An Introduction to Modeling Dynamic Behavior With Time Series Analysis

Joseph L. Hellerstein

IBM Thomas J. Watson Research Center, Yorktown Heights, NY

Abstract. The need to model dynamic behavior in information systems arises in many contexts, such as characterizing the locality of file access patterns, evaluating the dynamic behavior of scheduling algorithms, and identifying performance problems by their time serial behavior. This paper provides an introduction to time series analysis (a statistical technique), and applies it to analyzing the performance of information systems. The autoregressive, moving average (ARMA) model is discussed in detail, with an emphasis on identifying time series models from measurement data using the autocorrelation and partial autocorrelation functions. The paper concludes with a case study in which time series analysis is used to diagnosis a performance problem in a large computer system.

1 Introduction

The need to model dynamic behavior in information systems arises in many contexts. Some examples include:

1. Designing a disk cache is facilitated by having a characterization of the time serial behavior of file access patterns.
2. Evaluating the dynamic behavior of a scheduling algorithm requires assessing its response to transients in arrival rates and service times.
3. Identifying performance problems in computer systems can often be accomplished by relating the dynamic behavior of the problem to the dynamic behavior of applications running on the computer system.

This paper describes time series analysis, a statistical approach to modeling dynamic behavior, and applies it to measurements of information systems. Considered are the autoregressive and moving average (ARMA) models, with an emphasis on model identification and evaluation. The paper concludes with a case study that applies time series analysis to diagnosing a performance problem in a large computer system.

A time series consists of serial measurements of a process, such as sequences of response times of computer system interactions. Herein, we assume that time is discrete (e.g., thirty second intervals) and that measurement values are continuous (i.e., we can, in theory, obtain an unlimited number of digits to the right of the decimal point). Throughout, t is used to denote the time index (or observation), and y_t denotes the value of the t-th observation. A time series can be displayed in either a tabular or a graphical manner. For example, Table 1 displays the first four observations of response times over a nine hour shift at a large computer installation, and Fig. 1 plots the response time data for the entire nine hour shift.

Table 1. Illustration of a Time Series

t	Time	Response Time
1	8:01:00	1.5
2	8:01:30	1.6
3	8:02:00	.75
4	8:02:30	1.8

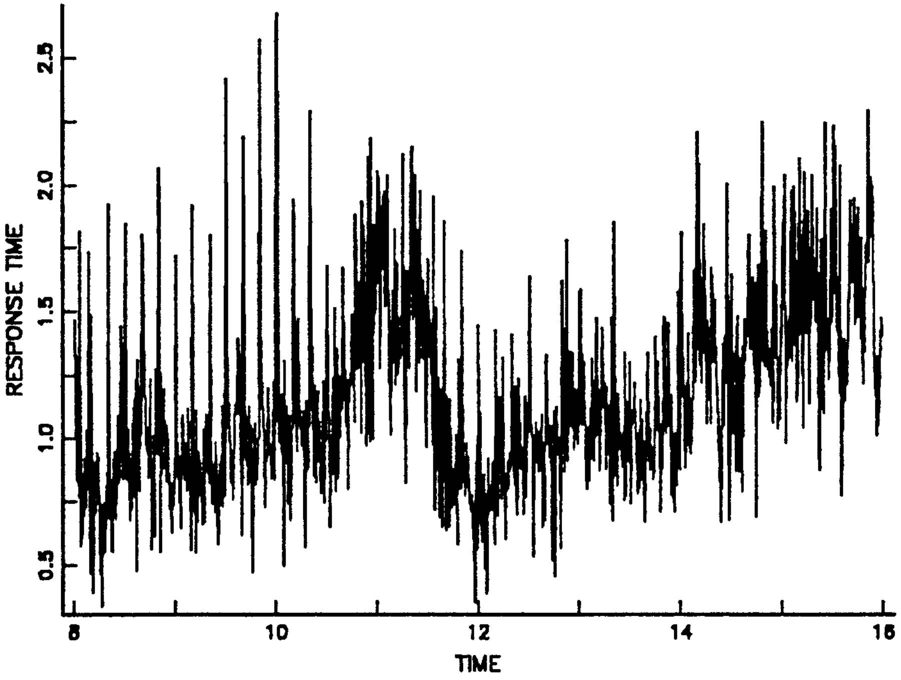

Fig. 1. Time Series Example

Constructing a time series model involves expressing y_t in terms of (i) previous observations (e.g., y_{t-1}, y_{t-2}) and (ii) shocks to the system, which are unobserved random variables that represent external events (e.g., changes in arrival rates and/or service times). In order to construct a time series model, y_t must be **stationary**. This means that (a) all terms in the series have the same mean and variance, (b) the covariance between terms in the series only depends on the number of time units between them (not their absolute position in the series), and (c) all of the foregoing are finite. The random shocks, which are denoted by a_t, are assumed to be independent and identically distributed **(i.i.d.)** with $E(a_t) = 0$ and $Var(a_t) = {\sigma_a}^2$.

A **time series model** specifies an algebraic relationship between random variables representing terms in the series; the t-th such random variable is denoted by $\tilde{y}_t$. (In contrast, the y_t are measured values, which are constants.) Since all terms in the series have the same mean, it is convenient to view $\tilde{y}_t$ as the deviation from the population mean; that is, $E(\tilde{y}_t) = 0$. Herein, we consider linear time series models that have the general form

$$\tilde{y}_t = \phi_1\tilde{y}_{t-1} + \cdots + \phi_p\tilde{y}_{t-p} + a_t - \theta_1 a_{t-1} - \cdots - \theta_q a_{t-q}. \qquad (1)$$

(It is a convention that a_{t-k} $(k > 0)$ be subtracted.) This equation states that the t-th term depends on the preceding p terms and on the preceding q random shocks.

Linear time series models are classified by the values of p and q in Eq. (1). Specifically,

- $p > 0$ and $q = 0$ defines a p parameter **autoregressive model**, which is denoted by AR(p).
- $p = 0$ and $q > 0$ specifies a q parameter **moving average model**, which is denoted by MA(q).
- $p, q > 0$ designates a **mixed model** that has p autoregressive parameters and q moving-average parameters; this is denoted by ARMA(p,q).
- $p = 0 = q$ is a model in which there no time serial dependency; this is referred to as the **white noise model**, and is denoted by either AR(0) or ARMA(0,0).

Our discussion of time series analysis is based largely on the classical Box-Jenkins approach [3]. This approach employs a five-step methodology for constructing time series models:

1. handling non-stationarities in the series
2. identification: determine the values of p and q in Eq. (1)
3. estimation: estimate the unknown constants $\phi_1, \cdots, \phi_p$, $\theta_1, \cdots, \theta_q$ in Eq. (1)
4. evaluation (diagnostics): assess the model constructed
5. forecasting: predict y_{t+k} $(k > 0)$ given previous values in the series

In practice, steps (1)-(4) are applied repeatedly before proceeding to step (5). Herein, we focus steps (1)-(4), with particular emphasis on the first two steps.

Applying time series analysis in practice requires that the analyst obtain values of a metric of interest (i.e., the y_t) and then apply the foregoing methodology to construct a time series model. How the model is used depends on the task at hand. In the case of workload characterization, y_t is a workload parameter, such as CPU consumption or input/output rates; the time series model can be used to generate a synthetic workload for a more complex system by using a random number generator to obtain the a_t. When evaluating the transient behavior of a scheduling algorithm, y_t is the performance of the system studied when the algorithm is employed; the time series model is used to evaluate the effect of transients by predicting y_{t+j} when the a_t are varied. For diagnosing performance problems, y_t is a metric that is used to detect performance problem; the cause of the problem can sometimes be deduced from the terms in the time series model.

Numerous books and articles have been written on the theory of time series analysis (e.g., [3], [12]). Unfortunately, it has been rare to apply time series analysis to information systems. One case in which it has been employed is forecasting the growth of workloads in computer systems (e.g., [6] and [8]), which is an important part of capacity planning. Another case in which time serial behavior is important is characterizing packet interarrival times in communications networks. Frequently, these interarrivals are not i.i.d. due to "train" effects induced by large transmissions (e.g., file transfers). For the most part, dependencies in interarrival times have been addressed using Markov modulated processes (e.g., [1], [4], [10], and [11]). However, time series techniques have been employed occasionally (e.g., [7]).

The remainder of this paper is organized as follows. Section 2 discusses how to identify and evaluate time series models given measurements of a stationary stochastic process. Section 3 addresses how to handle non-stationary data, which are common in information systems because of variations in workload. Section 4 contains a case study of applying time series analysis to diagnosing a performance problem. Our conclusions are contained in section 5.

2 Time Series Models

This section discusses how to construct ARMA(p,q) models, with an emphasis on AR(1) because of its importance in modeling dynamic behavior in queueing systems. We focus on the identification step in time series analysis, although the estimation and evaluation steps are considered as well. Throughout this section it is assumed that the underlying stochastic process is stationary. (Section 3 addresses non-stationary processes.)

A simple and very intuitive way to express time serial dependencies is to state that the current observation depends only on the previous observation and an i.i.d. random shock. This is the one parameter autoregressive model, or AR(1), which is expressed algebraically as

$$\tilde{y}_t = \phi \tilde{y}_{t-1} + a_t, \tag{2}$$

where $\phi = \phi_1$ in Eq. (1). (Readers familiar with the theory of stochastic processes will recognize the AR(1) model as a discrete-time, continuous-state Markov chain.) Key to this equation is the parameter ϕ, which determines how related successive observations are. If $\mid \phi \mid \approx 1$, we know much more about $\tilde{y}_t$ given $\tilde{y}_{t-1}$ than is the case if $\mid \phi \mid \approx 0$.

More insight into AR(1) processes can be obtained by expanding the recurrence relationship in Eq. (2).

$$\begin{aligned}\tilde{y}_t &= \phi\tilde{y}_{t-1} + a_t \\ &= \phi^2\tilde{y}_{t-2} + \phi a_{t-1} + a_t \\ &= \textstyle\sum_{k=0}^{\infty} \phi^k a_{t-k}\end{aligned}$$

This equation suggests that ϕ should be constrained so that $\mid \phi \mid < 1$. If this were not the case, shocks to the system that occurred in the distant past would have a larger effect on $\tilde{y}_t$ than shocks in the recent past (due to the exponent k of ϕ).

The importance of the $\mid \phi \mid < 1$ constraint can be demonstrated analytically by computing the variance of an AR(1) process. Since $E(\tilde{y}_t) = 0$, we have

$$\begin{aligned}Var(\tilde{y}_t) &= E(\tilde{y}_t\tilde{y}_t) \\ &= E[(\textstyle\sum_{i=0}^{\infty} \phi^i a_{t-i})(\sum_{j=0}^{\infty} \phi^j a_{t-j})] \\ &= \textstyle\sum_{k=0}^{\infty} \phi^{2k} E(a_{t-k}a_{t-k}) \\ &= \frac{\sigma_a{}^2}{1-\phi^2}\end{aligned} \tag{3}$$

(The third equation follows from the a_t being i.i.d.) Thus, unless $\mid \phi \mid < 1$, the variance of $\tilde{y}_t$ is infinite, which means the process is nonstationary.

The AR(1) model is often effective at characterizing the dynamic behavior of queueing systems. To illustrate this, we develop an approximation for the dynamic behavior of a single server, first-come first-served (FCFS) queueing system. Let R_t, A_t, and S_t denote (respectively) the response time, interarrival time, and service time of the t-th customer arriving at the queueing system. For a lightly loaded system, we have

$$R_t = S_t,$$

and for a saturated system

$$R_t = R_{t-1} - A_t + S_t.$$

Thus, the general situation can be approximated by

$$R_t = \phi(R_{t-1} - A_t) + S_t,$$

where $0 \leq \phi < 1$ (to ensure finite response times). Put differently,

$$R_t - E[R_t] = \phi(R_{t-1} - E[R_t]) - \phi A_t + S_t + C, \tag{4}$$

where C is a constant. Letting $\tilde{y}_t = R_t - E(R_t)$ and $a_t = -\phi A_t + S_t + C$, Eq. (4) becomes $\tilde{y}_t = \phi\tilde{y}_t + a_t$, which is an AR(1) model.

How accurately does Eq. (4) model the dynamic behavior of a single server, FCFS queueing system? Commonly used statistics such as the sample mean, variance, and distribution do not answer this question since they provide no insight into time serial dependencies. An alternative is to compare plots of time serial FCFS response times with realizations of potential AR(1) models. Figure 2 contains such plots for an M/M/1, FCFS queueing system (with an arrival rate of .7 and a service time of 1) and three AR(1) processes with different values of ϕ; in all cases, initial transients have been deleted and so the values plotted constitute a stationary series. Unfortunately, it is unclear how these plots should be compared, and so it is unclear which AR(1) model (if any) adequately characterizes the dynamic behavior of the FCFS queueing system.

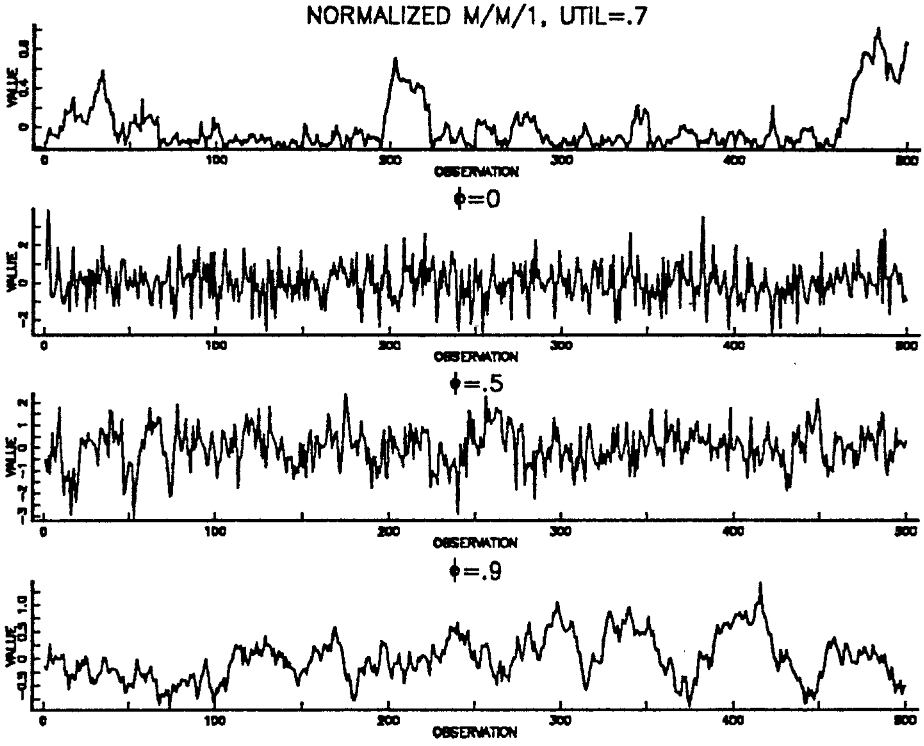

Fig. 2. Comparison of Several Time Series

The foregoing motivates the first step in time series analysis – model identification. The objective of this step is to use measurements of a stochastic process to determine a good choice for p and q in Eq. (1). Doing so requires having a way to characterize the time serial behavior of a stochastic process. One approach is to quantify the relationship between all observations separated by the same number of time units or **lags**. For example, lag 1 "relatedness" can be assessed from the pairs $(\tilde{y}_t, \tilde{y}_{t-1}), (\tilde{y}_{t-1}, \tilde{y}_{t-2}), \cdots$ and lag 2 "relatedness" from the pairs $(\tilde{y}_t, \tilde{y}_{t-2}), (\tilde{y}_{t-1}, \tilde{y}_{t-3}), \cdots$. Relatedness can be quantified by the covariance function; since this is applied to elements of the same series, it is referred to as autocovariance. The lag k **autocovariance** is denoted by γ_k, and is defined as

$$\gamma_k = E(\tilde{y}_t\tilde{y}_{t-k}).$$

(Since y_t is assumed to be stationary, γ_k does not depend on t.) γ_k can be computed directly from a time series model. For AR(1), this calculation is

$$\begin{aligned}\gamma_k[AR(1)] &= E[\tilde{y}_t\tilde{y}_{t-k}] \\ &= E\left[\left(\sum_{j=0}^{k-1}\phi^j a_{t-j} + \phi^k\tilde{y}_{t-k}\right)\tilde{y}_{t-k}\right] \\ &= \phi^k Var(\tilde{y}_t)\end{aligned}$$

γ_k is a number in the interval $(-\infty, \infty)$. The **lag** k **autocorrelation** normalizes γ_k to a value in the interval $[-1, 1]$. Denoted by ρ_k (an unfortunate conflict with the notation used for utilizations in queueing theory), the lag k autocorrelation is defined as follows:

$$\rho_k = \frac{\gamma_k}{Var(\tilde{y}_t)}$$

Thus,

$$\rho_k[AR(1)] = \phi^k \tag{5}$$

The **autocorrelation function (ACF)** of a series is a mapping from a set of lags $1, 2, \cdots, K$ to $\rho_1, \rho_2, \cdots, \rho_K$.

So for we have described how to characterize the dynamic behavior of a time series model by using the ACF. Our strategy is to characterize the time series data in a similar manner and then compare this characterization to the ACFs of several time series models. As before, we are interested in autocovariances from which we obtain autocorrelations. However, these metrics are *estimated* from the data instead of being computed analytically from a time series model. A commonly used estimator for γ_k is

$$c_k = \frac{1}{N}\sum_{i=1}^{N-k}(y_k - \bar{y})(y_{i+k} - \bar{y}), \tag{6}$$

where $\bar{y}$ is the sample mean. An estimator for ρ_k is

$$r_k = \frac{c_k}{c_0}.$$

(Note that c_0 is the sample variance.) Since r_k only *estimates* ρ_k for the stochastic process that produced the time series data, it is important to know when r_k values are truly significant and when, due to randomness in the measurements collected, an r_k is approximately 0. Such concerns are addressed by the Bartlett bound [3], which tests the hypothesis (at each lag) that $r_k = 0$ at a specified significance level. Figure 3 displays autocorrelations (vertical bars) versus lags for the time series data in Fig. 1. The Bartlett bounds (at the 5% significance level) are depicted by the dotted lines; bars that lie within the bounds are not considered different from 0. Note that the M/M/1 data have an ACF that has

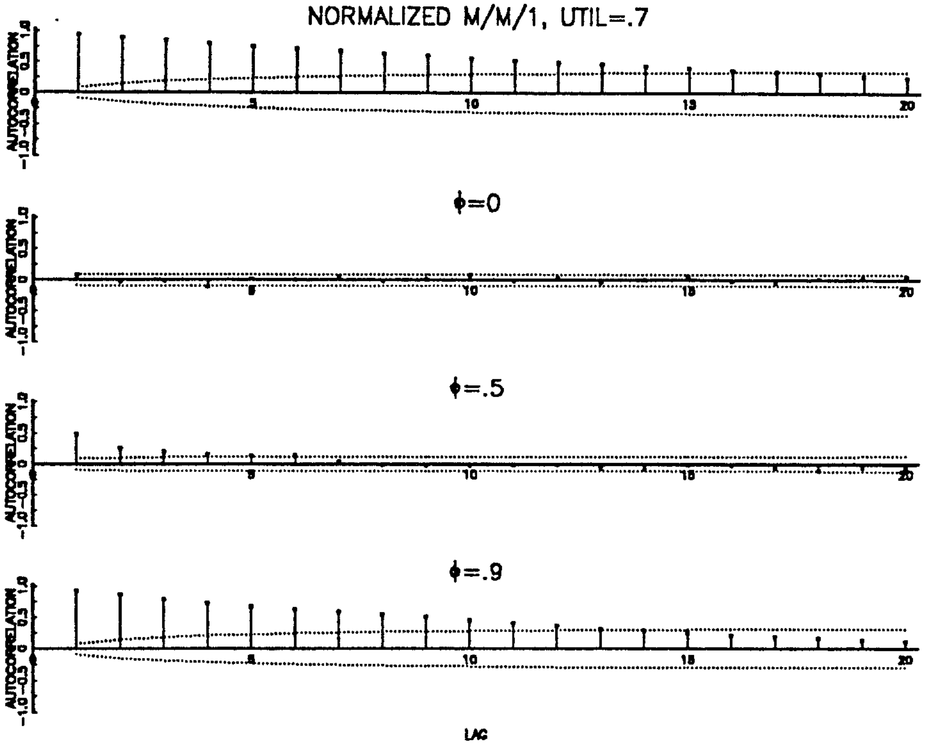

Fig. 3. Empirical ACFs with Bartlett Bounds

the form of a damped exponential, and all three AR(1) models have an ACF that decays in a similar manner (as suggested by Eq. (5) with $0 < \phi < 1$). However, none of the AR(1) models has a decay that is as long as that for the M/M/1 data, which suggests that $\phi > .9$.

One drawback of using the ACF for model identification is that terms in the ACF are highly correlated. For example, the AR(1) model expresses a *direct* relationship between $\tilde{y}_t$ and $\tilde{y}_{t-1}$. However, depending on ϕ, $\tilde{y}_t$ may also have a large correlation with $\tilde{y}_{t-2}$ $\tilde{y}_{t-3}$, and so on. This situation can be remedied by using the **partial autocorrelation function (PACF)**, which computes the lag k autocorrelation after having removed autocorrelations for lags $1, 2, \cdots, k-1$. As with the ACF, bounds can be computed for the PACF; herein, a 5% significance level is used. The PACF is commonly displayed in combination with the original time series and the series' ACF; together, we refer to these as the **identification plots.** Figure 4 contains the identification plots for a realization of an AR(1) process with $\phi = .9$. Note that the PACF is a single spike; this follows from the fact that once the lag 1 autocorrelation is removed from $\tilde{y}_t$, only a_t remains and the a_t are i.i.d..

After model identification, we proceed to the estimation step. Algorithms

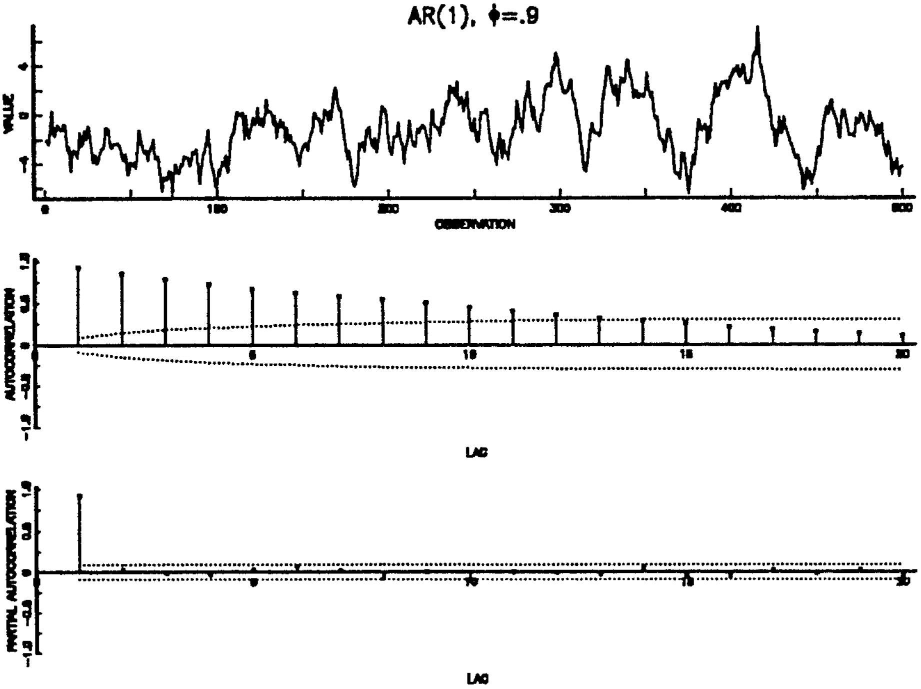

Fig. 4. Time Series, ACF, and PACF (Identification Plots) for an AR(1) Process

used for estimating model parameters are discussed in [3], although such details are typically not required in practice since many software packages implement these algorithms (e.g., [9], [13]). In the case of AR(1), there is only one parameter to estimate. The estimator, which is denoted by $\hat{\phi}$, is computed as follows:

$$\hat{\phi} = \frac{c_1}{c_0}.$$

Applying this calculation to the M/M/1 data in Fig. 2, we determine that $\hat{\phi} = .94$, which confirms our suspicion that $\phi > .9$.

Before proceeding to the evaluation step, we introduce a key concept: the model **residuals**. Denoted by e_t, the model residuals are the difference between the observed and estimated values of the y_t. For example, AR(1) residuals are computed as follows:

$$e_t = (y_t - \bar{y}) - \hat{\phi}(y_{t-1} - \bar{y})$$

The residuals provide a way to assess what is *not* explained by the model.

Model evaluation requires some negative logic. A good statistical model explains all patterns in the data. Thus, in a good model, the e_t should have no time serial dependencies since removing the effect of the time series model from

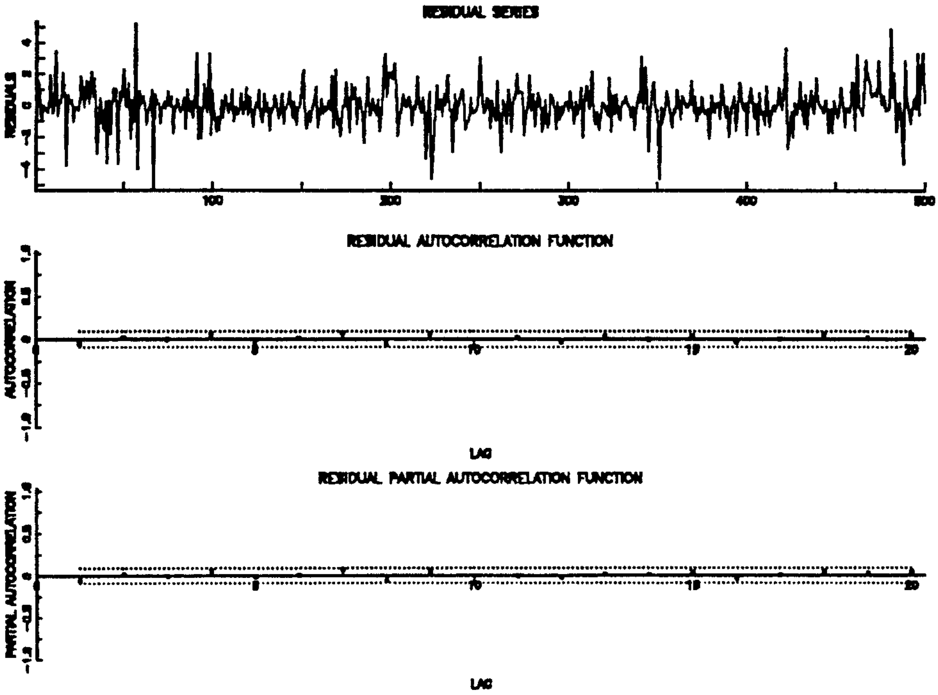

Fig. 5. Evaluation of the AR(1) Model of the M/M/1 Data

the data should leave no patterns in the residuals. Put differently, the residuals should be white noise, which means that both the ACF and the PACF should be 0 (since in a white noise model the terms are i.i.d.). A common way of assessing if the residuals are white noise is to display the identification plots for the e_t. For example, Fig. 5 contains the identification plots for the residuals of the M/M/1 data using an AR(1) model with $\hat{\phi} = .94$ (as obtained from model estimation). Note that both the ACF and the PACF are close to 0, which suggests that the residuals have no time serial dependency. Hence, we conclude that an AR(1) model with $\phi = .94$ provides a fairly good approximation to the dynamic behavior of the M/M/1 time series in Fig. 2. Had the residuals not been white noise, we would have revised the model to include the time serial behavior present in the residuals.

Our focus has been AR(1) models. Other models are often of interest as well. In particular, the one parameter moving average model, or MA(1), sometimes arises. The time series equation for MA(1) is

$$\tilde{y}_t = a_t - \theta a_{t-1}.$$

The ACF of an MA(1) model has a single spike at lag 1, and the PACF has values that decay exponentially. This is the opposite of AR(1) and is a consequence of the fact that an AR model can be expressed as an MA model and vice versa. ARMA(1,1) models have the form

$$\tilde{y}_t = \phi \tilde{y}_{t-1} + a_t - \theta a_{t-1}.$$

Here, both the ACF and PACF consist of values that decay exponentially. Table 2 summaries the characteristics of several time series models. This table is used in following sections to relate the empirical ACF and PACF of measurements to the ACF and PACF of time series models.

Table 2. Characterizations of Several Time Series

Model	ACF	PACF
AR(1)	d.e.	s.s.
MA(1)	s.s.	d.e.
ARMA(1,1)	d.e.	d.e.
White Noise	0	0

– d.e. - damped exponential beginning at lag 1
– s.s. - single spike at lag 1

To summarize, the key to model identification is characterizing the time series in terms of its autocorrelation function (ACF) and its partial autocorrelation function (PACF); these functions describe the relationship between terms in the series that are separated by the same number of time units (or lags). The identification step of time series analysis chooses a model whose theoretical ACF and PACF (as computed from the equation for the model) most closely matches the empirical ACF and PACF of the data. The evaluation step involves looking at the residuals obtained for the model chosen. A good model has residuals that show no evidence of time serial behavior; that is, the residuals are white noise. Time serial behavior in the residuals is detected by applying the identification step to the residuals.

3 Handling Non-Stationary Data

Constructing an ARMA model requires that the underlying process be stationary. Often, this is not the case, especially for measurements of information systems. For example, in time-shared computer systems, usage tends to peak in the mid-morning and just after lunch; as a result, the mean response time is larger at these times. A commonly used approach for handling non-stationary data is to develop two separate models. The first models the non-stationarity. The residuals from this model (i.e., what remains after the effects of the non-stationarity have been removed) should be stationary; otherwise the model is inadequate.

The second model applies the techniques described in section 2 to the residuals of the first model.

One approach to modelling non-stationary behavior is the **integrated (I)** model. To motivate this approach, consider an AR(1) process for which $\phi = 1$. That is, $\tilde{y}_t = \tilde{y}_{t-1} + a_t$. From Eq. (3), we see that this process has an infinite variance and so is non-stationary. However, the *difference* between successive terms in $\tilde{y}_t$ is stationary. Specifically, if $\tilde{w}_t = \tilde{y}_t - \tilde{y}_{t-1}$, then $\tilde{w}_t = a_t$. The $\tilde{w}_t$ series is called the first difference of the $\tilde{y}_t$ series. In theory, a series can be differenced an arbitrary number of times. Recovering the original series requires the inverse operation – summation or integration, which motivates the name integrated model.

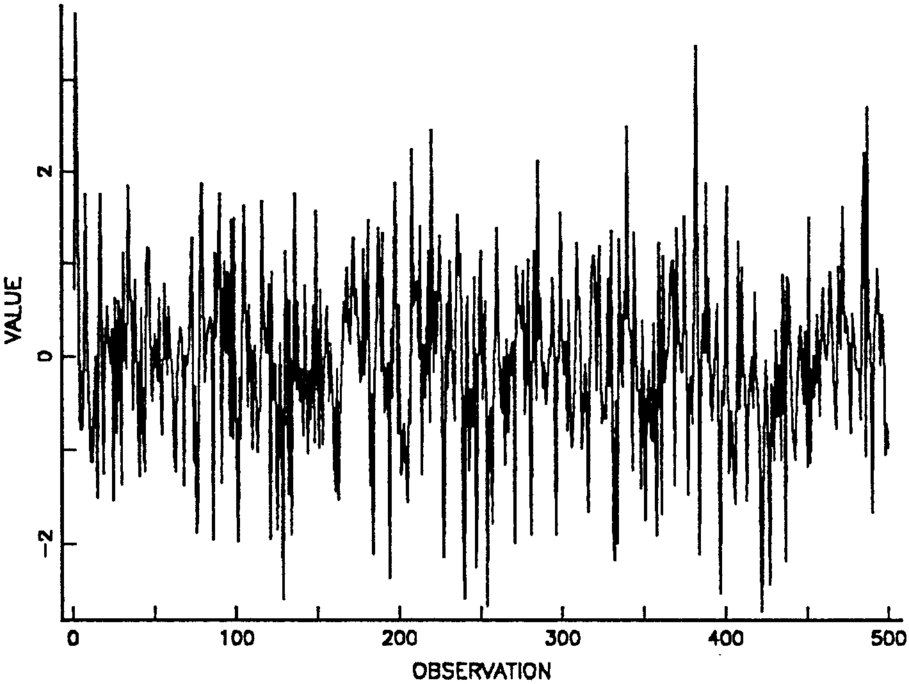

Fig. 6. Illustrative Data

How effective is differencing for handling non-stationarities? To answer this question, consider the data in Fig. 6. A cursory glance raises doubts about the stationarity of these data since there are several sequences that seem to be well above or well below the overall mean. Figure 7 plots the first difference of this data; the result lacks the long sequences of values that tend away from

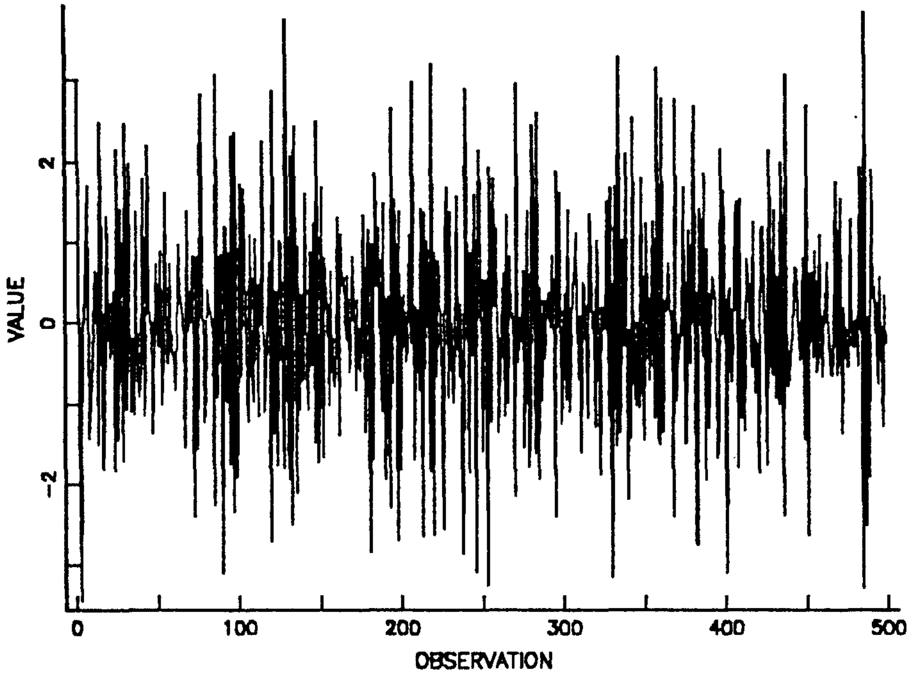

Fig. 7. Differenced Series

the sample mean, which suggests that differencing produced a stationary series. Next, we construct an ARMA model for the differenced series; its identification plots are displayed in Fig. 8. Note that the ACF has a single spike at lag 1, and the PACF is a damped exponential. From Table 2, such characteristics are consistent with an MA(1) model. Indeed, the residuals obtained by applying MA(1) to the differenced series have an ACF and PACF that are white noise (although due to space limitations these plots are not included.) Thus, it appears that the data in Fig. 6 come from a one parameter integrated, one parameter moving average process, which is denoted by IMA(1,1).

The foregoing illustrates a common mistake in time series analysis – applying differencing before non-stationarity has been confirmed. To confirm that a series is non-stationary, its ACF and PACF should be plotted; the data are non-stationary if the ACF and/or PACF do not stay within the significance bounds at large lags. Figure 9 contains the identification plots for the data in Fig. 6. We see that the original series is stationary; in fact, this series is white noise since both the ACF and PACF are in essence 0. In other words, differencing *created* an MA(1) process! To see why, let $\tilde{w}_t$ denote the differenced series. Since the

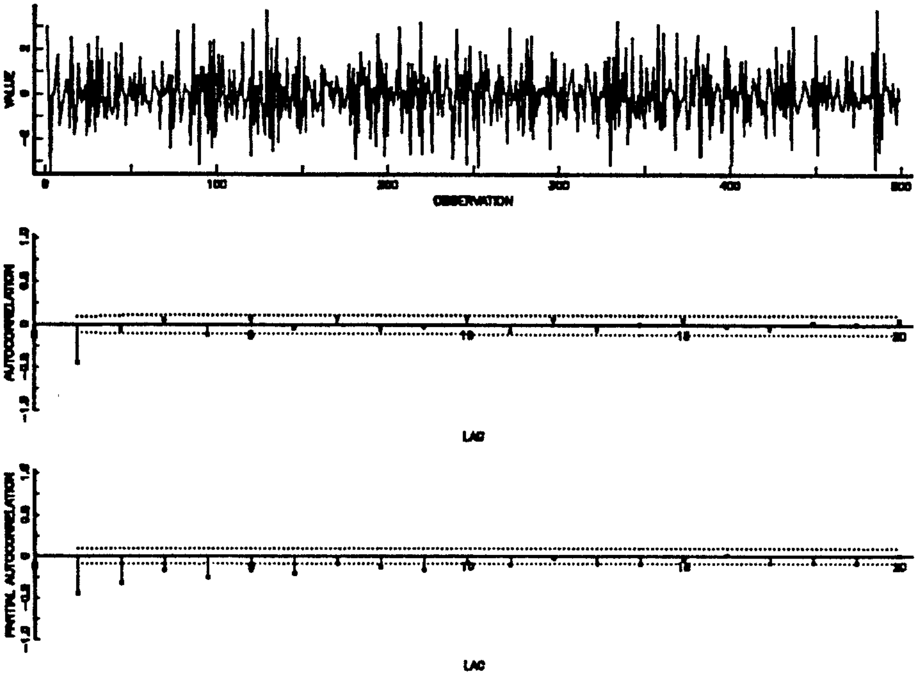

Fig. 8. Identification Plots For Differenced Series

original series is white noise, $y_t = a_t$, and so

$$\begin{aligned}\tilde{w}_t &= \tilde{y}_t - \tilde{y}_{t-1}\\ &= a_t - a_{t-1}\\ &= a_t - \theta_1 a_{t-1}.\end{aligned}$$

Considerable judgement is required when interpreting the ACF and PACF plots to determine if a series is stationary. Is there a way to eliminate nonstationarities without inadvertently creating an MA(1) model? One approach is to partition the series into multiple sub-series, each of which represents a different operating region. This too requires judgement, and so the ACF and PACF of each sub-series should be examined to confirm that the sub-series is stationary. This could be done by constructing identification plots for each sub-series. An alternative is to use least squares regression [5] to fit a moderate degree polynomial of time to the data. If this fit accounts for a small fraction of the variability in the data (say under 5%), no trend is present and so we feel more comfortable that the data are stationary. Both of these techniques are illustrated in the next section.

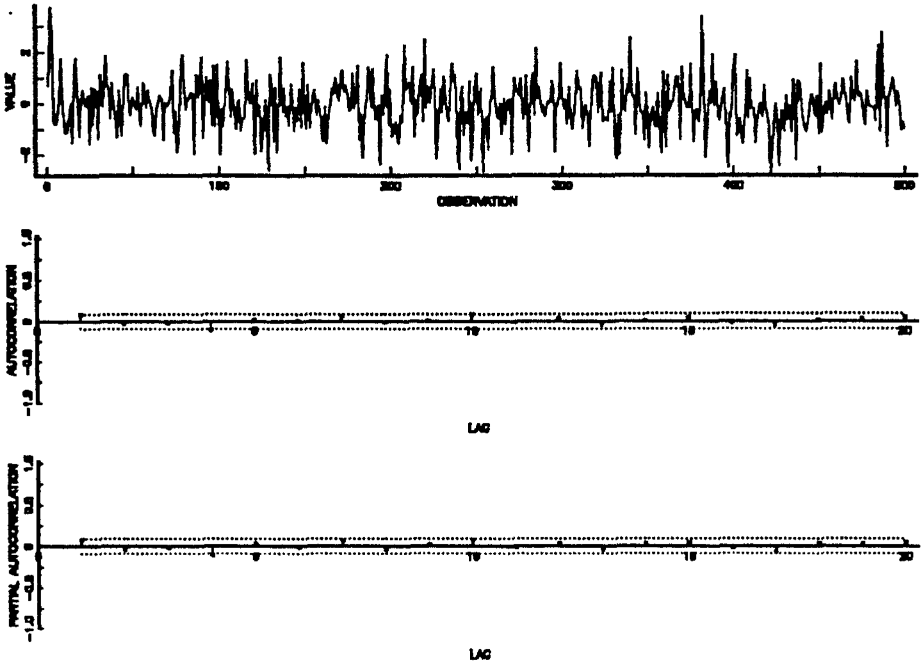

Fig. 9. Identification Plots for Illustrative Data

4 Case Study

This section illustrates time series analysis by applying it to the diagnosis of a performance problem in a large time sharing system at a major utility company. Users of this system complained of intermittently poor performance. In order to diagnosis the underlying problem, response times were measured every thirty seconds over a nine hour shift when there were performance complaints; Fig. 1 plots the measurements obtained. Herein is developed a time series model of these measurements with the objective of characterizing the cause of the performance problem.

The first step in developing a time series model is to detect and resolve non-stationarities in the data. Stationarity is an unreasonable assumption for the data in Fig. 1 in that there appear to be multiple operating regions: a relatively stationary (although highly variable) region from 8:00 AM until 10:30 AM, an abrupt increase in response time from 10:30 to 11:30, and an upward trend that starts at 12:00 PM and continues for the rest of the series. The identification plots in Fig. 10 confirm that these data are non-stationary since the ACF remains outside the significance bounds at large lags.

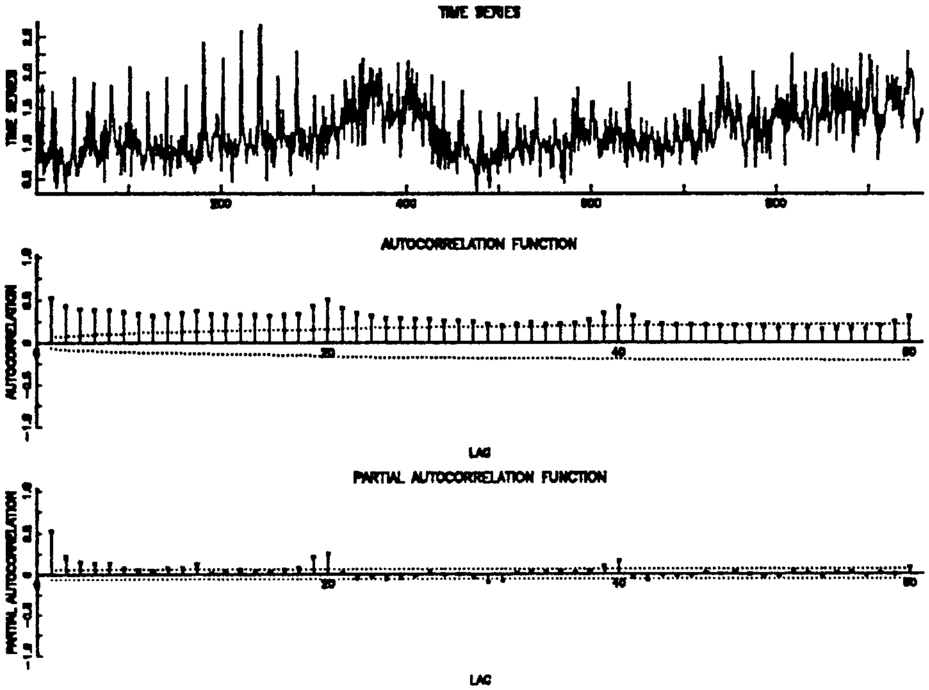

Fig. 10. Identification Plots for Full Time Series

One approach to resolving non-stationarities is to partition the data. This is particularly appropriate when there are multiple operating regions, as in Fig. 1 However, selecting a stationary sub-series requires some judgement. We focus on the sub-series from 8:00 AM to 10:30 AM (180 observations); Fig. 11 contains its identification plots. For the most part, the ACF and PACF lie within the significance bounds at larger lags. So, we could proceed with the identification step. Doing so might lead us to conclude that there is an AR(1) component in the time series since the first lag of the ACF is just above the significance bound. On the other hand, a lag 1 autocorrelation that is significant might be due to the sub-series being non-stationary.

To determine if the sub-series chosen is stationary, a second technique is applied: fitting a moderate degree polynomial of time to the data. Figure 12 displays a fifth degree polynomial of time (the dashed line) superimposed on the sub-series that we are modeling. The fitted curve, which we denote by $f(t)$, accounts for approximately 10% of the variation in the sub-series, which suggests that the sub-series is not stationary. If $f(t)$ adequately models the non-stationary behavior, the residuals of this model are stationary. Thus, our focus is these residuals. Denoted by w_t, the residuals are computed as $w_t = y_t - f(t)$.

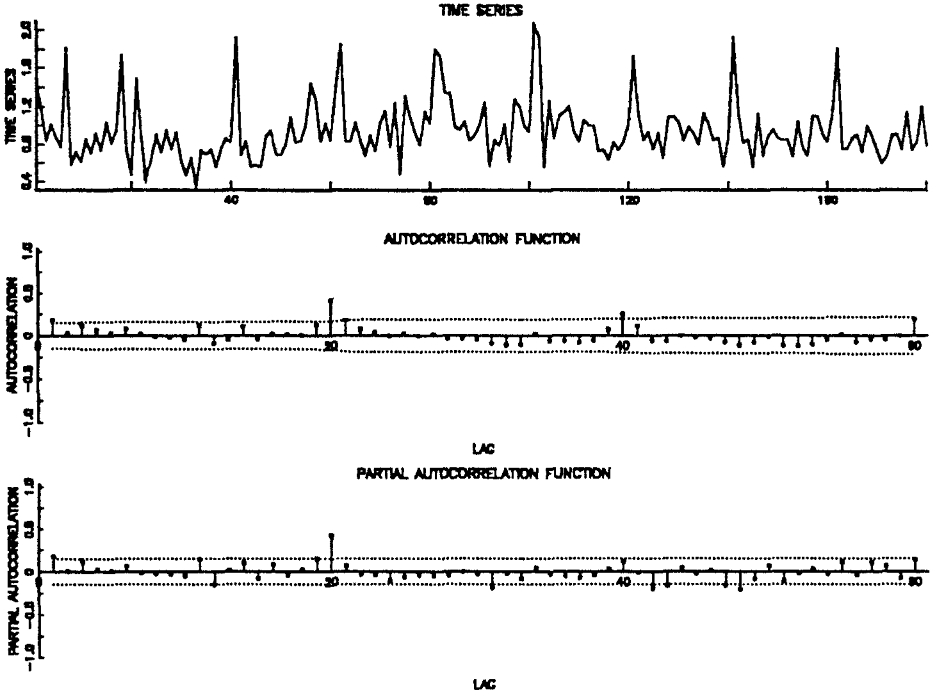

Fig. 11. Identification Plots for the Sub-series (First 180 Observations)

Figure 13 displays the identification plots for w_t. The lag one autocorrelation now lies within the Bartlett bounds, and so we conclude that there is no AR(1) component in w_t. However, there are several partial autocorrelations that lie just outside the significance bounds. Here some judgement is required. Since we have already taken several steps to eliminate non-stationary behavior and the offending values are just barely significant, we only consider the partial autocorrelation at lag twenty to be non-zero.

We now proceed to the identification step, which requires matching the ACF and PACF of the data with that of a time series model. The ACF plot in Fig. 13 has non-zero values at lags 20, 40, and 60; further, these autocorrelations show a gradual decline as the lag increases. The PACF consists of a single spike at lag 20. None of the models in table Fig. 2 have this kind of pattern. However, if we delete the non-zero lags from the ACF and PACF, the identification plots would look like an AR(1) model. In fact, what we have is an **AR(1) seasonal** model; seasonal models indicate the presence of a periodicity. Algebraically, this is expressed as:

$$\tilde{y}_t = \Phi \tilde{y}_{t-s} + a_t,$$

where s is the seasonality parameter that specifies the number of lags between a

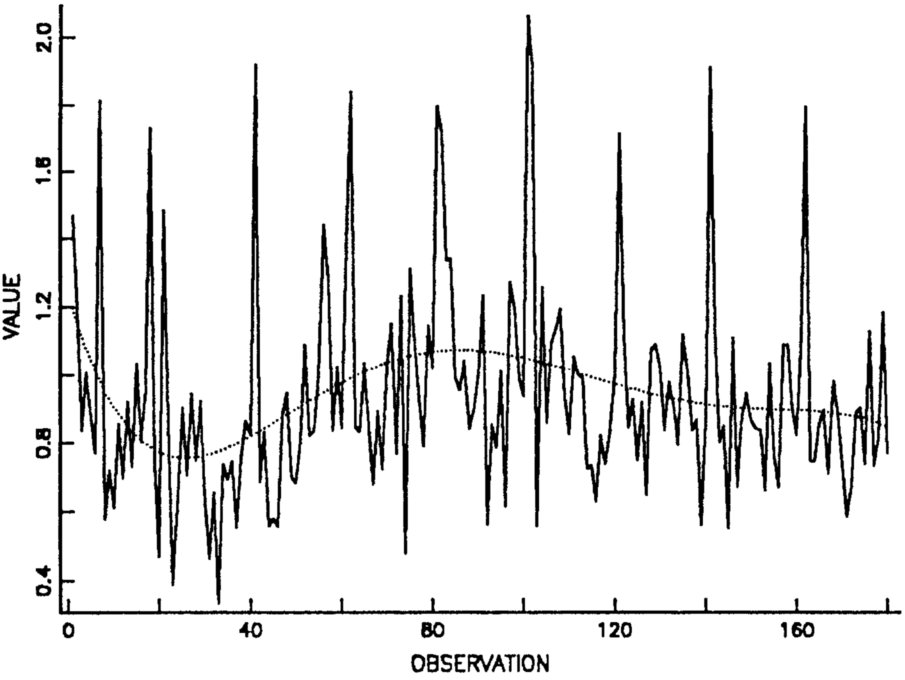

Fig. 12. Curve Fit (5-th degree polynomial) to Sub-series

periodic event. In our case, $s = 20$; that is, there is an event every twenty time units (or ten minutes) that has a significant effect on performance. Using the facilities of the AGSS statistical package [13], an estimate of .497 was obtained for Φ. Thus, we have the following model for the first 180 observations of the data in Fig. 1:

$$(\tilde{y}_t - f'(t)) = (.497)(\tilde{y}_{t-20} - f'(t-20)) + a_t, \tag{7}$$

where $f'(t) = f(t) - \bar{y}$. We evaluate this model by using the identification plots for its residuals, where

$$e_t = (y_t - f(t)) - (.497)(y_{t-20} - f(t-20)).$$

From Fig. 14 we see that the residuals are white noise; so Eq. (7) seems to be a reasonable model.

Eq. (7) indicates that performance is degraded significantly by a process that executes every ten minutes. This information allowed the operations staff to focus on a small subset of their applications; relatively quickly they discovered an inefficiently written application that executed every ten minutes. After changing a search routine in this application, system performance improved substantially.

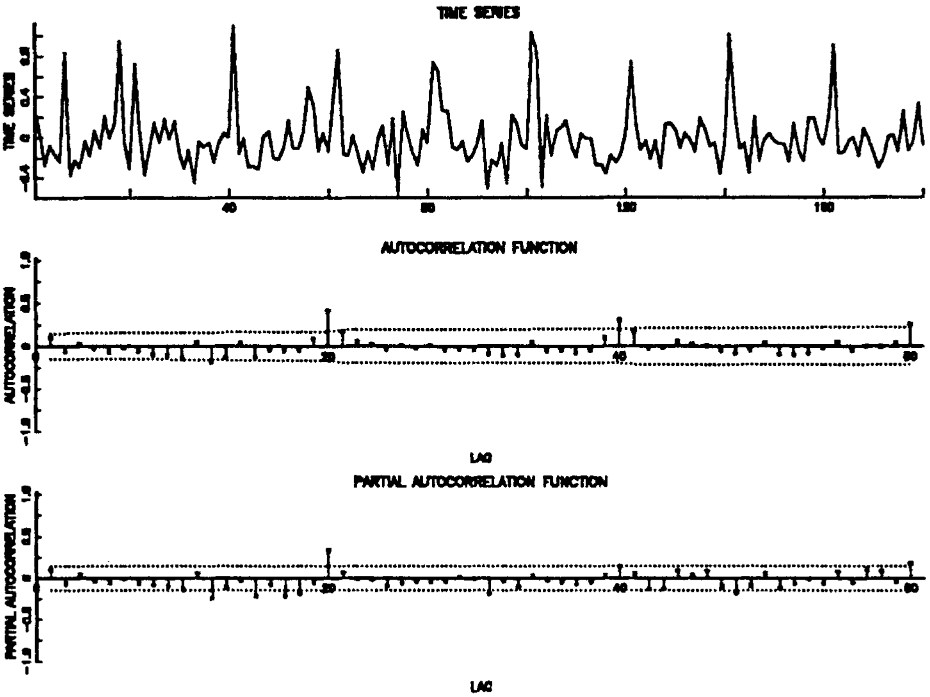

Fig. 13. Identification Plots for the Residuals of the Curve Fit

5 Summary

Modeling the dynamic behavior of information systems is of importance in many situations, such as characterizing the locality of disk accesses, evaluating the dynamic behavior of scheduling algorithms, and diagnosing intermittent performance problems. Time series analysis is a statistical approach to modeling dynamic behavior; a time series model is an algebraic expression that relates the t-th term to the proceeding p terms and to q random shocks (which represent random events that cannot be measured). Developing a time series model involves the following steps: (1) resolving non-stationarities in the data, (2) identifying the values for the parameters p and q, which determine the type of model such as autoregressive (AR) or moving average (MA), (3) estimating unknown constants, (4) evaluating the model constructed, and (5) forecasting future values. In general, model development is an iterative process in which steps (1) through (4) are applied repeatedly before proceeding to step (5).

Key to constructing a time series model is characterizing dynamic behavior. One commonly used approach employs the autocorrelation function (ACF) and partial autocorrelation function (PACF). For example, the second step in time

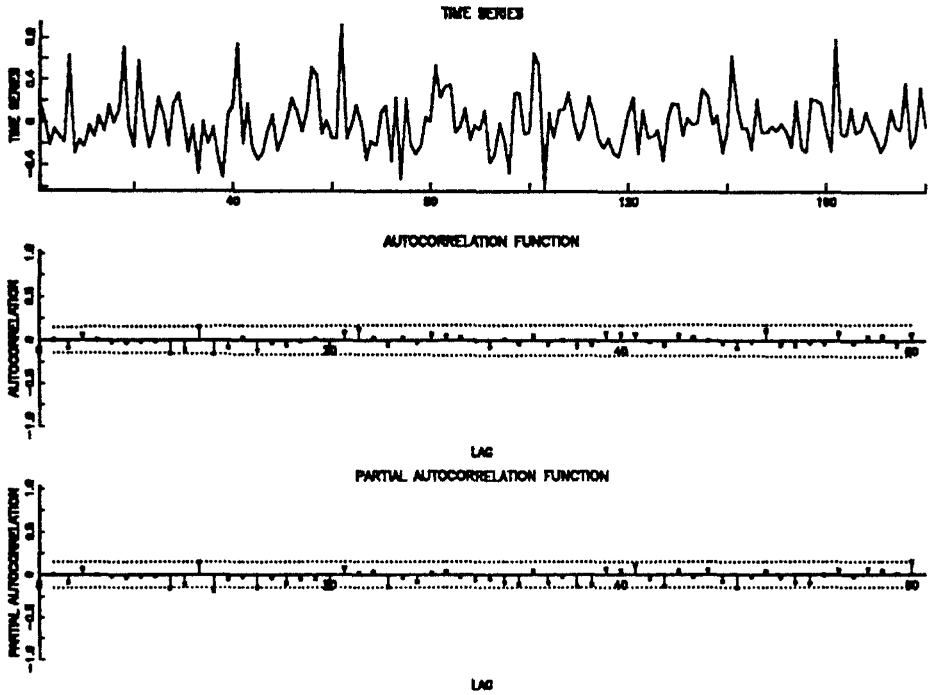

Fig. 14. Identification Plots for the Residuals of the Final Model

series analysis can be accomplished by comparing the empirical ACF and PACF of the time series data with the ACF and PACF of several time series models. (The latter are computed from the algebraic expression of the time series model).

There are several related topics that are worthwhile pursing. The case study in section 4 touched on AR(1) seasonal models. Seasonality can be incorporated into any ARMA model, and may appear in either (or both) the autoregressive or the moving average components. Another topic is transfer function models in which one time series is modelled in terms of one or more other time series (e.g., modelling response times in terms of interarrival and service times). Lastly, the area of stochastic control may be of particular interest to designers of information systems since it provides a formal approach to constructing optimal controls in the presence of random noise. The first two topics are discussed in depth in [3]; [3] touches on the third topic, but more details are contained in [2].

References

1. **Hamid Ahmadi and Parviz Kermani**: "Real Time Network Load Estimation in Packet Switched Networks," *Data Communication Systems and Their Performance*, Guy Pujolle and Ramon Ruigjaner, Ed., 367-380, 1991.
2. **Karl J. Astrom**: *Introduction to Stochastic Control Theory*, Academic Press, 1970.
3. **George E. P. Box and Gwilym M. Jenkins**: *Time Series Analysis Forecasting and Control*, Prentice Hall, 1976.
4. **John N. Diagle and Joseph D. Langford**: "Models for Analysis of Packet Voice Communications Systems," *IEEE Journal on Selected Areas in Communications*, **vol. 4, no. 6**, 847-855, 1986.
5. **N.R. Draper and H. Smith**: *Applied Regression Analysis*, John Wiley and Sons, 1968.
6. **EDP**: "Workload Demand Forecasting," *EDP Performance Review* 6-7, 1986.
7. **Daniel P. Heyman, Ali Tabatabai, and T.V. Lakshman**: "Statistical Analysis and Simulation Study of Video Teleconference Traffic in ATM Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, **vol. 2, no. 1**, 49-59, 1992.
8. **Helen Letmanyi**: "Guide on Workload Forecasting," *NBS Special Publication 500-123*, PB85-177632, 1985.
9. **SAS**: "SAS User's Guide," SAS Institute Inc., 1985.
10. **Prodip Sen, Basil Maglaris, Nasser-Eddine Rikli, and Dimitris Anastassiou**: "Models for Packet Switching of Variable-Bit-Rate Video Sources," *IEEE Journal on Selected Areas in Communications*, 865-869, 1989.
11. **Ioannis Stavrakakis**: "An Analysis Approach to Multi Level Networking," *International Conference on Communications*, 274-301, 1990.
12. **Walter Vandaele**: *Applied Time Series and Box-Jenkins Models*, Academic Press, Inc., 1983.
13. **Peter Welch and Thomas Lane**: "The Integration of a Menu-Oriented Graphical Statistical System with its Underlying General Purpose Language," *Computer Science and Statistics: Proceedings of the 19th Symposium on the Interface*, 267-273, Philadelphia, February 1987.