REPORT NO. IA-11-0005

DEVELOPMENT OF TIME-SERIES BASED TRANSIT PATRONAGE MODELS

VOLUME 2

HANDBOOK FOR APPLYING TIME SERIES MODELS TO TRANSIT RIDERSHIP FORECASTS

MICHAEL KYTE

JAMES STONER

JONATHAN CRYER

The University of Iowa
College of Engineering
Iowa City, Iowa 52240

MARCH 1985

FINAL REPORT

Prepared for

U.S. DEPARTMENT OF TRANSPORTATION
URBAN MASS TRANSPORTATION ADMINISTRATION
OFFICE OF TECHNICAL ASSISTANCE
UNIVERSITY RESEARCH AND TRAINING
WASHINGTON, D.C. 20590

A HANDBOOK FOR APPLYING TIME SERIES TRANSIT RIDERSHIP FORECASTS

Prepared By:

Jonathan Cryer
James Stoner
Michael Kyte

The University of Iowa

March 1985

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
UMTA-IA-11-0005-85-1		

4. Title and Subtitle	5. Report Date
Development of Time Series Based Transit Patronage Models. Volume II - Handbook for Applying Time Series Models to Transit Ridership Forecasts	March 1985
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
M. Kyte, J. Stoner, and J. Cryer	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
The University of Iowa College of Engineering Iowa City, iowa 52242	11. Contract or Grant No. IA-11-0005
12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
U.S. Department of Transportation Urban Mass Transportation Administration Office of Technical Assistance University Research & Training Program Washington, D.C. 20590	Final Report July 1983 - March 1985
	14. Sponsoring Agency Code URT-33

15. Supplementary Notes

UMTA project Director - Judy Z. Meade
Technical Advisor - Thomas J. Hillegas

16. Abstract

This report describes the statistical procedures for applying Box-Jenkins time series models to the forecasting of transit ridership. Techniques are described for: (1) generating data sets having appropriate characteristics for time series analysis, (2) evaluating the data for appropriate model forms, (3) estimating parameters for the model, and (4) applications of the model to forecasting and evaluation of previous changes in transit level of service or fare policy.

17. Key Words	18. Distribution Statement
Transit, Time Series, Modeling, Forecasting	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
unclassified	unclassified		

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

METRIC CONVERSION FACTORS

Approximate Conversions to Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
in	inches	*2.5	centimeters	cm
ft	feet	30	centimeters	cm
yd	yards	0.9	meters	m
mi	miles	1.6	kilometers	km
		AREA		
in^2	square inches	6.5	square centimeters	cm^2
ft^2	square feet	0.09	square meters	m^2
yd^2	square yards	0.8	square meters	m^2
mi^2	square miles	2.6	square kilometers	km^2
	acres	0.4	hectares	ha
		MASS (weight)		
oz	ounces	28	grams	g
lb	pounds	0.45	kilograms	kg
	short tons (2000 lb)	0.9	tonnes	t
		VOLUME		
tsp	teaspoons	5	milliliters	ml
Tbsp	tablespoons	15	milliliters	ml
fl oz	fluid ounces	30	milliliters	ml
c	cups	0.24	liters	l
pt	pints	0.47	liters	l
qt	quarts	0.95	liters	l
gal	gallons	3.8	liters	l
ft^3	cubic feet	0.03	cubic meters	m^3
yd^3	cubic yards	0.76	cubic meters	m^3
		TEMPERATURE (exact)		
°F	Fahrenheit temperature	5/9 (after subtracting 32)	Celsius temperature	°C

*1 in = 2.54 (exactly). For other exact conversions and more detailed tables, see NBS Misc. Publ. 286, Units of Weights and Measures, Price $2.25, SD Catalog No. C13.10:286.

Approximate Conversions from Metric Measures

Symbol	When You Know	Multiply by	To Find	Symbol
		LENGTH		
mm	millimeters	0.04	inches	in
cm	centimeters	0.4	inches	in
m	meters	3.3	feet	ft
m	meters	1.1	yards	yd
km	kilometers	0.6	miles	mi
		AREA		
cm^2	square centimeters	0.16	square inches	in^2
m^2	square meters	1.2	square yards	yd^2
km^2	square kilometers	0.4	square miles	mi^2
ha	hectares (10,000 m^2)	2.5	acres	
		MASS (weight)		
g	grams	0.035	ounces	oz
kg	kilograms	2.2	pounds	lb
t	tonnes (1000 kg)	1.1	short tons	
		VOLUME		
ml	milliliters	0.03	fluid ounces	fl oz
l	liters	2.1	pints	pt
l	liters	1.06	quarts	qt
l	liters	0.26	gallons	gal
m^3	cubic meters	35	cubic feet	ft^3
m^3	cubic meters	1.3	cubic yards	yd^3
		TEMPERATURE (exact)		
°C	Celsius temperature	9/5 (then add 32)	Fahrenheit temperature	°F

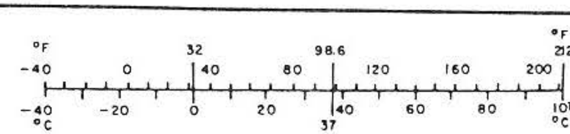


FIGURE 3. METRIC CONVERSION FACTORS

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EXHIBITS

I. INTRODUCTION

OVERVIEW

This report is the second of two technical reports that describe the methodology for analyzing and forecasting public transit ridership. The first report, Development and Application of Time-Series Ridership Models for Portland, Oregon, focuses on the development of the methodology and its application to data from Portland, Oregon. That report describes findings relating to the structure of the models, the impacts of past service level and travel cost changes on transit ridership, and the effectiveness of the models in forecasting transit ridership. Because the statistical techniques used in this work have not had wide application by transportation analysts, it was determined that a Handbook that explained these techniques was needed. The purpose of this second report is to provide such a step-by-step procedure for applying time-series analysis techniques to the problem of estimating transit ridership models. These techniques have come to be known as the Box-Jenkins Method.

The remainder of this Handbook addresses the following topics:

1. Why time-series analysis represents a needed improvement over standard regression models.
2. The techniques for identifying, estimating, and checking time-series models.
3. The three basic kinds of time-series models: Univariate Models, Transfer Function Models and Intervention Models.
4. How to prepare forecasts using these models.

It is assumed that the analyst who will use this handbook has some statistical background, including experience with multiple-regression models. The analyst should also have access to a standard computer statistical package with time-series modeling capability. (See Appendix A for a list of available packages.) The examples in the Handbook utilize the SAS and SAS-ETS statistical programs, one of the more readily available software packages. It is suggested that while reviewing those examples, the analyst should have access to the SAS/ETS Users Guide: Econometrics and Time Series Library for a complete description of job control language and program set-ups. The reader is also referred to Box and Jenkins (1976) and Hoff (1983) for a complete discussion of time-series methodology.

METHODOLOGY

The basic methodology used in this project includes three phases (see Figure 1). In the Model Development Phase, a model form is postulated that includes a description of the variables that are assumed to effect transit ridership. The structural relationships between transit ridership and the input variables are then identified, and the model is estimated and checked.

The model can then be used in the Impact Analysis Phase to analyze the impact on transit ridership of past changes in service level, fare, or other factors. An intervention variable is introduced to the basic model to account for a specific change. The intervention variable is a binary variable which assumes a value of one when the change is in effect and zero at all other times.

The model can also be used in the Forecasting Phase when an assessment of a proposed future change is desired.

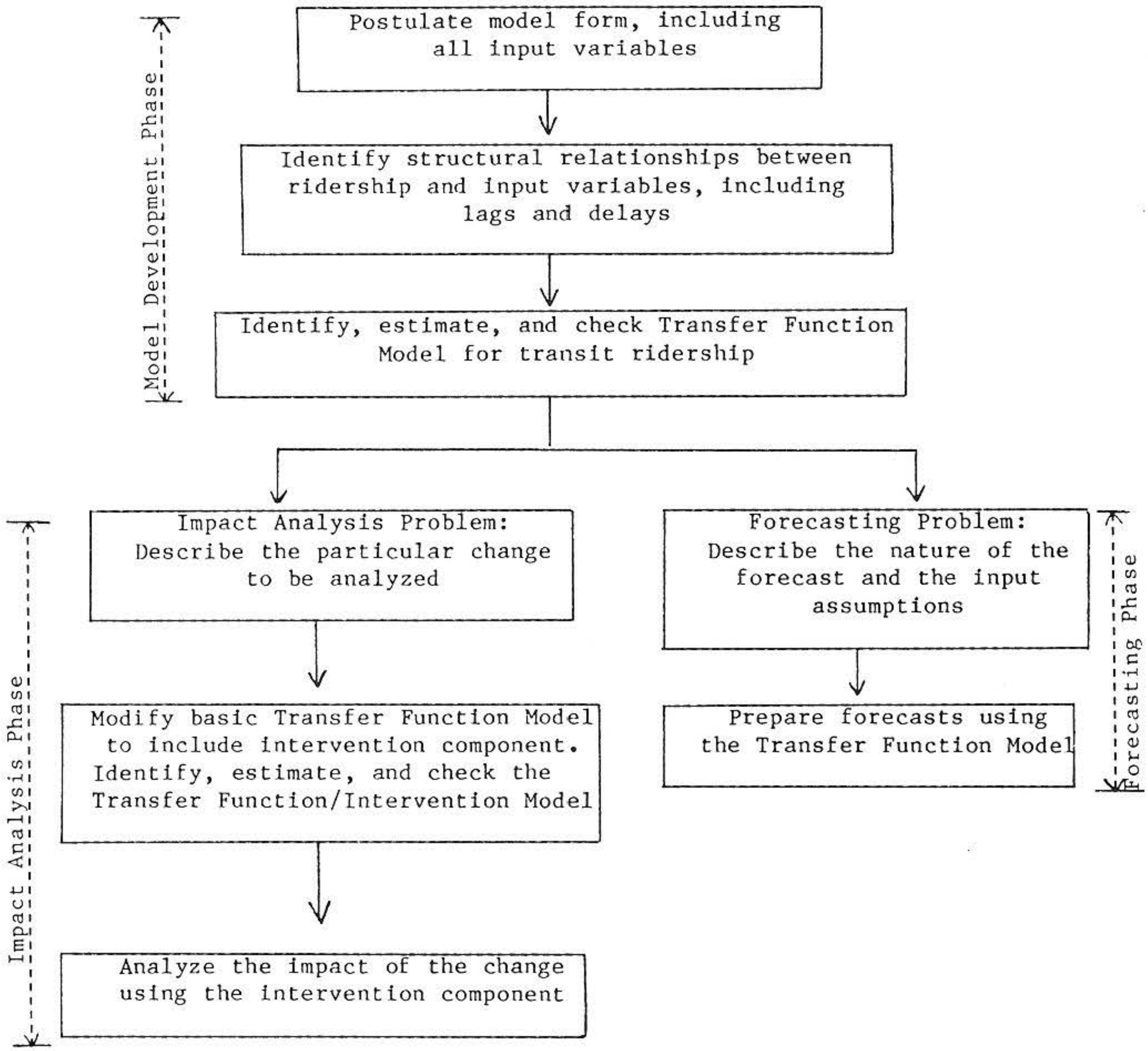
THE MODEL STRUCTURE

The basic model structure used here represents a compromise between theoretical considerations and a practical sense of the data available to the transportation analyst. Specifically, it is assumed that the current level of transit ridership is a function of present and past values of level of service, travel costs, market size, and special events or interventions. Data available to the transit analyst will guide the manner in which these general variables can be specifically described. Table 1 lists some of the specific time-series that can be used for each variable. The example used in this Handbook uses the following independent variables: platform hours of bus service, transit fare, gasoline price and employment.

FIGURE 1

METHODOLOGY FOR ANALYSIS AND FORECASTING

OF PUBLIC TRANSIT RIDERSHIP



REGRESSION MODELS

A number of researchers have used the basic model form described above, or some variant of it, to develop transit ridership regression models using time-series data. For example:

$$(1) \qquad R_t = \beta_0 + \beta_1 SL_t + \beta_2 F_t + a_t$$

relates ridership to service level (SL_t) and transit fare (F_t). In standard regression methods, β_0, β_1, and β_2 are unknown regression parameters to be estimated from the historical data and a_t is the "error" term due to all other variables affecting ridership other than transit fare or service level.

There are several difficulties with such models. First, regression models assume that the error terms, a_1, a_2, ... are statistically independent -- an assumption that will rarely be true of variables measured over time. Secondly, the effects of a fare or service change may not be instantaneously felt in ridership. For example, this month's ridership may be affected by a transit fare change that occurred two months ago, so that variables lagged over time are needed in the model. Furthermore, lagged effects which decay exponentially into the past as

$$(2) \qquad R_t = \beta_0 + \beta_1 F_t + \beta_1^2 F_{t-1} + \beta_1^3 F_{t-2} + \ldots + a_t$$

are not easily incorporated into regression models. Finally, it is often true that two or more of the independent variables increase or decrease together over a long period of time. For example, service level and gasoline price both significantly increased during the 1970's in many cities. This similarity in variation often leads to multicollinearity between variables and reduces the ability to separately estimate variable coefficients. For all of these reasons, other statistical techniques have been developed to handle time-series data.

TABLE 1

TIME SERIES VARIABLES

General Variable	Specific Data to Represent Variable
Level of Service	Platform hours or miles Route miles Platform miles per route mile
Travel Costs	Transit fare Gasoline price per gallon
Market Size	Population of service area Employment of service
Seasonal Factors	Monthly temperature, rainfall, or snowfall School days per month
Special Events or Interventions	Gasoline shortages Marketing or promotional programs Opening of new facilities Weather extremes

TIME-SERIES MODELS

The statistical models described in this Handbook have several attributes that make them more appropriate for use with time-series data than standard regression models:

1. The models include both present and past values of the input series (i.e., independent variables) at both present and past time periods, which reflect the lag structure relating the input and output series. See equation (2), for example.

2. The error structure usually includes the effects of previous time periods. Instead of the single a_t term, additional terms that reflect correlation with the previous time period (a_{t-1}) and with the same time period, a year ago (a_{t-12}) are often included.

3. The model usually relates the differences of the variables from one time period to the next, rather than the nominal values of the variable. That is, instead of the nominal value of fare at time t (F_t), the differenced variable ($F_t - F_{t-1}$) might be used. This usually resolves such statistical problems as multicollinearity.

Three different kinds of time-series models will be developed in this Handbook: Univariate Models, Transfer Function Models, and Intervention Models.

Univariate Models. Univariate models relate the current value of a time-series to its own past values. As an example of a univariate time-series model, consider the gasoline price series denoted G_t. It will be shown later that an adequate model for gasoline price relates its current value (G_t) to its price in previous months plus an error term:

$$G_t = (1 + \phi) G_{t-1} - \phi G_{t-2} + a_t \quad (3)$$

Though simple in form, such a model is often sufficient to make accurate forecasts of future values of a given time-series.

A more complicated univariate model was developed for transit ridership:

$$(4) \quad R_t = R_{t-1} - R_{t-12} + R_{t-13} + a_t - \theta_1 a_{t-12} - \theta_2 a_{t-24}$$

This model relates the current value of transit ridership to three of its previous values (lagged 1, 12, and 13 months) plus three error terms. Note that the model includes terms that are precisely twelve time periods apart, indicating the importance of the seasonal component in transit ridership.

Univariate models will serve only as a means to an end here. They are the first step in developing Transfer Function Models, which interrelate several time-series.

Transfer Function Models. Transfer Function Models use several "input" variables to explain the behavior of the "output" variable, transit ridership. Describing the methodology for developing transfer functions is the main goal of this Handbook.

A simple example of such a model that relates ridership to one input variable (transit fare) would be:

$$(5) \quad R_t = R_{t-12} + w_o(F_t + \delta F_{t-1} + \delta^2 F_{t-2} + \ldots) + a_t - \theta_1 a_{t-12} - \theta_2 a_{t-24}$$

Here, current ridership is related to last year's ridership for the same month (R_{t-12}), current and past values of the transit fare (F_t), and several error terms. The transit fare has been weighted in such a way that its previous values have less of an influence than its current values. That is, the fare coefficients (w_o and δ) describe a decay function. The error terms are correlated between twelve month periods reaching back two years in time.

Intervention Models. Intervention Models include the effects of one or more special events, such as a marketing program, severe weather, or a gasoline supply shortage, that influence transit ridership. These events are represented by a binary variable that has a value of one during the period that the event is occurring, and zero otherwise. A gasoline supply shortage that occurred during months 17, 18 and 19 of the study period would be represented as follows:

$$(6) \quad R_t = R_{t-12} + w_o \xi_t + a_t - \theta_1 a_{t-12} - \theta_2 a_{t-24}$$

$$\text{where} \quad \xi_t \begin{cases} = 1 \text{ for months 17, 18, and 19} \\ = 0 \text{ otherwise} \end{cases}$$

An intervention variable can also be used to represent a specific service or fare change so that its effects may be sorted out from other service or fare changes that occurred during the study period.

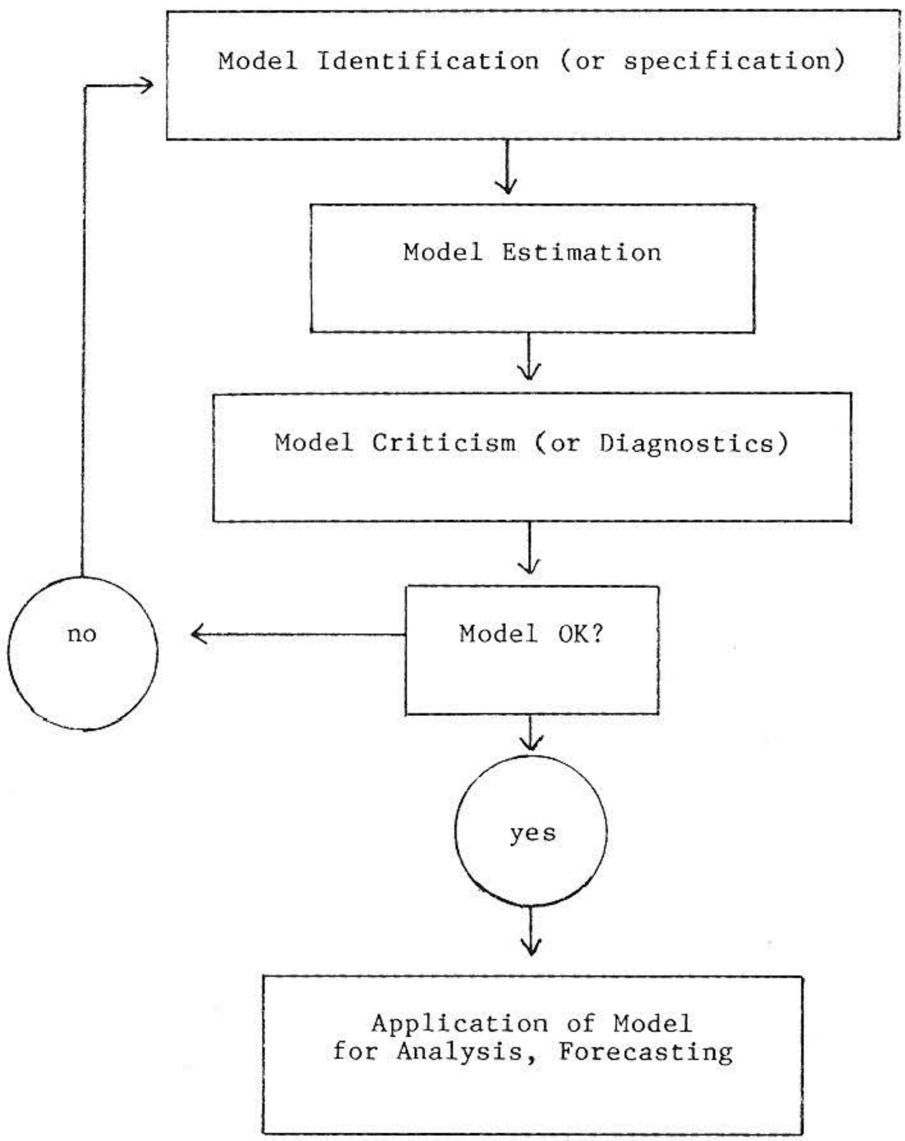
MODEL BUILDING STRATEGY

To develop the time-series models, the model building strategy of Box and Jenkins (1976) will be used. See Figure 2. The purpose of model identification is to separate out the particular time series models which may be appropriate for a given observed series. In this step, the time plot of each series is analyzed as well as a number of statistics computed from each series. In addition, the transit analysts' knowledge of the data is brought to bear.

It should be emphasized that the model specified at this point is tentative and always subject to revision later in the analysis. It should also be pointed out that the structure of the model is often not assumed in advance. The analyst lets the data "speak for itself".

With regard to model identification, the Principle of Parsimony is followed, which asserts that the models should contain the smallest possible number of parameters consistent with an adequate representation of the data. Albert Einstein is quoted in Parzen (1982) as remarking that "everything should be made as simple as possible, but not simpler".

FIGURE 2 - MODEL BUILDING



Also, the model will surely contain one or more unknown parameters whose values must be estimated from the observed series. Model estimation consists of finding the best possible estimates of those unknown parameters within a given model.

Model criticism is concerned with investigation of the quality of the model that has been specified and estimated in explaining the observed time series. How well does the model fit the data? If no inadequacies are found, the model may be used for analysis and forecasting. Otherwise, the nature of the inadequacies are used to respecify the model, and the analyst returns to step 1 of the process. In this way, a cycle is made through the three steps until, hopefully, an acceptable model is found.

EXAMPLE DATA

The data that are used in the examples in this Handbook are from Portland, Oregon. The data set includes 114 monthly data points covering the period January 1973 through June 1982. This period represented one of substantial growth in ridership for Portland's transit system. The changes in service level, transit fare, gasoline price, and employment during this period make it an example of practical interest. Each of the time series are presented in graphical form on the following pages. The detailed list of the data is given in Appendix B.

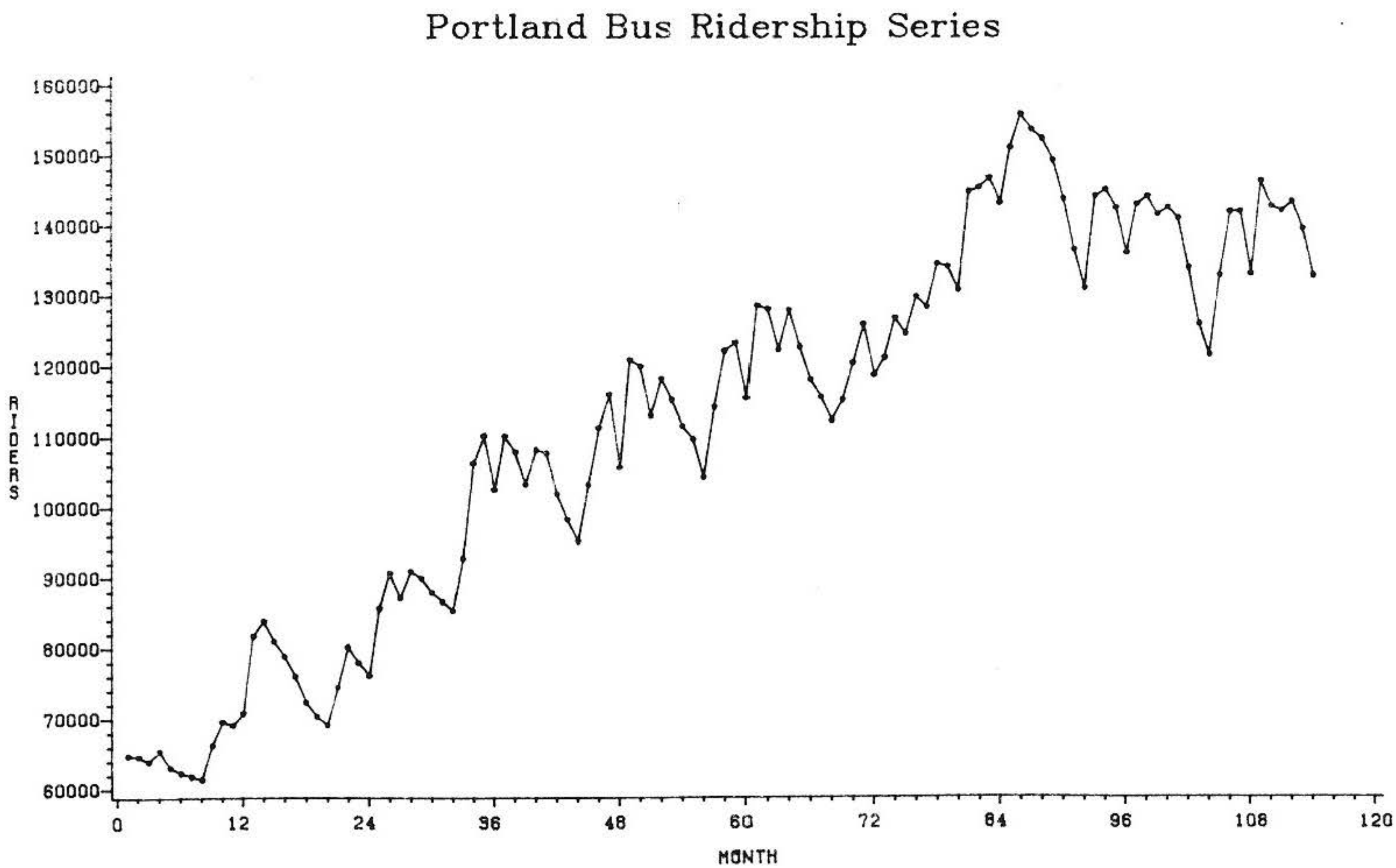


FIGURE 3
PORTLAND BUS RIDERSHIP SERIES

FIGURE 4

PORTLAND PLATFORM HOURS SERIES

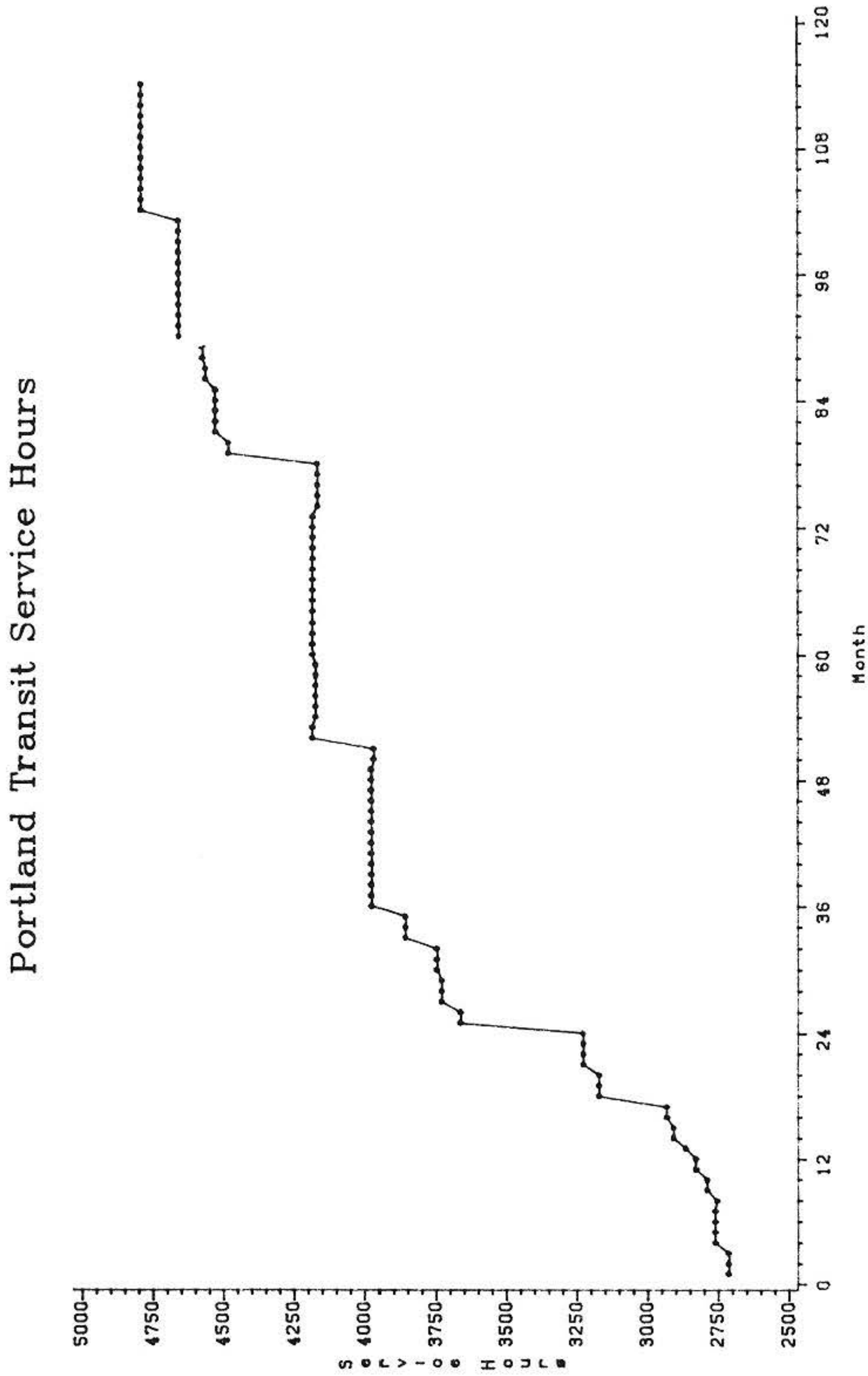


FIGURE 5

PORTLAND TRANSIT FARE SERIES

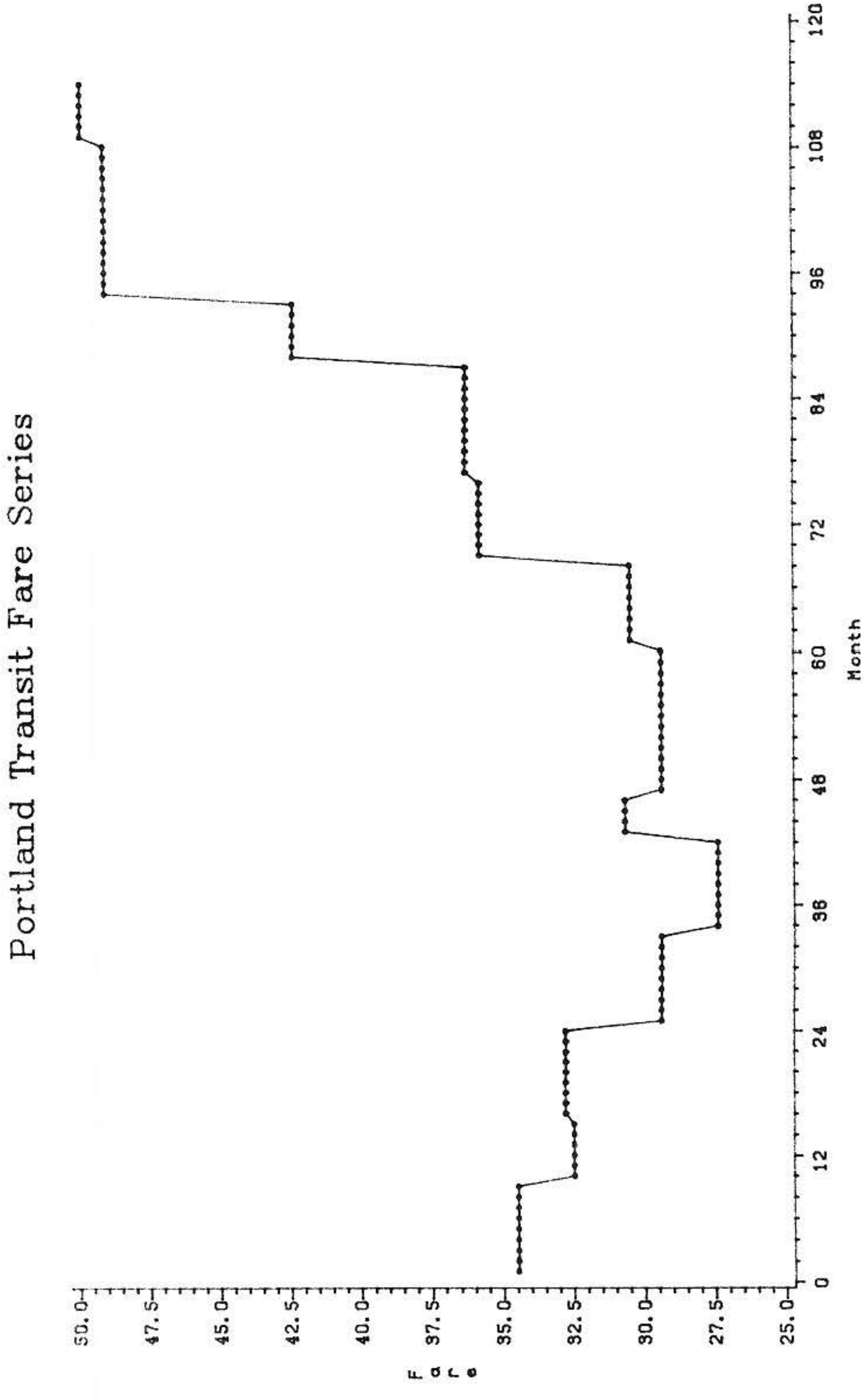


FIGURE 6

PORTLAND GASOLINE PRICE SERIES

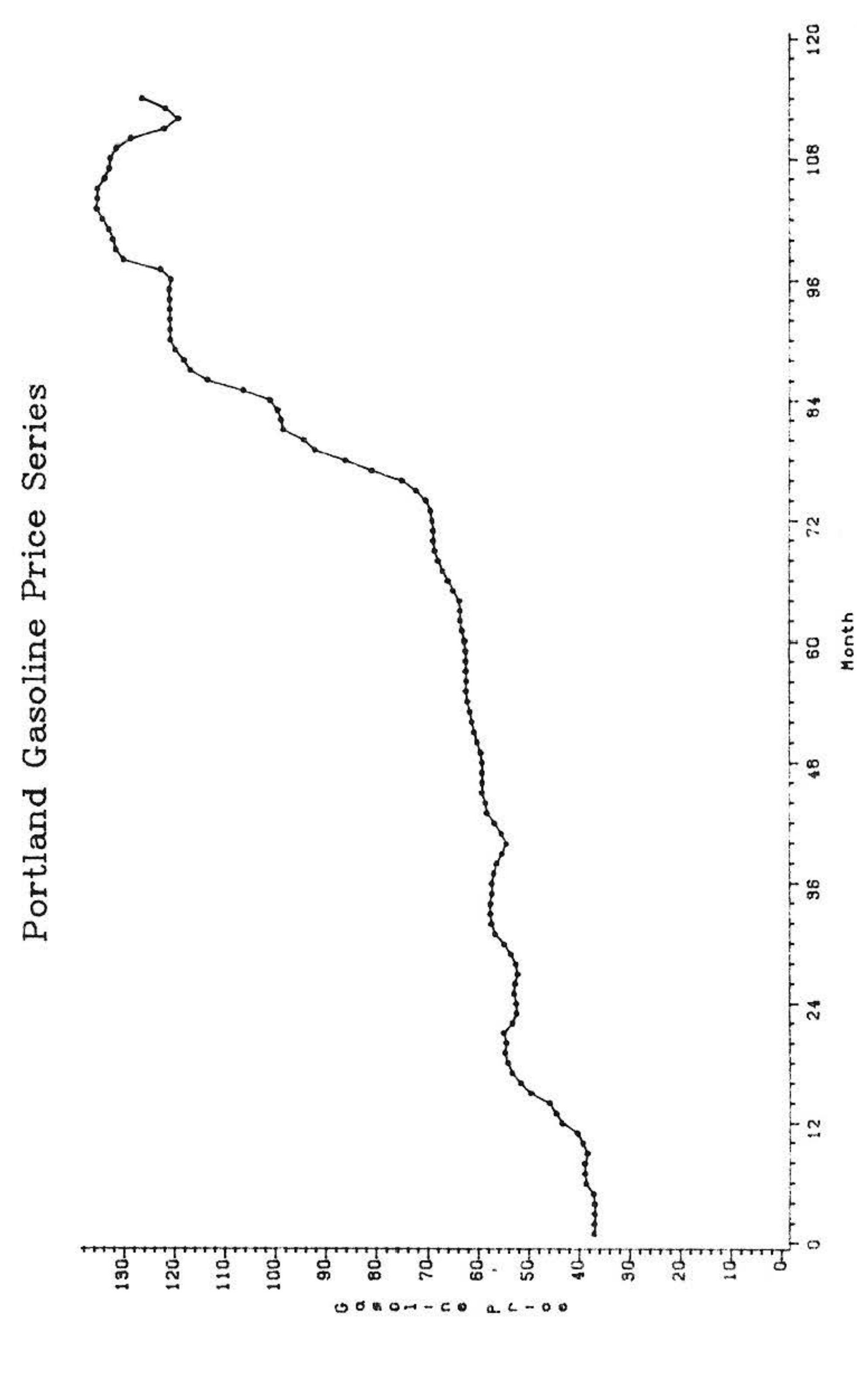
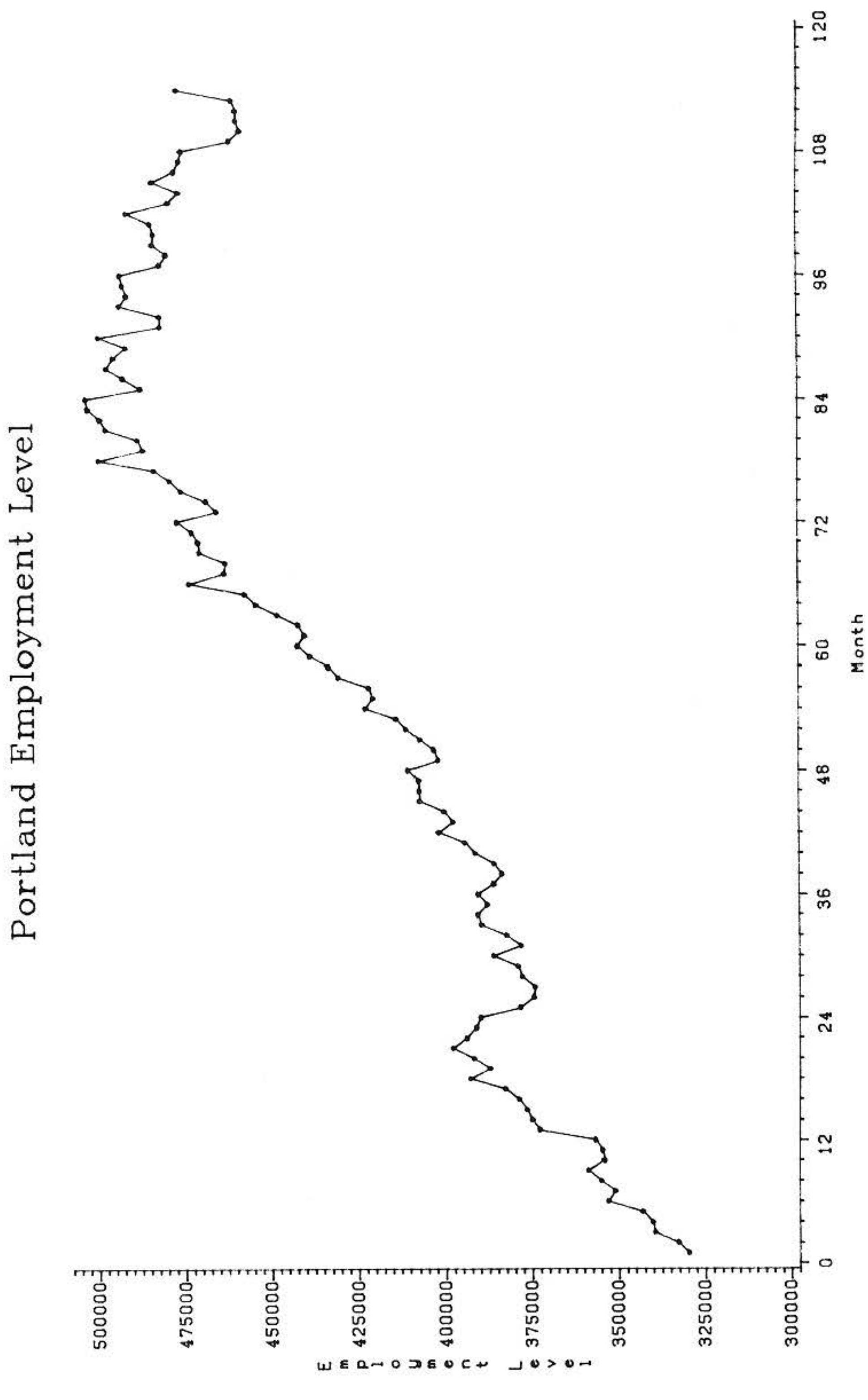


FIGURE 7

PORTLAND EMPLOYMENT SERIES



II. UNIVARIATE TIME-SERIES MODELS

This chapter covers the development of univariate time-series models. The topics covered include:

1. Stationary and Non-Stationary Models
2. The Autocorrelation Function
3. Autoregressive Models
4. Moving-Average Models
5. The Backshift Operator
6. Seasonal Models
7. The Model Building Process

A number of terms and concepts are introduced here. They are listed and defined briefly on the next page.

1. Univariate Models are models which depend only on their own past values.
2. Stationarity is a notion of statistical stability. A time-series is said to be stationary if it varies about some mean value and if the correlation between any two points in the series depends only on the relative time lag between the two points.
3. The autocorrelation function describes the relative dependence or correlation between two values in the time-series that are a given number of time periods apart. It is one of the major tools available in identifying the form of time-series models.
4. The lag between two points in a time-series is the number of time-periods between the two points.
5. Autoregressive Models are a class of models which relate the current value of a variable to past values of the variable.
6. Moving-Average Models are a class of models which relate the current value of a variable to past values of the random error terms.
7. Differences are the difference between a current value of a series and its value one month or one year previous.
8. The Backshift Operator B takes a time series Z_t and shifts all values back one time unit to produce a new time series: $B(Z_t) = Z_{t-1}$.
9. Seasonal Models are a class of time-series models with a seasonal component.

STATIONARITY

The first technical concept which we need to discuss is the notion of Stationarity. For our general discussion we will let the term Z_t denote the observed series at time t. Time units could be in months, quarters, or years; but we do assume that the series is observed at equally spaced time units.

The most important aspect of time series models is that we do not assume that the observations at different time points are statistically independent. It is precisely this dependence which we wish to capture. In our models, although we do not assume independence, it is important that the time series have some degree of statistical stability. We say that the series Z_t is stationary if the following two conditions hold:

1. The mean value of Z_t is the same for all times t.
2. The correlation between a series value at time t (Z_t) and a series value at time t-k (Z_{t-k}) depends only on the time lag k and not on t.

Figure 8 illustrates characteristics of stationary and non-stationary series. Particularly simple stationary models which have been found to be useful in practice are the autoregressive and moving-average models.

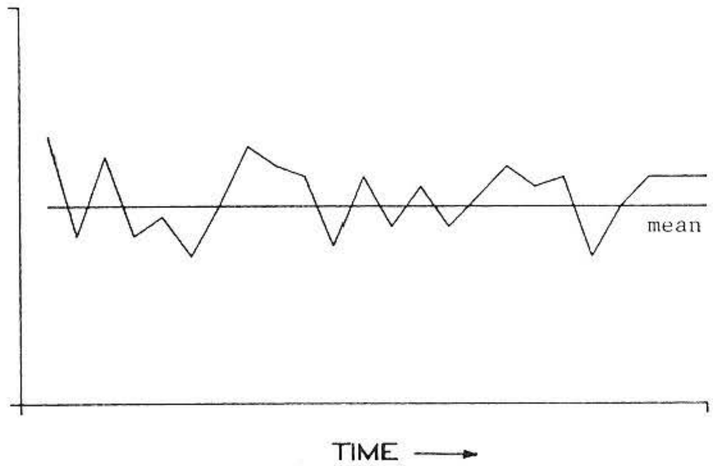
THE AUTOCORRELATION FUNCTION

The autocorrelation function (ACF) is one of the primary tools available to aid the analyst in identifying time series models. The form of the ACF provides clues regarding the nature and form of the models. For a stationary time-series, the autocorrelation function at lag k is defined as the correlation between Z_t and Z_{t-k}. The values for the ACF will always lie between -1 and +1 and indicate the strength of linear dependence between series values which are k units apart in time.

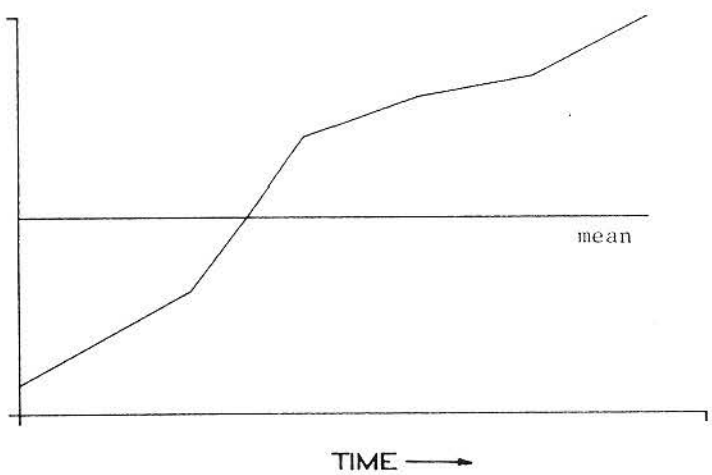
FIGURE 8

CONDITIONS FOR STATIONARITY

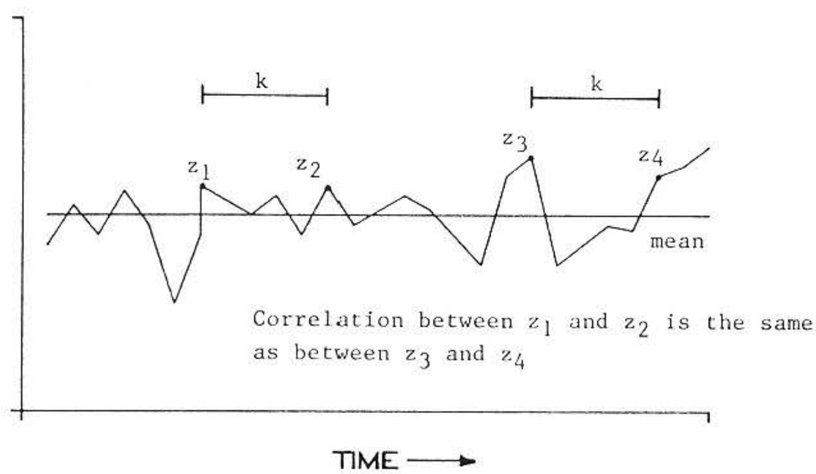
(a) A stationary time-series varying about a mean value



(b) A non-stationary time-series



(c) Correlation depends only on lag k, not on t



AUTOREGRESSIVE MODELS

An analogy with ordinary regression models can be made if we consider the relationship:

$$(7) \qquad Z_t = \phi Z_{t-1} + a_t \ .$$

Here we are regressing the series at time t on itself (hence auto) but at time t-1. For example, with monthly series we are saying that April's value is a ϕ proportion of March's value plus, of course, an "error term" a_t which does not depend on past values. Such a model is called an autoregressive model of order 1. Such a model forces Z_t to "depend on" Z_{t-1} and does capture a certain kind of dependence.

It may be shown that such a model will be stationary if and only if $-1<\phi<1$. Also if ϕ is so restricted the autocorrelation function ρ_k is:

$$(8) \qquad \rho_k = \phi^k, \ k = 1, 2, 3, \ldots$$

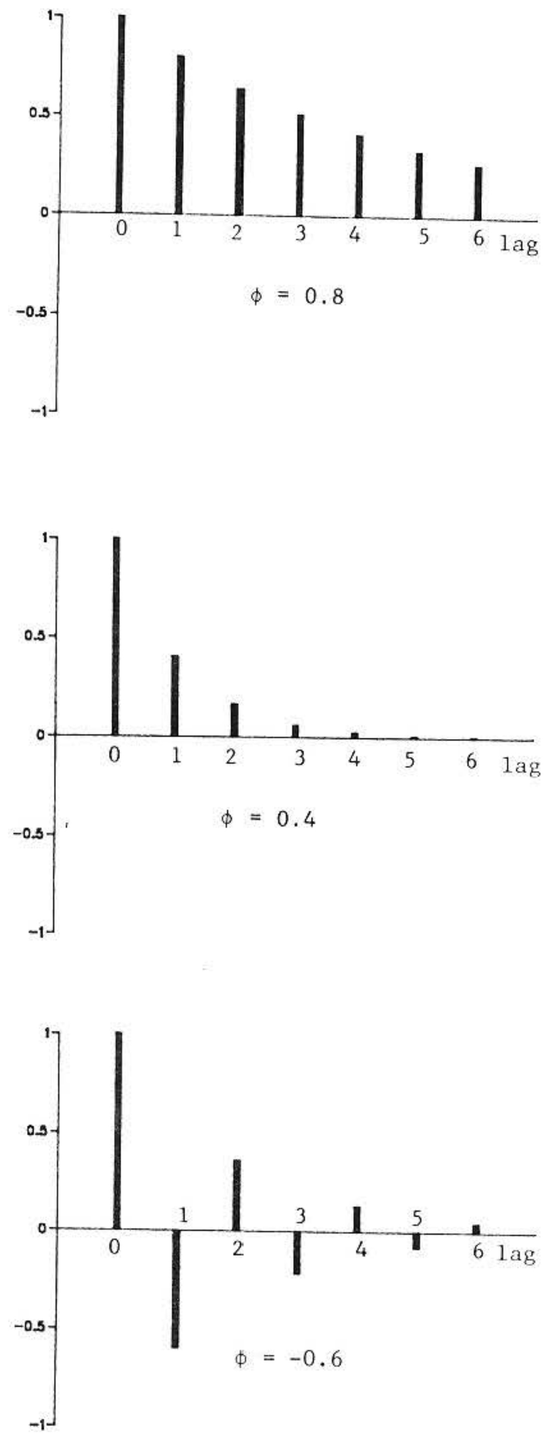
The dependence dies out in an exponential fashion as the lag k increases. The closer ϕ is to ± 1 the slower the decay. If ϕ is near zero, the decay will be quite rapid. Note that if ϕ is positive all autocorrelation values are positive while if ϕ is negative then ρ_k is negative for odd lags k and positive for even lags k. Typical shapes for ρ_k are given in Figure 9.

Exhibit 1(c), page 38 shows the sample autocorrelation function for the monthly changes in gasoline prices in Portland, Oregon over the period 1971-1972. This is an estimated or sample autocorrelation function (SACF) but its general shape shows the tendency for exponential decay which is characteristic of the first order autoregressive, or AR(1). Thus we can see the ACF has definite patterns or shapes that relate directly to a specific class of models.

FIGURE 9

AUTOCORRELATION FUNCTIONS

FOR SEVERAL AR(1) MODELS



More complex autocorrelation patterns apply to higher order autoregressive models. The second order autoregressive model is given by

$$(9) \qquad Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t \quad .$$

Here the current value Z_t depends linearly on the previous two values Z_{t-1} and Z_{t-2} plus a random error a_t.

Explicit formulas for the autocorrelation function in the second order case are more difficult to express. However under certain conditions on the parameters ϕ_1 and ϕ_2, ρ_k will follow a damped sine curve. Some illustrative shapes are given in Figure 10.

Clearly there is no difficulty in now defining an autoregressive model of order p:

$$(10) \qquad Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \dots + \phi_p Z_{t-p} + a_t \quad .$$

We say that Z_t follows an AR(p) model.

MOVING-AVERAGE MODELS

A second type of time series model relates the current value Z_t to the random error terms present and past a_t, a_{t-1}, a_{t-2}, ... instead of the past values of Z.

A moving average model of order one or MA(1) for Z_t is given by

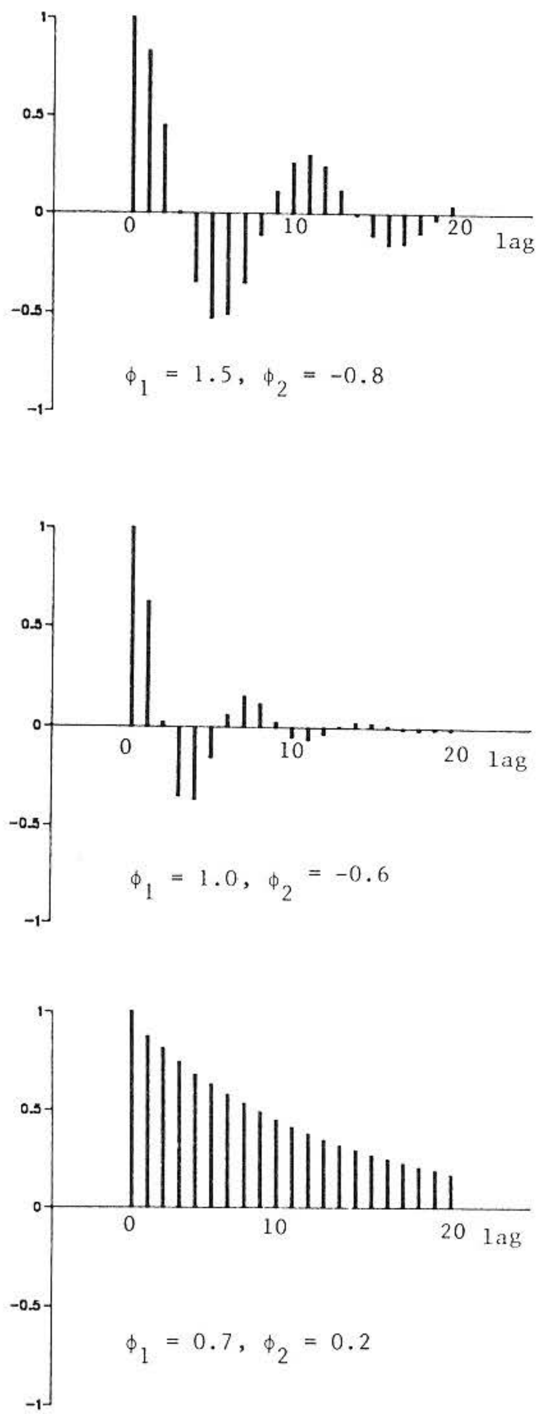
$$(11) \qquad Z_t = a_t - \theta_1 a_{t-1}$$

(The negative sign on θ_1 is purely conventional.)

FIGURE 10

AUTOCORRELATION FUNCTIONS

FOR SEVERAL AR(2) MODELS



For this model the autocorrelation function is non-zero only at lag 1. That is, for an MA(1) process, there is no correlation between any values that are greater than one lag apart.

A general qth order moving average process satisfies

$$(12) \qquad Z_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-1} - \cdots - \theta_q a_{t-q} \quad .$$

For the MA(q) model we can show that ρ_1, ρ_2, . . ., p_q are non-zero but for all lags k greater than q, $\rho_k = 0$. The dependence in the series extends only to q time lags.

MIXED AUTOREGRESSIVE MOVING AVERAGE MODELS

Quite general time series models may be formed by putting the two concepts of autoregressive and moving average together into one model. For example:

$$(13) \qquad Z_t = \phi_1 Z_{t-1} + a_t - \theta_1 a_{t-1} \quad .$$

We call this an ARMA(1,1) model. This model is stationary if $-1<\phi_1<1$. The autocorrelation function (ρ_k) exhibits exponential decay but, unlike the AR(1) case, from an "initial value" of ρ_1 rather than $\rho_0=1$.

A general ARMA (p,q) model is given by

$$(14) \quad Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \cdots - \theta_q a_{t-q} .$$

Fortunately, for real applications, p and q will usually be quite small, typically less than 4 or 5.

NON-STATIONARY ARIMA MODELS

Many, if not most, time series encountered in practice fail to appear stationary. They tend to show substantial growth or decay over time rather than stable statistical fluctuations around a fixed mean level. Fortunately many such series can be transformed to stationary series by differencing. For a series Z_t the first difference, ΔZ_t , is defined by

$$(15) \qquad \Delta Z_t = Z_t - Z_{t-1}.$$

That is, ΔZ_t is the series of changes in Z_t over successive times. Figure 6 shows a plot of the average monthly gasoline price in Portland, Oregon over the period 1971-1982. The growth in the series over that period is not compatible with the assumption of a stationary series. However, Figure 11 shows the first difference of the gasoline prices over the same period. Here an assumption of stationarity is quite tenable.

In some cases a second difference will need to be taken to achieve stationarity. The second difference of Z_t, denoted $\Delta^2 Z_t$, is just

$$(16) \qquad \begin{aligned} \Delta^2 Z_t &= \Delta(\Delta Z_t) = \Delta(Z_t - Z_{t-1}) \\ &= Z_t - 2Z_{t-1} + Z_{t-2}. \end{aligned}$$

Most (nonseasonal) "real world" time series can be transformed to stationarity with one, or at most two, differences.

Another aspect of nonstationarity may be reflected in the common occurrance of series whose magnitude of variability is directly related to the level of the series itself - the higher the level of the series, the larger the variability of the series around that level. The Portland Transit Ridership Series in Figure 3 illustrates this point. Under such circumstances, the series is usually transformed by taking logarithms of the values. Figure 12 shows the logs of the ridership series. Notice that now the variability is roughly the same for low ridership periods and high ridership periods. We would now consider differencing to attempt to make the series stationary.

Once we have transformed a series to stationarity either with differencing and/or logarithms we may assume that the transformed series may be modelled by a parsimonious ARMA (p,q) model. A model which after suitable differencing satisfies an ARMA model is called an integrated autoregressive moving average model. If d differences are necessary to achieve stationarity we say that we have an ARIMA (p,d,q) model.

As an example, consider an ARIMA (0,1,1) model. This says that the first difference $Z_t - Z_{t-1}$ satisfies an MA(1) model. So

(17)
$$Z_t - Z_{t-1} = a_t - \theta_1 a_{t-1}$$
or
$$Z_t = Z_{t-1} + a_t - \theta_1 a_{t-1} \quad .$$

An ARIMA (1,1,0) model satisfies

(18)
$$Z_t - Z_{t-1} = \phi(Z_{t-1} - Z_{t-2}) + a_t$$
or
$$Z_t = (1+\phi)Z_{t-1} - \phi Z_{t-2} + a_t \quad .$$

ARIMA (p,d,q) models form a large flexible class of time series models but they do not cover models which exhibit seasonal characteristics. Before our discussion of seasonal models it is convenient to introduce a shorthand notation.

FIGURE 11

GASOLINE PRICE SERIES
FIRST DIFFERENCE

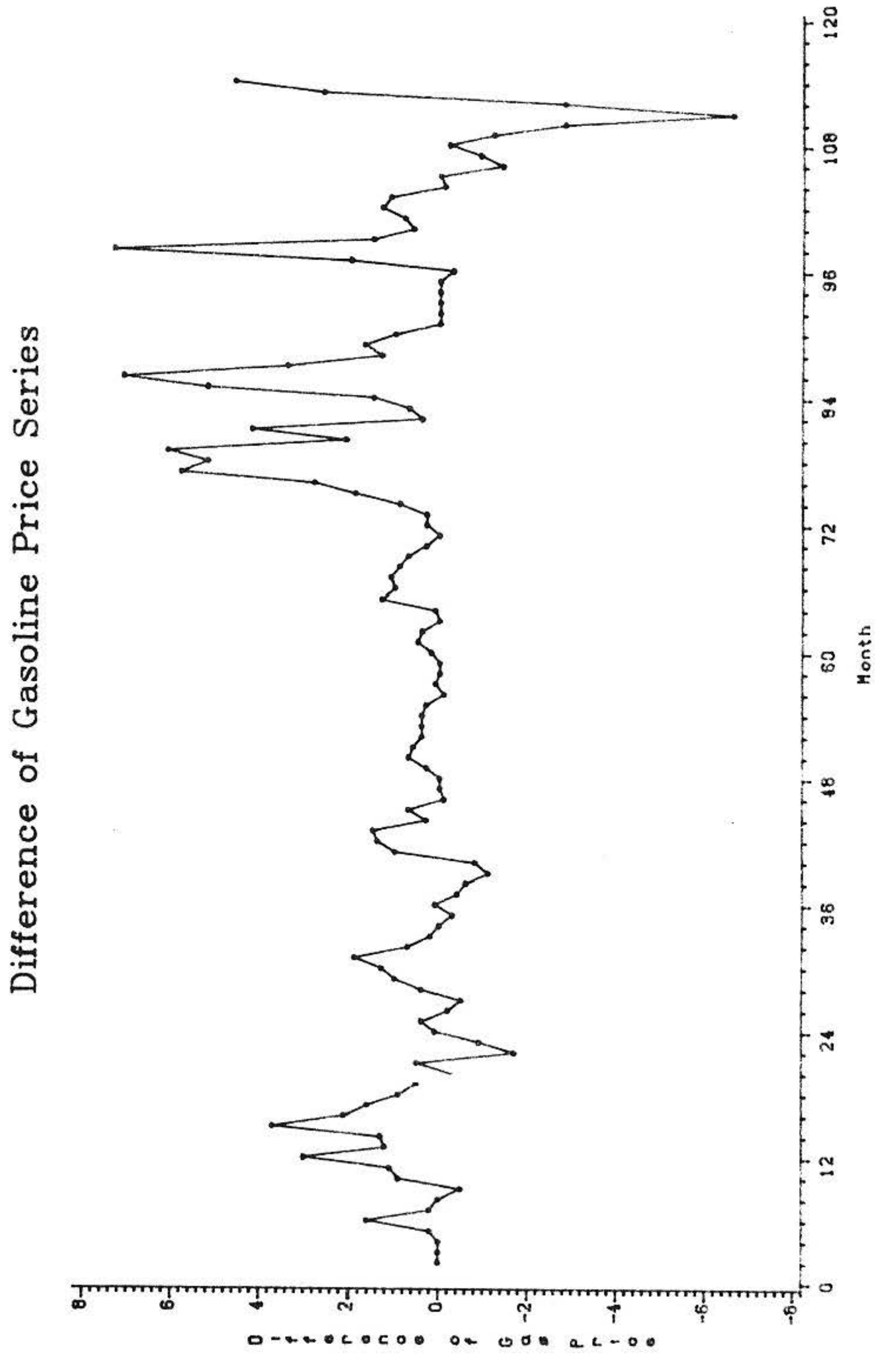
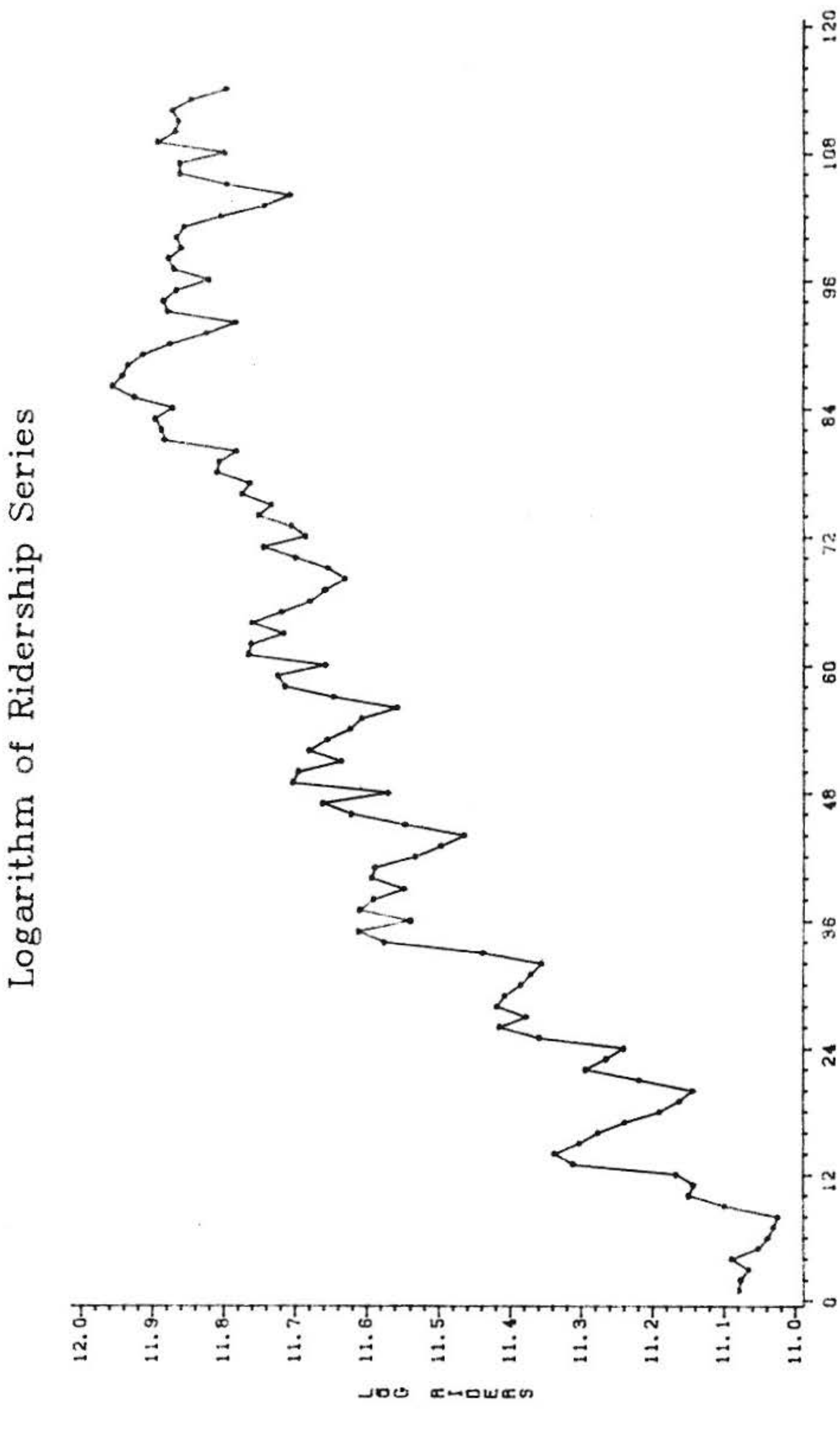


FIGURE 12

LOGARITHMS OF RIDERSHIP SERIES



THE BACKSHIFT OPERATOR

ARIMA models may be expressed very compactly in terms of the backshift operator B. The backshift operator B takes a time series Z_t and shifts time back one time unit to produce a new series. In particular

(19) $$B(Z_t) = Z_{t-1} \; .$$

Since $B(Z_t)$ is a new series, we could use B once more to obtain

(20) $$B(B(Z_t)) = B(Z_{t-1}) = Z_{t-2}$$

and we write

(21) $$B^2(Z_t) = Z_{t-2} \; .$$

Similarly then

(22) $$B^k(Z_t) = Z_{t-k} \; .$$

Using the backshift operator, the autoregressive and moving average models described earlier may be more compactly written.

MA(2) MODEL

(23) $$\begin{aligned} Z_t &= a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \\ &= a_t - \theta_1 B(a_t) - \theta_2 B^2(a_t) \\ &= (1 - \theta_1 B - \theta_2 B^2) a_t \end{aligned}$$

AR(2) Model

(24)
$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + a_t$$
$$Z_t - \phi_1 BZ_t - \phi_2 B^2 Z_t = a_t$$
$$(1 - \phi_1 B - \phi_2 B^2) Z_t = a_t$$

Any ARMA(p, q) model may then be expressed as

(25)
$$\phi(B)Z_t = \theta(B)a_t$$

where

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

and

$$\theta(B) = 1-\theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q \quad .$$

Differences may also be written as

(26)
$$\Delta Z_t = Z_t - Z_{t-1} = Z_t - B(Z_t) = (1-B)Z_t$$

and the dth order difference of Z_t is then $(1-B)^d(Z_t)$. Therefore the general ARIMA (p,d,q) model may be expressed as

(27)
$$\phi(B)(1-B)^d Z_t = \theta(B)a_t$$

The polynomials $\phi(B)$ and $\theta(B)$ are called the autoregressive and moving average characteristic polynomials.

SEASONAL MODELS

Seasonal behavior is very common in time series measured on a monthly or quarterly basis. The Portland Transit Ridership Series in Figure 3 shows a rather strong seasonal tendency - high during the winter months and relatively lower during the summer months. This is not a completely regular or fixed pattern, but only a general tendency which undergoes continual change over time.

Can ARIMA type models account for such seasonal effects? Consider the simple model

(28)	$Z_t = a_t - \Theta_1 a_{t-12}$.

Such a series will have non-zero correlation only at lag 12. That is, only values that are twelve months apart will be correlated. The correlation at all other lags will be zero. This model is known as a seasonal MA(1) model, with period 12.

A seasonal AR(1) model with period 12 is similarly defined as

(29)	$Z_t = \Phi_1 Z_{t-12} + a_t$

The only lags with non-zero correlation are those that are multiples of 12: 12, 24, 36, 48, etc.

Most realistic models will involve both seasonal and short-term non-seasonal effects such as

(30)	$Z_t = \Phi_1 Z_{t-12} + a_t - \theta_1 a_{t-1}$.

In backshift notation this is expressed as

(31)	$(1-\Phi_1 B^{12})Z_t = (1-\theta_1 B)a_t$.

In addition, seasonal differencing

$$(32) \qquad \Delta_{12}Z_t = Z_t - Z_{t-12} = (1-B^{12})Z^t \ ,$$

will frequently be employed to obtain a stationary series.

A model which we will find to be quite satisfactory for the Portland transit ridership series is given by

$$(33) \qquad (1-B)(1-B^{12})Z_t = (1-\Theta_1 B^{12} - \Theta_2 B^{24})a_t$$

where $\Theta_1 = .43$ and $\Theta_2 = .20$. The transit ridership series required both regular and seasonal differencing to induce stationarity. Correlation twelve and twenty-four months back was accounted for with the a_{t-12} and a_{t-24} terms.

A completely general multiplicative Seasonal ARIMA $(p,d,q) \times (P,D,Q)_s$ model with period s may be expressed as

$$(34) \qquad \phi(B)\Phi(B)(1-B)^d(1-B^s)^D Z_t = \theta(B)\Theta(B)a_t$$

where

d is the order of non-seasonal differencing

D is the order of seasonal differencing

$\phi(B) = 1-\phi_1 B- \dots - \phi_p B^p$ is the non seasonal autoregression characteristic polynomial

$\Phi(B) = 1-\Phi_1 B^s - \dots - \Phi_p B^{sP}$ is the seasonal autoregresive characteristic polynomial

$\theta(B) = 1-\theta_1 B - \dots - \theta_p B^q$ is the non-seasonal moving average characteristic polynomial

$\Theta(B) = 1 - \Theta_1 B^2 - \dots - \Theta_Q B^{sQ}$ is the seasonal moving average characteristic polynomial.

Once more in practical applications p,d,q,P,D and Q will all be small--

typically 2 or less.

MODEL BUILDING

We have now a large, flexible class of parametric models for both stationary and nonstationary series which can account for short-term nonseasonal dependence and also seasonality. Our model building task may be conveniently broken into four steps:

Step 1.	Choose appropriate orders p, d, and q for the autoregressive, differencing, and moving-average components of the model. (P, D, and Q also if we are considering a seasonal series.

Step 2.	Efficiently estimate the parameters (ϕ's, θ's) for the model selected in step 1.

Step 3.	Criticize the model estimated in step 2 to check its appropriateness.

Step 4.	Is the Model Adequate? If yes, this task is completed. If not, Step 1 should be repeated.

If the model appears inadequate in some way, we use the nature of the inadequacy to hypothesize an alternative model and proceed to estimate that new model and check it for adequacy. With a few iterations of this "model building strategy" we hope to arrive at the best model for a given series.

Step 1. Identification of the Model. The major tool for analyzing an observed time-series and identifying the nature of the model, is the sample autocorrelation function (SACF). The SACF is an estimate of the theoretical autocorrelation function (ρ_k). On the basis of the SACF, we look for patterns which are characteristic of known patterns in ρ_k for common ARMA models. For example, we know that $\rho_k = 0$ for $k > q$ in an MA(q) model so that if the SACF is close to zero for $k > 2$, say, then an MA(2) model is indicated.

Our first task however, is to decide on d, the number of differences, if any, needed to achieve stationarity and to decide if logarithms should be taken. An inspection of the plot of the series versus time should be made. Figure 6 shows the monthly gasoline prices in Portland for the period 1971 to 1982. Figure 13 gives the logarithms of the same series. The logged series seems to have better stability with respect to variability but could not be assumed stationary. Figure 14 plots the first difference of the logarithmic gas price series. This series could reasonably be assumed stationary. Additional indication of stationarity is given by computing the SACF of the logged series and of its first difference. Exhibit 1(b) shows the SACF for the logged gas prices. Note that it dies out very slowly. This is a strong indication of a nonstationary series. Exhibit 1(c) shows the SACF after we have differenced the series. We now see the nice exponential decay characteristic of an AR(1) model. Apparently, the logarithms of gasoline prices may be modeled as an ARIMA (1, 1, 0) series.

Another useful tool for identifying the order of AR(p) processes is the partial autocorrelation function. The <u>partial autocorrelation function</u>, ϕ_{kk}, at lag k, measures the correlation between z_t and z_{t+k} after removing the effect of the intervening variables $z_{t+1}, z_{t+2}, \cdots, z_{t+k-1}$. The theory for ϕ_{kk} is somewhat difficult and we will only note that it may be shown that for an AR(p) process ϕ_{kk} will be zero for $k > p$. The lag k just before ϕ_{kk} drops to zero will then indicate the order of the AR model. Exhibit 1(d) shows the estimated partial autocorrelation function for the first difference of the logged gasoline price series. Note that it is nearly zero except for lag 1 reinforcing our earlier selection of $p = 1$.

The plots of both the sample autocorrelation and partial autocorrelation functions give marks at "two standard errors". Based on large sample size theory, these marks allow us to easily see the "significance" of the correlations. If a sample correlation is outside these marks, we would reject the hypothesis of zero correlation at that particular lag, and assume that there is significant correlation at that lag.

FIGURE 13

LOGARITHMS OF GASOLINE PRICE SERIES

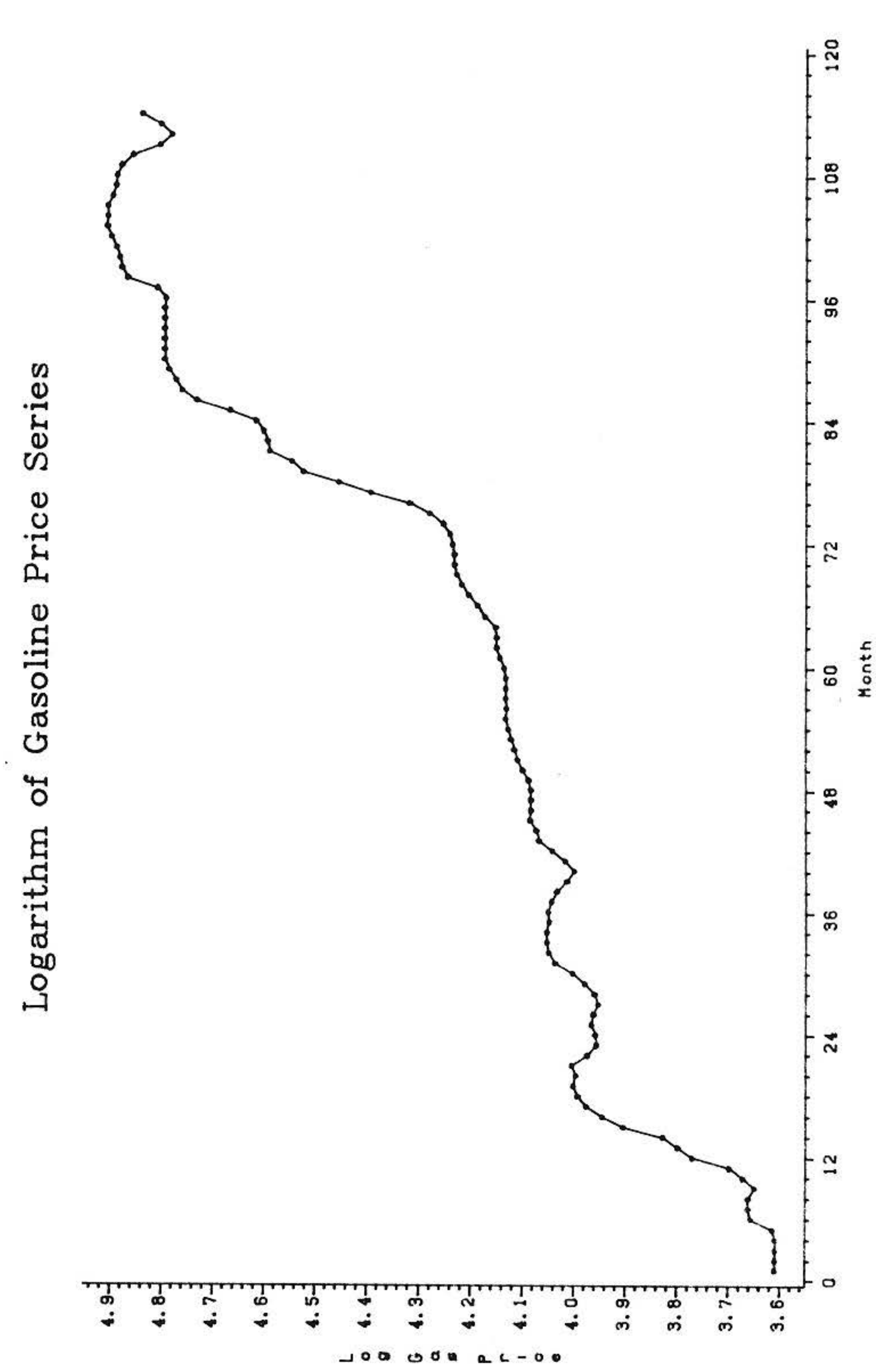


FIGURE 14

LOGARITHMS OF GASOLINE PRICE SERIES
FIRST DIFFERENCES

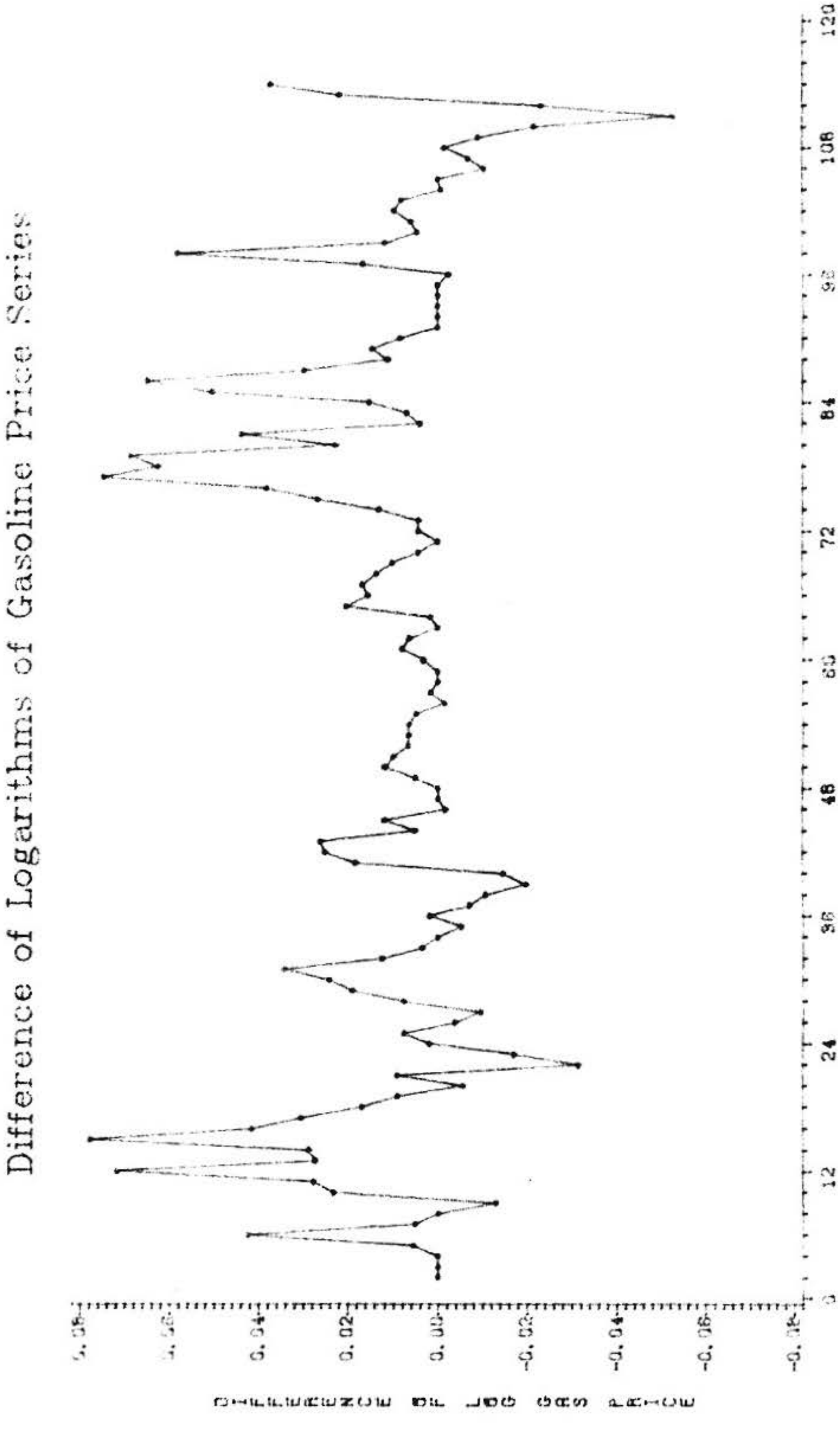


EXHIBIT 1

MODEL IDENTIFICATION FOR GASOLINE PRICE SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]

DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
     IDENTIFY VAR=LGAS;
     IDENTIFY VAR=LGAS(1);
```

1 List of input variables:
RIDERS = transit ridership
HOURS = platform hours
EMPLOY = employment
GAS = gasoline price
FARE = transit fare

2 Data transformation:
logarithms are taken of all input data.

3 IDENTIFY statement computes the simple autocorrelation functions to assist in the model identification for the gas price series. Here the SACF and SPACF are requested for both the lag series and the first difference of the lag series.

(b) Autocorrelation function, undifferenced series

```
NAME OF VARIABLE        = LGAS

MEAN OF WORKING SERIES=     4.2624
STANDARD DEVIATION    =   0.394877
NUMBER OF OBSERVATIONS=        114

AUTOCORRELATIONS
```

4 The slow linear decay of the SACF indicates that this series is non-stationary.

```
LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      STD
  0   0.155928     1.00000 |                    |********************|         0
  1   0.152313     0.97682 |                .   |********************| 0.0936586
  2    0.14849     0.95230 |              .     |******************* |  0.159724
  3     0.1445     0.92671 |            .       |******************* |  0.203524
  4   0.140199     0.89913 |          .         |******************  |  0.237673
  5   0.135482     0.86888 |         .          |*****************   |  0.265841
  6   0.130747     0.83851 |        .           |*****************   |  0.289683
  7    0.12587     0.80723 |        .           |****************    |  0.310244
  8   0.120888     0.77528 |       .            |****************    |  0.328151
  9   0.115697     0.74199 |      .             |***************     |  0.343843
 10   0.110452     0.70836 |      .             |**************      |  0.357613
 11   0.105251     0.67500 |     .              |************* .     |  0.369716
 12   0.100332     0.64345 |     .              |************* .     |  0.380372
 13  0.0954785     0.61233 |    .               |************   .    |  0.389804
 14  0.0907297     0.58187 |    .               |************   .    |  0.398152
 15  0.0863691     0.55390 |    .               |***********    .    |  0.405542
 16  0.0821727     0.52699 |    .               |***********    .    |  0.412125
 17  0.0781007     0.50088 |   .                |**********      .   |  0.417995
 18  0.0743094     0.47656 |   .                |**********      .   |  0.423227
 19  0.0705535     0.45248 |   .                |*********       .   |  0.427908
 20  0.0666428     0.42740 |   .                |*********       .   |  0.432085
 21  0.0626967     0.40209 |   .                |********        .   |  0.435777
 22  0.0585039     0.37520 |  .                 |********         .  |   0.43902
 23  0.0541732     0.34743 |  .                 |*******          .  |  0.441824
 24   0.049821     0.31951 |  .                 |******           .  |  0.444214
                           '.' MARKS TWO STANDARD ERRORS
```

(c) Autocorrelation function, one difference

```
NAME OF VARIABLE        = LGAS
PERIODS OF DIFFERENCING= 1.

MEAN OF WORKING SERIES= 0.0108889
STANDARD DEVIATION     = 0.0212049
NUMBER OF OBSERVATIONS=        113

                     AUTOCORRELATIONS

LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      STD
  0 .000449649   1.00000   |                    |********************|        0
  1 .00028238 9  0.62802   |                .   |*************       | 0.094072
  2 .000164635   0.36614   |               .    |*******             | 0.125818
  3 .000075567   0.16806   |               .    |*** .               | 0.134919
  4 .000059131   0.13150   |               .    |*** .               | 0.136759
  5 0.00002937   0.06532   |              .     |*    .              | 0.137873
  6  2.908E-06   0.00647   |              .     |     .              | 0.138147
  7 -2.730E-05  -0.06072   |              .    *|     .              | 0.138149
  8 -2.789E-05  -0.06202   |              .    *|     .              | 0.138385
  9 .000007262   0.01615   |              .     |     .              | 0.138631
 10 -1.042E-05  -0.02317   |              .     |     .              | 0.138648
 11  4.557E-06   0.01013   |              .     |     .              | 0.138682
 12 .000014798   0.03291   |              .     |*    .              | 0.138688
 13 -5.116E-06  -0.01138   |              .     |     .              | 0.138758
 14 -2.107E-05  -0.04687   |              .    *|     .              | 0.138766
 15 -7.810E-06  -0.01737   |              .     |     .              | 0.138906
 16  7.000E-06   0.01557   |              .     |     .              | 0.138925
 17 .000012502   0.02780   |              .     |*    .              | 0.138941
 18  6.432E-06   0.01430   |              .     |     .              |  0.13899
 19 .000014749   0.03280   |              .     |*    .              | 0.139003
 20 -8.221E-06  -0.01828   |              .     |     .              | 0.139071
 21 -3.426E-05  -0.07619   |              .   **|     .              | 0.139093
 22 -.00006348  -0.14118   |              .  ***|     .              | 0.139461
 23 -7.180E-05  -0.15969   |              .  ***|     .              |  0.14072
 24 -6.518E-05  -0.14495   |              .  ***|     .              | 0.142315
                           '.' MARKS TWO STANDARD ERRORS
```

5 The SACF dies out fairly rapidly (after two lags) indicating that the first differenced series is stationary. The pattern also indicates that an AR model would probably fit the data.

(d) Partial autocorrelation function, one difference

```
                         PARTIAL AUTOCORRELATIONS

LAG CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1    0.62802   |                 .  |*************      |
  2   -0.04668   |                 . *|  .                |
  3   -0.07166   |                 . *|  .                |
  4    0.11053   |                 .  |**.                |
  5   -0.06483   |                 . *|  .                |
  6   -0.04623   |                 . *|  .                |
  7   -0.04979   |                 . *|  .                |
  8    0.01746   |                 .  |  .                |
  9    0.11316   |                 .  |**.                |
 10   -0.13791   |                .***|  .                |
 11    0.09885   |                 .  |**.                |
 12    0.04244   |                 .  |* .                |
 13   -0.14785   |                .***|  .                |
 14    0.01364   |                 .  |  .                |
 15    0.06858   |                 .  |* .                |
 16    0.00951   |                 .  |  .                |
 17   -0.00283   |                 .  |  .                |
 18   -0.02568   |                 . *|  .                |
 19    0.10238   |                 .  |**.                |
 20   -0.15526   |                .***|  .                |
 21   -0.10210   |                 .**|  .                |
 22    0.00852   |                 .  |  .                |
 23   -0.06315   |                 . *|  .                |
 24   -0.02338   |                 .  |  .                |
```

6 The SPACF is non-zero only at lag 1, confirming an AR model. A model of order 1 is suggested.

As a second example of identifying p, d and q consider the Portland Transit Fare series plotted in Figure 5. Again we will take logs before any further analysis. Exhibit 2 displays the sample ACF and PACF for the logged fare series and also for the first difference of the logged fare series. The autocorrelations with no differencing indicate nonstationarity. After differencing virtually no autocorrelation is observed. Thus, a model with $p = q = 0$ and $d = 1$ appears to be adequate.

We turn now to an example involving seasonality. Most ridership series measured monthly or quarterly will show some seasonality. We look specifically at the logged monthly Portland bus ridership series of Figure 12. Anticipating nonstationarity and seasonal nonstationarity we compute the sample ACF and PACF out to 36 lags for the original series but also for the first differenced series and for the series differenced twice-one ordinary difference and one seasonal difference. The results are given in Exhibit 3. Results for the original series reinforce our prior belief that the series is nonstationary. Exhibit 3(c) shows the strong seasonal autocorrelations which remains after one ordinary difference has been taken. An additional seasonal difference produces the autocorrelations in Exhibit 3(d) where now minor correlations remain at lags 11 and 12. A model with $d = 1$, $D = 1$ and $Q = 1$ appears tenable for the ridership series.

EXHIBIT 2

MODEL IDENTIFICATION FOR FARE SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]

DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
    IDENTIFY VAR=LFARE;
    IDENTIFY VAR=LFARE(1);
```

> 1 The IDENTIFY statement computes SACF and SPACF of the Fare series to assist in the model identification.

(b) Autocorrelation function, undifferenced series

```
                        NAME OF VARIABLE       = LFARE

                        MEAN OF WORKING SERIES=   3.55483
                        STANDARD DEVIATION    =  0.192958
                        NUMBER OF OBSERVATIONS=       114

                                AUTOCORRELATIONS

LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1       STD
  0  0.0372329    1.00000 |                    |********************|         0
  1  0.0362028    0.97233 |                .   |******************* | 0.0936586
  2  0.0351727    0.94467 |              .     |******************* |  0.159243
  3  0.0341425    0.91700 |            .       |******************  |   0.20252
  4  0.0331124    0.88933 |           .        |******************  |   0.23615
  5  0.0321254    0.86282 |         .          |*****************   |  0.263899
  6  0.0311385    0.83632 |        .           |*****************   |  0.287582
  7  0.0300068    0.80592 |        .           |****************    |  0.308178
  8  0.0288751    0.77552 |       .            |****************    |  0.326142
  9  0.0276722    0.74322 |      .             |***************     |  0.341936
 10  0.0264736    0.71103 |      .             |**************      |  0.355824
 11   0.025275    0.67883 |     .              |**************.     |  0.368077
 12  0.0240575    0.64613 |     .              |************* .     |    0.3789
 13    0.02284    0.61343 |    .               |************   .    |  0.388445
 14  0.0215957    0.58002 |    .               |************   .    |  0.396851
 15  0.0203654    0.54697 |    .               |***********    .    |  0.404219
 16  0.0190581    0.51186 |    .               |**********     .    |   0.41066
 17  0.0177463    0.47663 |   .                |**********      .   |  0.416219
 18  0.0164166    0.44092 |   .                |*********       .   |   0.42098
 19  0.0151652    0.40731 |   .                |********        .   |  0.425011
 20  0.0136954    0.36783 |   .                |*******         .   |  0.428422
 21  0.0122257    0.32836 |   .                |*******         .   |  0.431183
 22   0.011191    0.30057 |   .                |******          .   |  0.433371
 23  0.0101756    0.27330 |   .                |*****           .   |  0.435196
 24 0.00916024    0.24603 |   .                |*****           .   |  0.436699
                           '.' MARKS TWO STANDARD ERRORS
```

> 2 The slow linear decay of the SACF indicates that this series is non-stationary.

(c) Autocorrelation function, one difference

```
NAME OF VARIABLE        =  LFARE
PERIODS OF DIFFERENCING=  1.

MEAN OF WORKING SERIES=0.00326603
STANDARD DEVIATION     = 0.0308169
NUMBER OF OBSERVATIONS=        113

                      AUTOCORRELATIONS

LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1        STD
  0 .000949683    1.00000  |                    |*******************|          0
  1 -1.076E-05   -0.01133  |               .   |   .               |   0.094072
  2 -1.086E-05   -0.01143  |               .   |   .               |  0.0940841
  3 -.00001095   -0.01153  |               .   |   .               |  0.0940964
  4 -5.459E-05   -0.05748  |               .  *|   .               |  0.0941089
  5 -1.114E-05   -0.01173  |               .   |   .               |  0.0944191
  6 .000186047    0.19590  |               .   |****               |   0.094432
  7 -1.086E-05   -0.01144  |               .   |   .               |  0.0979626
  8 .000060793    0.06401  |               .   |*  .               |  0.0979744
  9 -2.435E-05   -0.02564  |               .  *|   .               |  0.0983438
 10 -.00001287   -0.01355  |               .   |   .               |  0.0984029
 11  6.101E-06    0.00642  |               .   |   .               |  0.0984195
 12 -1.306E-05   -0.01375  |               .   |   .               |  0.0984232
 13 .000013823    0.01456  |               .   |   .               |  0.0984402
 14 -2.731E-05   -0.02876  |               .  *|   .               |  0.0984592
 15 .000065742    0.06922  |               .   |*  .               |  0.0985335
 16 -8.662E-06   -0.00912  |               .   |   .               |   0.098963
 17  4.742E-06    0.00499  |               .   |   .               |  0.0989704
 18 -9.225E-05   -0.09714  |               . **|   .               |  0.0989727
 19 .000206787    0.21774  |               .   |****               |  0.0998128
 20 -1.355E-05   -0.01427  |               .   |   .               |   0.103931
 21 .000012814    0.01349  |               .   |   .               |   0.103949
 22 -2.894E-05   -0.03048  |               .  *|   .               |   0.103964
 23 -9.591E-06   -0.01010  |               .   |   .               |   0.104043
 24 .000024384    0.02568  |               .   |*  .               |   0.104052
                           '.' MARKS TWO STANDARD ERRORS
```

3 The SACF is non-zero only at lag zero, indicating that the first difference of the fare series is stationary.

(d) Partial autocorrelation function, one difference

```
PARTIAL AUTOCORRELATIONS

LAG CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1    -0.01133  |               .   |   .               |
  2    -0.01156  |               .   |   .               |
  3    -0.01180  |               .   |   .               |
  4    -0.05790  |               .  *|   .               |
  5    -0.01343  |               .   |   .               |
  6     0.19484  |               .   |****               |
  7    -0.00892  |               .   |   .               |
  8     0.06559  |               .   |*  .               |
  9    -0.02267  |               .   |   .               |
 10     0.00925  |               .   |   .               |
 11     0.01079  |               .   |   .               |
 12    -0.04758  |               .  *|   .               |
 13     0.01807  |               .   |   .               |
 14    -0.05837  |               .  *|   .               |
 15     0.08494  |               .   |** .               |
 16    -0.01778  |               .   |   .               |
 17     0.00827  |               .   |   .               |
 18    -0.09439  |               . **|   .               |
 19     0.23204  |               .   |*****              |
 20    -0.00085  |               .   |   .               |
 21    -0.01408  |               .   |   .               |
 22    -0.03316  |               .  *|   .               |
 23     0.00151  |               .   |   .               |
 24     0.08232  |               .   |** .               |
```

> 4 The SPACF is non-zero at all lags indicating that the first difference of the fare series is simply white noise.

EXHIBIT 3

MODEL IDENTIFICATION FOR THE RIDERSHIP SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
```

[Data]

```
DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
     IDENTIFY VAR=LRIDERS NLAG=36;
     IDENTIFY VAR=LRIDERS(1) NLAG=36;
     IDENTIFY VAR=LRIDERS(1,12) NLAG=36;
```

1 The IDENTIFY statement computes the SACF and SPACF for three versions of the Rider series: the original series, the series with one difference, and the series with one regular difference and with one seasonal difference of period 12.

(b) Autocorrelation function, undifferenced series

```
NAME OF VARIABLE         =  LRIDERS

MEAN OF WORKING SERIES=   11.5939
STANDARD DEVIATION    =  0.264912
NUMBER OF OBSERVATIONS=       114
```

> 2 The slow linear decay of the SACF indicates that this series is non-stationary.

```
                              AUTOCORRELATIONS

LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      STD
  0  0.0701782     1.00000 |                    |********************|         0
  1  0.0675337     0.96232 |                .   |******************* | 0.0936586
  2  0.0647607     0.92280 |              .     |******************  |  0.158172
  3  0.0624568     0.88997 |            .       |******************  |  0.199896
  4  0.0600704     0.85597 |           .        |*****************   |  0.232065
  5  0.0574136     0.81811 |          .         |****************    |  0.258279
  6  0.0549172     0.78254 |         .          |****************    |   0.28009
  7  0.0528261     0.75274 |        .           |***************     |  0.298653
  8  0.0507576     0.72327 |       .            |**************      |  0.314856
  9  0.0490081     0.69834 |       .            |**************      |  0.329107
 10  0.0474532     0.67618 |      .             |**************      |  0.341859
 11  0.0465316     0.66305 |      .             |*************.      |  0.353396
 12  0.0458357     0.65313 |     .              |************* .     |  0.364145
 13  0.0438539     0.62489 |     .              |************  .     |   0.37428
 14  0.0418506     0.59635 |     .              |************  .     |  0.383323
 15  0.0401426     0.57201 |    .               |***********    .    |  0.391376
 16  0.0381593     0.54375 |    .               |***********    .    |  0.398642
 17  0.0359825     0.51273 |    .               |**********     .    |  0.405096
 18  0.0338178     0.48188 |    .               |**********     .    |  0.410749
 19  0.0319164     0.45479 |   .                |*********       .   |  0.415679
 20  0.0299143     0.42626 |   .                |*********       .   |  0.420021
 21  0.0281326     0.40087 |   .                |********        .   |  0.423799
 22  0.0264864     0.37742 |   .                |********        .   |  0.427112
 23  0.0251603     0.35852 |   .                |*******         .   |  0.430027
 24  0.0237515     0.33845 |   .                |*******         .   |  0.432641
 25  0.0213625     0.30440 |   .                |******          .   |  0.434958
 26  0.0191401     0.27274 |   .                |*****           .   |  0.436822
 27  0.0170756     0.24332 |  .                 |*****            .  |  0.438314
 28  0.0150347     0.21424 |  .                 |****             .  |  0.439497
 29  0.0128981     0.18379 |  .                 |****             .  |  0.440412
 30  0.0108736     0.15494 |  .                 |***              .  |  0.441084
 31 0.00916851     0.13065 |  .                 |***              .  |  0.441561
 32 0.00741502     0.10566 |  .                 |**               .  |    0.4419
 33 0.00597763     0.08518 |  .                 |**               .  |  0.442122
 34 0.00512737     0.07306 |  .                 |*                .  |  0.442266
 35 0.00492594     0.07019 |  .                 |*                .  |  0.442372
 36 0.00440624     0.06279 |  .                 |*                .  |  0.442469
                           '.' MARKS TWO STANDARD ERRORS
```

(c) Autocorrelation function, one regular difference

```
NAME OF VARIABLE        = LRIDERS
PERIODS OF DIFFERENCING= 1.

MEAN OF WORKING SERIES=0.00634323
STANDARD DEVIATION     = 0.0508814
NUMBER OF OBSERVATIONS=       113
```

> 3 The slow linear decay between consecutive time periods has been eliminated with a first difference. But, there is still a slow decay in the SACF for the seasonal component (lags 12, 24, 36, etc.).

```
                                 AUTOCORRELATIONS

LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1          STD
  0 0.00258892     1.00000  |                    |********************|          0
  1 -1.616E-05    -0.00624  |                .   |   .                |   0.094072
  2 -.00061679    -0.23824  |               *****|   .                |  0.0940757
  3 .000160693     0.06207  |                .   |*  .                |  0.0992715
  4 .000047622     0.01839  |                .   |   .                |  0.0996144
  5 -.00031429    -0.12140  |                . **|   .                |  0.0996444
  6 -.00032165    -0.12424  |                . **|   .                |   0.100945
  7 -.00021659    -0.08366  |                . **|   .                |   0.102289
  8 -2.396E-05    -0.00925  |                .   |   .                |   0.102893
  9 .000111387     0.04302  |                .   |*  .                |     0.1029
 10 -.00054945    -0.21223  |                ****|   .                |   0.103059
 11 -.00021661    -0.08367  |                . **|   .                |   0.106857
 12 0.00181535     0.70120  |                .   |**************      |   0.107435
 13 .000016603     0.00641  |              .     |     .              |   0.142284
 14 -.00051501    -0.19893  |              . ****|     .              |   0.142286
 15 .000132518     0.05119  |              .     |*    .              |   0.144727
 16 -4.952E-05    -0.01913  |              .     |     .              |   0.144887
 17 -.00026375    -0.10188  |              .   **|     .              |   0.144909
 18 -0.0003483    -0.13454  |              .  ***|     .              |   0.145541
 19 -0.0001128    -0.04357  |              .    *|     .              |   0.146638
 20 .000043603     0.01684  |              .     |     .              |   0.146752
 21 .000175733     0.06788  |              .     |*    .              |    0.14677
 22 -.00032199    -0.12437  |              .   **|     .              |   0.147047
 23 -.00014334    -0.05537  |              .    *|     .              |   0.147975
 24 0.00140071     0.54104  |              .     |***********         |   0.148158
 25 -1.374E-05    -0.00531  |             .      |      .             |   0.164718
 26 -.00041049    -0.15856  |             .   ***|      .             |   0.164719
 27 .000111027     0.04289  |             .      |*     .             |   0.166064
 28 -1.998E-05    -0.00772  |             .      |      .             |   0.166162
 29 -.00020008    -0.07728  |             .    **|      .             |   0.166165
 30 -.00033296    -0.12861  |             .   ***|      .             |   0.166483
 31 -.00010577    -0.04086  |             .     *|      .             |    0.16736
 32 0.00003902     0.01507  |             .      |      .             |   0.167448
 33  9.131E-06     0.00353  |             .      |      .             |    0.16746
 34 -.00033052    -0.12767  |             .   ***|      .             |   0.167461
 35 -8.064E-05    -0.03115  |             .     *|      .             |    0.16832
 36 0.00117906     0.45543  |             .      |*********           |   0.168371
                            '.' MARKS TWO STANDARD ERRORS
```

(d) Autocorrelation function, one regular and one seasonal difference

```
NAME OF VARIABLE        = LRIDERS
PERIODS OF DIFFERENCING= 1,12.

MEAN OF WORKING SERIES=-.00236887
STANDARD DEVIATION     = 0.0329903
NUMBER OF OBSERVATIONS=       101
```

> 4 Linear decay patterns have been eliminated from the SACF, indicating stationarity. Some correlation is still present at lags 11 and 12 suggesting that a model with a seasonal component might be appropriate.

```
                              AUTOCORRELATIONS

LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1      STD
  0 0.00108836    1.00000 |                    |********************|         0
  1 .000122615    0.11266 |                .   |** .                | 0.0995037
  2 -1.501E-05   -0.01379 |                .   |   .                |  0.100759
  3 .000114163    0.10489 |                .   |** .                |  0.100777
  4 .000023759    0.02183 |                .   |   .                |  0.101853
  5 -5.973E-05   -0.05488 |                .  *|   .                |  0.101899
  6 .000129027    0.11855 |                .   |** .                |  0.102191
  7 .000067103    0.06166 |                .   |*  .                |  0.103544
  8 -3.952E-05   -0.03632 |                .  *|   .                |  0.103907
  9 -.00015909   -0.14618 |                .***|   .                |  0.104032
 10 -0.0001725   -0.15849 |                .***|   .                |  0.106047
 11 -.00027786   -0.25530 |               *****|   .                |  0.108366
 12 -.00028133   -0.25849 |               *****|    .               |  0.114166
 13 -3.666E-05   -0.03369 |               .   *|    .               |  0.119821
 14 -1.827E-05   -0.01678 |               .    |    .               |  0.119915
 15 -3.803E-05   -0.03494 |               .   *|    .               |  0.119938
 16 -.00023396   -0.21497 |               .****|    .               |  0.120039
 17 -5.435E-05   -0.04994 |               .   *|    .               |  0.123791
 18 -9.997E-05   -0.09185 |               .  **|    .               |  0.123991
 19 -2.734E-05   -0.02512 |               .   *|    .               |  0.124663
 20 .000026388    0.02425 |               .    |    .               |  0.124713
 21 .000169355    0.15561 |               .    |*** .               |  0.124759
 22 .000125866    0.11565 |               .    |**  .               |  0.126666
 23 0.00009559    0.08783 |               .    |**  .               |  0.127708
 24 -.00010132   -0.09310 |               .  **|    .               |  0.128304
 25 .000047785    0.04391 |               .    |*   .               |  0.128971
 26 .000053558    0.04921 |               .    |*   .               |  0.129119
 27 .000113092    0.10391 |               .    |**  .               |  0.129305
 28 .000068965    0.06337 |               .    |*   .               |  0.130129
 29 .000108066    0.09929 |               .    |**  .               |  0.130434
 30 .000010903    0.01002 |               .    |    .               |   0.13118
 31 .000040585    0.03729 |               .    |*   .               |  0.131188
 32 .000035208    0.03235 |               .    |*   .               |  0.131293
 33 -.00011796   -0.10838 |               .  **|    .               |  0.131372
 34 -7.199E-05   -0.06614 |               .   *|    .               |  0.132254
 35  6.783E-06    0.00623 |               .    |    .               |  0.132581
 36 -.00008344   -0.07667 |               .  **|    .               |  0.132584
                           '.' MARKS TWO STANDARD ERRORS
```

(e) Partial autocorrelation function, one regular and one seasonal difference

> 5 The SPACF indicates significant correlation at the seasonal lags. This would suggest a seasonal MA(1) or MA(2) model.

```
                        PARTIAL AUTOCORRELATIONS

LAG CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1
  1    0.11266   !               .   !** .               !
  2   -0.02682   !               .  *!   .               !
  3    0.11100   !               .   !** .               !
  4   -0.00376   !               .   !   .               !
  5   -0.05269   !               .  *!   .               !
  6    0.12382   !               .   !** .               !
  7    0.02846   !               .   !*  .               !
  8   -0.03077   !               .  *!   .               !
  9   -0.16546   !               .***!   .               !
 10   -0.14977   !               .***!   .               !
 11   -0.23076   !              *****!   .               !
 12   -0.23733   !              *****!   .               !
 13   -0.01398   !               .   !   .               !
 14   -0.00453   !               .   !   .               !
 15    0.05118   !               .   !*  .               !
 16   -0.20594   !               ****!   .               !
 17    0.01387   !               .   !   .               !
 18   -0.07886   !               . **!   .               !
 19   -0.01163   !               .   !   .               !
 20   -0.08554   !               . **!   .               !
 21    0.01360   !               .   !   .               !
 22    0.03762   !               .   !*  .               !
 23   -0.01155   !               .   !   .               !
 24   -0.19525   !               ****!   .               !
 25   -0.02779   !               .  *!   .               !
 26   -0.04432   !               .  *!   .               !
 27   -0.00350   !               .   !   .               !
 28   -0.13170   !               .***!   .               !
 29    0.02363   !               .   !   .               !
 30   -0.01998   !               .   !   .               !
 31    0.10080   !               .   !** .               !
 32    0.01843   !               .   !   .               !
 33   -0.09395   !               . **!   .               !
 34   -0.10027   !               . **!   .               !
 35   -0.09541   !               . **!   .               !
 36   -0.20167   !               ****!   .               !
```

Step 2. Model Estimation. Once a tentative specification of the various orders (p, d, q, P, D, Q) has been made, we proceed to estimate the various coefficients in the model. The statistical and computational details of the estimation are beyond the scope of this handbook but may be found in most time series textbooks such as Box and Jenkins (1976). We simply note that nonlinear least squares estimates are obtained and we will rely on standard computer software (SAS/ETS in our examples) to carry out the necessary calculations.

Consider again the gasoline price series which we identified as $p = 1$ and $d = 1$. Exhibit 4 shows the estimation results: $\hat{\phi}_1 = 0.716558$ The estimated model could be written as

$$\text{(35)} \qquad \log G_t - \log G_{t-1} = 0.72\ (\log G_{t-1} - \log G_{t-2}) + a_t$$

where the estimated variance of the error term a_t is 0.000285. Rewriting we have

$$\text{(36)} \qquad \log G_t = 1.72 \log G_{t-1} - 0.72 \log G_{t-2} + a_t$$

for the dynamic model describing the development of monthly gasoline prices.

Note that the printout also gives us a standard error of the estimate and a T-ratio (est/std error). Based on large sample size theory, the T-ratio may be easily used to test the hypothesis that the theoretical coefficient (in this case ϕ_1) is zero. A T-ratio larger than 2 in magnitude would lead us to reject that hypothesis with a significance level of about 5%. In this example we have a T-ratio of 10.63 so there is no doubt as to the significance of the estimated coefficient.

EXHIBIT 4

MODEL ESTIMATION FOR GASOLINE PRICE SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE;
LRIDERS=LOG(RIDERS);
LHOURS=LOG(HOURS);
LFARE=LOG(FARE);
LGAS=LOG(GAS);
LEMPLOY=LOG(EMPLOY);
CARDS;
```

[Data]

```
PROC ARIMA;
   IDENTIFY VAR=LGAS(1) NOPRINT;
   ESTIMATE P=1 NOCONSTANT;
```

1 Here we are estimating, an AR(1) model [P=1] as suggested by the earlier analysis of the SACF and SPACF (see Exhibit 1).

(b) Model estimation

2 The estimation process yields a model with a significant constant term and AR(1) component.

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG
AR1,1	0.716558	0.0673776	10.63	1

VARIANCE ESTIMATE =.000285242
STD ERROR ESTIMATE = 0.0168891
NUMBER OF RESIDUALS= 113

CORRELATIONS OF THE ESTIMATES

	AR1,1
AR1,1	1.000

3 The SACF for the residuals do not show any correlation that could be further modeled. That is, there is essentially no correlation "left" in the residuals, and the model appears adequate.

AUTOCORRELATION CHECK OF RESIDUALS

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	2.70	5	0.746	0.002	0.028	-0.123	0.071	0.016	0.040
12	8.91	11	0.630	-0.049	-0.053	0.167	-0.039	0.034	0.116
18	10.76	17	0.869	0.037	-0.056	0.020	0.038	0.087	-0.001
24	13.89	23	0.930	0.113	0.023	0.031	-0.037	-0.027	0.076

Exhibit 5 displays the estimation results for the logged ridership series with d = 1, D = 1 (season of 12 months) and Q = 1 (season of 12 months). We see that $\hat{\theta}_1 = 0.494653$ with $\hat{\sigma}_a^2 = 0.000976$. Note also that the T-ratio of 5.46 readily shows the significance of the estimate $\hat{\theta}_1$. The estimated model may be written as:

(37) $$\Delta\Delta_{12} \log R_t = a_t - 0.49\ a_{t-12}$$

or

$$\Delta(\log R_t - \log R_{t-12}) = a_t - 0.49\ a_{t-12}$$

or

$$(\log R_t - \log R_{t-12}) - (\log R_{t-1} - \log R_{t-13}) = a_t - 0.49\ a_{t-12}$$

Note how current ridership R_t depends upon past months ridership R_{t-1} but also on ridership 12 and 13 months ago, R_{t-12} and R_{t-13}. In addition, there are effects from 12 months ago captured in the 0.49 a_{t-12} term which are not explained by ridership alone.

EXHIBIT 5

MODEL ESTIMATION FOR RIDERSHIP SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
```

[Data]

```
DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
     IDENTIFY VAR=LRIDERS(1,12) NOPRINT;
     ESTIMATE Q=(12) NOCONSTANT;
```

1 Here, a seasonal MA(1) model is estimated, as suggested by the earlier analysis (see Exhibit 3).

2 The estimation process yields a model with a significant seasonal MA(1) component.

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG
MA1,1	0.494653	0.0905199	5.46	12

VARIANCE ESTIMATE =0.00097642
STD ERROR ESTIMATE = 0.0312477
NUMBER OF RESIDUALS= 101

CORRELATIONS OF THE ESTIMATES

	MA1,1
MA1,1	1.000

3 The SACF for the residuals do not show any correlation that could be further modeled. That is, there is essentially no correlation "left" in the residual, and the model appears adequate.

AUTOCORRELATION CHECK OF RESIDUALS

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	6.08	5	0.298	0.038	-0.083	0.092	-0.068	-0.046	0.180
12	14.58	11	0.203	0.048	-0.029	-0.047	-0.144	-0.197	0.095
18	20.11	17	0.269	0.038	-0.044	0.030	-0.196	-0.049	-0.012
24	24.60	23	0.371	0.031	0.047	0.118	0.072	0.017	-0.107

Step 3. Model Criticism or Diagnostics. The third step in the model building process is that of model criticism or model diagnostics. We will present two complementary approaches: analysis of the residuals from the fitted model and analysis of overparameterized models, that is, models which are more general than the specified model but which contain the specified model as a special case.

As always, residuals are defined as the difference between what is actually observed and what the model predicts:

$$\text{(38)} \qquad \text{residual} = \text{actual-prediction.}$$

With an adequate model the residuals should "behave like" independent random errors. Thus, an important diagnostic procedure is to inspect the autocorrelations of the residuals to check for lack of independence. A guide to assessing the magnitude of these autocorrelations is given by an approximate standard error of $1/\sqrt{n}$ where n is the number of residuals. Only autocorrelations falling outside of $\pm 2/\sqrt{n}$ would be considered significantly different from zero and indicate model inadequacy. In this way, we can look at the autocorrelations individually. A statistic constructed to consider the magnitude of several autocorrelations simultaneously is given by

$$\text{(39)} \qquad Q = n(n+2) \sum_{k=1}^{m} r_k^2/(n-k)$$

where r_k is the sample autocorrelation of the residuals at lag k. The statistic Q was originally proposed by Box and Pierce (1970) and modified by Ljung and Box (1978). They showed that if the model is adequate, then Q has approximately a chi-square distribution with m-p-q degrees of freedom. If the model is inadequate then Q would be inflated and we would reject the specified model if Q exceeds the critical value from the upper tail of the appropriate chi-square distribution.

Exhibit 4(b) shows these residuals checks for the ARIMA (1, 1, 0) fit to the logs of the gasoline price series. In this case $2/\sqrt{113} = 0.188$ and none of the autocorrelations are this large. In addition, none of the chi-square statistics are unusually large. To reject the model the values under PROB should be small, say, 0.05 or smaller.

Exhibit 5(b) displays similar checks on the residuals of the ridership model. Here $2/\sqrt{101} = 0.199$ and only lag 11 comes close to being significant. The chi-square tests also indicate model adequacy.

We now turn to overfitting. Consider, in particular, overfitting an AR(1) model with an AR(2) model. If the AR(1) is acceptable we should see two things:

(a) the estimate of the new parameter ϕ_2 should not be significantly different from zero.

and

(b) the new estimate of the parameter is common between the models, ϕ_1, should not change very much.

As an example, consider the gasoline price series with p = 2. Exhibit 6(b) presents the estimation results. Notice that $\hat{\phi}_2$ is not significant (T-ratio = -0.03) and that $\hat{\phi}_1 = 0.7185$ has not changed much from our AR(1) model estimate of 0.7166.

An overfit for the ridership series is displayed in Exhibit 7(b). Here we have added an extra term at lag 24. The new parameter is almost significant with a T-ratio of 1.91 so we might be willing to keep this parameter in the model. The estimate of the common parameter is now 0.4256 compared with 0.4947 under the original model.

Having completed our discussion of models for single series we move on to the so-called transfer function models which interrelate two or more series.

EXHIBIT 6

MODEL ESTIMATION (OVERFITTING) FOR GASOLINE PRICE SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]
DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
  IDENTIFY VAR=LGAS(1) NOPRINT;
  ESTIMATE P=2 NOCONSTANT PLOT;
```

1 Here an AR(2) model [P=2] is estimated.

2 The estimation process indicates that an AR(2) is not warranted and that the AR(1) model originally proposed is adequate.

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG
AR1,1	0.718526	0.0956231	7.51	1
AR1,2	-.00279639	0.09598	-0.03	2

VARIANCE ESTIMATE =.000287809
STD ERROR ESTIMATE = 0.0169649
NUMBER OF RESIDUALS= 113

CORRELATIONS OF THE ESTIMATES

	AR1,1	AR1,2
AR1,1	1.000	-0.706
AR1,2	-0.706	1.000

3 The SACF of the residuals indicate an adequate model.

AUTOCORRELATION CHECK OF RESIDUALS

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	2.71	4	0.607	-0.000	0.029	-0.122	0.072	0.017	0.040
12	8.94	10	0.538	-0.048	-0.053	0.168	-0.039	0.034	0.116
18	10.79	16	0.822	0.037	-0.056	0.020	0.038	0.087	-0.001
24	13.93	22	0.904	0.113	0.023	0.031	-0.037	-0.027	0.077

EXHIBIT 7

MODEL ESTIMATION (OVERFITTING) FOR RIDERSHIP SERIES

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]
DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);
PROC ARIMA;
  IDENTIFY VAR=LRIDERS(1,12) NOPRINT;
  ESTIMATE Q=(12,24) NOCONSTANT PLOT;
```

1 Here a seasonal MA(2) model is estimated.

(b) Model estimation

```
ARIMA: LEAST SQUARES ESTIMATION

PARAMETER    ESTIMATE    STD ERROR   T RATIO  LAG

MA1,1        0.425622    0.101148      4.21   12
MA1,2        0.204935    0.107039      1.91   24

VARIANCE  ESTIMATE =.000946629
STD ERROR ESTIMATE = 0.0307673
NUMBER OF RESIDUALS=  101

        CORRELATIONS OF THE ESTIMATES

                MA1,1    MA1,2
        MA1,1   1.000   -0.514
        MA1,2  -0.514    1.000


AUTOCORRELATION CHECK OF RESIDUALS

TO   CHI                                 AUTOCORRELATIONS
LAG  SQUARE DF   PROB
 6    6.09   4  0.192      0.087 -0.049  0.121 -0.023 -0.032  0.174
12   14.18  10  0.165      0.064 -0.050 -0.067 -0.143 -0.197  0.019
18   19.29  16  0.254      0.017 -0.067 -0.004 -0.188 -0.041 -0.003
24   24.43  22  0.325      0.021  0.039  0.152  0.101  0.050  0.039
```

2 The estimation process shows that a seasonal MA(2) model is perhaps warranted, so that the "overfitting" was justified.

3 The SACF of the residuals indicate an adequate model.

III. TRANSFER FUNCTION MODELS

This chapter covers the development of transfer function models. The topics covered include:

1. Notation for Transfer Function Models
2. The Model Building Process

A number of terms and concepts are introduced here. They are described below.

The univariate time series models of the previous chapter relate a series to its own past. Transfer function time series models allow us to, in addition, relate a series to present and past values of other series which explain the behavior of the series in question. For example, transit ridership will be related to fare level, service level, market size and so forth. We call ridership the output series and the other series which influence ridership are called input series. It will usually be the case that changes in input series will produce delayed response in the output series, that is, the relationship is dynamic in nature. It takes time for current or potential riders to assess the impact of fare or service changes and modify their ridership patterns accordingly. Thus, our models must include lagged effects in addition to contemporaneous effects.

NOTATION FOR TRANSFER FUNCTION MODELS

A simple transfer function model can be written:

$$(40) \quad Y_t = w_o \left(X_{t-1} + \delta X_{t-2} + \delta^2 X_{t-3} + \delta^3 X_{t-4} + \ldots\right) + N_t$$

This model says that there is a pure delay of one month, say, before the input series X_t influences the output Y_t. Subsequently, the effect of the input decays exponentially at rate δ in the following months. The series N_t, which is unobservable, is called the noise or disturbance series and includes all the other influences on the output Y_t not captured in the input series X_t and its past.

In general we may represent the model as

$$(41) \qquad Y_t = \nu_o X_t + \nu_1 X_{t-1} + \nu_2 X_{t-2} + \ldots + N_t$$

where the coefficients ν_o, ν_1, ν_2, ... define the transfer function of the model. Using the backshift operator we may write

$$(42) \qquad Y_t = \nu(B) X_t + N_t$$

where $\nu(B) = \nu_o + \nu_1 B + \nu_2 B^2 + \ldots$. However, it will usually be unsatisfactory to parametrize the model in terms of the ν's. Rather a parsimonious representation of $\nu(B)$ as a ratio of low order polynomials will be sought.

For example, the ratio $\nu(B) = \frac{w_o B}{1-\delta B}$ containing just two parameters, w_o and δ, can be rewritten as

$$(43) \qquad \nu(B) = w_o(B + \delta B^2 + \delta^2 B^3 + \delta^3 B^4 + \ldots)$$

which, when used in (42), produces our example of (40).

THE MODEL BUILDING PROCESS

As with univariate time series modeling our first task is to specify the nonzero coefficients in the polynomials defining $\nu(B)$. Several alternative schemes have been presented in the literature to do this in general circumstances. However, our experience leads us to believe that the following method will usually suffice with typical transit data. With monthly ridership data, one ordinary difference and/or one seasonal difference is necessary to induce stationarity. In addition there will be seasonal effects in the disturbance term N_t but not in the input series such as fare, gasoline price or service level. Based on these considerations, we fit a transfer function model of the form:

$$(44) \quad \Delta\Delta_{12} \log R_t = \nu_o X_t + \nu_1 X_{t-1} + \dots + \nu_m X_{t-m} + a_t - \theta_1 a_{t-1} - \theta_{12} a_{t-12}$$

where $X_t = \Delta\Delta_{12} I_t$, I_t is the input series in question and m is of the order of 10 or 12. The value of m should be selected so that the maximum lag for which effects might be produced will be included in the model. Previous research indicates that all effects should be evident within 10 to 12 months after a change in the input variables. We then look for significant values among the estimates of $\nu_o, \nu_1, \dots, \nu_m$ and also for exponential decay patterns indicative of denominator polynomials. Finally, we hypothesize a parsimonious form for $\nu(B)$ and estimate the parameters of that model. As in Chapter 2 we then criticize the model and change it as appropriate.

To illustrate we consider using the transit fare series as an input to the ridership series. With m = 10 we estimate the model of (44) the results being presented in Exhibit 8. We see that $\hat{\theta}_1 = -0.006$ is insignificant but $\hat{\theta}_{12} = 0.268$ and $\hat{\nu}_o = -0.294$ are significant. All other $\hat{\nu}_j$ are insignificant but $\hat{\nu}_1$ has a T-ratio of 1.56 and might be further investigated. We also note the residual autocorrelation at lag 24 is -0.181 which, though not significant, is nearly so. On this evidence we decide to fit a model of the form:

$$(45) \quad \Delta\Delta_{12} \log R_t = \frac{w_o}{1-\delta B} \Delta\Delta_{12} \log F_t + (1 - \theta_{12}B^{12} - \theta_{24}B^{24})a_t$$

Exhibit 9 shows the results of this estimation. Note that all parameter estimates are judged significant with $\hat{w}_o = -0.240$, $\hat{\delta} = 0.625$, $\hat{\theta}_{12} = 0.342$, $\hat{\theta}_{24} = -.279$ and $\hat{\sigma}_a = 0.0287$. Furthermore, the autocorrelations of the residuals from this model look quite good. We also note that use of fare as

EXHIBIT 8

TRANSFER FUNCTION MODEL ESTIMATION

SINGLE INPUT (TRANSIT FARE)

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;

[Data]

DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);
PROC ARIMA;
  IDENTIFY VAR=LRIDERS(1,12) CROSSCOR=(LFARE(1,12)) NOPRINT;
  ESTIMATE Q=(1,12) INPUT=((1,2,3,4,5,6,7,8,9,10)LFARE) NOCONSTANT;
```

1 The input series (or independent variable) in this transfer function model is Transit Fare. To determine the lag structure for Transit Fare, all lags between 0 and 10 are included in this preliminary model. The initial noise model includes both regular and seasonal MA(1) terms.

EXHIBIT 8 (continued)

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG	VARIABLE
MA1,1	-.00598438	0.112062	-0.05	1	LRIDERS
MA1,2	0.267983	0.124224	2.16	12	LRIDERS
NUM1	-0.294143	0.079337	-3.71	0	LFARE
NUM1,1	0.123917	0.0793927	1.56	1	LFARE
NUM1,2	0.00962421	0.0797815	0.12	2	LFARE
NUM1,3	0.113822	0.0792045	1.44	3	LFARE
NUM1,4	0.0989929	0.079461	1.25	4	LFARE
NUM1,5	0.0417513	0.0778711	0.54	5	LFARE
NUM1,6	-0.0639779	0.0795591	-0.80	6	LFARE
NUM1,7	0.0283007	0.0788501	0.36	7	LFARE
NUM1,8	0.0155406	0.0791369	0.20	8	LFARE
NUM1,9	-0.0699369	0.0819997	-0.85	9	LFARE
NUM1,10	-0.0374005	0.0819071	-0.46	10	LFARE

2 The estimation process shows that the seasonal MA(1) form and the Fare terms at lags 0 and 1 are the most significant. The other forms can be dropped from the next model iteration.

VARIANCE ESTIMATE =0.00098055
STD ERROR ESTIMATE = 0.0313137
NUMBER OF RESIDUALS= 91

CORRELATIONS OF THE ESTIMATES

	MA1,1	MA1,2	NUM1	NUM1,1	NUM1,2	NUM1,3	NUM1,4	NUM1,5	NUM1,6	NUM1,7	NUM1,8	NUM1,9	NUM1,10
MA1,1	1.000	0.185	-0.029	-0.005	-0.047	-0.055	-0.051	-0.024	0.034	0.042	0.013	0.037	0.009
MA1,2	0.185	1.000	-0.030	-0.049	-0.153	-0.118	-0.148	0.053	0.142	0.031	0.046	0.063	-0.039
NUM1	-0.029	-0.030	1.000	-0.019	-0.018	-0.038	-0.074	-0.028	0.157	-0.084	0.088	-0.018	0.024
NUM1,1	-0.005	-0.049	-0.019	1.000	0.028	0.027	0.053	0.074	0.017	-0.160	0.078	-0.095	0.017
NUM1,2	-0.047	-0.153	-0.018	0.028	1.000	0.042	0.044	0.040	0.047	0.019	-0.167	0.071	-0.086
NUM1,3	-0.055	-0.118	-0.038	0.027	0.042	1.000	0.053	0.004	0.024	0.077	0.018	-0.152	0.089
NUM1,4	-0.051	-0.148	-0.074	0.053	0.044	0.053	1.000	0.043	-0.027	0.036	0.065	0.014	-0.142
NUM1,5	-0.024	0.053	-0.028	0.074	0.040	0.004	0.043	1.000	0.057	0.009	0.046	0.082	0.026
NUM1,6	0.034	0.142	0.157	0.017	0.047	0.024	-0.027	0.057	1.000	0.038	0.028	0.042	0.075
NUM1,7	0.042	0.031	-0.084	-0.160	0.019	0.077	0.036	0.009	0.038	1.000	0.025	0.026	0.029
NUM1,8	0.013	0.046	0.088	0.078	-0.167	0.018	0.065	0.046	0.028	0.025	1.000	0.026	0.023
NUM1,9	0.037	0.063	-0.018	-0.095	0.071	-0.152	0.014	0.082	0.042	0.026	0.026	1.000	0.027
NUM1,10	0.009	-0.039	0.024	0.017	-0.086	0.089	-0.142	0.026	0.075	0.029	0.023	0.027	1.000

AUTOCORRELATION CHECK OF RESIDUALS

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	3.24	4	0.518	-0.016	-0.030	0.034	-0.130	-0.091	0.075
12	6.45	10	0.776	0.010	-0.042	-0.034	0.017	-0.153	0.061
18	15.18	16	0.512	0.056	0.118	0.090	-0.179	0.143	-0.006
24	20.70	22	0.539	-0.086	0.032	0.019	-0.048	-0.023	-0.181

3 The SACF of the residuals indicates that the addition of a seasonal MA(2) might be warranted because of the correlation remaining at lag 24.

EXHIBIT 9

TRANSFER FUNCTION MODEL ESTIMATION

SINGLE INPUT (TRANSIT FARE)

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]
DATA BUS2;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);
```

1 The model has been modified based upon the first iteration shown in Exhibit 8. An exponential decay was indicated for the Fare series and a seasonal MA(2) form was suggested for the noise model.

```
PROC ARIMA;
   IDENTIFY VAR=LRIDERS(1,12) CROSSCOR=(LFARE(1,12)) NOPRINT;
   ESTIMATE Q=(12,24) INPUT=(/(1)LFARE) NOCONSTANT PLOT;
```

EXHIBIT 9 (continued)

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG	VARIABLE
MA1,1	0.341691	0.10218	3.34	12	LRIDERS
MA1,2	0.279424	0.107482	2.60	24	LRIDERS
NUM1	-0.240462	0.0656156	-3.66	0	LFARE
DEN1,1	0.624719	0.144273	4.33	1	LFARE

2 The estimation process shows that this model form can be supported as all coefficients are significant.

VARIANCE ESTIMATE =.000824233
STD ERROR ESTIMATE = 0.0287095
NUMBER OF RESIDUALS= 100

CORRELATIONS OF THE ESTIMATES

	MA1,1	MA1,2	NUM1	DEN1,1
MA1,1	1.000	-0.463	-0.025	-0.116
MA1,2	-0.463	1.000	0.088	0.206
NUM1	-0.025	0.088	1.000	0.527
DEN1,1	-0.116	0.206	0.527	1.000

AUTOCORRELATION CHECK OF RESIDUALS

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	4.71	4	0.319	0.045	-0.087	0.013	-0.083	-0.100	0.132
12	6.16	10	0.802	0.059	-0.059	0.000	0.006	-0.072	0.025
18	12.40	16	0.716	0.033	-0.052	-0.003	-0.170	0.111	0.080
24	13.00	22	0.933	-0.043	0.022	0.036	0.005	0.010	0.029

(c) Autocorrelation plot of residuals

```
LAG COVARIANCE CORRELATION -1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1        STD
  0 .000824233     1.00000 |                    |********************|        0
  1 .000037157     0.04508 |                .   |*  .                |      0.1
  2 -7.134E-05    -0.08656 |                . **|   .                | 0.100203
  3 .000010977     0.01332 |                .   |   .                | 0.100948
  4 -6.847E-05    -0.08307 |                . **|   .                | 0.100965
  5 -8.265E-05    -0.10028 |                . **|   .                | 0.101647
  6 .000108711     0.13189 |                .   |***.                | 0.102631
  7 .000048892     0.05932 |                .   |*  .                | 0.104312
  8 -4.867E-05    -0.05905 |                .  *|   .                | 0.104649
  9  3.386E-07     0.00041 |                .   |   .                | 0.104982
 10  5.065E-06     0.00615 |                .   |   .                | 0.104982
 11 -5.964E-05    -0.07235 |                .  *|   .                | 0.104986
 12 .000020531     0.02491 |                .   |   .                | 0.105483
 13 .000027247     0.03306 |                .   |*  .                | 0.105542
 14 -4.297E-05    -0.05213 |                .  *|   .                | 0.105645
 15 -2.291E-06    -0.00278 |                .   |   .                | 0.105902
 16 -.00014006    -0.16992 |                .***|   .                | 0.105903
 17 .000091078     0.11050 |                .   |** .                | 0.108595
 18 .000065559     0.07954 |                .   |** .                | 0.109714
 19 -3.571E-05    -0.04333 |                .  *|   .                | 0.110289
 20 .000018246     0.02214 |                .   |   .                | 0.110459
 21 .000029829     0.03619 |                .   |*  .                | 0.110503
 22  4.024E-06     0.00488 |                .   |   .                | 0.110622
 23  8.375E-06     0.01016 |                .   |   .                | 0.110624
 24 .000023534     0.02855 |                .   |*  .                | 0.110633
                           '.' MARKS TWO STANDARD ERRORS
```

3 The SACF plot for the residuals indicates an adequate model fit.

an input variable has reduced the residual standard deviation about 8% from the value for the univariate model based on ridership alone.

We proceed in this same way to investigate the effects of various lags of other potential input variables: gasoline price, service hours and employment level. It was found the gasoline price and employment level affect ridership only contemporaneously, (i.e., at lag zero) whereas service hours contains only a delayed effect at lag eight months. Thus, from a multiple input setting we consider the model:

$$(46) \qquad \Delta\Delta_{12} \log R_t = \frac{w_o}{1-\delta B} \Delta\Delta_{12} \log F_t + w_1 \Delta\Delta_{12} \log G_t + w_2 \Delta\Delta_{12} \log E_t + w_3 B^8 \Delta\Delta_{12} \log HR_t + (1-\theta_{12}B^{12} - \theta_{24}B^{24})a_t$$

In Exhibit 10 we display the results from fitting this model. The $\hat{\delta}$ being only marginally significant we drop it from the model and re-estimate obtaining the results shown in Exhibit 11. The residual analysis substantiates the adequacy of this model which we summarize as:

$$(47) \qquad \Delta\Delta_{12} \log R_t = w_o \Delta\Delta_{12} \log F_t + w_1 \Delta\Delta_{12} \log G_t + w_2 \Delta\Delta_{12} \log E_t + w_3 B^8 \Delta\Delta_{12} \log HR_t + (1 - \theta_{12}B^{12} - \theta_{24}B^{24})a_t$$

or

$$\Delta\Delta_{12} \log R_t = -.272 \, \Delta\Delta_{12} \log F_T + .277 \, \Delta\Delta \log G_t + .541 \, \Delta\Delta_{12} \log E_t + .259 \, B^8 \Delta\Delta_{12} \log HR_t + (1 - .324B^{12} - .291B^{24})a_t$$

EXHIBIT 10

TRANSFER FUNCTION MODEL ESTIMATION

FULL MODEL

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]
DATA BUS1;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);
PROC ARIMA;
 IDENTIFY VAR=LRIDERS(1,12) CROSSCOR=(LGAS(1,12) LFARE(1,12) LHOURS(1,12)
                                     LEMPLOY(1,12)) NOPRINT;
 ESTIMATE Q=(12,24) INPUT=(/(1)LFARE LEMPLOY 8$LHOURS LGAS) NOCONSTANT;
```

1	The full transfer function is estimated. The Fare series includes an exponential delay function while the Hours series has a delay of 8 months. The Employ and Gas series are contemporaneously correlated with Riders.

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG	VARIABLE
MA1,1	0.30141	0.114904	2.62	12	LRIDERS
MA1,2	0.296777	0.116384	2.55	24	LRIDERS
NUM1	-0.25775	0.0709758	-3.63	0	LFARE
DEN1,1	0.412635	0.211345	1.95	1	LFARE
NUM2	0.476623	0.285166	1.67	0	LEMPLOY
NUM3	0.232658	0.131283	1.77	0	LHOURS
NUM4	0.265879	0.119932	2.22	0	LGAS

2 Three of the terms are only marginally significant, and some modification of the model form is warranted for the next iteration. A good rule of thumb is to always simplify the model if possible, so the denominator term of the Fare series will be dropped for the next iteration.

VARIANCE ESTIMATE =0.00079058
STD ERROR ESTIMATE = 0.0281173
NUMBER OF RESIDUALS= 93

CORRELATIONS OF THE ESTIMATES

	MA1,1	MA1,2	NUM1	DEN1,1	NUM2	NUM3	NUM4
MA1,1	1.000	-0.474	-0.171	-0.179	0.263	0.002	-0.193
MA1,2	-0.474	1.000	0.214	0.156	-0.179	-0.014	0.159
NUM1	-0.171	0.214	1.000	0.425	-0.234	-0.029	0.137
DEN1,1	-0.179	0.156	0.425	1.000	-0.216	-0.163	-0.047
NUM2	0.263	-0.179	-0.234	-0.216	1.000	0.070	-0.240
NUM3	0.002	-0.014	-0.029	-0.163	0.070	1.000	0.081
NUM4	-0.193	0.159	0.137	-0.047	-0.240	0.081	1.000

(c) Autocorrelation check of residuals

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	3.51	4	0.476	-0.027	-0.145	0.006	-0.023	-0.041	0.108
12	8.92	10	0.540	0.057	-0.128	-0.070	0.094	-0.131	0.027
18	16.29	16	0.433	0.047	-0.016	0.025	-0.202	0.128	0.063
24	17.99	22	0.707	-0.089	-0.036	0.000	0.045	0.038	0.034

EXHIBIT 11

TRANSFER FUNCTION MODEL ESTIMATION

FULL MODEL

(a) SAS input listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
CARDS;
[Data]
DATA BUS1;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);
PROC ARIMA;
 IDENTIFY VAR=LRIDERS(1,12) CROSSCOR=(LGAS(1,12) LFARE(1,12) LHOURS(1,12)
                                     LEMPLOY(1,12)) NOPRINT;
 ESTIMATE Q=(12,24) INPUT=(LFARE LEMPLOY 8$LHOURS LGAS) NOCONSTANT;
```

1 This next iteration includes a lag of 8 months for the Hours series and no lag structure for the other input series.

2 The model estimation indicates that all coefficients are statistically significant or nearly so.

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG	VARIABLE
MA1,1	0.324207	0.112925	2.87	12	LRIDERS
MA1,2	0.290834	0.115107	2.53	24	LRIDERS
NUM1	-0.27236	0.0722462	-3.77	0	LFARE
NUM2	0.541324	0.279904	1.93	0	LEMPLOY
NUM3	0.258969	0.131279	1.97	0	LHOURS
NUM4	0.277233	0.11924	2.33	0	LGAS

VARIANCE ESTIMATE =.000796547
STD ERROR ESTIMATE = 0.0282232
NUMBER OF RESIDUALS= 93

CORRELATIONS OF THE ESTIMATES

	MA1,1	MA1,2	NUM1	NUM2	NUM3	NUM4
MA1,1	1.000	-0.484	-0.126	0.230	-0.010	-0.199
MA1,2	-0.484	1.000	0.152	-0.145	-0.006	0.158
NUM1	-0.126	0.152	1.000	-0.202	-0.012	0.138
NUM2	0.230	-0.145	-0.202	1.000	0.047	-0.251
NUM3	-0.010	-0.006	-0.012	0.047	1.000	0.062
NUM4	-0.199	0.158	0.138	-0.251	0.062	1.000

(c) Autocorrelation check of residuals

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	2.36	4	0.670	-0.018	-0.098	0.031	-0.005	-0.027	0.110
12	8.16	10	0.613	0.058	-0.116	-0.079	0.062	-0.164	0.030
18	16.15	16	0.443	0.057	-0.020	0.012	-0.242	0.079	0.030
24	18.88	22	0.653	-0.094	-0.040	-0.017	0.068	0.076	0.035

3 The SACF of the residuals shows no unaccounted for correlation.

IV. INTERVENTION ANALYSIS

The assessment of the impact of special events, either controlled or otherwise, on time series such as transit ridership is an important endeavor. What is the effect of a particular (controllable) fare or service change? How did ridership react to fuel shortages or severe winter weather (uncontrollable)? The methodology of time series intervention analysis has been developed to answer these questions -- the original reference being Box and Tiao (1975). The theory developed allows for delayed appearance of effects after the intervention, tapered responses and gradual diminishing of the effect rather than abrupt on-off behavior.

As a simple example of intervention modeling, we consider the effect on bus ridership of a very severe ice storm in January 1979 in Portland. Such an event should impact ridership during the storm but presumably have no lasting effect. Therefore, to model the intervention we create a variable which equals zero except for January 1979 when it equals one:

$$I_t = \begin{cases} 1 \text{ if } t = \text{January 1979}, \\ 0 \text{ otherwise.} \end{cases} \qquad (48)$$

We then add I_t as an additional input variable in our full transfer function model and re-estimate the model checking for significance on the coefficient associated with the indicator variable I_t.

Exhibit 12 displays the results of this estimation. The intervention variable (called TEST in the computer output) shows a significant coefficient of -0.04769 (with a T-RATIO of -3.12) indicating the substantial negative impact of the storm on ridership.

EXHIBIT 12

INTERVENTION MODEL

EXAMPLE: SEVERE WEATHER

1 The intervention variable test is created with the IF statements.

(a) SAS data listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
  IF YEAR<79 THEN TEST=0;
  IF YEAR=79 AND MONTH=1 THEN TEST=1;
  IF YEAR=79 AND MONTH>1 THEN TEST=0;
  IF YEAR>79 THEN TEST=0;
CARDS;
[Data]

DATA BUS1;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
 IDENTIFY VAR=LRIDERS(1,12) CROSSCOR=(LGAS(1,12) LFARE(1,12) LHOURS(1,12)
                                   TEST(1,12) LEMPLOY(1,12)) NOPRINT;
 ESTIMATE Q=(12,24) INPUT=(LFARE LEMPLOY B$LHOURS LGAS TEST) NOCONSTANT;
```

2 The model is estimated with the original inputs (see Exhibit 11) and the intervention variable test.

3 The variable TEST is statistically significant indicating that the intervention has an effect on ridership.

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG	VARIABLE
MA1,1	0.269899	0.110505	2.44	12	LRIDERS
MA1,2	0.300655	0.113088	2.66	24	LRIDERS
NUM1	-0.277854	0.0672102	-4.13	0	LFARE
NUM2	0.428418	0.267451	1.60	0	LEMPLOY
NUM3	0.257509	0.12195	2.11	0	LHOURS
NUM4	0.284698	0.113364	2.51	0	LGAS
NUM5	-0.0476929	0.0152871	-3.12	0	TEST

VARIANCE ESTIMATE =.000724982
STD ERROR ESTIMATE = 0.0269255
NUMBER OF RESIDUALS= 93

CORRELATIONS OF THE ESTIMATES

	MA1,1	MA1,2	NUM1	NUM2	NUM3	NUM4	NUM5
MA1,1	1.000	-0.392	-0.123	0.181	-0.002	-0.167	-0.008
MA1,2	-0.392	1.000	0.137	-0.080	-0.018	0.109	0.097
NUM1	-0.123	0.137	1.000	-0.201	-0.011	0.144	0.050
NUM2	0.181	-0.080	-0.201	1.000	0.051	-0.243	0.065
NUM3	-0.002	-0.018	-0.011	0.051	1.000	0.063	0.012
NUM4	-0.167	0.109	0.144	-0.243	0.063	1.000	0.034
NUM5	-0.008	0.097	0.050	0.065	0.012	0.034	1.000

(c) Autocorrelation check of residuals

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	3.49	4	0.479	0.010	-0.041	-0.033	-0.107	0.055	0.132
12	7.55	10	0.673	0.003	-0.022	-0.091	0.006	-0.168	0.026
18	15.41	16	0.495	0.056	-0.068	0.070	-0.229	0.040	0.048
24	20.09	22	0.577	-0.079	-0.077	0.009	0.019	0.154	0.036

As a second example, we wish to assess the impact of the (average) fare increase of 5.3 cents which occurred in September 1978. To isolate the fare increase from the rest of the fare variable we first remove the increase from the fare series.

$$(49) \quad NEWFARE_t = \begin{cases} OLDFARE_t, \text{ if t is before Sept. 1978} \\ OLDFARE_t - 5.3, \text{ if t is Sept. 1978 or after.} \end{cases}$$

Then we define a new variable S_t as

$$(50) \quad S_t = \begin{cases} 0 \text{ if t is before Sept. 1978} \\ 1 \text{ if t is Sept. 1978 or after.} \end{cases}$$

(S_t is a <u>step function</u> with step at September 1978.) We then re-estimate the transfer function model with our new fare and step function as input variables.

The results are shown in Exhibit 13. Notice that the intervention variable (called TEST in the computer output) has significant coefficient of -0.06197 (with a T-RATIO of -2.67) and shows the negative effect of the fare increase of September 1978.

EXHIBIT 13

INTERVENTION MODEL

EXAMPLE: FARE INCREASE

(a) SAS data listing

```
DATA BUS;
  INPUT YEAR  MONTH RIDERS HOURS EMPLOY GAS FARE CPI
    HRC HRX HRU HRW HRSW HRSE;
IF YEAR=78 AND MONTH >8 THEN TEST=1;
IF YEAR=78 AND MONTH<9 THEN TEST=0;
IF YEAR<78 THEN TEST=0;
IF YEAR>78 THEN TEST=1;
IF YEAR=78 AND MONTH>8 THEN FARE=FARE-5.3;
IF YEAR>78 THEN FARE=FARE-5.3;
CARDS;
```

> 1 The intervention variable test is created and the Fare series is modified.

[Data]

```
DATA BUS1;
  SET BUS;
  LRIDERS=LOG(RIDERS);
  LHOURS=LOG(HOURS);
  LFARE=LOG(FARE);
  LGAS=LOG(GAS);
  LEMPLOY=LOG(EMPLOY);

PROC ARIMA;
 IDENTIFY VAR=LRIDERS(1,12) CROSSCOR=(LGAS(1,12) LFARE(1,12) LHOURS(1,12)
                              TEST(1,12) LEMPLOY(1,12)) NOPRINT;
 ESTIMATE Q=(12,24) INPUT=(LFARE LEMPLOY 8$LHOURS LGAS TEST) NOCONSTANT;
```

> 2 The model is estimated with the original model inputs and the intervention variable test.

(b) Model estimation

ARIMA: LEAST SQUARES ESTIMATION

PARAMETER	ESTIMATE	STD ERROR	T RATIO	LAG	VARIABLE
MA1.1	0.313609	0.113397	2.77	12	LRIDERS
MA1.2	0.288749	0.115513	2.50	24	LRIDERS
NUM1	-0.211938	0.0766304	-2.77	0	LFARE
NUM2	0.477803	0.283064	1.69	0	LEMPLOY
NUM3	0.2455	0.131772	1.86	0	LHOURS
NUM4	0.282291	0.119267	2.37	0	LGAS
NUM5	-0.0619734	0.0231873	-2.67	0	TEST

3 The variable test is statistically significant

VARIANCE ESTIMATE =.000801104
STD ERROR ESTIMATE = 0.0283038
NUMBER OF RESIDUALS= 93

CORRELATIONS OF THE ESTIMATES

	MA1.1	MA1.2	NUM1	NUM2	NUM3	NUM4	NUM5
MA1.1	1.000	-0.470	-0.134	0.233	-0.008	-0.186	-0.019
MA1.2	-0.470	1.000	0.150	-0.137	-0.011	0.145	0.028
NUM1	-0.134	0.150	1.000	-0.222	-0.058	0.117	-0.001
NUM2	0.233	-0.137	-0.222	1.000	0.059	-0.246	0.019
NUM3	-0.008	-0.011	-0.058	0.059	1.000	0.062	0.081
NUM4	-0.186	0.145	0.117	-0.246	0.062	1.000	0.060
NUM5	-0.019	0.028	-0.001	0.019	0.081	0.060	1.000

(c) Autocorrelation check of residuals

TO LAG	CHI SQUARE	DF	PROB	AUTOCORRELATIONS					
6	2.25	4	0.689	0.003	-0.081	0.006	-0.038	0.006	0.120
12	7.71	10	0.657	0.050	-0.108	-0.062	0.048	-0.173	0.034
18	15.05	16	0.521	0.032	-0.028	0.030	-0.240	0.062	0.004
24	17.77	22	0.720	-0.083	-0.019	-0.035	0.081	0.072	0.042

V. FORECASTING

One of the primary objectives of building a model for a time series is to be able to forecast or predict the values for that series at future times. We also need to be able to asses the precision of those forecasts. Modern time series analysis software allows us to readily make these forecasts and measure their precision.

With our transfer function models, the input series can be put into two different classes - controllable and uncontrollable. Fare and service level are controllable whereas gasoline price and employment levels are not. With respect to forecasting future ridership levels clearly uncontrollable inputs must themselves be forecast into the future but fare and service level could be set to whatever we would like them to be. For example, the impact of proposed future fare or service level changes could be assessed by incorporating those changes into the future input values for fare and/or service level and then note the implied ridership forecasts.

Alternatively, forecasting techniques may be applied to judge the overall usefulness of the time series model. For example, we could "back up" one year from the end of our data set. Then use the model to forecast that last year using the actual known values for all the input series. We then compare the ridership forecast by the model to the known ridership for that last year.

Exhibit 14 illustrates this procedure. Our final transfer function model has been used to forecast the last year of observed ridership using the actual last year values for all of the input series. The computer output lists the forecasts for ridership (in logarithm terms), their standard errors, 95% prediction limits, the actual ridership values and the residuals or prediction errors.

Exhibit 15 presents forecasting results when we forecast the uncontrollable input series, gasoline price and employment level but set fare and service level to their actual values. We again have forecast ridership from one year back so that we may compare the forecasts to actual values. In comparison with the results in Exhibit 14 the forecast standard errors are somewhat larger reflecting the additional uncertainty associated with having to forecast future gasoline price and employment levels.

EXHIBIT 14

FORECASTS

(a) SAS input listing

```
DATA;
 INPUT YEAR MONTH RIDERS HOURS EMPLOY GAS FARE;
LRIDERS=LOG(RIDERS); LHOURS=LOG(HOURS); LEMPLOY=LOG(EMPLOY);
LGAS=LOG(GAS); LFARE=LOG(FARE);
N=_N_;LABEL N=Month;
CARDS;
[Data]

PROC ARIMA;
    IDENTIFY VAR=LRIDERS(1,12)
     CROSSCOR=(LFARE(1,12) LEMPLOY(1,12) LGAS(1,12) LHOURS(1,12)) NOPRINT;
  ESTIMATE Q=(12,24) INPUT=(B$LHOURS LFARE LGAS LEMPLOY) NOCONSTANT NOPRINT;
   FORECAST LEAD=12 BACK=12
    OUT=B ID=N ;


PROC PLOT DATA=B(FIRSTOBS=103);
  PLOT FORECAST*N='F' LRIDERS*N='*' L95*N='L' U95*N='U'/ OVERLAY
       VPOS=40 HPOS=60;
  TITLE Ridership Forecast, Actual and Confidence Limits;
```

1 Forecasts for the Rider series are made using historical data for the input series.

2 Forecasts are plotted.

EXHIBIT 14 (continued)

(b) Forecasts

FORECASTS FOR VARIABLE LRIDERS

OBS	FORECAST	STD ERROR	LOWER 95%	UPPER 95%	ACTUAL	RESIDUAL
--------FORECAST BEGINS--------						
103	11.7648	0.0282	11.7095	11.8202	11.7424	-0.0224
104	11.7225	0.0399	11.6443	11.8007	11.7068	-0.0157
105	11.8052	0.0489	11.7094	11.9010	11.7951	-0.0101
106	11.8491	0.0564	11.7384	11.9597	11.8615	0.0124
107	11.8409	0.0631	11.7172	11.9646	11.8615	0.0206
108	11.7867	0.0691	11.6512	11.9222	11.7974	0.0106
109	11.8302	0.0747	11.6838	11.9765	11.8920	0.0619
110	11.8273	0.0798	11.6709	11.9838	11.8671	0.0398
111	11.7816	0.0847	11.6156	11.9475	11.8629	0.0813
112	11.7852	0.0892	11.6103	11.9601	11.8720	0.0868
113	11.7703	0.0936	11.5868	11.9537	11.8452	0.0749
114	11.7497	0.0978	11.5580	11.9413	11.7958	0.0462

3 Forecasts for the Rider series are shown with the confidence band, the actual values, and the residuals.

(c) Forecast plots

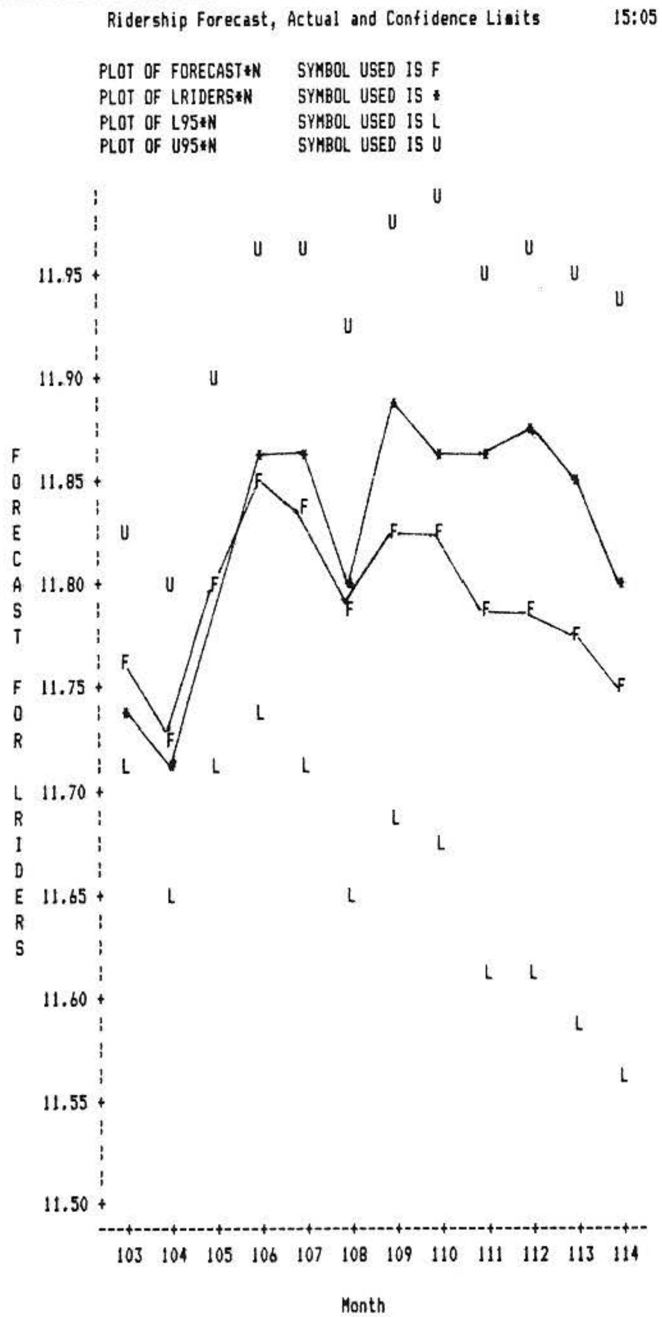


EXHIBIT 15

FORECAST

(a) SAS input listing

```
DATA;
 INPUT YEAR MONTH RIDERS HOURS EMPLOY GAS FARE;
LRIDERS=LOG(RIDERS); LHOURS=LOG(HOURS); LEMPLOY=LOG(EMPLOY);
LGAS=LOG(GAS); LFARE=LOG(FARE);
N=_N_;LABEL N=Month;
CARDS;
[Data]
PROC ARIMA;
  IDENTIFY VAR=LGAS(1,12) NOPRINT;
  ESTIMATE P=1 NOCONSTANT NOPRINT;
  IDENTIFY VAR=LEMPLOY(1,12) NOPRINT;
  ESTIMATE Q=(1)(12) NOCONSTANT NOPRINT;
    IDENTIFY VAR=LRIDERS(1,12)
     CROSSCOR=(LFARE(1,12) LEMPLOY(1,12) LGAS(1,12) LHOURS(1,12)) NOPRINT;
  ESTIMATE Q=(12,24) INPUT=(8$LHOURS LFARE LGAS LEMPLOY) NOCONSTANT NOPRINT;
   FORECAST LEAD=12 BACK=12
    OUT=B ID=N ;
```

> 1 Univariate models for Gas and Employ series are identified and estimated

> 2 Forecasts for the Rider series are made using the forecasted values for Gas and Employ and the assumed values for Hours and Fare as inputs. Back=12 establishes the forecast origin as June 1981.

```
PROC PLOT DATA=B(FIRSTOBS=103);
  PLOT FORECAST*N='F' LRIDERS*N='*' L95*N='L' U95*N='U'/ OVERLAY
       VPOS=40 HPOS=60;
  TITLE Ridership Forecast, Actual and Confidence Limits;
```

> 3 Forecasts are plotted.

EXHIBIT 15 (continued)

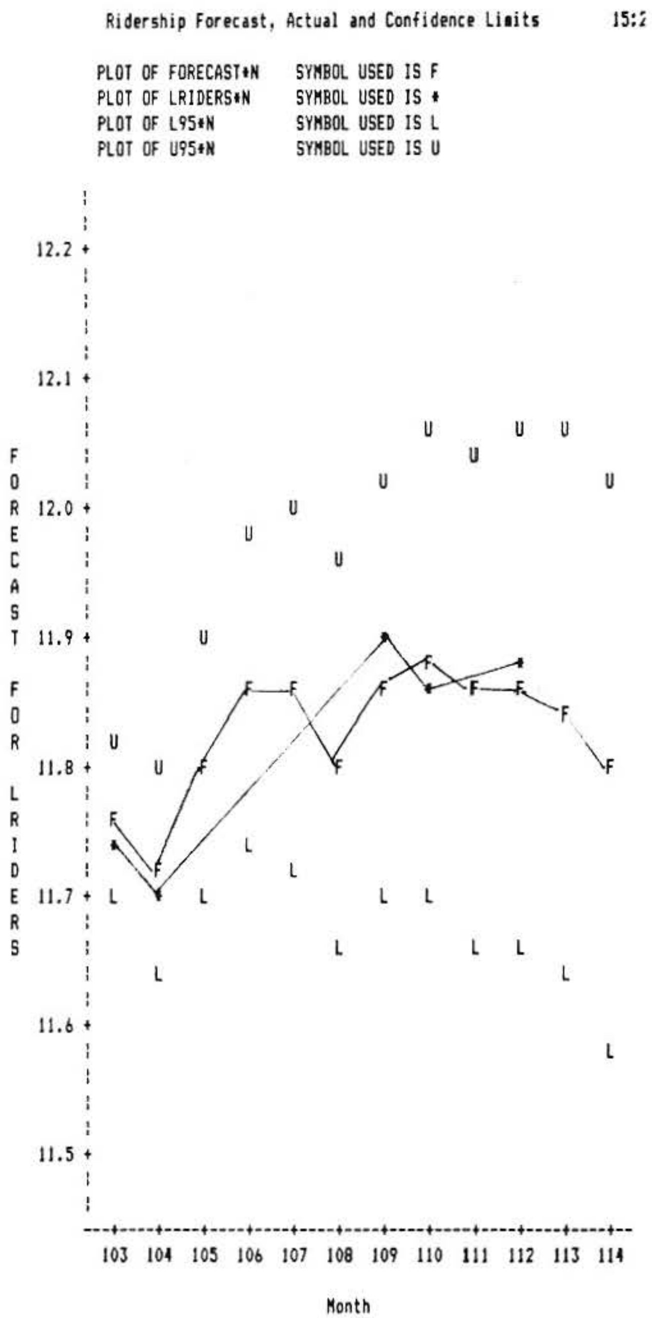
(b) Forecast-ridership series

4 Forecasts for the Rider series are made.

FORECASTS FOR VARIABLE LRIDERS

OBS	FORECAST	STD ERROR	LOWER 95%	UPPER 95%	ACTUAL	RESIDUAL
--------FORECAST BEGINS--------						
103	11.7576	0.0294	11.7000	11.8151	11.7424	-0.0151
104	11.7195	0.0425	11.6362	11.8027	11.7068	-0.0126
105	11.8056	0.0530	11.7018	11.9094	11.7951	-0.0105
106	11.8583	0.0621	11.7366	11.9800	11.8615	0.0032
107	11.8556	0.0703	11.7179	11.9933	11.8615	0.0059
108	11.8030	0.0778	11.6506	11.9555	11.7974	-0.0057
109	11.8561	0.0847	11.6901	12.0221	11.8920	0.0359
110	11.8785	0.0912	11.6998	12.0573	11.8671	-0.0114
111	11.8543	0.0973	11.6636	12.0450	11.8629	0.0086
112	11.8650	0.1030	11.6630	12.0670	11.8720	0.0070
113	11.8447	0.1085	11.6320	12.0574	11.8452	0.0005
114	11.8066	0.1137	11.5837	12.0295	11.7958	-0.0107

(c) Forecast plots



VI. REFERENCES

Box, G.E.P. and Jenkins, G.M. (1976), Time Series Analysis: Forecasting and Control, Revised Edition, San Francisco: Holden-Day.

Box, G.E.P. and Pierce, D.A. (1970). "Distributions of Residual Autocorrelations in Autoregressive-Integrated Moving Average Models", Jour. Amer. Stat. Assoc., 64, pp. 1509-1526.

Box, G.E.P. and Tiao, G.C. (1975), "Intervention Analysis with Application to Economic and Environmental Problems," Journal of the American Statistical Association, 70, 70-79.

Ljung, G.M. and Box, G.E.P. (1978), "On a Measure of Lack of Fit in Time Series Models", Biometrika, 65, pp. 67-72.

Hoff, J.C. (1983), A Practical Guide to Box-Jenkins Forecasting, Lifetime Learning, Belmont, CA.

Parzen, E. (1982). "ARAMA Models for Time Series Analysis and Forecasting", Jour. Forecasting, 1.

APPENDIX A
TIME SERIES COMPUTER PACKAGES

There are a number of time-series computer packages currently available for both main-frame and microcomputer environments. Three packages that have been used by the authors of this report are listed below.

1. SAS (Statistical Analysis System)
   Available from:
   SAS Institute, Inc.
   Box 8000
   Cary, NC

2. SCA System
   Available from:
   Scientific Computing Associates
   P.O. Box 625
   De Kalb, IL 60115

3. BMDP
   Available from:
   BMDP Statistical Software, Inc.
   1964 Westwood Blvd.
   Suite 202
   Los Angeles, CA 90025

APPENDIX B
LISTING OF PORTLAND DATA

TABLE 2

Year	Month	Riders	Hours	Employment	Gasoline Price	Fare
73	1	64800	2714	329800	36.9	34.5
73	2	64600	2714	332700	36.9	34.5
73	3	63900	2714	339500	36.9	34.5
73	4	65400	2761	340300	36.9	34.5
73	5	63000	2761	343100	37.1	34.5
73	6	62200	2761	353000	38.7	34.5
73	7	61700	2761	351000	38.9	34.5
73	8	61300	2755	355000	38.9	34.5
73	9	66100	2791	358800	38.4	34.5
73	10	69500	2791	354100	39.3	32.5
73	11	69000	2830	354700	40.4	32.5
73	12	70700	2830	356800	43.4	32.5
74	1	81700	2866	372800	44.6	32.5
74	2	83900	2909	374900	45.9	32.5
74	3	81000	2909	376500	49.6	32.5
74	4	78900	2933	378800	51.7	32.8
74	5	76000	2933	382700	53.3	32.8
74	6	72400	3172	392700	54.2	32.8
74	7	70400	3172	387100	54.7	32.8
74	8	69100	3172	391700	54.4	32.8
74	9	74500	3229	397800	54.9	32.8
74	10	80300	3229	393700	53.2	32.8
74	11	78000	3229	391000	52.3	32.8
74	12	76100	3229	389800	52.4	32.8
75	1	85700	3664	378200	52.8	29.4
75	2	90700	3664	374300	52.6	29.4
75	3	87300	3731	374000	52.1	29.4
75	4	91000	3731	377800	52.5	29.4
75	5	90000	3731	379000	53.5	29.4
75	6	88000	3748	385900	54.8	29.4
75	7	86700	3748	378000	56.7	29.4
75	8	85400	3748	382300	57.4	29.4
75	9	92800	3860	389400	57.6	29.4
75	10	106400	3860	390500	57.6	27.4
75	11	110300	3860	387800	57.3	27.4
75	12	102600	3976	390400	57.4	27.4
76	1	110200	3976	385900	57.0	27.4
76	2	108000	3976	383500	56.4	27.4
76	3	103400	3976	385700	55.3	27.4
76	4	108300	3976	391300	54.5	27.4
76	5	107800	3976	394100	55.5	27.4
76	6	102000	3976	401600	56.9	27.4
76	7	98400	3976	397500	58.4	30.7
76	8	95200	3976	400100	58.7	30.7
76	9	103300	3976	407100	59.4	30.7
76	10	111400	3976	407300	59.3	30.7
76	11	116000	3976	407400	59.3	29.4
76	12	105800	3976	410200	59.3	29.4
77	1	120900	3976	401700	59.6	29.4
77	2	120000	3967	403000	60.3	29.4
77	3	113000	3967	407000	60.9	29.4
77	4	118200	4183	410700	61.3	29.4
77	5	115200	4183	413600	61.7	29.4
77	6	111600	4171	422400	62.1	29.4
77	7	109800	4171	420200	62.4	29.4
77	8	104400	4171	421500	62.3	29.4
77	9	114200	4171	430200	62.4	29.4

Year	Month	Riders	Hours	Employment	Gasoline Price	Fare
77	10	122200	4171	433100	62.4	29.4
77	11	123400	4171	438400	62.4	29.4
77	12	115500	4183	442000	62.6	29.4
78	1	128600	4183	439900	63.1	30.5
78	2	128100	4183	441700	63.5	30.5
78	3	122400	4183	447800	63.5	30.5
78	4	128000	4183	453900	63.6	30.5
78	5	122800	4183	457200	64.9	30.5
78	6	118100	4183	473300	65.9	30.5
78	7	115600	4183	463100	67.0	30.5
78	8	112400	4183	462800	67.9	30.5
78	9	115200	4183	470200	68.6	35.8
78	10	120500	4183	470600	68.9	35.8
78	11	126000	4183	472500	68.9	35.8
78	12	118800	4183	476800	69.2	35.8
79	1	121200	4183	465300	69.5	35.8
79	2	126900	4165	468200	70.4	35.8
79	3	124600	4165	475500	72.3	35.8
79	4	129900	4165	478800	75.1	35.8
79	5	128400	4165	483200	80.9	36.3
79	6	134500	4165	499300	86.1	36.3
79	7	134100	4481	486400	92.2	36.3
79	8	130800	4481	488000	94.3	36.3
79	9	144800	4527	497200	98.5	36.3
79	10	145400	4527	498900	98.9	36.3
79	11	146700	4527	502400	99.6	36.3
79	12	143100	4527	502900	101.1	36.3
80	1	151000	4527	487100	106.3	36.3
80	2	155800	4562	492100	113.4	36.3
80	3	153600	4562	497000	116.8	36.3
80	4	152300	4573	494900	118.1	42.4
80	5	149200	4573	491200	119.8	42.4
80	6	143700	4657	499200	120.8	42.4
80	7	136500	4657	481500	120.8	42.4
80	8	131000	4657	481600	120.8	42.4
80	9	144100	4657	493100	120.8	42.4
80	10	145000	4657	491100	120.8	49.1
80	11	142400	4657	492200	120.8	49.1
80	12	136000	4657	493000	120.5	49.1
81	1	142900	4657	481600	122.5	49.1
81	2	144000	4657	479600	129.8	49.1
81	3	141400	4657	483600	131.3	49.1
81	4	142400	4657	483300	131.9	49.1
81	5	140800	4657	484300	132.7	49.1
81	6	133700	4788	491000	134.0	49.1
81	7	125800	4788	478900	135.1	49.1
81	8	121400	4788	475900	135.0	49.1
81	9	132600	4788	483600	135.0	49.1
81	10	141700	4788	477293	133.6	49.1
81	11	141700	4788	475793	132.7	49.1
81	12	132900	4788	475080	132.5	49.1
82	1	146100	4788	461279	131.3	49.9
82	2	142500	4788	458140	128.5	49.9
82	3	141900	4788	459172	121.9	49.9
82	4	143200	4788	459347	119.1	49.9
82	5	139411	4788	460563	121.7	49.9
82	6	132700	4788	476336	126.3	49.9