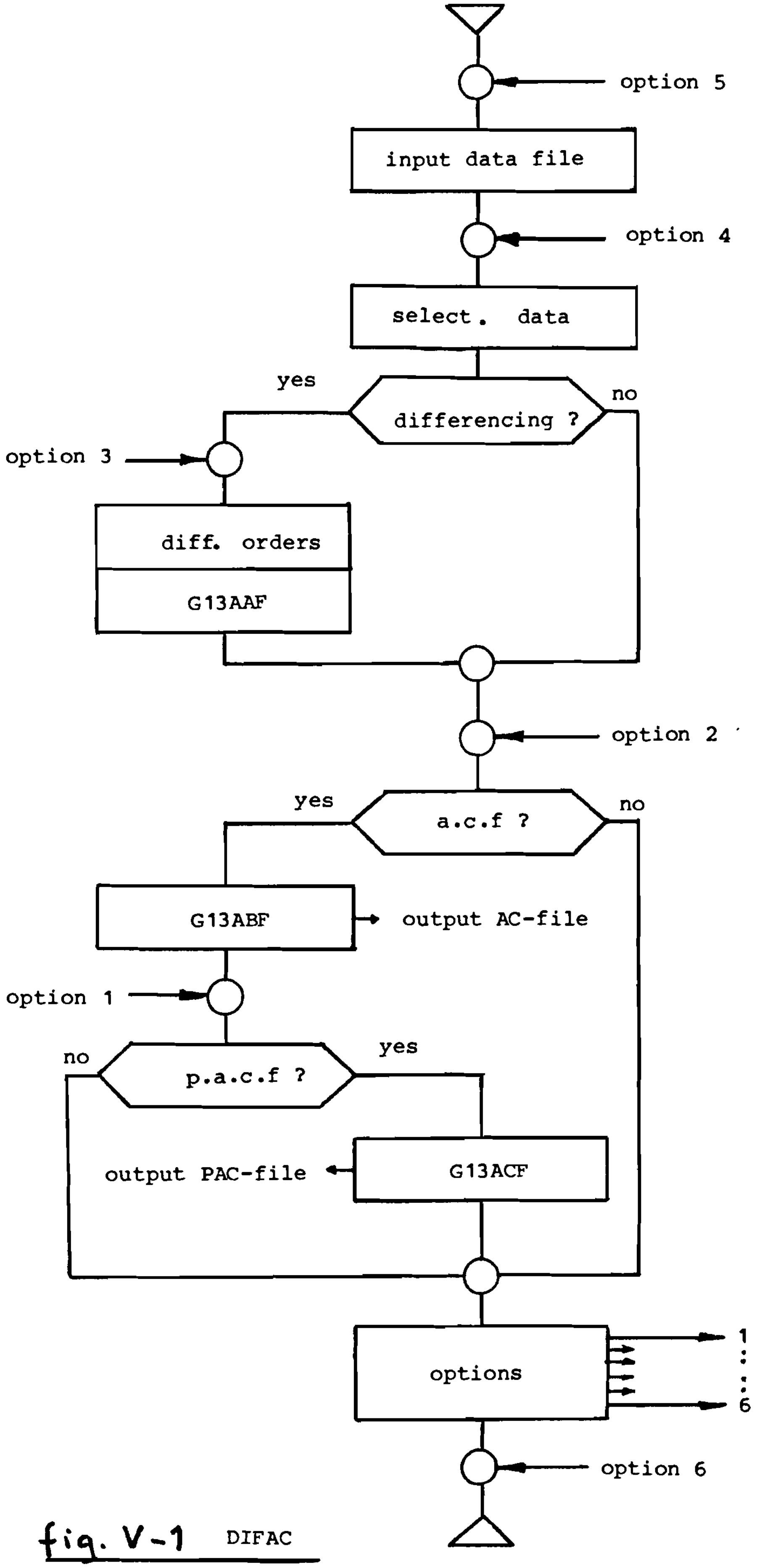

option 5
input data file
option 4
select. data
yes
differencing ?
no
option 3
diff. orders
G13AAF
option 2
yes
a.c.f ?
no
G13ABF
output AC-file
option 1
no
p.a.c.f ?
yes
output PAC-file
G13ACF
options
1
6
option 6

fig. V-1 DIFAC

The flow-diagram in fig. V-2 shows the main structure of the program MEP. It is started by the call : r[un] MEP.

First the program offers the possibility to calculate preliminary estimates of the parameters of an ARMA-model from an autocorrelation function. This calculating is carried out by routine G13ADF.

To do so first the name of the AC-file and second the model order has to be entered. The results of the preliminary estimation are displayed on the screen.

If one has performed preliminary estimation, the name (if smaller than 8 characters; see file structure) and the the part of the samples, used for calculating the a.c.f, will be known from the AC-file. Otherwise one has to enter the name of the data file containing the time series.

After entering or changing of the desired model orders the estimation of the parameters of the ARIMA-model is carried out by the routine G13AEF. The results of the estimation are displayed on the screen.

The model-fit can be tested by diagnostic checking. G13ABF is used to compute the a.c.f of the residuals and the Box-Pierce statistic Q. Routine G01BCF computes the tail-probability of the chi-square distribution.

The residual series and the autocorrelation function of the residuals can be written into files.

The fitted model can be used to produce forecasts. Routine G13AHF calculates the forecasts and routine G13AGF is used for updating. Forcasts can be written into a so called PRED-file.

Finally, by selecting one of the options the user can either stop the program or repeat any part of the program.

File structure.

File type : unformatted, direct access, fixed recordlength.
The recordlength = 8, except for the PRED-file recl. = 12.

The structure of the file satisfies the requirements of the plotprogram GRA (see : "Gebruikershandleiding voor het plotprogramma GRA", by E.M.M. Ploemen).

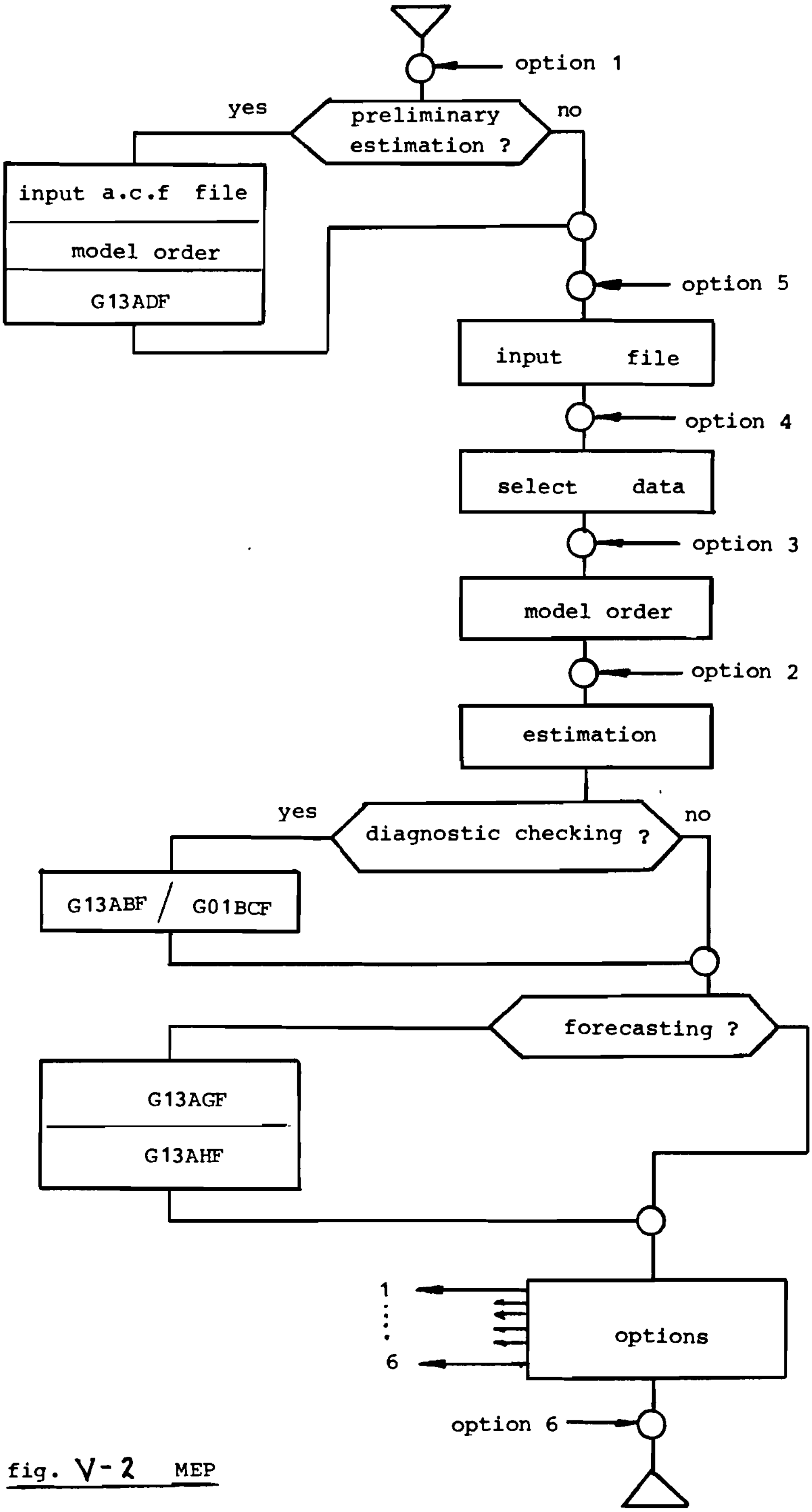

option 1
yes
preliminary estimation ?
no
input a.c.f file
model order
G13ADF
option 5
input file
option 4
select data
option 3
model order
option 2
estimation
yes
diagnostic checking ?
no
G13ABF / G01BCF
forecasting ?
G13AGF
G13AHF
1
6
options
option 6

fig. V-2 MEP

data (time series) file :

record 1 : nsam, nvar, nstart
rec. 2 : 0, 0, 0
where : nsam = number of data samples
nvar = 1
nstart = start record (contains first sample of the series)

By restricting nvar to 1 (only one series in the file) it is possible to permit the data to be either single precision or double precision.

AC-file (output from DIFAC).

rec. 1 : nk, nvar, nstart, nsdat, nedat
rec. 2 : 0, 0, 0
rec. 3 : xv, xm, stat
rec. 4 : nk, nd, nds, nsea
rec. 5 : name
where : nk = number of autocor. coeff. calculated
nvar = 1
nstart = 6
nsdat = first sample of series used to calculate the a.c.f
nsdat = last sample of input time series used
xv = sample variance input time series
xm = sample mean
stat = Box-Pierce value Q
nd = order non-seasonal differencing before computing the a.c.f
nds = order seasonal differencing
nsea = periodicity
name = first 8 characters of the name of the data file.

PAC-file (output from DIFAC)

rec. 1 : nvl, 3, 3
rec. 2 : 0, 0, 0
where : nvl = number of calculated coefficients.
There are 3 variables, that is: 1 p.a.c.f
2 p.e.v.r | see routine
3 parameters AR(nvl) model | document

PRED-file (forecasts, output from MEP).

"model-fitting" : rec 1 : npred, nvar, nstart

rec 2 : 0, 0, 0

where : npred = number of forecasts

nvar = 6, that is : 1 true value of the sample

2 forecast

3 forecast + σ

4 forecast - σ

5 σ = standard error forecasts

6 difference (true - forecast)

"sea forecasting : rec. 1 : ntot, nvar, nstart, npre

rec. 2 : 0, 0, 0

whrer : ntot = total number of samples

nvar = 6, (see "model-fitting")

nstart = 3

npre = number of forecasts

explanation : if ntot = 300 and npre = 100, this means that the first 200 (ntot - npre) samples are used for estimation and the last 100 for one step ahead forecasting with updating.

Npre is used by the program PRER to distinguish between "model-fitting" and "real forecasting".

"ℓ - steps ahead" see "real forecasting"

The program PRER can be used to calculate the forecast errors.

note : The program PRER and other programs, which can be used to read files, for data adjusting, etc, are comprehensively explained by comments in the program-listings.

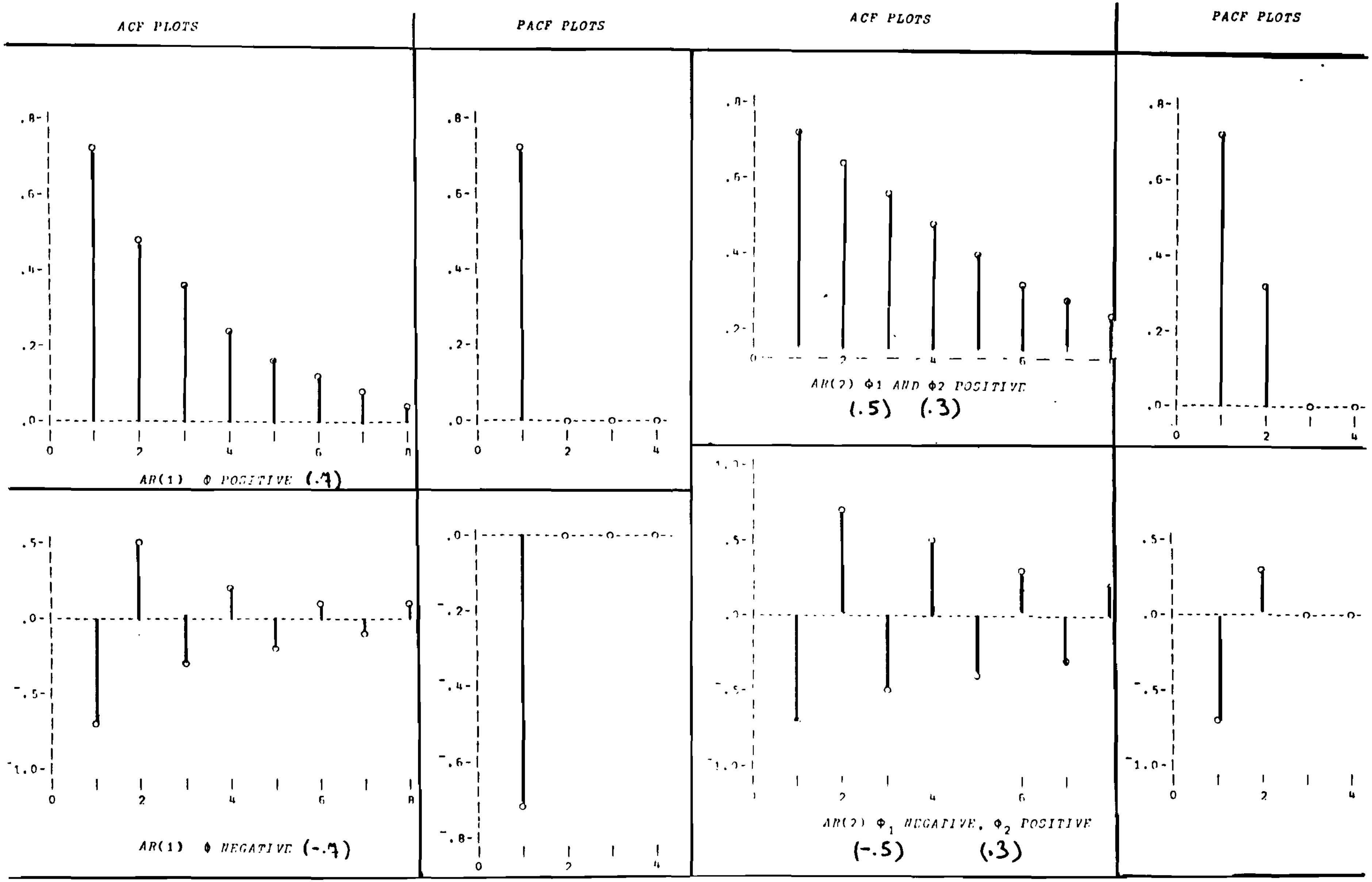
APPENDIX VI Identification tools.

As identification tools in this appendix are listed :

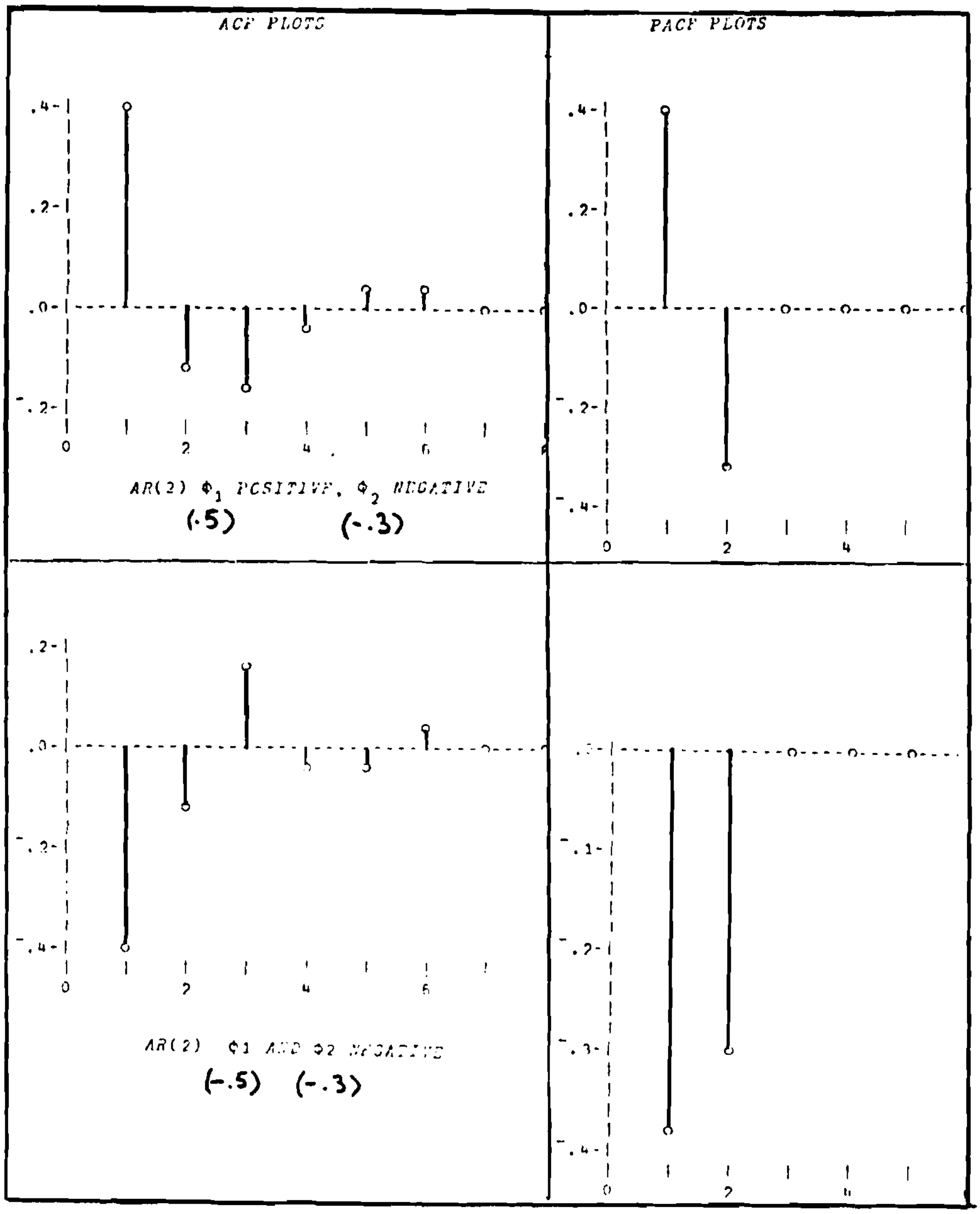
1 : theoretical correlation functions of non-seasonal models,
(page : 108, 109, 110).

2 : covariance structures for a number of seasonal models,
(page : 111, 112).

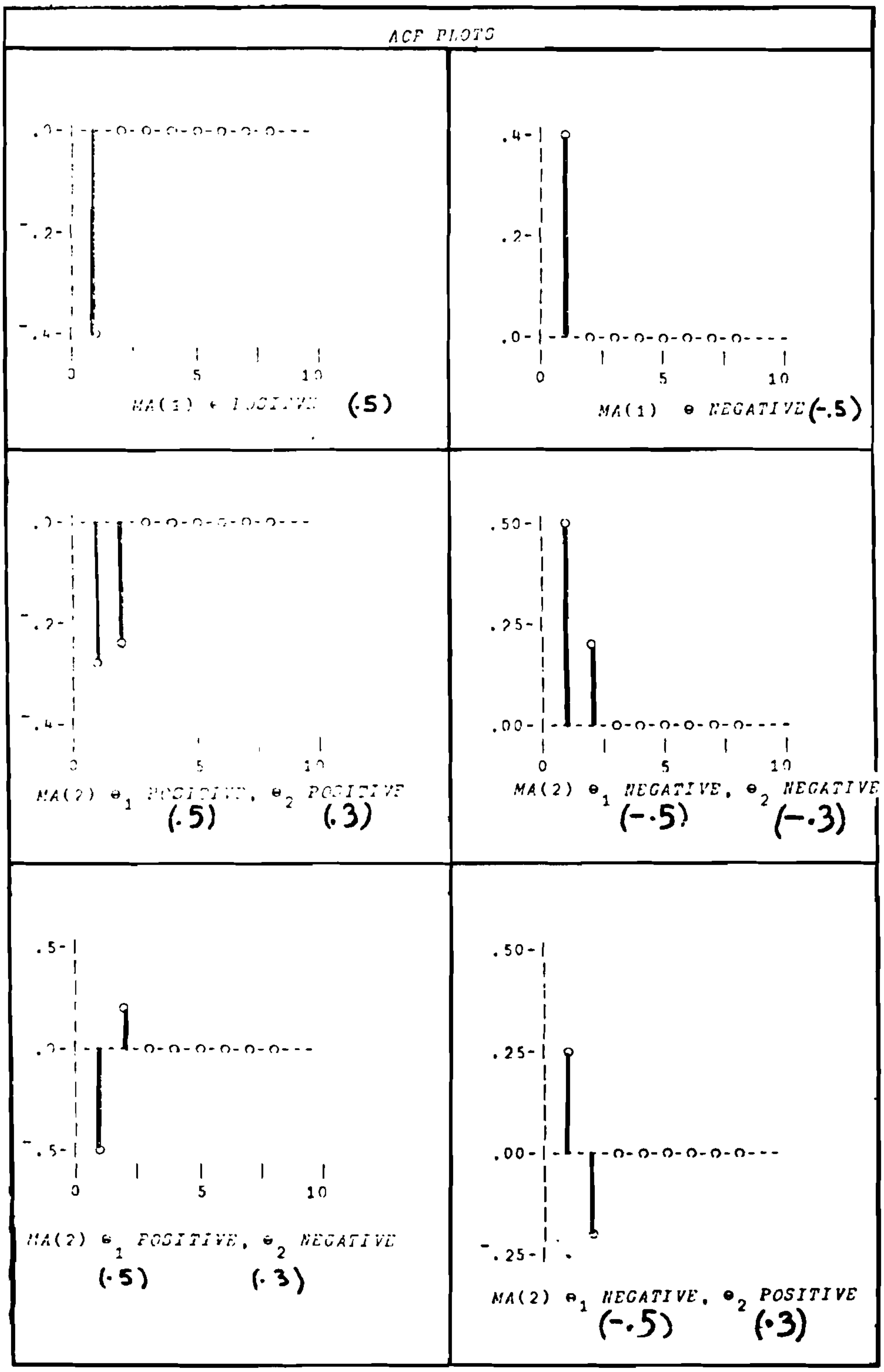
3 : some autocorrelation functions estimated from simulated data,
(page : 113 - 114).



Theoretical a c f and p a c f plots for an AR(p) model, p = 1, 2.



Theoretical a.c.f and p.a.c.f plots for an AR(2) model.



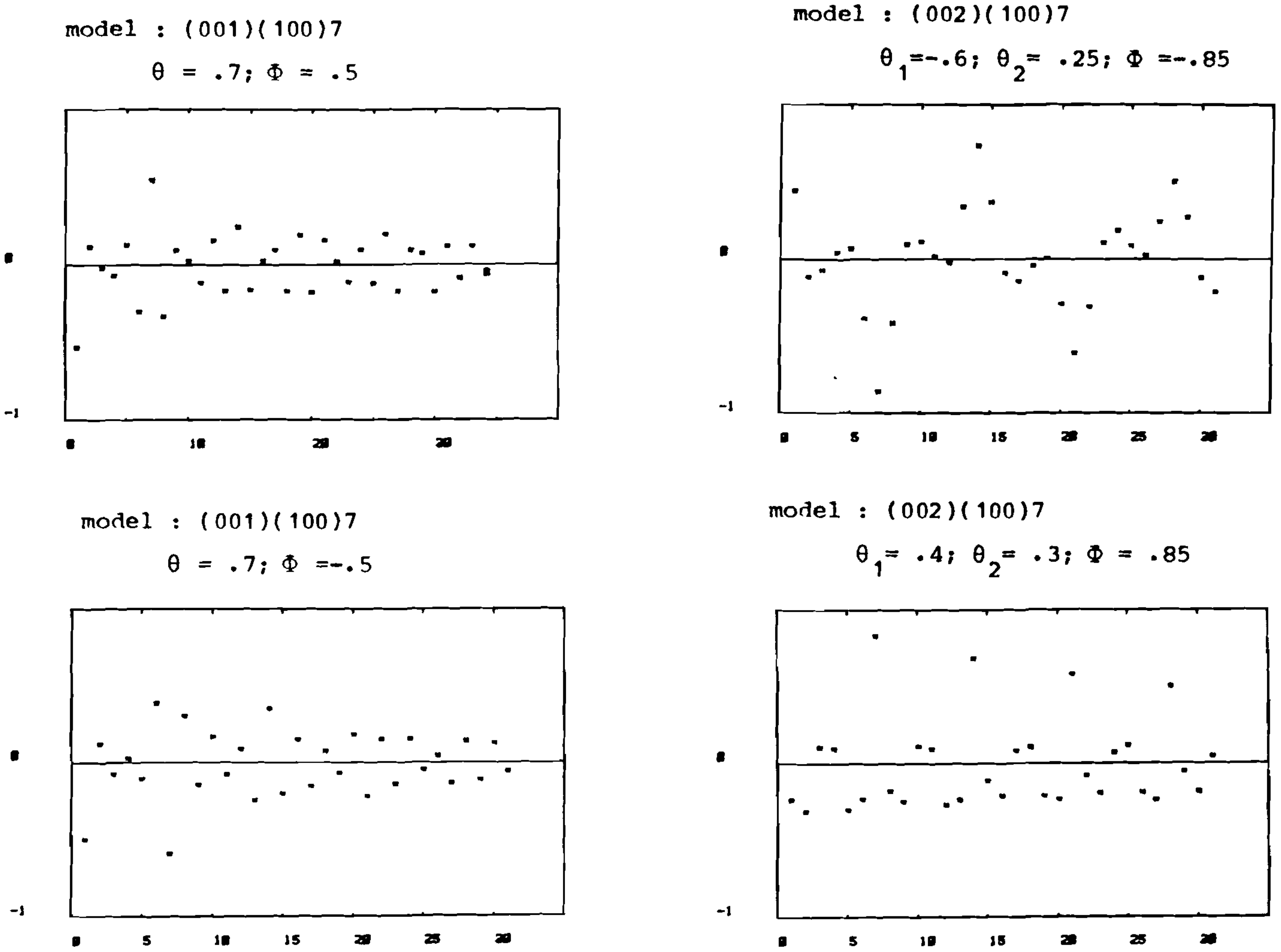
Theoretical a.c.f and p.a.c.f plots for a MA(q) model, q = 1, 2.

Model	(Autocovariances of w_t) σ_a^2	Special characteristics
(1) $w_t = (1 - \theta B)(1 - \Theta B^s)a_t$ $w_t = a_t - \theta a_{t-1} - \Theta a_{t-s} + \theta\Theta a_{t-s-1}$ $s \geqslant 3$	$\gamma_0 = (1 + \theta^2)(1 + \Theta^2)$ $\gamma_1 = -\theta(1 + \Theta^2)$ $\gamma_{s-1} = \theta\Theta$ $\gamma_s = -\Theta(1 + \theta^2)$ $\gamma_{s+1} = \gamma_{s-1}$ All other autocovariances are zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\rho_{s-1} = \rho_{s+1} = \rho_1\rho_s$
(2) $(1 - \Phi B^s)w_t = (1 - \theta B)(1 - \Theta B^s)a_t$ $w_t - \Phi w_{t-s} = a_t - \theta a_{t-1} - \Theta a_{t-s}$ $+ \theta\Theta a_{t-s-1}$ $s \geqslant 3$	$\gamma_0 = (1 + \theta^2)\left[1 + \frac{(\Theta - \Phi)^2}{1 - \Phi^2}\right]$ $\gamma_1 = -\theta\left[1 + \frac{(\Theta - \Phi)^2}{1 - \Phi^2}\right]$ $\gamma_{s-1} = \theta\left[\Theta - \Phi - \frac{\Phi(\Theta - \Phi)^2}{1 - \Phi^2}\right]$ $\gamma_s = -(1 + \theta^2)\left[\Theta - \Phi - \frac{\Phi(\Theta - \Phi)^2}{1 - \Phi^2}\right]$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_j = \Phi\gamma_{j-s} \qquad j \geqslant s + 2$ For $s \geqslant 4$, $\gamma_2, \gamma_3, \ldots, \gamma_{s-2}$ are all zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\gamma_j = \Phi\gamma_{j-s} \qquad j \geqslant s + 2$
(3) $w_t = (1 - \theta_1 B - \theta_2 B^2)$ $\cdot (1 - \Theta_1 B^s - \Theta_2 B^{2s})a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \Theta_1 a_{t-s}$ $+ \theta_1\Theta_1 a_{t-s-1} + \theta_2\Theta_1 a_{t-s-2}$ $- \Theta_2 a_{t-2s} + \theta_1\Theta_2 a_{t-2s-1}$ $+ \theta_2\Theta_2 a_{t-2s-2}$ $s \geqslant 5$	$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_1 = -\theta_1(1 - \theta_2)(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_2 = -\theta_2(1 + \Theta_1^2 + \Theta_2^2)$ $\gamma_{s-2} = \theta_2\Theta_1(1 - \Theta_2)$ $\gamma_{s-1} = \theta_1\Theta_1(1 - \theta_2)(1 - \Theta_2)$ $\gamma_s = -\Theta_1(1 + \theta_1^2 + \theta_2^2)(1 - \Theta_2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{s+2} = \gamma_{s-2}$ $\gamma_{2s-2} = \theta_2\Theta_2$ $\gamma_{2s-1} = \theta_1\Theta_2(1 - \theta_2)$ $\gamma_{2s} = -\Theta_2(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{2s+1} = \gamma_{2s-1}$ $\gamma_{2s+2} = \gamma_{2s-2}$ All other autocovariances are zero.	(a) $\gamma_{s-2} = \gamma_{s+2}$ (b) $\gamma_{s-1} = \gamma_{s+1}$ (c) $\gamma_{2s-2} = \gamma_{2s+2}$ (d) $\gamma_{2s-1} = \gamma_{2s+1}$
(3a) *Special Case of Model 3* $w_t = (1 - \theta_1 B - \theta_2 B^2)(1 - \Theta B^s)a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \Theta a_{t-s}$ $+ \theta_1\Theta a_{t-s-1} + \theta_2\Theta a_{t-s-2}$ $s \geqslant 5$	$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)(1 + \Theta^2)$ $\gamma_1 = -\theta_1(1 - \theta_2)(1 + \Theta^2)$ $\gamma_2 = -\theta_2(1 + \Theta^2)$ $\gamma_{s-2} = \theta_2\Theta$ $\gamma_{s-1} = \theta_1\Theta(1 - \theta_2)$ $\gamma_s = -\Theta(1 + \theta_1^2 + \theta_2^2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{s+2} = \gamma_{s-2}$ All other autocovariances are zero.	(a) $\gamma_{s-2} = \gamma_{s+2}$ (b) $\gamma_{s-1} = \gamma_{s+1}$

Covariance structures for seasonal models.

Model	(Autocovariances of $w_t)/\sigma_a^2$	Special characteristics
(3b) *Special Case of Model 3* $w_t = (1-\theta B)(1-\Theta_1 B^s - \Theta_2 B^{2s})a_t$ $w_t = a_t - \theta a_{t-1} - \Theta_1 a_{t-s} + \theta\Theta_1 a_{t-s-1}$ $- \Theta_2 a_{t-2s} + \theta\Theta_2 a_{t-2s-1}$ $s \geqslant 3$	$\gamma_0 = (1+\theta^2)(1+\Theta_1^2+\Theta_2^2)$ $\gamma_1 = -\theta(1+\Theta_1^2+\Theta_2^2)$ $\gamma_{s-1} = \theta\Theta_1(1-\Theta_2)$ $\gamma_s = -\Theta_1(1+\theta^2)(1-\Theta_2)$ $\gamma_{s+1} = \gamma_{s-1}$ $\gamma_{2s-1} = \theta\Theta_2$ $\gamma_{2s} = -\Theta_2(1+\theta^2)$ $\gamma_{2s+1} = \gamma_{2s-1}$ All other autocovariances are zero.	(a) $\gamma_{s-1} = \gamma_{s+1}$ (b) $\gamma_{2s-1} = \gamma_{2s+1}$
(4) $w_t = (1-\theta_1 B - \theta_s B^s - \theta_{s+1}B^{s+1})a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $- \theta_{s+1}a_{t-s-1}$ $s \geqslant 3$	$\gamma_0 = 1+\theta_1^2+\theta_s^2+\theta_{s+1}^2$ $\gamma_1 = -\theta_1 + \theta_s\theta_{s+1}$ $\gamma_{s-1} = \theta_1\theta_s$ $\gamma_s = \theta_1\theta_{s+1} - \theta_s$ $\gamma_{s+1} = -\theta_{s+1}$ All other autocovariances are zero.	(a) In general, $\gamma_{s-1} \neq \gamma_{s+1}$ $\gamma_1\gamma_s \neq \gamma_{s+1}$
(4a) *Special Case of Model 4* $w_t = (1-\theta_1 B - \theta_s B^s)a_t$ $w_t = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $s \geqslant 3$	$\gamma_0 = 1+\theta_1^2+\theta_s^2$ $\gamma_1 = -\theta_1$ $\gamma_{s-1} = \theta_1\theta_s$ $\gamma_s = -\theta_s$ All other autocovariances are zero.	(a) Unlike model 4, $\gamma_{s+1} = 0$
(5) $(1-\Phi B^s)w_t = (1-\theta_1 B - \theta_s B^s$ $- \theta_{s+1}B^{s+1})a_t$ $w_t - \Phi w_{t-s} = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $- \theta_{s+1}a_{t-s-1}$ $s \geqslant 3$	$\gamma_0 = 1 + \theta_1^2 + \frac{(\theta_s-\Phi)^2}{1-\Phi^2} + \frac{(\theta_{s+1}+\theta_1\Phi)^2}{1-\Phi^2}$ $\gamma_1 = -\theta_1 + \frac{(\theta_s-\Phi)(\theta_{s+1}+\theta_1\Phi)}{1-\Phi^2}$ $\gamma_{s-1} = (\theta_s-\Phi)\left[\theta_1 + \Phi\frac{(\theta_{s+1}+\Phi\theta_1)}{1-\Phi^2}\right]$ $\gamma_s = -(\theta_s-\Phi)\left[1-\Phi\frac{(\theta_s-\Phi)}{1-\Phi^2}\right]$ $+ (\theta_{s+1}+\theta_1\Phi)\left[\theta_1+\Phi\frac{(\theta_{s+1}+\theta_1\Phi)}{1-\Phi^2}\right]$ $\gamma_{s+1} = -(\theta_{s+1}+\theta_1\Phi)\left[1-\Phi\frac{(\theta_s-\Phi)}{1-\Phi^2}\right]$ $\gamma_j = \Phi\gamma_{j-s} \quad j \geqslant s+2$ For $s \geqslant 4$, $\gamma_2, \ldots, \gamma_{s-2}$ are all zero.	(a) $\gamma_{s-1} \neq \gamma_{s+1}$ (b) $\gamma_j = \Phi\gamma_{j-s} \quad j \geqslant s +$
(5a) *Special Case of Model 5* $(1-\Phi B^s)w_t = (1-\theta_1 B - \theta_s B^s)a_t$ $w_t - \Phi w_{t-s} = a_t - \theta_1 a_{t-1} - \theta_s a_{t-s}$ $s \geqslant 3$	$\gamma_0 = 1 + \frac{\theta_1^2 + (\theta_s-\Phi)^2}{1-\Phi^2}$ $\gamma_1 = -\theta_1\left[1-\Phi\frac{(\theta_s-\Phi)}{1-\Phi^2}\right]$ $\gamma_{s-1} = \frac{\theta_1(\theta_s-\Phi)}{1-\Phi^2}$ $\gamma_s = \frac{\Phi\theta_1^2 - (\theta_s-\Phi)(1-\Phi\theta_s)}{1-\Phi^2}$ $\gamma_j = \Phi\gamma_{j-s} \quad j \geqslant s+1$ For $s \geqslant 4$, $\gamma_2, \ldots, \gamma_{s-2}$ are all zero.	(a) Unlike model 5, $\gamma_{s+1} = \Phi\gamma_1$

Covariance structures for seasonal models.



model : (001)(100)7

$\theta = .7;\ \Phi = .5$

model : (002)(100)7

$\theta_1 = -.6;\ \theta_2 = .25;\ \Phi = -.85$

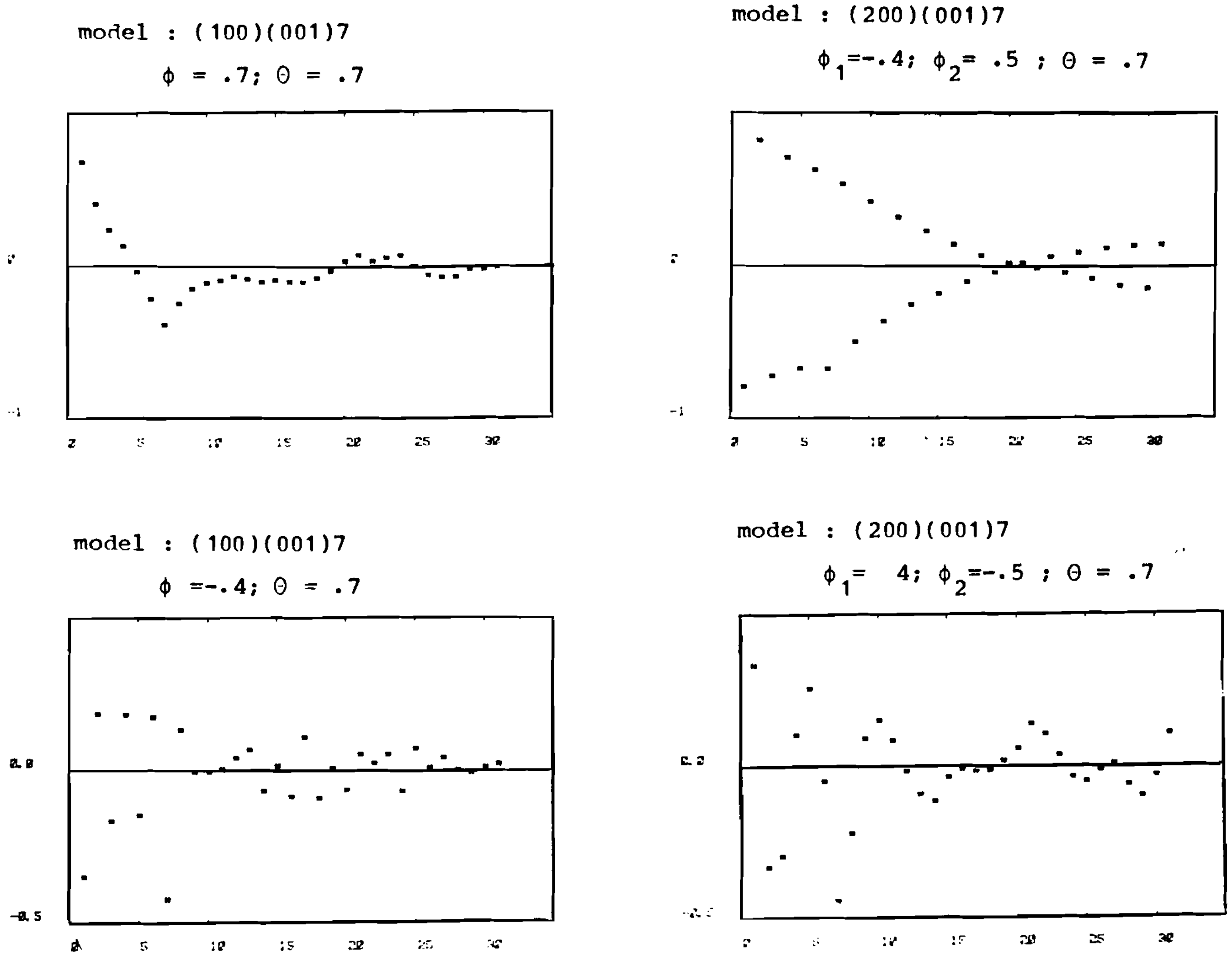
model : (001)(100)7

$\theta = .7;\ \Phi = -.5$

model : (002)(100)7

$\theta_1 = .4;\ \theta_2 = .3;\ \Phi = .85$

some autocorrelation functions estimated from simulated data.



model : (100)(001)7

ϕ = .7; Θ = .7

model : (200)(001)7

ϕ_1=-.4; ϕ_2= .5 ; Θ = .7

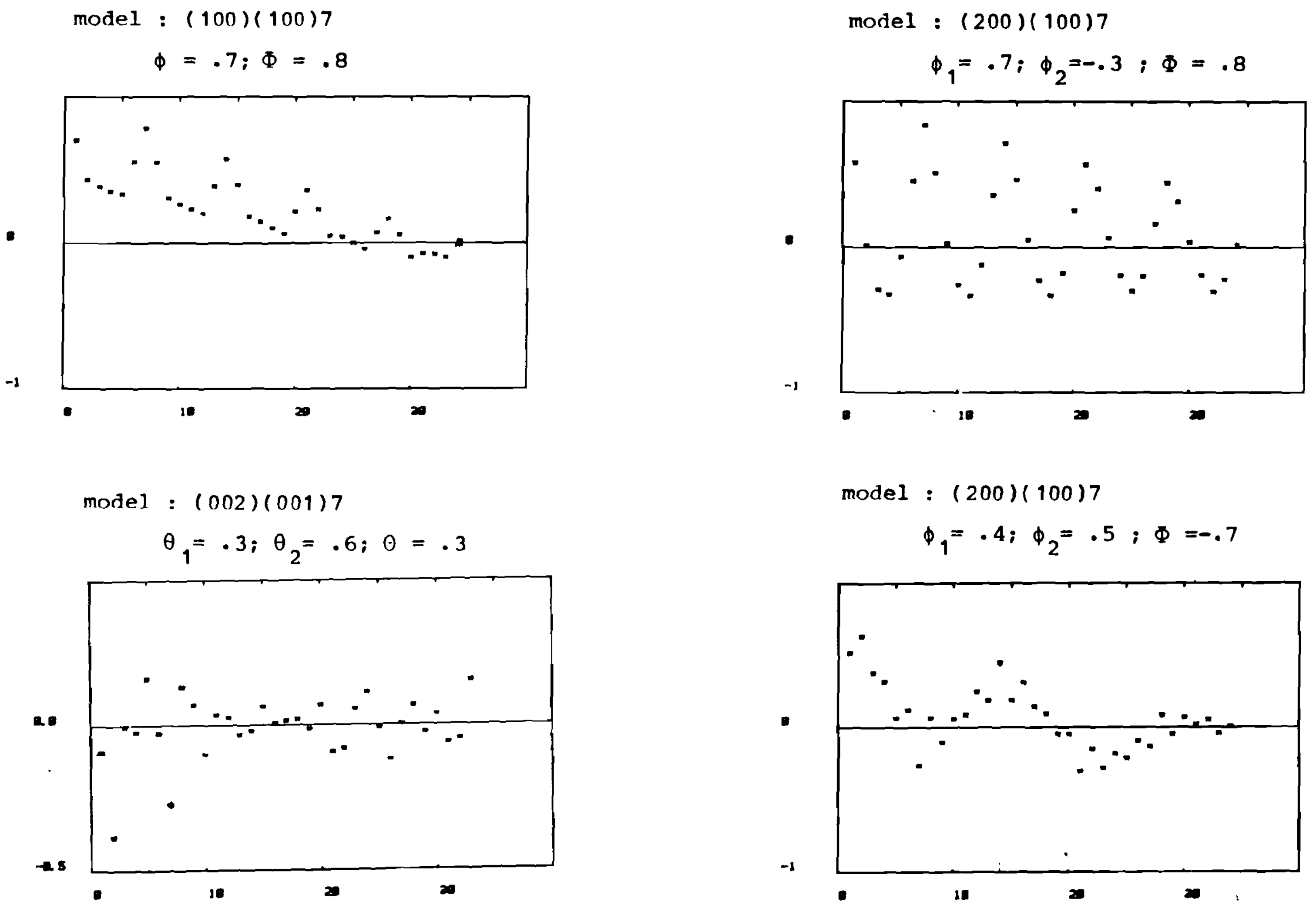
model : (100)(001)7

ϕ =-.4; Θ = .7

model : (200)(001)7

ϕ_1= 4; ϕ_2=-.5 ; Θ = .7

some autocorrelation functions estimated from simulated data.



some autocorrelation functions estimated from simulated data.

REFERENCES.

[1] G.E.P Box and G.M. Jenkins
"Time series analysis : forecasting and control."
Holden-Day,San Francisco.(Revised edition 1976)

[2] M.S. Bartlett
"On the theoretical specification of sampling properties of autocorrelated time series"
Jour. Royal. Stat. Soc., B8, 27 (1946)

[3] G.U. Yule
"On a method of investigating periodicities in disturded series, with special reference to Wolfer's sunspot numbers"
Phil. Trans. ,A226, 267 (1927)

[4] E.G.F van Winkel
"ARIMA-processen : de betekenis van univariate methoden voor de econometrie"
Dr.Thesis, VU Amsterdam, Netherlands (1979)

[5] J. de Beer and H. v.d. Hoven
"Specificatie en voorspelkracht van tijdreeksmodellen"
Kwant. Meth. 15, 1984, pp 58-75

[6] S. Makridakis e.o.
"The accuracy of extrapolation (time series) methods : Results of a forecasting competition"
Jour. of Forecasting, vol 1, 1982, pp 111-153

[7] S. Makridakis and M. Hibon
"Accuracy of forecasting : An emperical investigation"
J.R.Statist.Soc, 142 part 2, 1979, pp 97-145

[8] G.E.P. Box and D.A. Pierce
"Distribution of residual autocorrelations in autoregressive-moving average time series models"
Jour.Amer.Stat.Assoc., 64, 1509, 9170

[9] D.W. Marquardt
"An algorithm for least squares estimation of non-linear parameters"
Jour.Soc.Ind.Appl.Math., 11, 431, 1963

[10] K.J. Åstrom
Maximum likelihood and prediction errors methods"
Automatica, vol. 6, 1980, pp 551-574