

THE PREDICTIVE DISCRIMINATION OF
AUTOREGRESSIVE TIME SERIES
WITH UNKNOWN ORDER

By

HON RICHARD TACHIA

Bachelor of Science
University of Nigeria
Nsukka, Nigeria
1978

Master of Science
Iowa State University
Ames, Iowa
1983

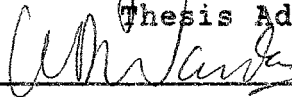
Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 1993

THE PREDICTIVE DISCRIMINATION OF
AUTOREGRESSIVE TIME SERIES
WITH UNKNOWN ORDER

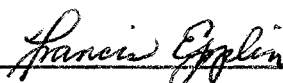
Thesis Approved:



Thesis Adviser









Dean of the Graduate College

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis advisor, Dr. Gary Stevens for believing in me and for his support and encouragement throughout the phase of the dissertation. I am also particularly thankful to Dr. William D. Warde, the chairman of my Ph.D Committee, for his advice, guidance and above all his patience in helping me through the research. I also appreciate the advice and suggestions of Dr. Sarkar, a member of my Ph.D Committee. My special thanks to Dr. Francis Epplin, also a Committee member for being there for me.

The Academic Computing Resources, Wright State University, Dayton, Ohio made the simulation study possible by permitting me the use of the University's IBM Model 3090 computers. For this and other hardware assistance, I say Thank you.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
Statement of the Problem.....	1
Stationary Stochastic Processes.....	2
Research Objectives.....	5
The Methodology of the Proposed Solution.....	7
The Analysis of Time Series.....	9
(i) Time Series Models.....	10
(ii) Describing a Time Series.....	11
(iii) Inference.....	14
(iv) Forecasting and Control.....	17
Frequency Domain Analysis.....	16
II. LITERATURE REVIEW	21
Predictive Analysis.....	25
III. TIME SERIES DISCRIMINATION.....	27
Preliminaries	27
Estimative Discrimination.....	30
Burg Estimation.....	31
Yule-Walker Estimation.....	33
Estimative Discriminant Functions.....	35
Predictive Discrimination.....	36
Probability Distributions for the AR Order.....	36
Joint Prior Probability Density for ϕ and τ	39
Time Domain Discriminant Functions.....	41
Frequency Domain Discriminant Functions.....	45
IV. NUMERICAL STUDY.....	48
V. SUMMARY AND CONCLUSIONS.....	53
Graphical Analyses	55
Conclusions	64
Further Work	66

BIBLIOGRAPHY.....	67
APPENDIXES.....	72
APPENDIX A - SOME FORMULA DERIVATIONS.....	73
APPENDIX B - SAS SOURCE CODE FOR THE SIMULATION STUDY.....	77
APPENDIX C - SUBROUTINES USED IN THE SIMULATION STUDY.....	91
APPENDIX D - TABLES.....	94

Table

I. Error Rates and J-Divergence from AR(1) Versus AR(1) Classification...	96
II. Error Rates and J-Divergence from AR(1) Versus AR(2) Classification...	98
III. Error Rates and J-Divergence from AR(2) Versus AR(2) Classification...	100

LIST OF FIGURES

Figures for AR(1) Versus AR(1) Discrimination	102
Figure	Page
1. Spectral Densities of AR(1) Series	103
2. Box and Whisker Plot	104
3. Quantile Plot in Frequency Domain	105
4. Quantile Plot in Time Domain	106
5. Error Rates and Lengths of Training Data and Test Realizations in Frequency Domain	107
6. Error Rates and Lengths of Training Data and Test Realizations in Time Domain	108
7. Error Rates and J-Divergence in Frequency Domain..	109
8. Error Rates and J-Divergence in Time Domain	110
9. Improvement of Frequency over Time Domain	111
10. Improvement Rate and J-Divergence in Frequency Domain	112
11. Improvement Rate and J-Divergence in Time Domain	113
Figures for AR(2) versus AR(1) Classification	114
12. Spectral Densities of AR(1) Series	115
13. Box and Whisker Plot	116
14. Quantile Plot in Frequency Domain	117
15. Quantile Plot in Time Domain	118

16.	Error Rates and Lengths of Training Data and Test Realizations in Frequency Domain	119
17.	Error Rates and Lengths of Training Data and Test Realizations in Time Domain	120
18.	Error Rates and J-Divergence in Frequency Domain..	121
19.	Error Rates and J-Divergence in Time Domain	122
20.	Improvement of Frequency over Time Domain	123
21.	Improvement Rate and J-Divergence in Frequency Domain	124
22.	Improvement Rate and J-Divergence in Time Domain	125
	Figures for AR(2) versus AR(2) Classification	126
23.	Spectral Densities of AR(1) Series	127
24.	Box and Whisker Plot	128
25.	Quantile Plot in Frequency Domain	129
26.	Quantile Plot in Time Domain	130
27.	Error Rates and Lengths of Training Data and Test Realizations in Frequency Domain	131
28.	Error Rates and Lengths of Training Data and Test Realizations in Time Domain	132
29.	Error Rates and J-Divergence in Frequency Domain..	133
30.	Error Rates and J-Divergence in Time Domain	134
31.	Improvement of Frequency over Time Domain	135
32.	Improvement Rate and J-Divergence in Frequency Domain.....	136
33.	Improvement Rate and J-Divergence in Time Domain.....	137

CHAPTER I

INTRODUCTION

Statement of the Problem

We are often faced with the problem of having to decide what classification to assign to a set of observations given that the observations are derived from one of several classes or groups. In Geophysics for example Shumway (1982) and Tjostheim (1975) have considered the classification of a given seismic pattern as coming from an earthquake or a nuclear explosion. In Engineering, the problem is that of detecting a radar signal or equivalently discriminating between a pattern generated by signal plus noise and noise alone. Similar questions exist in such diverse areas as Medicine, Speech and Agriculture.

In all of the above circumstances we are required to use the available information about a given set of classes to identify the origin of a new data set according to some specified criterion. In this study we are interested in the classification of a time series that obeys the autoregressive model.

We are concerned with the problem of classifying a

given finite empirical time series $Y_N = (y(1), y(2), \dots, y(N))'$ into one of two mutually exclusive autoregressive (AR) populations (or classes) \mathcal{S}_1 and \mathcal{S}_2 having unknown and possibly different orders p_1 and p_2 respectively. Y_N is generated by a stochastic process from one of these classes and obeys the AR equation given by

$$\sum_{j=0}^p \varphi_j y(t-j) = \varepsilon(t), \quad t=p+1, \dots, N, \quad (1)$$

where

$\varphi_0=1$, $\{\varepsilon(t)\}$ is a sequence of independent and identically distributed normal random variables with mean zero and variance $1/\tau$, $\tau > 0$, and $\varphi = (\varphi_1, \dots, \varphi_p)'$, the set of AR coefficients, is a vector of solutions to the difference equation (1) such that the process $\{y(\cdot)\}$ is weakly stationary. We note that the difference equation in (1) may also be written as the following regression equation: $y(t) = \varphi_1 y(t-1) + \varphi_2 y(t-2) \dots \varphi_p y(t-p) + \varepsilon(t)$, $t=p+1, \dots, N$. The stationarity of a stochastic process is defined below.

Stationary Stochastic Processes

A real-valued stochastic process $\{y(t), t=0, \pm 1, \pm 2, \dots\}$ is said to be strictly or (strongly) stationary in distribution if for any $n=1, 2, \dots$ and any n -tuple (t_1, \dots, t_n) and k integers,

$$F(y(t_1), \dots, y(t_n)) = F(y(t_1+k), \dots, y(t_n+k)),$$

where

$F(y(t_1), \dots, y(t_n))$ is the n -dimensional distribution function of $y(t_1), \dots, y(t_n)$.

In this study, a weaker sense of stationarity will do. A stochastic process is said to be weakly or covariance stationary if the first two moments exist and are time invariant. This definition means that a covariance stationary process will have constant mean, variance and covariance. Henceforth the term stationary will refer to covariance stationary.

The stochastic process may thus be characterized by the parameter

$$\theta_i = (\phi_i', \tau_i, p_i)', \quad i=1,2.$$

That is, if Y_N comes from class \mathfrak{s}_i , this fact is symbolized by

$$Y_N \sim AR(p_i), \quad i=1,2.$$

The classification problem is to formulate a decision rule which divides the observation vector Y_N into two disjoint regions (C_1, C_2) so as to minimize the probability of misclassification of Y_N into one of the classes \mathfrak{s}_1 and \mathfrak{s}_2 .

The probability of misclassifying Y_N that originates from class \mathfrak{s}_1 is given by

$$\int_{C_2} P(Y_N | \theta_1) dY_N,$$

where $P(Y_N|\theta_1)$ is the conditional probability of observing Y_N given that Y_N originates from class \mathcal{S}_1 and $dY_N = dy(1)\dots dy(N)$. Further, suppose m_1 denotes the unconditional probability that Y_N comes from class \mathcal{S}_1 and $c(1|2)$ is the penalty or cost for misclassifying Y_N as coming from \mathcal{S}_1 when it in fact comes from \mathcal{S}_2 . Then the expected cost of misclassifying Y_N that originates from \mathcal{S}_1 is given by

$$c(1|2)m_1 \int_{C_2} P(Y_N|\theta_1) dY_N.$$

Defining m_2 and $c(2|1)$ similarly, the expected cost of misclassifying Y_N is

$$c(1|2)m_1 \int_{C_2} P(Y_N|\theta_1) dY_N + c(2|1)m_2 \int_{C_1} P(Y_N|\theta_2) dY_N.$$

For specified values of m_1 and m_2 , the Bayes solution to the classification problem is obtained (Anderson, 1984) by assigning Y_N to class \mathcal{S}_1 if Y_N falls in the region defined by

$$C_1^* = \left\{ Y_N : \frac{P(Y_N|\theta_1)}{P(Y_N|\theta_2)} \geq \frac{m_2 c(1|2)}{m_1 c(2|1)} \right\};$$

otherwise assign Y_N to \mathcal{S}_2 .

If $m_1 = m_2$, that is, the two classes are equally likely and if equal costs of misclassification are incurred, then the above assignment becomes

$$C_1^* = \{Y_N : \delta_{12}(Y_N; \theta_1, \theta_2) \geq 0\}, \quad (2)$$

where

$$\delta_{12}(Y_N ; \theta_1, \theta_2) = \ln \left(\frac{P(Y_N | \theta_1)}{P(Y_N | \theta_2)} \right) \quad (3)$$

is called the Discriminant Function.

Now $\delta_{12}(Y_N; \theta_1, \theta_2)$ is unknown since it depends on the unknown AR parameters θ_1 and θ_2 . The classical procedures estimate $\delta_{12}(Y_N; \theta_1, \theta_2)$ by first estimating θ_1 and θ_2 and then replacing θ_1 and θ_2 with their estimates. These estimation procedures are referred to as estimative methods and include the techniques of maximum likelihood, least squares, Burg, and Yule-Walker. This study proposes an alternative approach to the approximation of $\delta_{12}(Y_N; \theta_1, \theta_2)$.

The alternative procedure estimates $\delta_{12}(Y_N; \theta_1, \theta_2)$ by replacing $P(Y_N | \theta_i)$ with $P(Y_N | X_i)$, called the predictive probability density of Y_N given the training realization $X_i = (x(1), x(2), \dots, x(N_i))$ from class θ_i , $i=1,2$. This classification procedure is performed in both time and frequency domains and is the basis of this investigation whose objectives are given in the following section.

Research Objectives

1. Identify prior probability densities for the AR parameter $\theta = (\varphi, \tau, p)$. These priors are given in Chapter III. The priors proposed for the AR order p are new and to the author's knowledge, have not been used elsewhere in

the literature.

2. Use the priors in objective 1 above to derive time domain and frequency domain predictive discriminant functions for the classification of the test realization Y_N . The derived functions are unique and new because no value of the AR order is assumed; other forms of the predictive discriminant functions assume known values of the order p .

3. Conduct a simulation study to evaluate the relative performances of the estimative and predictive discrimination procedures. To the author's knowledge no such study has been done previously. The simulation study is done in Chapter IV.

4. Compute the J-divergence rate (Shumway and Unger, 1974). The J-divergence rate gives a measure of distance (the amount of information available for discriminating) between the two classes \mathcal{S}_1 and \mathcal{S}_2 . The J-divergence rate is defined as

$$J(1,2;Y_N)=I(1;Y_N)+I(2;Y_N)$$

where

$$I(1;Y_N)=E_1\left\{\frac{f_Y(\omega)}{f_1(\omega)} - \ln\left[\frac{f_Y(\omega)}{f_1(\omega)}\right] - 1\right\},$$

$$I(2;Y_N)=E_2\left\{\frac{f_Y(\omega)}{f_2(\omega)} - \ln\left[\frac{f_Y(\omega)}{f_2(\omega)}\right] - 1\right\},$$

and E_i is expectation with respect to class \mathcal{S}_i , $i=1,2$.

$I(i;Y)$ is the Kullback-Liebler discrimination information

(Parzen, 1982) for measuring the distance between the spectral densities $f_Y(\omega)$ of Y_N and $f_i(\omega)$ of X_i , the training realization from class \mathcal{S}_i , $i=1,2$.

Predictive discrimination in the time domain involves the use of the predictive density of Y_N in the evaluation of the discriminant function. In the frequency domain, the discriminant function is based on the spectral density of Y_N evaluated at the frequencies $0, 1/N, \dots, (N-1)/N$.

The Methodology of the Proposed Solution

The basis for this research is the notion of predictive discrimination from the works of Geisser (1964), Dunsmore (1966), and Aitchison and Dunsmore (1975). The essence of this procedure is that $P(Y_N|\theta_i)$ in the discriminant function in equation (2) is replaced by the predictive density $P(Y_N|X_i)$ defined by

$$\begin{aligned}
 P(Y_N|X_i) &= \frac{\int_{\theta_i} P(Y_N, \theta_i, X_i) d\theta}{P(X_i)} \\
 &= \frac{\int_{\theta_i} P(Y_N|\theta_i, X_i) P(X_i|\theta_i) P(\theta_i) d\theta_i}{P(X_i)} \\
 &= \frac{\int_{\theta_i} P(Y_N|\theta_i) P(X_i|\theta_i) P(\theta_i) d\theta_i}{P(X_i)}, \tag{4}
 \end{aligned}$$

where θ_i is the support for θ_i , the test realization Y_N is independent of the training realization X_i from class \mathcal{S}_i , $P(X_i)$ and $P(\theta_i)$ are respectively the marginal probability densities of X_i and θ_i ; given the AR parameter θ_i from class \mathcal{S}_i , $P(Y_N|\theta_i)$ and $P(X_i|\theta_i)$ are respectively the conditional probability densities of Y_N and X_i .

The primary purpose of this study is the classification rather than the identification of a given autoregressive time series of unknown order p . Hence p will be assigned an a priori probability density which will be eventually summed out in the course of the analysis. If the order p is known, or a reliable estimate exists, then the prior density becomes unnecessary and such knowledge would lend to a substantial reduction of the computational effort.

Hermans and Habbema (1975) have demonstrated that in discriminating between two normal populations, the estimative procedure of maximum likelihood and the predictive discrimination have approximately equal error rates when there are a large number of training realizations. In the case of small sample sizes, however, simulation studies by Aitchison et al (1977) and Moran and Murphy (1979) have shown that the predictive approach has a lower error rate than the maximum likelihood procedure in discriminating between two multivariate normal populations. Aitchison and Dunsmore (1975) explain this

discrepancy by the fact that whereas the estimative procedures ignore the sampling variability of $\hat{\theta}_i(X_i)$ (that is, θ_i estimated from the training data X_i generated from class \mathcal{S}_i), the predictive density weights the possible distributions of $P(Y_N|\mathcal{S}_i)$ on the plausible values of θ_i , $i=1,2$. In our case, we will examine the error rates from the predictive and estimative procedures for the classification of time series. In particular, we will be interested in finding out if the conclusions stated above hold when we discriminate time series data from an autoregressive process. It is also of interest to relate the classification of a time series to the overall scheme of time series analysis. This relationship is best explained by first stating the usual objectives of time series analysis.

The Analysis of Time Series

The study of time series usually starts with a determination of the model that best describes the series. The representations for time series are defined in terms of the stochastic processes, linear or nonlinear, that give rise to the series. This study considers only linear processes and may be represented by moving average (MA), autoregressive (AR) or a combination of these, referred to autoregressive moving average (ARMA) models. These models are defined in the next section.

(i) Time Series Models

Moving Average Processes. The moving average model for the time series $y(t)$ is a linear combination of a sequence of uncorrelated random variables given by

$$y(t) = \mu + \varepsilon(t) + \psi_1 \varepsilon(t-1) + \psi_2 \varepsilon(t-2) + \dots$$

$$= \mu \sum_{j=0}^{\infty} \psi_j \varepsilon(t-j) ,$$

where $\psi_0 = 1$, $\{\varepsilon(t)\}$ is a sequence of uncorrelated random variables from a fixed distribution with constant mean and variance. $\{\varepsilon(t)\}$ is called a white noise process.

If only q of the ψ weights are nonzero, that is, $\psi_k = 0$, if $k > q$, then the resulting process is called a moving average process of order q and is denoted as $MA(q)$.

Autoregressive (AR) Processes. The autoregressive model for $y(t)$ is obtained by regressing $y(t)$ on its past values and a white noise process, $\{\varepsilon(t)\}$. That is,

$$y(t) = \phi_1 y(t-1) + \phi_2 y(t-2) + \dots + \varepsilon(t)$$

$$= \sum_{j=1}^{\infty} \phi_j y(t-j) + \varepsilon(t) .$$

If only p of the ϕ weights in the above representation are nonzero, then the resulting process is said to be an autoregressive process of order p , and is denoted as $AR(p)$.

Autoregressive moving average (ARMA) Processes. A

difficulty that is often encountered in restricting a time series model to only the autoregressive or the moving average is that a very large number of parameters may be needed. An alternative to either model is the mixed autoregressive moving average ARMA(p,q) given by

$$y(t) = \phi_1 y(t-1) + \phi_2 y(t-2) + \dots + \phi_p y(t-p) + \psi_1 \varepsilon(t-1) + \psi_2 \varepsilon(t-2) + \dots + \psi_q \varepsilon(t-q) + \varepsilon(t).$$

The analysis of a given time series typically involves descriptive, inferential, forecasting and control procedures. The basic features of each of these analytical schemes are given next.

(ii) Describing a Time Series

Sample statistics and graphs are used to describe a time series in order to have a better understanding of the stochastic process that generated the series. In the time domain analysis, some of the statistics that are used most often include the autocorrelation coefficient, the partial autocorrelation coefficient and the autocovariance.

Given a time series $Y_n = (y(1), y(2), \dots, y(n))'$ of length n , the autocorrelation coefficient of lag ν is defined by

$$\rho(\nu) = \frac{\gamma(\nu)}{\gamma(0)}, \quad \nu = 1, 2, \dots, p,$$

where $\gamma(\nu) = \text{Cov}(y(t+\nu), y(t))$ is the autocovariance function of lag ν ; $\gamma(0)$, the autocovariance of lag 0 is

the variance of the series. The sample estimate of $\rho(\nu)$ is given by

$$\hat{\rho}(\nu) = \frac{\hat{\gamma}(\nu)}{\hat{\gamma}(0)}, \quad (4)$$

where

$$\hat{\gamma}(\nu) = \frac{1}{n} \sum_{t=1}^{n-\nu} (y(t+\nu) - \bar{y})(y(t) - \bar{y}), \quad \nu < n,$$

and

$$\bar{y} = \frac{1}{n} \sum_{t=1}^n y(t).$$

A plot of $\hat{\rho}(\nu)$ versus ν is called the correlogram. The correlogram is useful for identifying an MA(q) process. If the correlogram of a time series vanishes after lag q, then the series may be identified as having the moving average structure with lag q.

The partial autocorrelation coefficient of lag ν is the correlation coefficient between $y(t)$ and $y(t+\nu)$ after eliminating the linear effect of $y(t+1), \dots, y(t+\nu-1)$. Given the autoregressive series of order p, the sample partial autocorrelation coefficient, denoted $\hat{\pi}_j$, $j=1, 2, \dots, p$, is obtained by simultaneously solving the set of linear equations:

$$\hat{\rho}(k) = \sum_{j=1}^p \hat{\pi}_j \hat{\rho}(k-j), \quad k=1, \dots, p.$$

The graph of $\hat{\pi}_j$ versus j is called the sample partial

autocorrelogram and drops to zero for all $j > p$. The partial correlogram is extremely useful for a preliminary estimate of the AR order.

Other descriptive measures usually considered in a preliminary time series analysis are given in the frequency domain. The basic theory of the frequency domain (or spectral analysis) of time series is given in the last section of this chapter. In this section, we will only define and interpret the spectral density function, the (cumulative) spectral distribution function, the periodogram and the cumulative periodogram.

The sample spectral density of $Y_n = (y(1), \dots, y(n))'$ for the frequency ω_k is

$$\hat{f}(\omega_k) = \frac{1}{n} \left| \sum_{t=1}^n y(t) \exp(2\pi j(t-1)\omega_k) \right|^2,$$

where $j = \sqrt{-1}$, $\omega_k = \frac{k-1}{n}$, $k=1, 2, \dots, [n/2]+1$ and $[x]$ is the largest integer less than or equal to x . The graph of $\hat{f}(\omega_k)$ versus ω_k is called the periodogram of Y_n . The periodogram exhibits peaks at frequencies that correspond to periodicities in the series Y_n and is thus useful for determining the periodicities of an AR series. The periodogram of a white noise (purely random) series is flat and without any peaks. This flatness property may be utilized for a preliminary identification of a white noise

time series.

There are equivalent expressions for the ordinate of the periodogram that are often used. These include the standardized periodogram ordinates

$$\frac{\hat{f}(\omega_k)}{\hat{\sigma}^2} \quad \text{and} \quad \log \left\{ \frac{\hat{f}(\omega_k)}{\hat{\sigma}^2} \right\},$$

where

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^n (y(t) - \bar{y})^2 \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{t=1}^n y(t).$$

The sample spectral distribution function of Y_n is

$$\hat{F}(\omega_k) = \frac{\sum_{j=1}^k \hat{f}(\omega_j)}{\sum_{j=1}^q \hat{f}(\omega_j)}, \quad k=1,2,\dots,q \text{ and } q=[n/2]+1.$$

Jumps in the cumulative periodogram, the plot of $F(\omega_k)$ versus ω_k , at various values of ω_k correspond to periodicities at the respective frequencies.

(iii) Inference

The statistical inferences in time series analysis deal for the most part with the estimation and testing of time series models as well as the distributional properties of estimators. Both parametric and nonparametric techniques for model estimation are well

covered in most texts on time series analysis. These include Box and Jenkins (1976), Priestley (1981), Diggle (1990) and Wei (1990). In this study, we summarize, for fixed order p , the estimation of ϕ and τ for an $AR(\phi, \tau, p)$. The estimation procedures are due to Burg (1967, 1968) and Yule-Walker (Box and Jenkins (1976)) and are briefly stated next.

The Yule-Walker estimate for ϕ is obtained by replacing ρ and the matrix P with the sample estimates in the following Yule-Walker equations:

$$P\phi = \rho,$$

where

$$P = \begin{pmatrix} 1 & \rho(1) & \dots & \rho(p-1) \\ \rho(1) & 1 & \dots & \rho(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(p-1) & \rho(p-2) & \dots & 1 \end{pmatrix},$$

$\rho = (1, \rho(1), \rho(2), \dots, \rho(p-1))'$, $\hat{\rho}(\nu)$, $\nu=0, 1, \dots, p-1$ is as defined in equation (4). The estimated white noise variance $\hat{\tau}^{-1}$ is given by

$$\hat{\tau}^{-1} = \sum_{j=0}^p \hat{\phi}(j) \hat{\gamma}(j),$$

where $\hat{\gamma}(j)$ is the sample covariance function is as defined by equation (4). The other estimative procedure is due to Burg (1967, 1968).

The expression for $\hat{\tau}^{-1}$ above is readily obtained from multiplying

$$y(t) = \phi y(t-1) + \phi y(t-2) + \dots + \phi y(t-p) + \varepsilon(t)$$

by $y(t)$ and taking expectations, noting that $E(y(t)\varepsilon(t)) = \tau^{-1}$, the white noise variance.

If $Y_n \sim \text{AR}(p)$ and p is fixed, ϕ and τ may be estimated by an entropy-based procedure due to Burg (1967, 1968). Burg's procedure has been found to produce a superior estimator to that due to Yule-Walker in the case of small sample sizes, and also when Y_n is close to being nonstationary. The computational algorithms for the Burg and Yule-Walker procedures are given in Chapter III, section 3.1. If the order p of the AR series is fixed and unknown, numerous procedures exist for the estimation of p .

The estimation of the AR order p of a time series has received considerable attention from researchers in time series analysis. The vast literature on this topic deals with various estimation criteria, many of which are related. The AIC (Akaike Information Criterion) of Akaike (1971, 1974), and various forms of it, have become some of the most widely accepted of the criteria. The AIC for a time series of order j is defined as

$$\text{AIC}(j) = -2\log L(\hat{\alpha}) + 2j, \quad j=0,1,2,\dots$$

where $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_j)'$ are the maximum likelihood estimates of the model parameter α and $L(\hat{\alpha})$ is the likelihood function of α . The AIC determines the order by selecting the j for which $\text{AIC}(j)$ is a minimum.

Shibata (1976) has shown that the AIC is inconsistent and tends to overestimate the true order no matter how large the sample size. Schwarz (1978) and Hannan and Quinn (1979) have modified the AIC to ensure consistency by replacing $2j$ in $AIC(j)$ with $\log(n)$ and $\log(\log(n))$ respectively. Lutkepohl (1985) has used simulation studies of certain multivariate processes to compare most of the current order estimation criteria. Having understood the underlying stochastic process that generates a time series and having estimates of the model parameters, we would want to use such information to predict or forecast, and possibly control, the future values of the time series.

(iv) Forecasting and Control

An important component of time series analysis is forecasting, that is, predicting the future behavior of a time series. The ability to correctly forecast a process enables us to control, or at least prepare an appropriate response to, the future behavior of the process. The importance attached to forecasting is reflected in the considerable literature on forecasting techniques. Standard forecasting techniques such as the Box-Jenkins approach may be found in most texts on time series analysis such as Box and Jenkins (1976) and Anderson (1984). In this study, we are concerned with identifying

the origin of a given series rather than the forecasting of future values of that series. We are also interested in how time series discrimination relates to the classical analysis of time series.

The problem of time series discrimination is essentially an inferential procedure since, in contrast to forecasting, discriminant analysis and the conclusions therein pertain only to the current observation at hand. Predictive discrimination differs from the standard or classical discriminant analysis only in the way that the discriminant function in equation (3) is approximated.

Frequency Domain Analysis

Frequency domain analysis may be defined as inference regarding the spectral density function. The principal concept in the frequency domain (or spectral) analysis of a time series is that the series can be expressed in terms of independent sinusoids. A sinusoid is a combination of sines and cosines. In general a discrete stationary time series $y(t)$ measured at unit intervals has the spectral representation (Priestley (1981), Newton (1988)) given by the stochastic integral

$$y(t) = \int_0^1 \cos(2\pi t\omega) du(\omega) + \int_0^1 \sin(2\pi t\omega) dv(\omega),$$

where $u(\omega)$ and $v(\omega)$ are uncorrelated stochastic processes

with orthogonal increments. The expression of $y(t)$ as the sum of independent frequency components is analogous to the analysis of variance where the effect of a treatment on an experimental unit may be viewed as the sum of linear, quadratic and higher-order effects that are statistically independent. An important time domain measure that has a spectral representation is the autocovariance function $\gamma(\nu)$:

$$\gamma(\nu) = \int_0^1 \cos(2\pi\nu\omega) dF(\omega), \quad \nu=0,1,2,\dots,$$

where $F(\omega)$ is called the spectral distribution function and represents the contribution to the variance of the series by all the frequencies in the range $(0,\omega)$. The total variation of the series is thus $F(1)$ given by

$$F(1) = \gamma(0) = \text{variance of } y(t).$$

For a discrete stationary process, $F(\omega)$ is a continuous function on $(0,1)$ and may therefore be differentiated with respect to ω in $(0,1)$. Hence

$$f(\omega) = \frac{dF(\omega)}{d\omega},$$

where $f(\omega)$ is called the spectral density function or simply the spectrum. $\gamma(\nu)$ may therefore be expressed as

$$\gamma(\nu) = \int_0^1 \cos(2\pi\nu\omega) f(\omega) d\omega, \quad \nu=0,1,2,\dots$$

The quantity $f(\omega)$ represents the contribution to the

variance of components with frequencies in $(\omega, \omega + d\omega)$. Thus a peak in the periodogram implies an important contribution from the frequencies in that interval. It turns out that just as the autocovariance function $\gamma(\omega)$ of a stationary stochastic process can be expressed in terms of $f(\omega)$ as a cosine transform, an inverse relation exists whereby $f(\omega)$ is the following Fourier transform of $\gamma(\nu)$:

$$f(\omega) = \frac{1}{\pi} \sum_{\nu=-\infty}^{\infty} \gamma(\nu) e^{i\omega\nu}.$$

The autocovariance function and the spectral density function are thus a Fourier pair.

Chapter II provides a historical perspective to the classification problem with particular attention to time series and predictive analysis. The main results of this study are given in chapter III.

CHAPTER II

LITERATURE REVIEW

Time series has been studied since the 1920's but most of the literature is devoted almost exclusively to model estimation and forecasting techniques. A series of observations indexed in time often produces a pattern which may form a basis for discriminating among different classes of events. The Discriminant Analysis of Time Series may be studied in the time domain or the frequency domain. McLachlan's book (1992) offers not only a good account of the recent developments in discriminant analysis but also provides an extensive bibliography of the literature in the field.

The problem in time series discrimination in time domain is that one observes a discrete parameter time series $\{y(t), t=1, \dots, N\}$ at each of N points in time with the objective of classifying the observed series into one of two mutually exclusive and exhaustive categories \mathcal{S}_1 and \mathcal{S}_2 . The sampled time series is conveniently represented as an $N \times 1$ vector $\mathbf{Y}_N = (y(1), y(2), \dots, y(N))'$. The classification problem then reduces to one that is well covered in standard texts on Multivariate Statistical Analysis such

as Anderson (1984). The standard optimal classification rule (Anderson, 1984) is to assign Y_N to

$$\begin{cases} \mathfrak{g}_1, & \text{if } \frac{P(Y_N|\theta_1)}{P(Y_N|\theta_2)} \geq c ; \\ \mathfrak{g}_2, & \text{otherwise.} \end{cases}$$

When $P(Y_N|\theta_i)$ are Multivariate Normal, $MVN(\mu_i, \Sigma_i)$, $i=1,2$ the above rule is reduced to Wald's (1944) criterion which, after simplification, classifies Y_N into

$$\begin{cases} \mathfrak{g}_1, & \text{if } W(Y_N) \geq \log(c); \\ \mathfrak{g}_2, & \text{otherwise,} \end{cases}$$

where the discriminant function $W(Y_N)$ is given by

$$W(Y_N) = Y' \lambda - \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

and

$$\lambda = \Sigma^{-1} (\mu_1 - \mu_2).$$

The discriminant function $W(Y_N)$ will provide the basis for assessing the error rate of our rule and will be estimated from the training realizations. The error rate is the estimated probability of misclassification and in this study will be defined as the proportion of misclassifications to the total number of test realizations. The distributional properties of $W(Y_N)$ have been studied by Wald (1944) and Anderson (1951). An alternative classification rule based on the likelihood ratio criterion involves the discriminant function $Z(Y_N)$ given by

$$Z(Y_N) = \frac{n + \frac{N_2}{N_2+1} (\bar{Y}_N - \bar{X}_2)' S^{-1} (\bar{Y}_N - \bar{X}_2)}{n + \frac{N_1}{N_1+1} (\bar{Y}_N - \bar{X}_1)' S^{-1} (\bar{Y}_N - \bar{X}_1)}$$

where \bar{X}_i is the mean vector of the training realizations averaged over N_i units, $i=1,2$, \bar{Y}_N is the mean vector of the test realization averaged over N units, S is the covariance matrix of the test realization Y_N and $n=N_1+N_2$.

The distribution of $Z(Y_N)$ is known (Anderson, 1984) to be asymptotically equivalent to that of $W(Y_N)$. However the immense difficulty in calculating their exact distributions has turned efforts to their limiting distributions.

Okamoto (1963, 1973) has obtained the asymptotic expansions of $W(Y_N)$ and $Z(Y_N)$ to terms of order n^{-2} , while Siotani and Wang (1975, 1977) have extended the expansions to terms of order n^{-3} . The complexity in evaluating these probability distributions coupled with the cumbersome matrix calculations involved have severely limited the use of these time domain rules. An attractive alternative form of analysis is found in spectral (or frequency domain) analysis. The use of spectral approximations in statistical discrimination has been fairly standard since the early 1950's. Shumway (1982) has used approximations by Wahba (1968), Liggett (1971) and Shumway and Unger (1974) to discriminate between earthquakes and nuclear explosions. Tjostheim (1975) and Tylor and Marshall (1991) have

utilized autoregressive techniques to classify earthquakes and nuclear explosions. In Dargahi-Noubary and Laycock (1981) the Kullback-Leibler information measure has been used to identify the relevant frequency bands for the discrimination of stationary time series while an estimate of the error rate is derived from spectral ratios. Dargahi-Noubary (1992) has further shown that if classes have equal mean functions, discrimination based on the frequency with the largest Kullback-Liebler information is equivalent to classification based on the best linear discriminant.

Kashyap (1978) has however demonstrated that in discriminating among autoregressive processes the methodology of time domain analysis can be simplified to a useful form. Utilizing the autoregressive structure of the observation vector Y_N , he derived an optimal feature which is not only amenable to easy computation but more importantly possesses all the information contained in Y_N that is relevant for classification. The resulting optimal classification rule, unlike the procedure suggested by Wald (1944) and related forms of it, is not quadratic in all the feature components, even for large N . Krzysko (1983) has extended Kashyap's (1978) results to multivariate autoregressive processes and, has also investigated the true order of the multivariate autoregressive equation by minimizing the posterior risk.

The time domain analysis portion of this study will be done along the lines of Kashyap (1978) and Krzysko (1983) but will incorporate the notion of predictive analysis in discriminating between autoregressive classes. The spectral analysis will be treated principally along the lines of Cook (1985), Dargahi-Noubary and Laycock (1981) and Dargahi-Noubary (1992). This study incorporates predictive analysis in the discrimination of time series. A brief discussion of the basic concept of the predictive procedure is appropriate at this stage.

Predictive Analysis

The essential feature of Statistical Predictive Analysis is that from the information at our disposal we wish to make some reasoned statement about a future observation. Formally the predictive density of a future observation given the available data is obtained as follows:

$$P(\text{future}|\text{data}) = \int P(\text{future}|\text{data}, \text{parameters})$$

$$.P(\text{data}|\text{parameters}).d(\text{parameters})$$

Jeffreys (1961), by the formulation above and, Fisher (1935) from the fiducial argument, derived the predictive density for observations following a univariate normal density. Zellner and Chetty (1965) derived the predictive distributions for the multivariate regression model. Geisser (1964, 1966, 1982) has applied the multivariate

normal extension of Jeffrey's (1961) derivation to discriminant analysis.

The Bayesian predictive discrimination of time series overlaps with the general Bayesian analysis of time series as formulated by Zellner (1971). Harrison and Stevens (1976) and Chow (1975) derived Bayesian forecasting techniques that are based on the Bayesian predictive distribution of future observations. The use of the predictive density function for forecasting has been studied further by Shaarawy and Broemeling (1984) and by Broemeling and Land (1984).

In this study we are interested in the use of predictive densities for the discrimination of autoregressive processes of unknown and possibly different orders. The various analytical procedures in both the time and frequency domains are given in the next chapter.

CHAPTER III

TIME SERIES DISCRIMINATION

Preliminaries

The observed time series vector $Y_N = (y(1), y(2), \dots, y(N))'$ from the AR process of unknown order p may be partitioned into

$$Y_N = (Y_0, Z)',$$

where $Y_0 = (y(1), \dots, y(p))'$ serves as the set of initial conditions to the difference equation (1) in Chapter I and $Z = (y(p+1), \dots, y(N))'$. Hence the difference equation (1) may be expressed as

$$Z = W\varphi + \varepsilon, \tag{5}$$

where

$$W_{(N-p) \times p} = \begin{pmatrix} y(p) & y(p+1) & \dots & y(1) \\ y(p+1) & y(p) & \dots & y(2) \\ \vdots & \vdots & \ddots & \vdots \\ y(N-1) & y(N-2) & \dots & y(N-p) \end{pmatrix},$$

$$\varphi = (\varphi_1, \varphi_2, \dots, \varphi_p)'$$

and

$$\varepsilon = (\varepsilon(p+1), \varepsilon(p+2), \dots, \varepsilon(N))'.$$

Since Y_0 contains virtually no information about θ , the distribution of Y_N can be expressed as (Kashyap (1978)):

$$P(Y_N | \theta) = \left(\frac{\tau}{2\pi} \right)^{(N-p)/2} \exp \left[-\frac{\tau}{2} (Z - W\varphi)' (Z - W\varphi) \right] P(Y_0), \tag{6}$$

where $P(Y_0) = P(Y_0|\theta)$.

We introduce the following notation: The statistics with $\hat{\cdot}$ and $\tilde{\cdot}$ are computed respectively from the training data X_i from the class ϑ_i , $i=1,2$ and from the test realization Y_N to be classified. In particular we define the following:

$$\begin{aligned}\tilde{\varphi} &= (W'W)^{-1}W'Z, \\ \tilde{\tau}^{-1} &= (1/N)(Z-W\tilde{\varphi})'(Z-W\tilde{\varphi}).\end{aligned}$$

$$\begin{aligned}\text{Now, from } (Z-W\varphi)'(Z-W\varphi) &= [(Z-W\tilde{\varphi})-(W\varphi-W\tilde{\varphi})]'[(Z-W\tilde{\varphi})-(W\varphi-W\tilde{\varphi})] \\ &= (Z-W\tilde{\varphi})'(Z-W\tilde{\varphi}) + (\varphi-\tilde{\varphi})'W'W(\varphi-\tilde{\varphi}),\end{aligned}$$

we may equivalently express the probability density of Y_N from equation (4) as

$$P(Y_N|\theta) = P(Y_0) \left(\frac{\tau}{2\pi} \right)^{(N-p)/2} \exp \left\{ -\frac{\tau}{2} \{ N\tilde{\tau}^{-1} + (\varphi-\tilde{\varphi})'W'W(\varphi-\tilde{\varphi}) \} \right\} \quad (7)$$

We may similarly express the probability density function of $X_i = (x(1), x(2), \dots, x(N_i))'$ together with the corresponding statistics $\hat{\varphi}_i$ and $\hat{\tau}_i^{-1}$. The distributional form of Y_N and X_i defined above will be utilized in the derivation of the discriminant functions in the time domain. The rest of this chapter describes the estimative and predictive discrimination procedures.

In the time domain analysis of time series, we assume that observations are taken at discrete and equal intervals over a finite period. The purpose of time domain discrimination is thus to classify time series that are ordered according to the sequence in which they were

collected or observed. This time domain discrimination involves the evaluation of the discriminant function given in equation (2) of chapter II, and reproduced here as

$$\delta_{12}(Y_N ; \theta_1, \theta_2) = \ln \left(\frac{P(Y_N | \theta_1)}{P(Y_N | \theta_2)} \right),$$

where $P(Y_N | \theta_i)$ is defined in equation (4) and Y_N originates from class i .

In the frequency domain, the classification of the test realization Y_N into one of two AR classes \mathcal{S}_1 and \mathcal{S}_2 is essentially a test of the hypothesis

$$H_0: f_Y(\omega | \theta_1) = f_Y(\omega | \theta_2), \quad \omega \in (0, 1),$$

where $f_Y(\omega | \theta_i)$, as defined in section 3.2.4, is the spectral density of Y_N given that Y_N is generated by $AR(p_i)$, $i=1,2$.

At frequency ω , Parzen (1982) describes the following information divergence:

$$I(f_Y(\omega); f_i(\omega)) = \frac{1}{2} \int_0^1 \left\{ \frac{f_Y(\omega | \theta_i)}{f_i(\omega)} - \log \left(\frac{f_Y(\omega | \theta_i)}{f_i(\omega)} \right) - 1 \right\} d\omega,$$

where $f_i(\omega)$ is the spectral density of X_i , $i=1,2$.

The information divergence above is also referred to as a measure of the distance between the spectral densities defined. We may thus define the following discriminant function for the classification of Y_N into one of the classes \mathcal{S}_1 and \mathcal{S}_2 :

$$\begin{aligned}
\delta_{12}(Y_N; f_1, f_2) &= I(f_Y(\omega); f_1(\omega)) - I(f_Y(\omega); f_2(\omega)) \\
&= \frac{1}{2} \int_0^1 \left\{ \frac{f_Y(\omega)}{f_1(\omega)} - \frac{f_Y(\omega)}{f_2(\omega)} + \log \left(\frac{f_2(\omega)}{f_1(\omega)} \right) \right\} d\omega. \quad (8)
\end{aligned}$$

The discriminant rule in the frequency domain is given as follows:

Classify Y_N into \mathcal{S}_1 if $\delta_{12}(Y_N; f_1, f_2) \geq 0$;
otherwise classify Y_N into \mathcal{S}_2 .

Both the discriminant rules of equation (2) in the time domain and of equation (8) depend on ϕ , τ , and p , the parameters of the AR model. Estimative Discrimination and Predictive Discrimination in the sections that follow are attempts at the evaluation of the discriminant functions.

Estimative Discrimination

The estimative approach to discrimination estimates the discriminant function by replacing θ_i with $\hat{\theta}_i$, where $\hat{\theta}_i$ is the estimate from the training realization X_i , $i=1,2$. The discriminant function $2\delta_{12}$, from equations (3) and (7), becomes

$$\begin{aligned}
&(N-p_1) \log \left(\frac{\tau_1}{2\pi} \right) + \tau_2 (Z - W\phi_2)' (Z - W\phi_2) \\
&- (N-p_2) \log \left(\frac{\tau_2}{2\pi} \right) - \tau_1 (Z - W\phi_1)' (Z - W\phi_1).
\end{aligned} \quad (9)$$

For specified AR orders p_1 and p_2 , the problem is to

determine estimates of the AR parameters φ_i and τ_i , $i=1,2$.

If p is known and fixed, then φ_k ($k \leq p$) is defined as the k -th element of the p -dimensional vector φ . However to accomodate changing values of the dimension p , we introduce a second subscript as follows: $\varphi_{k,j}$ refers to the k -th element of the j -dimensional vector φ . This convention will be used in the recursive formulations of Burg and Yule-Walker in the next two sections.

Burg Estimation

The Maximum Entropy Method (MEM) for estimating φ and τ of an $AR(\theta)$ process, $\theta=(\varphi,\tau,p)$, and a given order p was first formulated by Burg (1967, 1968). For a series of limited length, this procedure has been shown to be superior to other estimation methods for spectral estimation. See for example Ulrich and Bishop (1975). MEM is essentially a recursive formulation that estimates the AR parameters (for known order p) by utilizing only the existing sample information. That is MEM, unlike other spectral estimators, such as Yule-Walker, makes no assumptions on the extension of the available sample information.

The recursion suggested by Burg (1967,1968) is very similar to that used in Yule-Walker estimation outlined in Box and Jenkins (1976, p.82). The principal difference between these two recursions is the way partial

autocorrelations are estimated by the Burg formulation. Suppose that $X_n = (x(1), \dots, x(n))'$ ~ AR(p) model and satisfies:

$$\sum_{k=0}^p \varphi_k x(t-k) = \varepsilon(t), \quad t=p+1, \dots, n,$$

where $\varepsilon(t)$ is white noise (or prediction error) with variance $1/\tau$. The p-point prediction filter is the set of coefficients $\varphi = (\varphi_1, \varphi_2, \dots, \varphi_p)'$ and p is the length of the prediction filter.

To estimate the prediction filter of length $p(\geq 1)$, the Burg procedure starts by estimating the coefficient for $p=1$ and then $p=2$ and so on. When $p=1$, the Burg estimate for $\varphi_{1,1}$, the first element of a 1-dimensional vector is

$$\hat{\varphi}_{1,1} = \frac{2 \sum_{t=1}^{n-1} x(t)x(t+1)}{\sum_{t=1}^{n-1} \{[x(t)]^2 + [x(t+1)]^2\}}.$$

The recursion formulation for an arbitrary filter of length p is given by Andersen (1974). The estimation procedure involves the following steps:

Step 1: For $m=1$, find $\hat{\varphi}_{1,1}$ as given above.

Step 2: Increase m by 1 and Compute

$$a(m,t) = a(m-1,t) - \hat{\varphi}_{m-1,m-1} b(m-1,t)$$

and $b(m,t) = b(m-1,t+1) - \hat{\phi}_{m-1,m-1} a(m-1,t+1),$

where $a(1,t) = x(t)$ and $b(1,t) = x(t+1).$

$$\text{Step 3: Compute } \hat{\phi}_{m,m} = \frac{2 \sum_{t=1}^{n-m} x(t)x(t+m)}{\sum_{t=1}^{n-m} \{[x(t)]^2 + [x(t+m)]^2\}}.$$

Step 4: Compute $\hat{\phi}_{m,k} = \hat{\phi}_{m-1,k} - (\hat{\phi}_{m,m})(\hat{\phi}_{m-1,m-k}), \quad k=2, \dots, p.$

Step 5: Estimate the white noise variance as follows:

$$\hat{\tau}_m^{-1} = \hat{\tau}_{m-1}^{-1} (1 - \hat{\phi}_{m,m}^2).$$

If $m < p$, return to step 2;

if $m = p$, stop. The Burg estimates are given by

$$\hat{\phi} = (\hat{\phi}_{1,p}, \hat{\phi}_{2,p}, \dots, \hat{\phi}_{p,p}),$$

$$\text{and } \hat{\tau}^{-1} = \hat{\tau}_{p-1}^{-1} (1 - \hat{\phi}_{p,p}^2).$$

Yule-Walker Estimation

The Yule-Walker estimates for the AR Coefficients (for known AR order p) are obtained by solving the Yule-Walker equations given in section 1.2, Chapter I. The recursion for performing the computation is as follows (Box and Jenkins, 1976):

Step 1: Start with $m=1$ and compute

$$\hat{\phi}_{1,1} = \frac{\sum_{t=1}^{n-1} x(t)x(t+1)}{\sum_{t=1}^n [x(t)]^2}$$

Step 2: Increase m by 1 and compute

$$\hat{\rho}(m) = \frac{\sum_{j=1}^{m-1} \hat{\phi}_{m-1,j} \hat{\rho}(m-j)}{1 - \sum_{j=1}^{m-1} \hat{\phi}_{m-1,j} \hat{\rho}(j)}$$

where $\hat{\phi}_{m,k} = \hat{\phi}_{m-1,k} - \hat{\phi}_{m,m} \hat{\phi}_{m-1,p-k}$, $k=1, \dots, p-1$,

and $\hat{\rho}(j)$, $j=1, \dots, p$ is the estimated autocorrelation of

lag j , defined by

$$\hat{\rho}(j) = \frac{\hat{\gamma}(j)}{\hat{\gamma}(0)},$$

where

$$\hat{\gamma}(j) = \sum_{t=1}^{n-1} (x(t+j) - \bar{x})(x(t) - \bar{x}),$$

$$\bar{x} = \frac{1}{n} \sum_{t=1}^n x(t),$$

$$\hat{\tau}_m^{-1} = \hat{\tau}_{m-1}^{-1} (1 - \hat{\phi}_{m,m}^2),$$

and

$$\hat{\tau}_0^{-1} = \frac{1}{N} \sum_{t=1}^n [x(t)]^2.$$

Step 3: If $m < p$, return to step 2;

if $m = p$, stop. The Yule-Walker estimates are given by

$$\hat{\phi} = (\hat{\phi}_{1,p}, \hat{\phi}_{2,p}, \dots, \hat{\phi}_{p,p}),$$

and $\hat{\tau}^{-1} = \hat{\tau}_{p-1}^{-1} (1 - \hat{\phi}_{p,p}^2)$.

Estimative Discriminant Functions

The Burg discriminant function in the time domain is obtained by replacing the parameters of equation (9) with the following estimates:

Substitute \hat{p}_i for p_i using the Schwarz (1978) or other estimation criterion;

replace ϕ_i and τ_i with the Burg estimates described in the previous section.

In the frequency domain, obtain the Burg discriminant function using the estimates in the above paragraph in equation (8).

The Yule-Walker discriminant functions are obtained in an equivalent way. The Schwarz (1978) or some other criterion is used for the estimation of p_i .

Predictive Discrimination

The starting point of predictive discrimination is the prior density for the AR parameter $\theta_i = (\phi_i', \tau_i, p_i)'$, $i=1, 2$. The joint probability density of θ_i may be expressed as

the product of the following probability densities:

$$P(\theta_i) = P(\phi_i | \tau_i, p_i) P(\tau_i | p_i) P(p_i), \quad i=1,2.$$

The predictive approach to statistical discrimination proposed by Geisser (1964, 1966) replaces $P(Y_N | \theta_i)$ with the predictive density, $P(Y_N | X_i)$ defined as

$$\frac{\sum_{p_i} \int_{\tau_i} \int_{\phi_i} P(Y_N | \theta_i) P(\phi_i | \tau_i, p_i) P(\tau_i | p_i) P(X_i | \theta_i) d\phi_i d\tau_i P(p_i)}{P(X_i)},$$

where

$$P(X_i) = \sum_{p_i} \int_{\tau_i} \int_{\phi_i} P(Y_N | \theta_i) P(\phi_i | \tau_i, p_i) P(\tau_i | p_i) P(X_i | \theta_i) d\phi_i d\tau_i P(p_i)$$

is the marginal probability density of the training realization X_i from class \mathcal{S}_i , $i=1,2$ and $P(p_i)$ is the probability distribution of the AR order p_i from class \mathcal{S}_i . The expressions for $P(\theta_i)$ and $P(Y_N | \theta_i)$ will be extremely useful for the determination of predictive densities in the time and frequency domains.

Probability Distributions for the AR order

(1) Subjective Prior Distribution. The motivation for this subjective prior arises from Akaike's (1971, 1974) Information Criterion (AIC) for the estimation of p . Schwarz (1978) modified the AIC to ensure consistency of the estimator of p while Shibata (1976) demonstrated by simulation that when $p \leq 9$, the AIC correctly identified p

about 75% of the time. The relative frequency distribution for the AR order p_i from class \mathcal{S}_i , $i=1, 2$ is defined as follows:

$$P(p_i) = \begin{cases} 0.75, & \text{if } p_i = \hat{p}_i; \\ \frac{0.25}{K-1}, & \text{if } p_i \neq \hat{p}_i, p_i = 1, \dots, K \leq 9; \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

where \hat{p}_i is estimated from the training sample from the population \mathcal{S}_i using the Schwarz (1978) criterion; and K is the maximum value of p_i , $i=1, 2$.

(2) The Maximum Entropy Distribution. A maximum entropy density incorporates the available partial information but is as noninformative as possible. In deriving the density for the AR order, the partial information is obtainable from the training sample X_i from class \mathcal{S}_i , $i=1, 2$.

The entropy of the prior density $\xi(p)$ of p is defined by Kullback (1958) and Parzen (1982) as

$$H(p) = \sum_P \xi(p) \log \xi(p)$$

where P is the set of allowable values of the AR order p . In this study we define the maximum entropy prior density $\xi_0(p)$ as the $\xi(p)$ that maximizes $H(p)$ such that $E_{\xi}(p_i) = \hat{p}_i$, where \hat{p}_i is estimated from the training realization and E_{ξ} is expectation with respect to the probability density ξ . The solution to this maximization problem is, from Berger (1985), given by

$$\xi_0(p_i) = \begin{cases} \frac{\exp(p_i \lambda_i)}{\sum_P \exp(p_i \lambda_i)}, & \text{if } p_i \in P; \\ 0, & \text{otherwise} \end{cases}$$

where $|\lambda_i| < 1$ is a constant determined from $E_\xi(p_i) = \hat{p}_i$, $i=1,2$, and $P = \{1, 2, \dots, K, K < \infty\}$.

Noting that

$$\sum_P \exp(p_i \lambda_i) = \exp(\lambda_i) \left\{ \frac{1 - \exp(K\lambda_i)}{1 - \exp(\lambda_i)} \right\},$$

we have

$$\xi_0(p_i) = \begin{cases} \exp(\lambda_i (p_i - 1)) \left\{ \frac{1 - \exp(K\lambda_i)}{1 - \exp(\lambda_i)} \right\}, & \text{if } p_i \in P; \\ 0, & \text{otherwise,} \end{cases}$$

where λ_i is determined from

$$\exp(\lambda_i) \left\{ \frac{1 - \exp(K\lambda_i)}{1 - \exp(\lambda_i)} \right\} \sum_P \exp(p_i \lambda_i) = \hat{p}_i.$$

If we drop the upper bound on K and let $K \longrightarrow \infty$, then from

$$\xi(p_i) = \begin{cases} \frac{\exp(p_i \lambda_i)}{\sum_P \exp(p_i \lambda_i)}, & \text{if } p_i \in P; \\ 0, & \text{otherwise,} \end{cases}$$

and

$$\sum_{p_i=1}^{\infty} \exp(p_i \lambda_i) = \frac{\exp(\lambda_i)}{1 - \exp(\lambda_i)},$$

$$\xi_0(p_i) = \begin{cases} \{\exp(\lambda_i)\}^{p_i-1} \{1 - \exp(\lambda_i)\}, & \text{if } p_i \in \mathbb{P}, |\lambda_i| < 1; \\ 0, & \text{otherwise.} \end{cases}$$

That is $p_i \sim \text{Geometric}(\exp(\lambda_i))$.

Hence $E_{\xi}(p_i) = \hat{p}_i$ gives

$$\hat{p}_i = \frac{1 - \exp(\lambda_i)}{\exp(\lambda_i)}$$

or
$$\exp(\lambda_i) = \frac{1}{1 + \hat{p}_i},$$

yielding the maximum entropy prior density in terms of \hat{p}_i :

$$P(p_i) = \begin{cases} \hat{p}_i (1 + \hat{p}_i)^{-p_i}, & \text{if } p_i \in \mathbb{P}; \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Joint Prior Probability Density for φ_i and τ_i

(1) The Multivariate Normal Gamma. For a specified white noise variance $1/\tau_i$ ($\tau_i > 0$) and AR order p_i , suppose that the a priori probability density of φ_i is the multivariate normal with mean μ_i and covariance matrix $(1/\tau_i)\Sigma_i^{-1}$ given by

$$(2\pi)^{-\frac{p_i}{2}} |\Sigma_i|^{1/2} \tau_i^{\frac{p_i}{2}} \exp \left[-\frac{\tau_i}{2} (\varphi_i - \mu_i)' \Sigma_i (\varphi_i - \mu_i) \right],$$

where $\mu_i \in R^{p_i}$ and Σ_i^{-1} , a $p_i \times p_i$ matrix of constants are the hyperparameters of φ_i and τ_i for fixed p_i .

Hyperparameters are the parameters of the a priori distribution of the original parameters.

If the scale parameter τ_i has the a priori Gamma density $P(\tau_i)$ with parameters $\alpha_i/2$ and $\beta_i/2$, then

$$P(\tau_i) = \frac{(\beta_i/2)^{\alpha_i/2}}{\Gamma(\alpha_i/2)} \tau_i^{\frac{\alpha_i}{2} - 1} \exp(-\beta_i \tau_i/2), \quad \alpha_i > 0, \beta_i > 0$$

where α_i and β_i are the hyperparameters of τ_i .

The a priori probability density of θ_i becomes

$$P(\theta_i) \propto \tau_i^{(p_i + \tau_i)/2} \exp\left\{\frac{1}{2}\tau_i((\varphi_i - \mu_i)' \Sigma_i (\varphi_i - \mu_i) + \beta_i)\right\} P(p_i), \quad (12)$$

where $P(p_i)$ is the a priori mass function for the order of the AR process from class ϑ_i .

(2) The noninformative prior density. If the distributional form of φ_i is unknown then we assume that φ_i is simply a vector of real constants in the p_i -dimensional plane. In that case, θ_i is distributed a priori as

$$P(\theta_i) \propto \tau_i^{-\frac{1}{2}q_i} P(p_i), \quad \tau_i > 0, p_i = 1, 2, \dots, \quad (13)$$

where $q_i \geq 0$ is the hyperparameter for τ_i and $P(p_i)$ is the prior probability mass function of the AR order p_i .

Time Domain Discriminant Functions

If ϕ_i and τ_i are jointly distributed as Multivariate Normal Gamma Prior Density, then $P(X_i)$, the marginal probability density of X_i is given (as shown in Appendix A) by

$$\sum_{p_i} P(X_0) \left(\frac{\pi}{2}\right)^{\frac{N_i - 2p_i}{2}} |A_i + \Sigma_i|^{-1/2} \left(k_{1i} + \hat{t}_i^{-1} + \beta_i\right)^{-\frac{N_i + \alpha_i - p_i}{2}} P(p_i) \Gamma_2, \quad (14)$$

where $A_i = W_i' W_i$ is derived from the training realization X_i ,

$$c_{1i} = (A_i + \Sigma_i)^{-1} (A_i \hat{\phi}_i + \Sigma_i \mu_i),$$

$$k_{1i} = (\phi_i - \mu_i)' A_i (A_i + \Sigma_i)^{-1} \Sigma_i (\phi_i - \mu_i),$$

and $\Gamma_2 = \Gamma((N_i + \alpha_i - p_i)/2)$.

Again from the derivation in Appendix A, the predictive density $P(Y_N | X_i)$ is given by

$$\frac{\sum_{p_i} \left(\frac{\pi}{2}\right)^{\frac{N + N_i - 2p_i}{2}} |A_{0i} + A_i + \Sigma_i|^{-1/2} (u_{0i})^{-\frac{\eta_i - q + \alpha_i}{2}} \Gamma\left(\frac{\eta_i - q + \alpha_i}{2}\right) P(p_i)}{\sum_{p_i} \left(\frac{\pi}{2}\right)^{\frac{N_i - 2p_i}{2}} |A_i + \Sigma_i|^{-\frac{1}{2}} \left(\frac{k_{1i} + \hat{t}_i^{-1} + \beta_i}{2}\right)^{-\frac{N_i + \alpha_i - p_i}{2}} P(p_i) \Gamma_2}, \quad (15)$$

where $A_{0i} = W_0' W_0$ is derived from the test realization Y_N from class \mathfrak{P}_i ,

$$k_{0i} = (\tilde{\varphi}_i - \hat{\varphi}_i)' A_{0i} (A_{0i} + A_i)^{-1} A_i (\tilde{\varphi}_i - \hat{\varphi}_i),$$

$$u_{0i} = \beta_i + N_i \hat{\tau}_i^{-1} + N \tilde{\tau}_i^{-1} + k_{0i} + k_{2i},$$

and $\eta_i = N + N_i + q - 2p_i + 2.$

If the AR orders p_1 and p_2 are known, then we may drop the summation with respect to p_i in equation (15) and the predictive density is reduced to

$$(1/\pi)^{(N-p_i)/2} \left(\frac{|A_i + \Sigma_i|}{|A_{0i} + A_i + \Sigma_i|} \right)^{\frac{1}{2}} \frac{(k_{1i} + \hat{\tau}_i^{-1} + \beta_i)^{(N_i + \alpha_i - p_i)/2}}{(u_{0i})^{(N_i + N + \alpha_i - 2p_i)/2}} \Gamma_3,$$

where

$$c_{0i} = (A_{0i} + A_i)^{-1} (A_{0i} \varphi_i + A_i \varphi_i),$$

$$k_{2i} = (c_{0i} - \mu_i)' (A_{0i} + A_i) (A_{0i} + A_i + \Sigma_i)^{-1} \Sigma_i (c_{0i} - \mu_i)$$

and $\Gamma_3 = \frac{\Gamma((N + N_i + \alpha_i - 2p_i)/2)}{\Gamma((N + \alpha_i - p_i)/2)}.$

Hence, when p_1 and p_2 are known, the predictive discriminant function simplifies to

$$2D_{12}(Y_N; \theta_1, \theta_2) = D_1 + D_2 + D_3 + D_4 + D_5 + D_6,$$

where

$$D_1 = \log \left(\frac{|A_{02} + A_2 + \Sigma_2|}{|A_{01} + A_1 + \Sigma_1|} \right),$$

$$D_2 = \log \left(\frac{|A_1 + \Sigma_1|}{|A_2 + \Sigma_2|} \right),$$

$$D_3 = (N + N_2 + \alpha_2 - 2p_2) \log(\beta_2 + N\tau_2^{-1} + N_2\hat{\tau}_2^{-1} + k_{02} + k_{22})$$

$$- (N + N_1 + \alpha_1 - 2p_1) \log(\beta_1 + N\tau_1^{-1} + N_1\hat{\tau}_1^{-1} + k_{01} + k_{21}),$$

$$D_4 = (N_1 + \alpha_1 - p_1) \log(k_{11} + \hat{\tau}_1^{-1} + \beta_1) - (N_2 + \alpha_2 - p_2) \log(k_{12} + \hat{\tau}_2^{-1} + \beta_2),$$

$$D_5 = \log \left(\frac{\Gamma((N + N_1 + \alpha_1 - 2p_1)/2)}{\Gamma((N + N_2 + \alpha_2 - 2p_2)/2)} \right),$$

and

$$D_6 = \log \left(\frac{\Gamma((N_2 + \alpha_2 - 2p_2)/2)}{\Gamma((N_1 + \alpha_1 - p_1)/2)} \right).$$

Suppose that the prior density for θ_i is the vague or noninformative prior defined in equation (13). Then from

$$P(X_i) = \sum_{p_i} \int_{\tau_i} \int_{\phi_i} P(X_i | \theta_i) P(\phi_i | \tau_i, p_i) P(\tau_i | p_i) d\phi_i d\tau_i P(p_i),$$

and Appendix A, $P(X_i)$, the marginal density of X_i is given by

$$\sum_{p_i} P(X_0) (\pi/2)^{(N_i - 2p_i)/2} |A_i|^{-1/2} (2\hat{\tau}_i/N_i)^{(\eta_i - N)/2} \Gamma((\eta_i - N)/2) P(p_i).$$

The corresponding predictive probability density, $P(Y_N | X_i)$, is given by

$$\frac{\sum_{p_i} (\pi/2)^{(N+N_i-3p_i)/2} |A_{0i}+A_i|^{-\frac{1}{2}} (d_{0i})^{-\frac{\eta_i-p_i}{2}} \Gamma\left(\frac{\eta_i-p_i}{2}\right) P(p_i)}{\sum_{p_i} (\pi/2)^{(N_i-2p_i)/2} |A_i|^{-1/2} (2\hat{\tau}_i/N_i)^{(\eta_i-N)/2} \Gamma((\eta_i-N)/2) P(p_i)}, \quad (16)$$

where

$$d_{0i} = \frac{N\hat{\tau}_i^{-1} + N_i\hat{\tau}_i^{-1} + k_{0i}}{2}.$$

The predictive discriminant function is defined as

$$D_{12}(Y_N; \theta_1, \theta_2) = \log \left(\frac{P(Y_N | X_1)}{P(Y_N | X_2)} \right).$$

If the AR orders p_1 and p_2 are known, that is $P(p_i)=1$, $i=1,2$, the summation signs in equation (16) become unnecessary and the predictive density $P(Y_N | X_i)$ becomes

$$(1/\pi)^{(N_i-p_i)/2} \left(\frac{|A_i|}{|A_i+A_{0i}|} \right)^{\frac{1}{2}} \frac{(N_i\hat{\tau}_i^{-1})(\eta_i-N)/2}{(d_{0i})^{(\eta_i-p_i)/2}} \Gamma_{1i},$$

$$\text{where } \Gamma_{1i} = \frac{\Gamma((\eta_i-p_i)/2)}{\Gamma((\eta_i-N)/2)}.$$

Hence, if p_1 and p_2 are known, the predictive discriminant function corresponding to the vague prior density is

$$\begin{aligned}
& \log\left(\frac{|A_{02}+A_2|}{|A_{01}+A_1|}\right) + \log\left(\frac{|A_1|}{|A_2|}\right) + (\eta_2 - p_2) \log(N\tilde{\tau}_2^{-1} + N_2\hat{\tau}_2^{-1} + k_{02}) \\
& - (\eta_1 - p_1) \log(N\tilde{\tau}_1^{-1} + N_1\hat{\tau}_1^{-1} + k_{01}) + (\eta_1 - N) \log(N_1\hat{\tau}_1^{-1}) \\
& - (\eta_2 - p_2) \log(N_2\hat{\tau}_2^{-1}).
\end{aligned}$$

Frequency Domain Discriminant Functions

The spectral density of an autoregressive process with parameter $\theta = (\varphi, \tau, p)$ is given by (Newton (1988), p.101)

$$f(\omega|\theta_i) = \frac{1}{p_i \tau_i \left| \sum_{k=0}^{p_i} \varphi_i(k) \exp(2\pi\omega jk) \right|^2}, \quad j=\sqrt{(-1)}, \quad \omega \in (0,1).$$

$f^{-1}(\omega|\theta_i)$, the multiplicative inverse of $f(\omega|\theta_i)$ is defined as

$$f^{-1}(\omega|\theta_i) = \tau_i \left| \sum_{k=0}^{p_i} \varphi_i(k) \exp(2\pi\omega jk) \right|^2, \quad j=\sqrt{(-1)}, \quad \omega \in (0,1).$$

The multiplicative inverse may be expressed (Cook (1985)) as the following quadratic form:

$$f^{-1}(\omega|\theta_i) = \tau_i \varphi_i' D_i(\omega) \varphi_i, \quad \omega \in (0,1), \quad (17)$$

where $D_i(\omega)$ is a $p_i \times p_i$ symmetric matrix given by

$$\begin{pmatrix}
1 & \cos\omega & \dots & \cos(p_i-1)\omega \\
\cos\omega & 1 & \dots & \cos(p_i-2)\omega \\
\vdots & \vdots & \ddots & \vdots \\
\cos(p_i-1)\omega & \cos(p_i-2)\omega & \dots & 1
\end{pmatrix}.$$

The inverted predictive spectral density may therefore be defined at frequency ω by

$$\begin{aligned} f^{-1}(\omega|X_i) &= \sum_{p_i} \int_{(\varphi_i, \tau_i)} f^{-1}(\omega|\theta_i) P(X_i|\theta_i) P(\theta_i) d\varphi_i d\tau_i \\ &= E_{\theta_i|X_i}(f(\omega|\theta_i)) \\ &= E_{\theta_i|X_i}(\tau_i \varphi_i' D_i(\omega) \varphi_i). \end{aligned}$$

If the joint prior density of φ_i and τ_i follows the multivariate normal gamma density given in equation (12), then following Cook (1985), for fixed p_i ,

$$E_{\theta_i|X_i}\{\tau_i \varphi_i' D_i(\omega) \varphi_i\} = \text{tr}(D_i(\omega) F_{1i}) + \frac{1}{b_i} (N_i + q) B_i' D_i(\omega) B_i$$

where

$\text{tr}(G)$ = trace of matrix G ,

$$F_{1i} = A_i + \Sigma_i,$$

$$F_{2i} = W_i' Z_i + \Sigma_i \mu_i,$$

$$B_i = F_{1i}^{-1} F_{2i},$$

and

$$b_i = Z_i' Z_i + \mu_i' \Sigma_i \mu_i + \beta_i - F_{2i}' F_{1i}^{-1} F_{2i}.$$

The inverted predictive spectral density becomes

$$f^{-1}(\omega|X_i) = \sum_{p_i} \{ \text{tr}(D_i(\omega) F_{1i}) + \frac{1}{b_i} (N_i + q) \varphi_i' D_i(\omega) \varphi_i \} P(p_i) \quad (18)$$

The predictive discriminant function in the frequency domain may be evaluated from the data by combining

equations (8) and (18) as follows:

$$\frac{1}{2N} \sum \left\{ \frac{f_1^{-1}(\hat{\omega})}{f_Y^{-1}(\hat{\omega})} - \frac{f_2^{-1}(\hat{\omega})}{f_Y^{-1}(\hat{\omega})} + \log \left(\frac{f_1^{-1}(\hat{\omega})}{f_2^{-1}(\hat{\omega})} \right) \right\}, \quad (19)$$

where $\hat{\omega} = (k-1)/N$, is the natural frequency at times $k=1, 2, \dots, N$ of the test realization Y_N .

The noninformative prior may be obtained from the multivariate normal gamma prior by the simple substitutions:

$$\Sigma_i = 0 \quad \text{and} \quad \beta_i = 0.$$

Then the inverted predictive spectral density corresponding to the noninformative prior is

$$f^{-1}(\omega | X_i) = \sum_{P_i} \{ \text{tr}(D_i(\omega) F_{1i}) + \frac{1}{b_i} (N_i + q) \varphi' D_i(\omega) \varphi \} P(P_i)$$

where

$\text{tr}(G)$ = trace of matrix G ,

$$F_{1i} = A_i,$$

$$F_{2i} = W_i' Z_i,$$

$$B_i = F_{1i}^{-1} F_{2i},$$

and

$$b_i = Z_i' Z_i - F_{2i}' F_{1i}^{-1} F_{2i}$$

discriminate each of the following pairs of AR classes: two AR(1) processes, AR(1) versus AR(2), and two AR(2) processes. That is, for simulation purposes, the minimum and maximum AR orders are respectively 1 and 2. This restriction on p_i is solely for computational convenience to accomodate the immense CPU time required for the evaluation of the predictive discriminant functions. The theory has been developed for any AR order in the range $1 \leq p_i \leq 9$, $i=1,2$. Further, because the AR series are generated by stationary processes, the mean and variance of each series assume the values 0 and 1 respectively without any loss of generality.

The AR coefficients, ϕ_i are chosen to ensure some resemblance of the spectra for the two AR series being discriminated. The problem of discriminating between series with distinct spectra is clearly a trivial exercise and is not reflective of the practical problems encountered in the applications of discriminant analysis. This similarity is readily verified from the superimposed spectral plots of Figures 1, 12 and 23.

The AR coefficients ϕ_i used in this simulation study were generated from a short computer program. The computer program provided the coefficients and the corresponding spectral densities from which those in this study were chosen. The three pairs of AR processes for classification are as follows (AR(p_i) is autoregression from process S_i of

order p_i , $i=1,2$):

1. CLASS 1 : AR(1)

$$p_1 = 1$$

$$\phi_1 = -.8258$$

AR Equation:

$$y(t) = .8258y(t-1) + \epsilon$$

CLASS 2 : AR(1)

$$p_2 = 1$$

$$\phi_2 = -.9238$$

AR Equation:

$$y(t) = .9238y(t-1) + \epsilon$$

2. CLASS 1 : AR(2)

$$p_1 = 2$$

$$\phi' = (.2944, .6503)$$

AR Equation: $y(t) =$

$$-.2944y(t-1) - .6503y(t-2) + \epsilon$$

CLASS 2 : AR(1)

$$p_2 = 1$$

$$\phi_2 = .9572$$

AR Equation: $y(t) =$

$$-.9238y(t-1) + \epsilon$$

3. CLASS 1 : AR(2)

$$p_1 = 2$$

$$\phi' = (-.8266, -.934)$$

AR Equation: $y(t) =$

$$.8266y(t-1) + .9340y(t-2) + \epsilon$$

CLASS 2 : AR(2)

$$p_2 = 2$$

$$\phi' = (-.4939, -.8207)$$

AR Equation: $y(t) =$

$$.4939y(t-1) + .8207y(t-2) + \epsilon$$

Having defined the AR processes for classification we must next obtain the training data from each class.

Training data consist of samples of existing data from each AR class. In the simulation situation these samples are obtained by generating from each class series of lengths (or sizes) 50, 100, 200, and 400. The AR series to be classified, referred to as the test realization, are generated randomly from each class. The lengths of these

realizations range from small to large as 50, 100, 200, and 400. The training data and test realizations are combined to determine the error rates.

For each combination of training data length and length of test realization, 100 test realizations are randomly generated from each class and classified according to each of the six discriminant functions in the time and frequency domains. The J-divergence rate is also evaluated for each combination. The entire source code for the simulation study was written in SAS (using the IML procedure) and is given in Appendix B. Several difficulties were encountered not only in writing and running the program but also in the graphical analysis of the results.

The greatest difficulty arose from the extensive CPU time needed to evaluate the discriminant functions. More than 100 minutes of CPU time (on the IBM computer model 3090) were required to run each of the three simulation cases. This extensive demand of CPU time also restricted the number of simulation cases to the three listed at the beginning of this chapter. The OVERLAY option of the plot procedure in SAS is available in two-dimensional not in three-dimensional plots. The lack of the OVERLAY option meant that in the three-dimensional plotting of (X, Y, Z) , several values of the third variable, Z could not be simultaneously plotted for the same values of X and Y .

Thus whereas SAS plots (without superimposed Z-values) can not provide a fast and visual comparison of Z-values, such plots are possible with EXECUSTAT, a software package for statistical analysis. The complexity of the functions that had to be evaluated also necessitated the use of subroutines and modules to perform the repetitive tasks. All of these subroutines are completely listed and fully described in Appendix C. Appendix B contains a listing of the SAS source for the simulation program.

CHAPTER V

SUMMARY AND CONCLUSIONS

The objective of this research was to propose a new procedure for the discrimination of univariate autoregressive time series of unknown order. The new procedure, called the predictive discriminant analysis, is based on the evaluation of the discriminant function using the predictive density of the observed series. Since no population parameters are estimated, the proposed approach to classification avoids the usual problems associated with discrimination procedures that are based on parameter estimation. One such problem is the variability of the estimates.

Chapter I defined the goals of the study and also introduced the basic notion of predictive analysis, the basis of the new classification method. Chapter II not only gave the historical perspective to the general classification problem but also outlined the special difficulties associated with the discrimination of time series. In Chapter III, the estimative and predictive forms of the discriminant function were obtained in the time and frequency domains. The estimative discriminant functions were based on the parameter estimation

methods due to Burg (1967, 1968) and due to Yule-Walker (Box and Jenkins, 1976). The predictive forms of the discriminant function were obtained using various prior densities for the autoregressive parameter $\theta = (\phi', \tau, p)$ defined in equation (1), Chapter 1.

Two prior densities for the unknown order p were given: the subjective prior and the maximum entropy prior. The joint prior distribution of the autoregressive coefficient ϕ and the innovation variance $1/\tau$ was modelled by the vague prior and by the multivariate-normal-gamma prior. Predictive densities and the corresponding predictive discriminant functions were derived for each joint prior density for ϕ and τ . These derivations were obtained in the time and in the frequency domains. The simulation study detailed in Chapter IV was based on the time and frequency domain evaluation of the estimative and predictive discriminant functions. The evaluated predictive discriminant functions were derived from the joint vague prior density for ϕ and τ and on the subjective prior for the order p . The purpose of the simulation was to investigate the accuracy of the proposed new classification technique for discriminating between two autoregressive processes. To conduct the simulation study, three separate and independent pairs of univariate autoregressive processes were chosen for classification.

The first pair consisted of two processes each of

order 1; the second involved an AR process of order 1 and an AR process of order 2; the third case compared AR processes both of order 2. The three pairs of processes were simulated and studied separately and independently. The results and conclusions of the simulation studies are demonstrated in the graphs and tables given respectively in the list of Figures and Tables. The description and analysis of each graph is given in the next section.

Graphical Analyses

The graphs in the list of Figures are arranged according to the classification cases. That is, Figures 1-11 pertain to the classification of AR(1) processes, Figures 12-22 refer to AR(1) and AR(2) and Figures 23-33 are for AR(2) processes. The analysis for each plot is done by a statement of the purpose or objective of the plot and the results or conclusions to be drawn from the plot. The plots are arranged and analyzed according to the following order: spectra, Box and Whisker, Quantile, 3-dimensional, J-Divergence, Frequency over Time Domain, and Relative Improvements and J-Divergence.

Spectral Plots

Figures 1, 12 and 23 contain respectively the superimposed spectral plots of the simulated series from the classification of two AR(1), AR(2) and AR(1), and two

AR(2) processes.

The purpose of the spectral plots is to highlight the similarity that exists between the spectra of the AR series being discriminated. The closeness of the spectra demonstrates the difficulty of attempting to classify the AR series solely by their spectra.

Box and Whisker Plots

The Box and Whisker plots in Figures 2, 13 and 24 provide not only the important descriptive features of the error rates for each classification scheme but also permit an easy comparison of these features among the six discrimination methods considered.

Each plot contains a central box that extends from the first quartile to the third quartile and hence contains 50% of the error rates. One whisker extends from the lower quartile to the minimum error rate observed; the other whisker extends from the upper quartile to the maximum error rate. The plus sign in the box gives the location of the mean error rate and the center line locates the median. These features are described for each pair of AR processes for classification.

AR(1) Processes. Figure 2 contains the Box and Whisker plots for the discrimination of two AR(1) processes. The six plots correspond to the six discriminant functions: Predictive, Burg and Yule-Walker, each evaluated in time

and frequency domains.

An overview of the six plots shows that predictive discrimination in the frequency domain yields the best (lowest) error rate of about 15% while the Yule-Walker procedure, also in the frequency domain, has the worst (highest) error rate of about 52%. The plot also shows that the predictive discrimination in the frequency domain has an average error rate of about 24.5%, about 25% of the rates are under 19%, one half of the rates fall between 19% and 30% and that one fourth are above 30%.

AR(1) and AR(2). The Box and Whisker plot for this analysis is in Figure 13. Predictive discrimination in the frequency domain, as in the previous case, appears to be superior to the other classification methods since it provides the smallest minimum error rate of about 11% and the smallest maximum of about 30.5%. The next best discrimination methods are the predictive in time domain, Burg in frequency domain, Burg in time domain, and the Yule-Walker procedures in the time and frequency domains in that order.

AR(2) Processes. Figure 24 shows that in discriminating between two autoregressive series of order two, the predictive methods (in time and frequency domains) are almost equally effective and perform significantly better than the other discrimination procedures. The

predictive methods have a mean error rate of 20% and approximate minimum and maximum error rates of 13% and 36% respectively. The predictive methods are followed by the Burg in the frequency domain with an average rate of about 25% ; the minimum and maximum rates are respectively 17% and 46%. The worst procedure in terms of error rates appears to be Yule-Walker in the frequency domain with the highest minimum rate of about 21% and the largest maximum rate of 58.5%.

Quantile Plots

The quantile plots in Figures 3-4, 14-15 and 25-26 provide percentiles of the error rates. That is, each point on the quantile plot approximates the fraction of the error rates that are below a particular value. For instance, consider Figure 3, the quantile plot for the frequency domain discrimination of $AR_1(1)$ and $AR_2(1)$. The proportion .475 (or 47.5%) indicates that 47.5% of the error rates fall below 17% from using predictive discrimination, 19.5% from Burg and 25% from Yule-Walker.

The Box and Whisker and the Quantile plots do not account for the effects of other variables on the error rates. For instance, the plots do not relate the sizes or lengths of the training data and of the test realizations to the error rates. These plots also fail to account for the relationship between the error rate and the

J-divergence rate. The next sections on 3-dimensional and J-divergence plots examine these relationships.

3-Dimensional Plots

The 3-dimensional plots in Figures 5-6, 16-17 and 27-28 show the effects of the lengths of training data and test realizations on the error rates. All of the plots show that the error rates tend to decrease (improve) as the lengths of the training data and test realizations increase. The question that remains to be answered is a determination of which classification is best at various sizes of the training data and test realizations. This question will be answered on a case by case basis according to the classification of $AR_1(1)$ and $AR_2(1)$, $AR_1(2)$ and $AR_2(1)$ and, $AR_1(2)$ and $AR_2(2)$ in that order.

AR(1) and AR(1). The plot of Figure 5 (in frequency domain) shows the Yule-Walker error rates to be consistently inferior to (higher than) the predictive and Burg rates. This inferiority is most prominent with the smaller sizes of the training and test data. The edge that the predictive enjoys over the Burg procedure diminishes with increasing data size and is virtually nonexistent at size 400.

Figure 6 (in the time domain) also shows the predictive procedure to produce error rates that are consistently lower than those by the Burg and Yule-Walker

methods. For large samples of the training and test realizations however, the three classification schemes appear to be equally effective.

AR(2) and AR(1). Figures 16 and 17 (in frequency and time domains respectively) show the superiority of predictive discrimination over Burg and Yule-Walker. Whereas Figure 16 shows the Burg procedure to be consistently better than Yule-Walker, Figure 17 shows these two estimative methods to produce virtually identical error rates.

AR(2) Processes. Figures 27 and 28 depict respectively the frequency and time domain plots for discriminating between two AR(2) processes. Figure 27 shows error rates that indicate a distinct and consistent superiority of the predictive discrimination over Burg and of Burg over Yule-Walker.

Figure 28 on the other hand, while maintaining the predictive as the best procedure, shows Burg as having the worst error rates especially for small samples. All of these methods however recover quickly and have almost identical rates for large samples. While the lengths of training data and test realizations have a causal effect on the error rates, the J-divergence rate attempts to account for the magnitude of the error rates. This non-causal relationship will be examined in the rest of

the Graphical Analyses.

Error Rates and J-Divergence Plots

We recall that the J-divergence rate measures the amount of available information in the training data and test realizations for the discrimination of AR processes. We would thus expect the classification errors to decrease as the J-divergence increases. This fact is validated by Figures 7-8, 18-19 and 29-30 which are derived respectively from the discrimination of two $AR(1)$, $AR_1(2)$ and $AR_2(1)$ and, two $AR(2)$ processes.

All of the plots show the predictive to be the best classification procedure. As the J-divergence rate increases, however, the gap in the error rates among the three procedures narrows.

Plots for Time and Frequency

Domain Comparison

A practical problem in time series discrimination is a determination of the more effective domain of analysis; that is, whether the time or the frequency domain provides the lower error rates for a given classification procedure. Figures 9, 20 and 31 attempt to answer this question by plotting the quantity $100(\text{TIME}-\text{FREQ})/\text{TIME}$ against the J-divergence rate. FREQ and TIME refer respectively to the error rates in the frequency and time domains from a given classification method. Hence for a

specified discrimination procedure, $100(\text{TIME-FREQ})/\text{TIME}$ measures the percent improvement of frequency over time domain discrimination at each value of the J-divergence rate.

AR(1) Processes. In figure 9, the use of the Burg procedure in the frequency domain to discriminate between two AR(1) series yields an improvement that lies in the 7% to 14% range. Figure 9 also shows that Yule-Walker is more effective in the time than in the frequency domain with an improvement between 7% and 12%. Predictive discrimination on the other hand appears to be equally effective in either domain.

AR(2) and AR(1). Figure 20 shows that the Yule-Walker and predictive procedures are more effective in time domain with relative improvement in the ranges 29% to 59% and 23% to 34% respectively. Burg discrimination is more efficient in the frequency domain with improvement between 11% and 22%.

AR(2) Processes. In Figure 31, while the Burg discrimination appears more efficient in the frequency domain (14% to 18% relative improvement), Yule-Walker and predictive procedures show no discernible difference between domains.

Plots for Improvement Rate and J-Divergence

Figures 10-11, 21-22 and 32-33 show respectively

pairwise comparisons of the three discriminant functions for the classification of AR(1) versus AR(1), AR(2) versus AR(1) and AR(2) versus AR(2). Each figure plots the ordinate $100(a-b)/a$ against the J-divergence rate. The quantities a and b denote the error rates from the two discriminant functions being compared.

These plots attempt to determine which of two discrimination criteria A and B is more accurate at a given rate of J-divergence. Points on the plots that fall above the zero mark correspond to method B having lower error rates than method A while points below zero indicate that B is a better classifier at the given value of J-divergence. We now compare each pair of classification procedures by an examination of each plot.

AR(1) Processes. Figure 10 shows the predictive and Burg procedures with significant improvements over Yule-Walker, with relative improvements ranging respectively from 17% to 21% and from 14% to 17.5%. The improvement of predictive discrimination over Burg on the other hand ranges between 0.5% and 7%.

The time domain comparisons are provided in Figure 11. In this figure, the greatest improvement is by the predictive over Burg, followed by predictive over Yule-Walker. This figure also shows Yule-Walker outperforming Burg in the amount .2% to 7.8%.

AR(2) and AR(1). Figures 21 and 22 are respectively the frequency and time domain plots. In the frequency domain, while the Burg and predictive procedures out perform Yule-Walker, Burg is seen to have improvements between 6% and 18% over predictive discrimination. In time domain, no substantial difference is noticed between the predictive and Yule-Walker procedures. However significant improvements can be seen of both the predictive and Yule-Walker over Burg.

AR(2) Processes. The frequency domain plot of Figure 32 shows that the Burg has an almost constant improvement of about 21.5% over Yule-Walker. The predictive procedure has significant improvements over Burg and Yule-Walker, the most dramatic improvement, ranging between 35% and 41%, being over Yule-Walker.

In the time domain, Figure 33 shows a constant but insignificant improvement of Burg over Yule-Walker. The most improvement comes from the predictive over Yule-Walker and Burg.

Conclusions

The principal objective of this study was to determine which of three discrimination procedures - Burg, Yule-Walker and predictive - provided the lowest error rates in the classification of two autoregressive

processes. The processes were assumed to be linear, stationary and of unknown order.

The simulation study has shown that of the three classifications, predictive analysis produces the lowest error rates in both time and frequency domains. The largest margin of this superiority is in the frequency domain. This frequency domain edge is probably due to the fact that whereas the discriminant function in the frequency domain is evaluated from all available information, the time domain function loses some information (from the training and test realizations) which is used as the initial condition to the difference equation (1). However, the asymptotic error rates are barely distinguishable among the three methods. The study also shows that the estimative procedures of Burg and Yule-Walker tend to produce higher error rates for increasing orders of the autoregressions. This tendency is probably due to the fact that higher AR orders correspond to more AR parameters that must be estimated thus increasing the risk of estimation errors.

The results of the study lead to a recommendation of the predictive procedure in the frequency domain especially for AR series of order more than one. It should be noted however that the accuracy of predictive discrimination comes with a high cost of CPU time and requires substantial computer programming.

Further Work

The technique of predictive discrimination may be extended to other areas such as

- (1) Nonlinear Time Series
- (2) Moving average processes
- (3) Multivariate time series
- (4) The consideration of priors other than those in this study for modeling the AR parameters.

REFERENCES

- Aitchison, J. and Dunsmore I. (1975). Statistical Prediction Analysis. Cambridge University Press, Cambridge, United Kingdom.
- Aitchison, J., Habbema, D. F. F. and Kay J. W. (1977). A critical comparison of two methods of statistical discrimination. Applied Statistics, 26, 15-25.
- Akaike, H. (1971). Information Theory and an Extension of the Maximum Likelihood Principle. Research Memorandum No. 46, Institute of Statistical Mathematics, Tokyo.
- Akaike, H. (1974). A New Look at the Statistical Model Identification. I.E.E.E. Trans. Auto. Control, 19, 716-723.
- Andersen, N. (1974). On the Calculation of Filter Coefficients for Maximum Spectral Analysis. Geophysics, Vol. 39, No. 1, 69-72.
- Anderson, T. W. (1951). Classification by Multivariate Analysis. Psychometrika, 16, 31-50.
- Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, Inc., New York, N. Y.
- Berger, O. B. (1985). Statistical Decision Theory and Bayesian Analysis (2nd ed.). Springer-Verlag, New York, N. Y.
- Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis, Forecasting and Control. Addison-Wesley, Reading, Massachusetts.
- Box, G. E. P. and Tiao, G. C. (1973). Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, Massachusetts.
- Broemeling, L. and Land, M. (1984). On Forecasting with Univariate Autoregressive Processes. Communications in Statistics, 13, 1305-20.

- Burg, J. P. (1967). Maximum Entropy Spectral Analysis. Paper Presented at the 13th Annual International Meeting, Soc. of Explor. Geophys. Oklahoma City, Ok.
- Burg, J. P. (1968). A New Analysis Technique for Time Series Data. Paper Presented at Advanced Study Institute on Signal Processing, NATO. Enschede, Netherlands.
- Chow, G. C. (1975). Multiperiod Predictions from Stochastic Difference Equations by Bayesian methods. Studies in Bayesian Econometrics and Statistics, edited by S. E. Fienberg and A. Zellner. North-Holland, Amsterdam, Netherlands.
- Cook, P. (1985). Bayesian Autoregressive Spectral Analysis. Commun. Statist.-Theor. Method, 14(5), 1001-10018.
- Dargahi-Noubary, G. R. and Laycock, P. J. (1981). Spectral Ratio Discrimination and Information Theory. Journal of Time Series Analysis, 2, 71-86.
- Dargahi-Noubary, G. R. (1992). Discrimination between Gaussian time series based on their spectral differences. Commun. Statist.-Theory Meth., 21(9), 2439-2458.
- Diggle, P. J. (1990). Time Series: A Biostatistical Introduction. Clarendon Press, London, United Kingdom.
- Dunsmore, I. R. (1966). A Bayesian Approach to Classification. J.R. Statist. Soc. B 28, 568-77.
- Fisher, R. A. (1935). The Fiducial Argument in Statistical Inference. Ann. Eugenics, 6, 391-8.
- Geisser, S. (1964). Posterior Odds in Multivariate Classification. J. R. Statist. Soc. B, 26, 69-76.
- Geisser, S. (1966). Predictive Discrimination. In P. R. Krishnaiah (Ed.), Multivariate Analysis. Academic Press, New York, N. Y.
- Geisser, (1982). Bayesian Discrimination. In P. R. Krishnaiah and L. N. Kanal (Ed.), Handbook of Statistics, Vol 2. North Holland, Amsterdam, Netherlands.
- Hannan, E. J. and Quinn, B. G. (1979). The Determination of the Order of an Autoregression. J. R. Statist. Soc. B, 41, No. 2, 190-195.

- Harrison, P. J. and Stevens, C. F. (1976). Bayesian Forecasting (with Discussion). J. R. Statist. Soc. B, 8, 205-47.
- Hermans, J. and Habbema, J. D. F. (1975). Comparison of Five Methods to Estimate Posterior Probabilities. EDV in Med. and Biol., 6, 14-19.
- Jeffreys, H. (1961). Theory of Probability. Oxford University Press, London.
- Kashyap, R. L. (1978). Optimal Feature Selection and Decision Rules in Classification Problems with Time Series. IEEE Trans. Information Theory, IT-24, 281-88.
- Kryzsko, M. (1983). The Discriminant Analysis of Multivariate Time Series. IEEE Trans. Information Theory, IT-29, 612-4.
- Kullback, S. (1958). Information Theory and Statistics. Dover Publications, Inc., New York, N.Y.
- Liggett, W. S. (1971). On the Asymptotic Optimality of Spectral Analysis for Testing Hypotheses about Time Series. Annals of Mathematical Statistics, 42, 1348-58.
- Lutkepohl, H. (1985). Comparison of Criteria for Estimating the Order of a Vector Autoregressive Process. Journal of Time Series Analysis, Vol. 6, No. 1, 35-52.
- McLachlan, G. J. (1992). Discriminant Analysis and Statistical Pattern Recognition. John Wiley and Sons, Inc., New York, N. Y.
- Moran, M. A. and Murphy, B. J. (1979). A Closer Look at Two Alternative Methods of Statistical Discrimination. Applied Statistics, 28, 223-232.
- Newton, H. J. (1988). Timeslab: A Time Series Analysis Laboratory. Wadsworth, Pacific Grove, California.
- Okamoto, M. (1963). An Asymptotic Expansion for the Distribution of the Linear Discriminant Function. Annals of Mathematical Statistics, 34, 1286-1301 (Correction, 39 (1968), 1358-1359).
- Okamoto, M. (1973). Distinctiveness of the Eigenvalues of a Quadratic Sample. Annals of Statistics, 1, 763-765.

- Parzen, E. (1982). Maximum Entropy Interpretation of Autoregressive Spectral Densities. Statistics and Probability Letters, 1, 2-6.
- Priestley, M. B. (1981). Spectral Analysis of Time Series, Vol 1 and 2. Academic Press, London, New York.
- SAS System for Personal Computers, Release 6.03, SAS Institute Inc., SAS Circle, Box 8000, Cary, North Carolina, 27512-8000.
- Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Mathematical Statistics, Vol. 6, No. 2, 461-464.
- Shaarawy, S. and Broemeling, L. (1984). Bayesian Inferences and Forecasting with Moving Average Processes. Comm. Statist. Theor. Math. 13, (150), 1871-88
- Shibata, R. (1976). Selection of the Order of an Autoregressive Model by Akaike's Criterion. Biometrika, 63, 117-126.
- Shumway, R. H. (1982). Discriminant Analysis for Time Series. In P. R. Krishnaiah, and L. N. Kanal, (Ed.), Handbook of Statistics, Vol 2. North Holland, Amsterdam, Netherlands.
- Shumway, R. H. and Unger, A. N. (1974). Linear Discriminant Functions for Stationary Time Series. J. Amer. Statist. Assoc., 69, 948-59.
- Siotani, M., and Wang, R. H. (1975). Further Expansion Formulae for Error Rates and Comparison of the W- and the Z-Procedures in Discriminant Analysis. Technical Report No. 33, Department of Statistics, Kansas State University, Manhattan, Kansas.
- Siotani, M., and Wang, R. H. (1977). Asymptotic Expansions for Error Rates and Comparison of the W- and the Z-Procedures in Discriminant Analysis. North-Holland, Amsterdam, Netherlands.
- Tjostheim, D. (1975). Autoregressive Representation of Seismic P-Wave Signals with an Application to the Problem of Short Period Discriminants. Geophys. J. Roy. Astrn. Soc., 43, 269-91.
- Tylor, S. R. and Marshall, P.D. (1991). Spectral discriminant between Soviet explosions and earthquake using short-period array data. Geophys. J. International, 106, 265-273.

- Ulrich, T. J. and Bishop, T. N. (1975). Maximum Entropy Spectral Analysis and Autoregressive Decomposition. Reviews of Geophysics and Space Physics, 13 (1), 183-200.
- Wahba, G. (1968). On the Distribution of Some Statistics Useful in the Analysis of Jointly Stationary Time Series. Annals of Mathematical Statistics, 39, 1849-62.
- Wald, A. (1944). On a Statistical Problem Arising in the Classification of an Individual into one of two groups. Annals of Mathematical Statistics, 15, 145-62.
- Wei, W. W. S. (1990). Time Series Analysis: Univariate and Multivariate Methods. Addison-Wesley, Reading, Massachusetts.
- Zellner, A. (1971). An Introduction to Bayesian Inference in Econometrics. John Wiley and Sons, Inc., New York, N. Y.
- Zellner, A. and Chetty, V. K. (1965). Prediction and Decision Problems in Regression Models from the Bayesian Point of View. J. Amer. Statist. Ass., 60, 608-616.

APPENDICES

APPENDIX A

SOME FORMULA DERIVATIONS

Suppose that a given series X_n obeys the autoregressive model with parameters φ , τ and p . Assume that for fixed p , the joint prior density of φ and τ is the multivariate normal gamma defined in Chapter III, section 3.2.2.

1. $P(X_i)$, the Marginal Density of X_i

From the definition of $P(X_i)$ in Chapter III, section 3.2, we have

$$P(X_N) = \frac{P(X_0) |\Sigma|^{1/2}}{(2\pi)^{(N-p)/2}} \int_{\tau} \tau^{\frac{N+\alpha}{2}-1} \exp\left\{-\frac{\tau}{2} (\hat{\tau}^{-1} + \beta)\right\} \\ \times \int_{\varphi} \exp\left\{-\frac{\tau}{2} [(\varphi - \hat{\varphi})' A (\varphi - \hat{\varphi}) + (\varphi - \mu)' \Sigma (\varphi - \mu)]\right\} d\varphi d\tau,$$

where $\hat{\tau}^{-1}$ and $\hat{\varphi}$ are respectively sample estimates from the training realization X and are defined on Page 28. From Box and Tiao (1973, p. 418), the quadratic forms under the second integral may be combined as follows:

$$(\varphi - \hat{\varphi})' A (\varphi - \hat{\varphi}) + (\varphi - \mu)' \Sigma (\varphi - \mu) = (\varphi - c)' (A + \Sigma) (\varphi - c) + k ,$$

where $c_1 = (A+\Sigma)^{-1}(A\phi+\Sigma\mu)$,

and $k_1 = (\phi-\mu)'A(A+\Sigma)^{-1}\Sigma(\phi-\mu)$.

Hence the integral with respect to ϕ becomes

$$\begin{aligned} & \exp\left(-\frac{\tau}{2}k_1\right) \int_{\phi} \exp\left\{-\frac{\tau}{2}(\phi-c_1)'(A+\Sigma)(\phi-c_1)\right\} d\phi \\ &= \exp\left(-\frac{\tau}{2}k_1\right) (2\pi/\tau)^{\frac{p}{2}} |A+\Sigma|^{-1/2}. \end{aligned}$$

Thus $P(X_N)$ reduces to

$$P(X_0)(2\pi)^{-(N-2p)/2} \left(\frac{1}{|A+\Sigma|}\right)^{1/2} \int_{\tau}^{\frac{N+\alpha-p}{2}-1} \exp\left\{-\tau\left(\frac{1}{2}(\hat{\tau}^{-1}+\beta+k_1)\right)\right\} d\tau$$

The integral part is evaluated as

$$\left[\frac{1}{2}(\hat{\tau}^{-1}+\beta+k_1)\right]^{-\frac{1}{2}(N+\alpha-p)} \Gamma\left(\frac{N+\alpha-p}{2}\right)$$

A combination of the last two expressions and taking expectation with respect to $P(p_i)$, the distribution of p_i yields equation (8), the expression for $P(X_N)$.

2. The Predictive Density, $P(Y_N|X_N)$

The numerator of $P(Y_N|X)$ (suppressing the n subscript is

$$\begin{aligned} & \frac{P(X_0)|\Sigma|^{1/2}}{(2\pi)^{(n-p)/2}} \int_{\tau}^{\frac{N+n+\alpha-p}{2}-1} \exp\left\{-\frac{\tau}{2}(\hat{\tau} + \tilde{\tau} + \beta)\right\} \\ & \times \int_{\phi} \exp\left\{-\frac{\tau}{2}[(\phi-\phi)'A(\phi-\phi) + (\phi-\mu)'\Sigma(\phi-\mu) + (\phi-\phi)'A_0(\phi-\phi)]\right\} d\phi d\tau. \end{aligned}$$

Combining the quadratic forms successively, the integral with respect to φ is reduced to

$$\begin{aligned} & \exp(-\tau/2) \int_{\varphi} \exp\left\{-\frac{\tau}{2}(\varphi-c_2)' \tau(A_0+A+\Sigma)(\varphi-c_2)\right\} d\varphi \\ &= \exp(-\tau/2) (2\pi/\tau)^{p/2} |A_0+A+\Sigma|^{-1/2} \end{aligned}$$

where

$$\begin{aligned} c_0 &= (A_0 + A)^{-1} (A_0 \varphi + A \varphi) \\ k_0 &= (\tilde{\varphi} - \hat{\varphi})' A_0 (A + A_0)^{-1} A (\tilde{\varphi} - \hat{\varphi}) \\ c_2 &= (A_0 + A + \Sigma)^{-1} [(A_0 + A) c_0 + \Sigma \mu] \\ k_2 &= (c_0 - \mu)' (A_0 + A) (A_0 + A + \Sigma)^{-1} \Sigma (c_0 - \mu) \end{aligned}$$

The numerator of $P(Y|X_n)$ then becomes

$$\frac{P(X_0)P(Y_0)}{(2\pi)^{(N+n-3p)/2}} \left(\frac{|\Sigma|}{|A+A_0+\Sigma|} \right)^{1/2} \int_{\tau} \tau^{\frac{\eta+\alpha-q}{2}-1} \exp\left(-\frac{\tau u}{2}\right) d\tau,$$

where $u = N\tilde{\tau}^{-1} + n\hat{\tau}^{-1} + \beta + k_2 + k_0$
and $\eta = N + n + q + 2 - 2p$.

Noting that the integral part of the last expression is

$$\left(\frac{u}{2}\right)^{-\frac{1}{2}(\eta+\alpha-q)} \Gamma\left(\frac{\eta+\alpha-q}{2}\right),$$

the predictive density in equation (11) is obtained upon division of the numerator of $P(Y|X_n)$ by $P(X_n)$.

APPENDIX B

SAS SOURCE CODE FOR THE SIMULATION STUDY

```

PROC IML;
  START PAIC(X,N,MXP);
    PH=J(MXP,MXP,0);
    PART=J(MXP,1,0);
    ERRVAR=PART;
    AC=PART;
    AIC=PART;
    SQ=SSQ(X);
    ERRV0=SQ/N;
    AC(1)=COVLAG(X,1)/SQ;
    PH(1,1)=AC(1);
    PART(1)=AC(1);
    ERRVAR(1)=ABS(ERRV0*(1-PART(1)**2));
    AIC(1)=N*LOG(ERRVAR(1))+2*LOG(N);
    DO I=2 TO MXP;
      COV=COVLAG(X,I);
      AC(I)=COV(1,I)/SQ;
      NUM=0.0;
      DEN=0.0;
      DO K=1 TO I-1;
        NUM=NUM+PH(I-1,K)*AC(I-K);
        DEN=DEN+PH(I-1,K)*AC(K);
      END;
      PART(I)=(AC(I)-NUM)/(1-DEN);
      ERRVAR(I)=ABS(ERRVAR(I-1)*(1-PART(I)**2));
      PH(I,I)=PART(I);
      AIC(I)=N*LOG(ERRVAR(I))+2*LOG(N);
      DO K=1 TO I-1;
        PH(I,K)=PH(I-1,K)-PART(I)*PH(I-1,I-K);
      END;
    END;
  END PAIC;

```

```

        END;

    END;

    P=AIC(|>:<|);
    RETURN (P);
    FINISH PAIC;

*BEGIN BGYW SUBROUTINE;
START BGYW(PHIBG,PHIYW,ERRVBG,ERRVYW,X,N,P);
    PH=J(P,P,0);
    PART=J(P,1,0);
    ERRVAR=PART;
    AC=PART;
    AIC=PART;
    SQ=SSQ(X);
    ERRVO=SQ/N;
    AC(|1|)=COVLAG(X,1)/SQ;
    PH(|1,1|)=AC(|1|);
    PART(|1|)=AC(|1|);
    ERRVAR(|1|)=ABS(ERRVO*(1-PART(|1|)**2));
    AIC(|1|)=N*LOG(ERRVAR(|1|))+2*LOG(N);
IF P>1 THEN
    DO I=2 TO P;
        COV=COVLAG(X,I)/SQ;
        AC(|I|)=COV(|1,I|);
        NUM=0.0;
        DEN=0.0;
        DO K=1 TO I-1;
            NUM=NUM+PH(|I-1,K|)*AC(|I-K|);
            DEN=DEN+PH(|I-1,K|)*AC(|K|);
        END;
        PART(|I|)=(AC(|I|)-NUM)/(1-DEN);
        ERRVAR(|I|)=ABS(ERRVAR(|I-1|)*(1-PART(|I|)**2));
        PH(|I,I|)=PART(|I|);
        AIC(|I|)=N*LOG(ERRVAR(|I|))+2*LOG(N);
        DO K=1 TO I-1;
            PH(|I,K|)=PH(|I-1,K|)-PART(|I|)*PH(|I-1,I-K|);

```

```

        END;
    END;
    ERRVYW=ERRVAR(|P|);
    PHIYW=PH(|P,|);
*END YULE-WALKER ESTIMATES;
*BEGIN BURG ESTIMATES;
    B1=J(P,N-1,0);
    B2=J(P,N-1,0);
    DEN=0;
    NUM=0;
    DO T=1 TO N-1;
        B1(|1,T|)=X(|T|);
        B2(|1,T|)=X(|T+1|);
        DEN=DEN+B1(|1,T|)**2+B2(|1,T|)**2;
        NUM=NUM+B1(|1,T|)*B2(|1,T|);
    END;
    AC(|1|)=2*NUM/DEN;
    PH(|1,1|)=AC(|1|);
    PART(|1|)=AC(|1|);
    ERRVAR(|1|)=ABS(ERRVO*(1-PART(|1|)**2));
    AIC(|1|)=N*LOG(ERRVAR(|1|))+2*LOG(N);
    IF P>1 THEN
        DO I=2 TO P;
            NUM=0.0;
            DEN=0.0;
            DO T=1 TO N-I;
                TEMP=PART(|I-1|);
                B1(|I,T|)=B1(|I-1,T|)-TEMP*B2(|I-1,T|);
                B2(|I,T|)=B2(|I-1,T+1|)-TEMP*B1(|I-1,T|);
                NUM=NUM+B1(|I,T|)*B2(|I,T|);
                DEN=DEN+B1(|I,T|)**2+B2(|I,T|)**2;
            END;
            PART(|I|)=2*NUM/DEN;
            ERRVAR(|I|)=ABS(ERRVAR(|I-1|)*(1-PART(|I|)**2));
            PH(|I,I|)=PART(|I|);
        END;
    END;

```

```

      AIC(|I|)=N*LOG(ERRVAR(|I|))+2*LOG(N);
      DO K=1 TO I-1;
        PH(|I,K|)=PH(|I-1,K|)-PART(|I|)*PH(|I-1,I-K|);
      END;
    END;
    PHIBG=PH(|P,|);
    ERRVBG=ERRVAR(|P|);
    *END BURG ESTIMATES;
  FINISH BGYW;
*END BGYW SUBROUTINE;
*BEGIN SUBROUTINE D-MATRIX AT FREQUENCY FREQ;
  START DMATRIX(D,P,FREQ,N);
    D=J(P,P,0);
    DO I=1 TO P;
      DO J=1 TO P;
        K=ABS(I-J)/N;
        D(|I,J|)=COS(K*FREQ);
      END;
    END;
  FINISH DMATRIX;
*END SUBROUTINE DMATRIX;
*BEGIN SUBROUTINE DISCFN;
  START DISCFN(TDISCFN,FDISCFN,JDIV,SER,IP1,IP2,N,PH1,PH2,
    VAR1,VAR2,PH01,PH02,VAR01,VAR02,D1,D2,PKFRQ,MAXPK);
    Z1=J(N-IP1,1,0);
    W1=J(N-IP1,IP1,0);
    Z2=J(N-IP2,1,0);
    W2=J(N-IP2,IP2,0);
    Z1=SER(|IP1+1:N|);
    W1=SER(|IP1:N-1|);
    IF IP1>1 THEN
      DO M=2 TO IP1;
        W1=W1||SER(|IP1-M+1:N-M|);
      END;
    Z2=SER(|IP2+1:N|);

```

```

W2=SER(|IP2:N-1|);
IF IP2>1 THEN
  DO M=2 TO IP2;
    W2=W2||SER(|IP2-M+1:N-M|);
  END;
TWOPI=8*ATAN(1);
E=(N-IP2)*LOG(VAR2*TWOPI)-(N-IP1)*LOG(VAR1*TWOPI);
TDISCFN=(SSQ(Z2-W2*PH2`))/VAR2-(SSQ(Z1-W1*PH1`))/VAR1+E;
FDISCFN=0;
JDIV=0;
DO I=1 TO MAXPK;
  FREQ=PKFRQ(|I|);
  CALL DMATRIX(D1,IP1,FREQ,N);
  CALL DMATRIX(D2,IP2,FREQ,N);
  F1=PH1*D1*PH1`/VAR1;
  F01=PH01*D1*PH01`/VAR01;
  F2=PH2*D2*PH2`/VAR2;
  F02=PH02*D2*PH02`/VAR02;
  FDISCFN=FDISCFN+F1/F01-F2/F02+LOG(F2/F1);
  JDIV=JDIV+(F1/F01-F2/F02+LOG(F2/F1)-2)/(2*N);
END;
FINISH DISCFN;
*END SUBROUTINE DISCFN;
*BEGIN SUBROUTINE STATS;
  START STATS(V,Z,W,A,AINV,WZ,PHI,Y,N,P);
  Z=Y(|P+1:N|);
  W=Y(|P:N-1|);
  DO M=2 TO P;
    W=W||Y(|P-M+1:N-M|);
  END;
  A=W`*W;
  AINV=INV(A);
  WZ=W`*Z;
  PHI=SOLVE(A,WZ);
  V=SSQ(Z-W*PHI);

```

```

        FINISH STATS;
*END SUBROUTINE STATS;
*BEGIN SUBROUTINE GAMM;
    START GAMM(E,A);
        GAM=1.0;
        DEN=2*A;
        DEND=4*A;
    IF MOD(E,2)=0.0 THEN
        DO J=2 TO E/2+1;
            GAM=GAM*(J-1)/DEN;
        END;
    ELSE
        DO J=1 TO (E-1)/2;
            GAM=GAM*(2*J-1)/(DEND);
        END;
    IF MOD(E,2)=0.5 THEN GAM=GAM*SQRT(22/7);
    RETURN (GAM);
FINISH GAMM;
*END SUBROUTINE GAMM;
;;
FINALPR=J(1,7,1);
CPR="LTRAIN" "LTEST" "JDIV" "ERROTMPR"
    "ERROFRPR" "SDTMPR" "SDFRPR";
CREATE PRED.DATA FROM FINALPR (|COLNAME=CPR|);
FINALEST=J(1,11,0);
C="LTRAIN" "LTEST" "JDIV" "ERROTMBG" "ERROTMYW" "ERROFRBG"
    "ERROFRYW" "SDTMBG" "SDTMYW" "SDFRBG" "SDFRYW";
CREATE EST.DATA FROM FINALEST (|COLNAME=C|);
SIGMA1= $\sigma_1$ ;
SIGMA2= $\sigma_2$ ;
AR1={1  $\varphi_1$ };
AR2={1  $\varphi_2$ };
SEED=563693504;
NTEST=100;
MAXP=MAXIMUM AR ORDER ALLOWED;

```

```

LTRAIN=25;
DO LTR=1 TO 4 BY 1;
  LTRAIN=LTRAIN*2;
  X1=ARMASIM(AR1,1,0,1,LTRAIN,SEED);
  X2=ARMASIM(AR2,1,0,1,LTRAIN,SEED);
  X1=X1-X1(!:!);
  X2=X2-X2(!:!);
  P1=PAIC(X1,LTRAIN,MAXP);
  P2=PAIC(X2,LTRAIN,MAXP);
  FREE PH ERRVAR AIC PART DEN NUM AC;
  RUN BGYW(PHIBG1,PHIYW1,ERRVBG1,ERRVYW1,X1,LTRAIN,P1);
  RUN BGYW(PHIBG2,PHIYW2,ERRVBG2,ERRVYW2,X2,LTRAIN,P2);
  FREE ERRVAR1 ERRVAR B1 B2 PART1 PH1 AC1 AIC1 PHIBG PHIYW;
  FREE ERRVAR B1 B2 PART PH AC AIC;
  LTEST=25;
  DO LTE=1 TO 4 BY 1;
    LTEST=LTEST*2;
    MAXPEAKS=.1*LTEST;
    *GENERATE NTEST REALIZATIONS TO BE CLASSIFIED.;
    NTBG1=0;
    NTYW1=0;
    NTBG2=0;
    NTYW2=0;
    NFBG1=0;
    NFW1=0;
    NFBG2=0;
    NFW2=0;
    NT1=J(3,1,0);
    NT2=NT1;
    NF1=NT1;
    NF2=NT1;
    ERROTMPR=J(3,1,0);
    ERROFRPR=ERROTMPR;
    SDTMPR=ERROTMPR;
    SDFRPR=ERROTMPR;
  
```



```

PKFREQ1=J(MAXPEAKS+1,1,0);
PKFREQ2=J(MAXPEAKS+1,1,0);
DO NT=1 TO NTEST;
  Y1=ARMASIM(AR1,1,0,1,LTEST,SEED);
  Y1=Y1-Y1(:,:);
  PKT=FFT(Y1);
  PK=PKT(:,##);
  DO I=1 TO MAXPEAKS;
    PKFREQ1(I)=PK(<:>);
    K=PKFREQ1(I);
    PK(K)=-1*PK(K);
  END;
  Y2=ARMASIM(AR2,1,0,1,LTEST,SEED);
  Y2=Y2-Y2(:,:);
  PKT=FFT(Y2);
  PK=PKT(:,##);
  DO I=1 TO MAXPEAKS;
    PKFREQ2(I)=PK(<:~>);
    K=PKFREQ2(I);
    PK(K)=-1*PK(K);
  END;
  FREE PK;
  RUN BGYW(PHIBG1,PHIYW1,ERRVBG1,ERRVW1,Y1,
    LTEST,P1);
CALL DISCFN(TIMEBG1,FREQBG1,JDIVBG,Y1,P1,P2,LTEST,PHIBG1,
  PHIBG2,ERRVBG1,ERRVBG2,PHIBG1,ERRVBG1,D1,
  D2,PKFREQ1,MAXPEAKS);
CALL DISCFN(TIMEYW1,FREQYW1,JDIVYW,Y1,P1,P2,LTEST,PHIYW1,
  PHIYW2,ERRVW1,ERRVW2,PHIYW1,ERRVW1,D1,
  D2,PKFREQ1,MAXPEAKS);
IF TIMEBG1<0 THEN NTBG1=NTBG1+1;
IF TIMEYW1<0 THEN NTYW1=NTYW1+1;
IF FREQBG1>0 THEN NFBG1=NFBG1+1;
IF FREQYW1>0 THEN NFW1=NFW1+1;

```

```

RUN BGYW(PHIBG02,PHIYW02,ERRVBG02,ERRVYW02,Y2,LTEST,P2);
CALL DISCFN(TIMEBG2,FREQBG2,JDIVBG,Y2,P1,P2,LTEST,PHIBG1,
            PHIBG2,ERRVBG1,ERRVBG2,PHIBG02,D1,D2,PKFREQ2,
            MAXPEAKS);
CALL DISCFN(TIMEYW2,FREQYW2,JDIVYW,Y2,P1,P2,LTEST,PHIYW1,
            PHIYW2,ERRVYW1,ERRVYW2,PHIYW02,D1,D2,PKFREQ2,
            MAXPEAKS);
    IF TIMEBG2>0 THEN NTB2=NTB2+1;
    IF TIMEYW2>0 THEN NTY2=NTY2+1;
    IF FREQBG2<0 THEN NFB2=NFB2+1;
    IF FREQYW2<0 THEN NFY2=NFY2+1;
TMNUM1=J(3,1,0);
TMNUM2=TMNUM1;
TMDEN1=TMNUM1;
TMDEN2=TMNUM1;
TMNUM11=TMNUM1;
TMNUM12=TMNUM1;
TMDEN21=TMNUM1;
TMDEN22=TMNUM1;
TMNUM21=TMNUM1;
TMNUM22=TMNUM1;
TMDEN11=TMNUM1;
TMDEN12=TMNUM1;
TIMEPR1=TMNUM1;
TIMEPR2=TMNUM1;
*FIND PREDICTIVE DENSITIES IN TIME AND FREQUENCY DOMAINS;
FR1=J(3,1,0);
FR2=FR1;
FR01=FR1;
FR02=FR1;
FREQPR1=FR1;
FREQPR2=FR1;
FR11=FR1;
FR12=FR1;

```

```

FR21=FR1;
FR22=FR1;
  DO K=1 TO MAXPEAKS;
    DO P=1 TO MAXP;
      OMEGA=PKFREQ1(|K|);
      RUN STATS(V1,Z1,W1,A1,A1INV,WZ1,PHI1,X1,LTRAIN,P);
      RUN STATS(V2,Z2,W2,A2,A2INV,WZ2,PHI2,X2,LTRAIN,P);
      RUN STATS(V0,Z0,W0,A0,A0INV,WZ0,PHI0,Y1,LTEST,P);
    *EVALUATE TIME DOMAIN PREDICTIVE DISCRIMINANT FUNCTIONS;
    IF K=1 THEN
      DO;
        SQ=SQRT(22/14);
        EN=LTRAIN+LTEST-3*P;
        E11=(SQ)**(EN);
        AA01=INV(A0+A1);
        E12=SQRT(ABS(DET(AA01)));
        K01=(PHI0-PHI1)`*A0*AA01*A1*(PHI0-PHI1);
        A01=V1+V0+K01;
        NUM11=GAMM(EN,A01)*E11*E12;
        ED=LTRAIN-2*P;
        D11=(SQ)**ED/SQRT(ABS(DET(A1)));
        DEN11=D11*GAMM(ED,V1);
        E12=1/SQRT(ABS(DET(A2+A0)));
        AA02=INV(A0+A2);
        K02=(PHI0-PHI2)`*A0*AA01*A2*(PHI0-PHI2);
        A02=V2+V0+K02;
        NUM12=GAMM(EN,A02)*E11*E12;
        D11=(SQ)**ED/SQRT(ABS(DET(A2)));
        DEN12=D11*GAMM(ED,V2);
        IF P=P1 THEN
          DO I=1 TO 3;
            TMNUM11(|I|)=TMNUM11(|I|)+NUM11*(0.65+I*0.1);
            TMDEN11(|I|)=TMDEN11(|I|)+DEN11*(0.65+I*0.1);
            TMNUM12(|I|)=TMNUM12(|I|)+NUM12*(0.65+I*0.1);
            TMDEN12(|I|)=TMDEN12(|I|)+DEN12*(0.65+I*0.1);
          END DO;
        END DO;
      END IF;
    END DO;
  END DO;

```

```

      END;
    ELSE
      DO I=1 TO 3;
        TMNUM11(I)=TMNUM11(I)+NUM11*(0.35-I*0.1);
        TMDEN11(I)=TMDEN11(I)+DEN11*(0.35-I*0.1);
        TMNUM12(I)=TMNUM12(I)+NUM12*(0.35-I*0.1);
        TMDEN12(I)=TMDEN12(I)+DEN12*(0.35-I*0.1);
      END;
    *EVALUATE TIME DOMAIN PREDICTIVE DISCRIMINANT FUNCTIONS;
    RUN STATS(V0,Z0,W0,A0,A0INV,WZ0,PHI0,Y2,LTEST,P);
    AA01=INV(A0+A1);
    E12=SQRT(ABS(DET(AA01)));
    K01=(PHI0-PHI1)*A0*AA01*A1*(PHI0-PHI1);
    A01=V1+V0+K01;
    NUM21=GAMM(EN,A01)*E11*E12;
    D11=(SQ)**ED/SQRT(ABS(DET(A1)));
    DEN21=D11*GAMM(ED,V1);
    AA02=INV(A0+A2);
    E12=SQRT(ABS(DET(AA02)));
    K02=(PHI0-PHI2)*A0*AA02*A2*(PHI0-PHI2);
    A02=V2+V0+K02;
    NUM22=GAMM(EN,A02)*E11*E12;
    D11=(SQ)**ED/SQRT(ABS(DET(A2)));
    DEN22=D11*GAMM(ED,V2);
    IF P=P2 THEN
      DO I=1 TO 3;
        TMNUM21(I)=TMNUM21(I)+NUM21*(0.65+I*0.1);
        TMDEN21(I)=TMDEN21(I)+DEN21*(0.65+I*0.1);
        TMNUM22(I)=TMNUM22(I)+NUM22*(0.65+I*0.1);
        TMDEN22(I)=TMDEN22(I)+DEN22*(0.65+I*0.1);
      END;
    ELSE
      DO I=1 TO 3;
        TMNUM21(I)=TMNUM21(I)+NUM21*(0.35-I*0.1);
        TMDEN21(I)=TMDEN21(I)+DEN21*(0.35-I*0.1);

```

```

        TMNUM22(I)=TMNUM22(I)+NUM22*(0.35-I*0.1);
        TMDEN22(I)=TMDEN22(I)+DEN22*(0.35-I*0.1);
    END;

END;

* ESTIMATE SPECTRAL DENSITY;
    RUN STATS(V0,Z0,W0,A0,AOINV,WZ0,PHI0,Y1,LTEST,P);
    CALL DMATRIX(D,P,OMEGA,LTEST);
    F1=TRACE(D*A1)+(LTRAIN+1)*PHI1`*D*PHI1/V1;
    F2=TRACE(D*A2)+(LTRAIN+1)*PHI2`*D*PHI2/V2;
    F0=TRACE(D*A0)+(LTRAIN+1)*PHI0`*D*PHI0/V0;
    IF P=P1 THEN
        DO I=1 TO 3;
            FR11(I)=FR11(I)+F1*(0.65+I*0.1);
            FR01(I)=FR01(I)+F0*(0.65+I*0.1);
            FR12(I)=FR12(I)+F2*(0.65+I*0.1);
        END;
    ELSE
        DO I=1 TO 3;
            FR11(I)=FR11(I)+F1*(0.35-I*0.1);
            FR01(I)=FR01(I)+F0*(0.35-I*0.1);
            FR12(I)=FR12(I)+F2*(0.35-I*0.1);
        END;
    OMEGA=PKFREQ2(K);
    RUN STATS(V0,Z0,W0,A0,AOINV,WZ0,PHI0,Y2,LTEST,P);
    F0=TRACE(D*A0)+(LTRAIN+1)*PHI0`*D*PHI0/V0;
    IF P=P2 THEN
        DO I=1 TO 3;
            FR21(I)=FR21(I)+F1*(0.65+I*0.1);
            FR02(I)=FR02(I)+F0*(0.65+I*0.1);
            FR22(I)=FR22(I)+F2*(0.65+I*0.1);
        END;
    ELSE
        DO I=1 TO 3;
            FR21(I)=FR21(I)+F1*(0.35-I*0.1);
            FR02(I)=FR02(I)+F0*(0.35-I*0.1);

```

```

        FR22(I)=FR22(I)+F2*(0.35-I*0.1);
    END;
END; *END LOOP FOR P=1 TO MAXIMUM P;
DO I=1 TO 3;
    TEMP1=TMNUM11(I)*TMDEN12(I);
    TEMP2=TMDEN11(I)*TMNUM12(I);
    TIMEPR1(I)=LOG(TEMP1/TEMP2);
    PT1=FR11(I)/FR01(I);
    PT2=FR12(I)/FR01(I);
    PT3=LOG(FR12(I)/FR11(I));
    FREQPR1(I)= FREQPR1(I)+PT1-PT2+PT3;
    TEMP1=TMNUM21(I)*TMDEN22(I);
    TEMP2=TMDEN21(I)*TMNUM22(I);
    TIMEPR2(I)=LOG(TEMP1/TEMP2);
    PT1=FR21(I)/FR02(I);
    PT2=FR22(I)/FR02(I);
    PT3=LOG(FR22(I)/FR21(I));
    FREQPR2(I)= FREQPR2(I)+PT1-PT2+PT3;
END;
END; *END LOOP FOR PEAK FREQUENCIES;
DO I=1 TO 3;
    IF TIMEPR1(I)< 0 THEN NT1(I)=NT1(I)+1;
    IF FREQPR1(I)> 0 THEN NF1(I)=NF1(I)+1;
    IF TIMEPR2(I)> 0 THEN NT2(I)=NT2(I)+1;
    IF FREQPR2(I)< 0 THEN NF2(I)=NF2(I)+1;
END; *END I PROB;
END;*GENERATION OF NTEST REALIZATIONS;
    TEMP=2*NTEST;
DO I=1 TO 3;
    ERROTMPR(I)=100*(NT1(I)+NT2(I))/TEMP;
    ERROFRPR(I)=100*(NF1(I)+NF2(I))/TEMP;
    SDTMPR(I)=SQRT(ERROTMPR(I)*(100-ERROTMPR(I))/TEMP);
    SDFRPR(I)=SQRT(ERROFRPR(I)*(100-ERROFRPR(I))/TEMP);
END;
    ERROTM75=ERROTMPR(1);

```

```

ERROTM85=ERROTMPR(|2|);
ERROFRPR=ERROFRPR(|1|);
  SDTMPR=SDTMPR(|1|);
  SDFRPR=SDFRPR(|1|);
  ERROTMYW=100*(NTYW1+NTYW2)/TEMP;
  ERROTMBG=100*(NTBG1+NTBG2)/TEMP;
  ERROFRYW=100*(NFWY1+NFWY2)/TEMP;
  ERROFRBG=100*(NFBG1+NFBG2)/TEMP;
  SDTMYW=SQRT(ERROTMYW*(100-ERROTMYW)/TEMP);
  SDTMBG=SQRT(ERROTMBG*(100-ERROTMBG)/TEMP);
  SDFRYW=SQRT(ERROFRYW*(100-ERROFRYW)/TEMP);
  SDFRBG=SQRT(ERROFRBG*(100-ERROFRBG)/TEMP);
FINALEST=LTRAIN||LTEST||JDIV||ERROTMBG||ERROTMYW||
      ERROFRBG||ERROFRYW||SDTMBG||SDTMYW||SDFRBG||SDFRYW;
SETOUT EST.DATA; APPEND FROM FINALEST;
;
FINALPR=LTRAIN||LTEST||JDIV||ERROTMPR||ERROFRPR||
      SDTMPR||SDFRPR;
SETOUT PRED.DATA; APPEND FROM FINALPR;
END; * END DO LOOP FOR LTE=1 TO LTEST;
END; * END TRAINING DATA GENERATION: DO LOOP LTR=1,LTRAIN;
CLOSE EST.DATA;
CLOSE PRED.DATA;
PROC PRINT DATA=EST.DATA;
PROC PRINT DATA=PRED.DATA;

```

APPENDIX C

SUBROUTINES USED IN THE SIMULATION STUDY

1. BGYW(PHBG,PHIYW,ERRVBG,ERRVYW,X,N,P);

PURPOSE: To estimate AR coefficients and error variance by the Burg and Yule-Walker estimation criteria.

INPUT: An AR series X of size N and order P.

OUTPUT: PHIBG, PHIYW, ERRVBG, ERRVYW, where

PHIBG = Burg estimate of ϕ , the AR coefficients,

PHIYW = Yule-Walker estimate of ϕ ,

ERRVBG = Burg estimate of $1/\tau$, the error variance,

ERRVYW = Yule-Walker estimate of $1/\tau$.

2. DISCFN(TDISCFN,FDISCFN,JDIV,SER,IP1,IP2,N,PH1,PH2,
VAR1,VAR2,PH01,PH02,VAR01,VAR02,D1,D2,PKFRQ,MAXPK)

PURPOSE: To evaluate the discriminant functions in time and frequency domains and also to evaluate the J-divergence rate.

INPUT: SER=test realization of length N,

IP1=estimated order of AR class 1

IP2=estimated order of AR class 2,

N=length of test realization for classification,

PH1=estimate of ϕ using training data from class 1,

PH2=estimate of ϕ using training data from class 2,
 VAR1=estimate of error variance $1/\tau$ from training data
 in class 1,
 VAR2=estimate of error variance $1/\tau$ from training data in
 class 2,
 PH01=estimate of ϕ using test realization SER ASSUMED
 to come from class 1,
 PH02=estimate of ϕ using test realization SER ASSUMED
 to come from class 2,
 VAR01=estimate of τ using test realization SER assumed
 to come from class 1,
 VAR02=estimate of τ using test realization SER assumed
 to come from class 2,
 D1=matrix from subroutine DMATRIX assuming class 1,
 D2=matrix from subroutine DMATRIX assuming class 2,
 PKFRK=vector of peak frequencies identified by the FFT,
 Fast Fourier Transform,
 MAXPK=the maximum length of the vector PKFRK.
 OUTPUT: TDISCFN=Time domain discriminant function,
 FDISCFN=Frequency domain discriminant function,
 JDIV=J-divergence rate.

3. DMATRIX(D,P,FREQ,N)

PURPOSE: To create the $p \times p$ matrix, $D(\omega)$ at frequency ω ,

$$D(\omega) = \begin{pmatrix} 1 & \cos\omega & \dots & \cos(p-1)\omega \\ \cos\omega & 1 & \dots & \cos(p-2)\omega \\ \vdots & \vdots & \ddots & \vdots \\ \cos(p-1)\omega & \cos(p-2)\omega & \dots & 1 \end{pmatrix}$$

INPUT: P, FREQ, N where

P = the order of a series,

FREQ = the estimated frequency, and

N = the length of the AR series.

OUTPUT: A p by p symmetric Matrix D of cosine values.

4. GAMM(E,A)

PURPOSE: To compute the quantity $\frac{\Gamma(E/2)}{(A/2)^{E/2}}$

INPUT: E is a positive integer and A is greater than zero.

OUTPUT: The computed value of $\frac{\Gamma(E/2)}{(A/2)^{E/2}}$

5. PAIC(X,N,MXP)

PURPOSE: To use the AIC (the Akaike Information Criterion) to estimate the order of a given time series.

INPUT: The time series X of length N and maximum order MXP.

OUTPUT: The estimated order of X.

6. STATS(V,Z,W,A,AINV,WZ,PHI,Y,N,P)

PURPOSE: To obtain some basic statistics for the evaluation of discriminant functions.

INPUT: Y = (y(1), . . . , y(P))', a time series vector of length N and order P.

OUTPUT: Z = (y(P+1), . . . , y(N))',

$$W = \begin{pmatrix} y(P) & y(P+1) & \dots & y(1) \\ y(P+1) & y(P) & \dots & y(2) \\ \vdots & \vdots & & \vdots \\ y(N-1) & y(N-2) & \dots & y(N-P) \end{pmatrix},$$

$$A = W'W,$$

$$AINV = A^{-1}$$

$$WZ = W'Z,$$

$$PHI = AINV*W'Z,$$

$$V = (Y-W*PHI)'*(Y-W*PHI).$$

APPENDIX D

TABLES

TABLE I
ERROR RATES AND J-DIVERGENCE FROM AR(1)
VERSUS AR(1) CLASSIFICATION

- Notes: (a) DATA refers to the length of the training data,
(b) TEST refers to the length of the
test realization,
(c) JDIV refers to the J-Divergence rate,
(d) The first entry for each (DATA,TEST) tuple is
the misclassification percentage; the second
entry is the corresponding standard error.
- (e) TIME DOMAIN RATES: 1 = PREDICTIVE;
3 = BURG;
4 = YULE-WALKER
FREQUENCY DOMAIN RATES: 2 = PREDICTIVE;
5 = BURG;
6 = YULE-WALKER

TABLE I (Continued)

DATA	TEST	JDIV	CLASSIFICATION METHOD					
			1	2	3	4	5	6
50	50	0.94	42.94	41.86	48.46	47.52	43.63	51.66
			4.95	4.93	5.00	4.99	4.96	5.00
50	100	0.99	35.86	35.41	41.02	39.46	36.28	43.34
			4.80	4.78	4.92	4.89	4.81	4.96
50	200	1.10	28.34	28.33	32.84	31.02	28.55	34.40
			4.51	4.51	4.70	4.63	4.52	4.75
50	400	1.38	21.46	21.64	25.10	23.38	21.55	26.13
			4.11	4.12	4.34	4.23	4.11	4.39
100	50	1.06	34.67	33.44	38.68	38.56	35.36	41.56
			4.76	4.72	4.87	4.87	4.78	4.93
100	100	1.12	30.65	29.88	34.59	33.92	31.14	36.88
			4.61	4.58	4.76	4.73	4.63	4.82
100	200	1.25	25.57	25.26	29.26	28.14	25.88	30.91
			4.36	4.34	4.55	4.50	4.38	4.62
100	400	1.56	20.19	20.18	23.39	22.10	20.34	24.51
			4.01	4.01	4.23	4.15	4.03	4.30
200	50	1.37	26.62	25.48	29.46	29.70	27.21	31.83
			4.42	4.36	4.56	4.57	4.45	4.66
200	100	1.45	24.69	23.81	27.54	27.45	25.17	29.59
			4.31	4.26	4.47	4.46	4.34	4.56
200	200	1.62	21.82	21.27	24.62	24.15	22.17	26.25
			4.13	4.09	4.31	4.28	4.15	4.40
200	400	2.02	18.20	17.97	20.82	20.02	18.41	21.99
			3.86	3.84	4.06	4.00	3.88	4.14
400	50	2.29	19.74	18.80	21.73	22.06	20.20	23.56
			3.98	3.91	4.12	4.15	4.02	4.24
400	100	2.42	18.92	18.10	20.93	21.10	19.34	22.62
			3.92	3.85	4.07	4.08	3.95	4.18
400	200	2.70	17.54	16.92	19.57	19.51	17.89	21.03
			3.80	3.75	3.97	3.96	3.83	4.07
400	400	3.37	15.50	15.11	17.49	17.16	15.75	18.65
			3.62	3.58	3.80	3.77	3.64	3.90

TABLE II

ERROR RATES AND J-DIVERGENCE FROM AR(2)
VERSUS AR(1) CLASSIFICATION

- Notes: (a) DATA refers to the length of the training data,
(b) TEST refers to the length of the
test realization,
(c) JDIV refers to the J-Divergence rate,
(d) The first entry for each (DATA,TEST) tuple is
the misclassification percentage; the second
entry is the corresponding standard error.
- (e) TIME DOMAIN RATES: 1 = PREDICTIVE;
 3 = BURG;
 4 = YULE-WALKER
FREQUENCY DOMAIN RATES: 2 = PREDICTIVE;
 5 = BURG;
 6 = YULE-WALKER

TABLE II (Continued)

DATA	TEST	JDIV	CLASSIFICATION METHOD					
			1	2	3	4	5	6
50	50	1.03	31.21	30.51	40.72	43.57	35.69	44.15
			4.63	4.60	4.91	4.96	4.79	4.97
50	100	1.15	27.06	27.15	34.95	37.05	31.02	37.64
			4.44	4.45	4.77	4.83	4.63	4.84
50	200	1.42	22.12	22.66	28.32	29.78	25.41	30.32
			4.15	4.19	4.51	4.57	4.35	4.60
50	400	2.16	17.16	17.83	21.84	22.83	19.74	23.29
			3.77	3.83	4.13	4.20	3.98	4.23
100	50	1.23	24.37	23.21	32.11	34.65	27.81	35.02
			4.29	4.22	4.67	4.76	4.48	4.77
100	100	1.37	22.27	21.78	29.07	31.10	25.47	31.51
			4.16	4.13	4.54	4.63	4.36	4.65
100	200	1.69	19.30	19.37	24.93	26.42	22.13	26.85
			3.95	3.95	4.33	4.41	4.15	4.43
100	400	2.58	15.76	16.15	20.18	21.21	18.10	21.60
			3.64	3.68	4.01	4.09	3.85	4.12
200	50	1.76	18.26	17.02	24.24	26.32	20.80	26.56
			3.86	3.76	4.29	4.40	4.06	4.42
200	100	1.96	17.35	16.52	22.86	24.67	19.80	24.93
			3.79	3.71	4.20	4.31	3.98	4.33
200	200	2.42	15.86	15.50	20.69	22.14	18.13	22.43
			3.65	3.62	4.05	4.15	3.85	4.17
200	400	3.68	13.73	13.78	17.74	18.80	15.74	19.10
			3.44	3.45	3.82	3.91	3.64	3.93
400	50	3.59	13.34	12.25	17.78	19.38	15.17	19.54
			3.40	3.28	3.82	3.95	3.59	3.96
400	100	3.98	12.98	12.10	17.22	18.70	14.78	18.87
			3.36	3.26	3.78	3.90	3.55	3.91
400	200	4.92	12.33	11.74	16.24	17.53	14.07	17.72
			3.29	3.22	3.69	3.80	3.48	3.82
400	400	7.49	11.26	11.01	14.70	15.73	12.88	15.94
			3.56	3.13	3.54	3.64	3.35	3.66

TABLE III

ERROR RATES AND J-DIVERGENCE FROM AR(2)
VERSUS AR(2) CLASSIFICATION

Notes: (a) DATA refers to the length of the training data,

(b) TEST refers to the length of the test realization,

(c) JDIV refers to the J-Divergence rate,

(d) The first entry for each (DATA,TEST) tuple is the misclassification percentage; the second entry is the corresponding standard error.

(e) TIME DOMAIN RATES: 1 = PREDICTIVE;

3 = BURG;

4 = YULE-WALKER

FREQUENCY DOMAIN RATES: 2 = PREDICTIVE;

5 = BURG;

6 = YULE-WALKER

TABLE III (Continued)

DATA	TEST	JDIV	CLASSIFICATION METHOD					
			1	2	3	4	5	6
50	50	1.11	36.24	35.96	54.88	56.54	45.80	58.42
			4.81	4.80	4.98	4.96	4.98	4.93
50	100	1.30	29.36	29.49	44.00	45.58	36.87	46.90
			4.55	4.56	4.96	4.98	4.82	4.99
50	200	1.77	22.65	22.97	33.64	35.02	28.29	35.90
			4.19	4.21	4.72	4.77	4.50	4.80
50	400	3.30	16.87	17.22	24.90	26.00	20.99	26.59
			3.75	3.78	4.32	4.39	4.07	4.42
100	50	1.36	30.24	29.67	46.25	47.41	38.45	49.18
			4.59	4.57	4.99	4.99	4.86	5.00
100	100	1.59	25.92	25.72	39.26	40.45	32.76	41.79
			4.38	4.37	4.88	4.91	4.69	4.93
100	200	2.17	20.97	21.06	31.43	32.56	26.33	33.50
			4.07	4.08	4.64	4.69	4.40	4.72
100	400	4.05	16.15	16.38	23.99	24.96	20.17	25.59
			3.68	3.70	4.27	4.33	4.01	4.36
200	50	2.05	23.85	23.19	36.75	37.53	30.47	39.05
			4.26	4.22	4.82	4.84	4.60	4.88
200	100	2.39	21.58	21.17	33.00	33.83	27.44	35.09
			4.11	4.08	4.70	4.73	4.46	4.77
200	200	3.26	18.48	18.34	27.99	28.84	23.36	29.80
			3.88	3.87	4.49	4.53	4.23	4.57
200	400	6.07	14.94	15.00	22.38	23.19	18.76	23.86
			3.56	3.57	4.17	4.22	3.90	4.26
400	50	4.61	18.01	17.41	27.88	28.41	23.08	29.61
			3.84	3.79	4.48	4.51	4.21	4.57
400	100	5.38	16.98	16.51	26.16	26.72	21.69	27.80
			3.75	3.71	4.40	4.42	4.12	4.48
400	200	7.34	15.36	15.06	23.48	24.07	19.53	24.97
			3.61	3.58	4.24	4.28	3.96	4.33
400	400	13.65	13.15	13.05	19.91	20.51	16.62	21.20
			3.38	3.37	3.99	4.04	3.72	4.09

FIGURES

SECTION 1**FIGURES FOR AR(1) VERSUS AR(1) CLASSIFICATION**

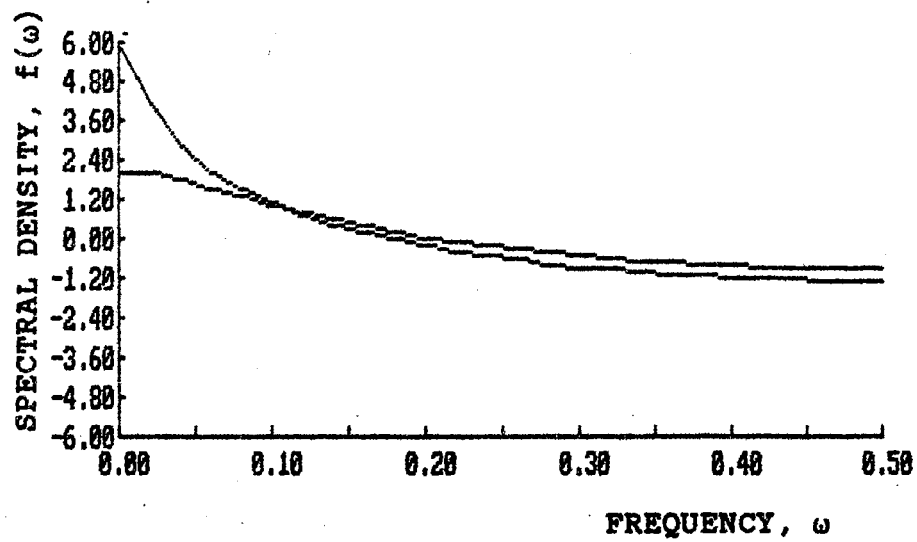


Figure 1. Spectral Plots for AR(1) and AR(1)

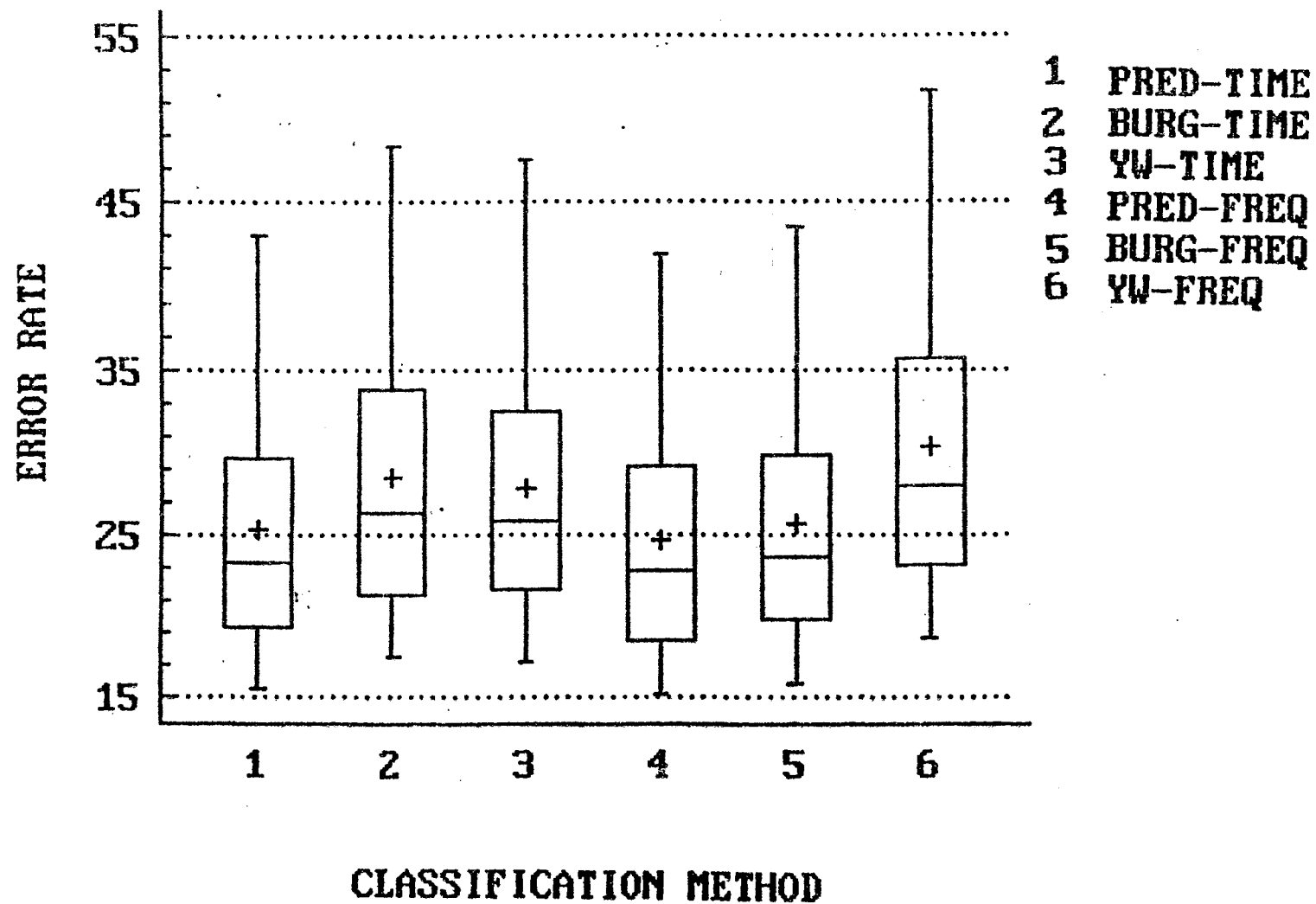


Figure 2. Box and Whisker Plot for AR(1) vs AR(1)

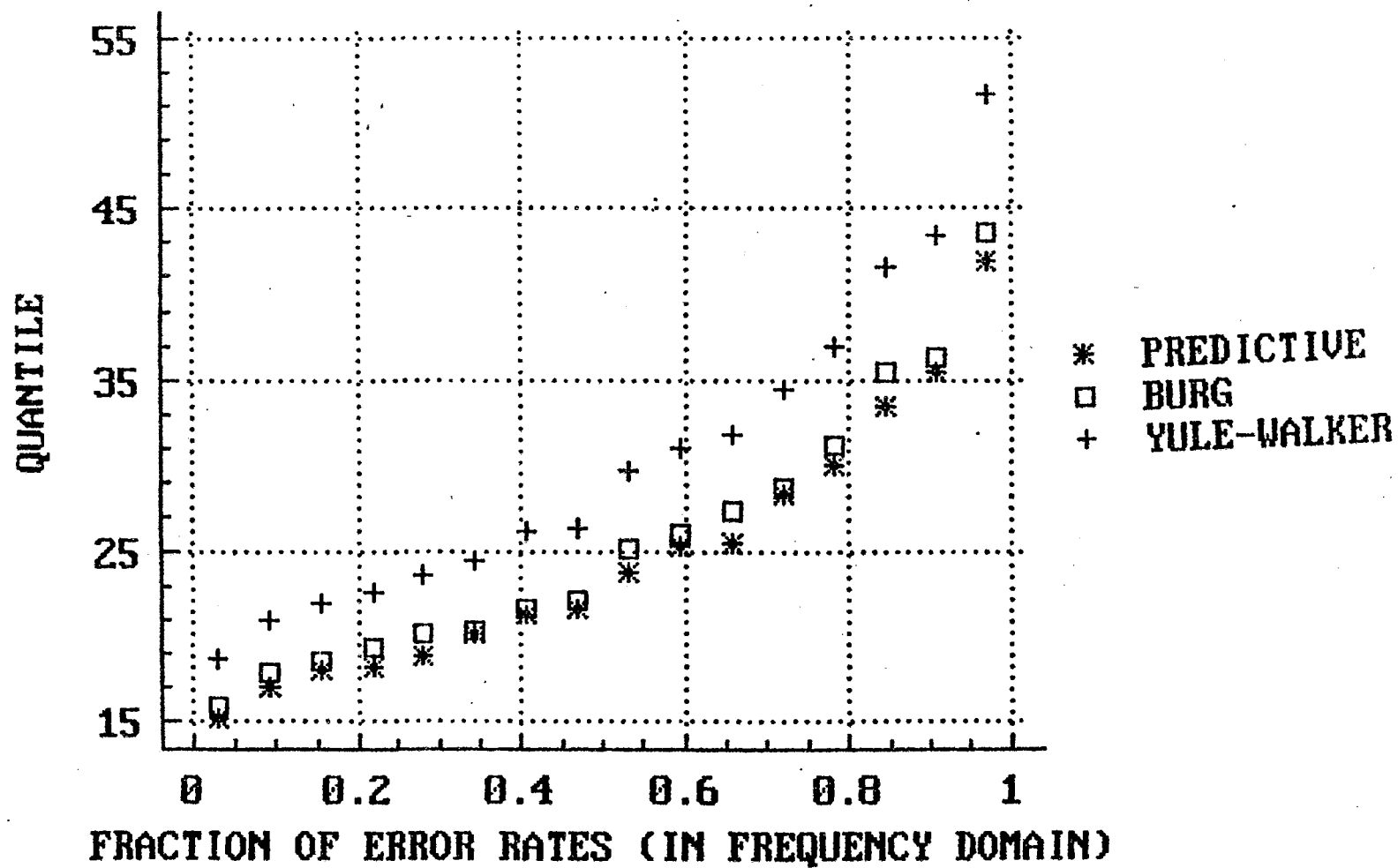


Figure 3. Quantile Plot for AR(1) vs. AR(1)

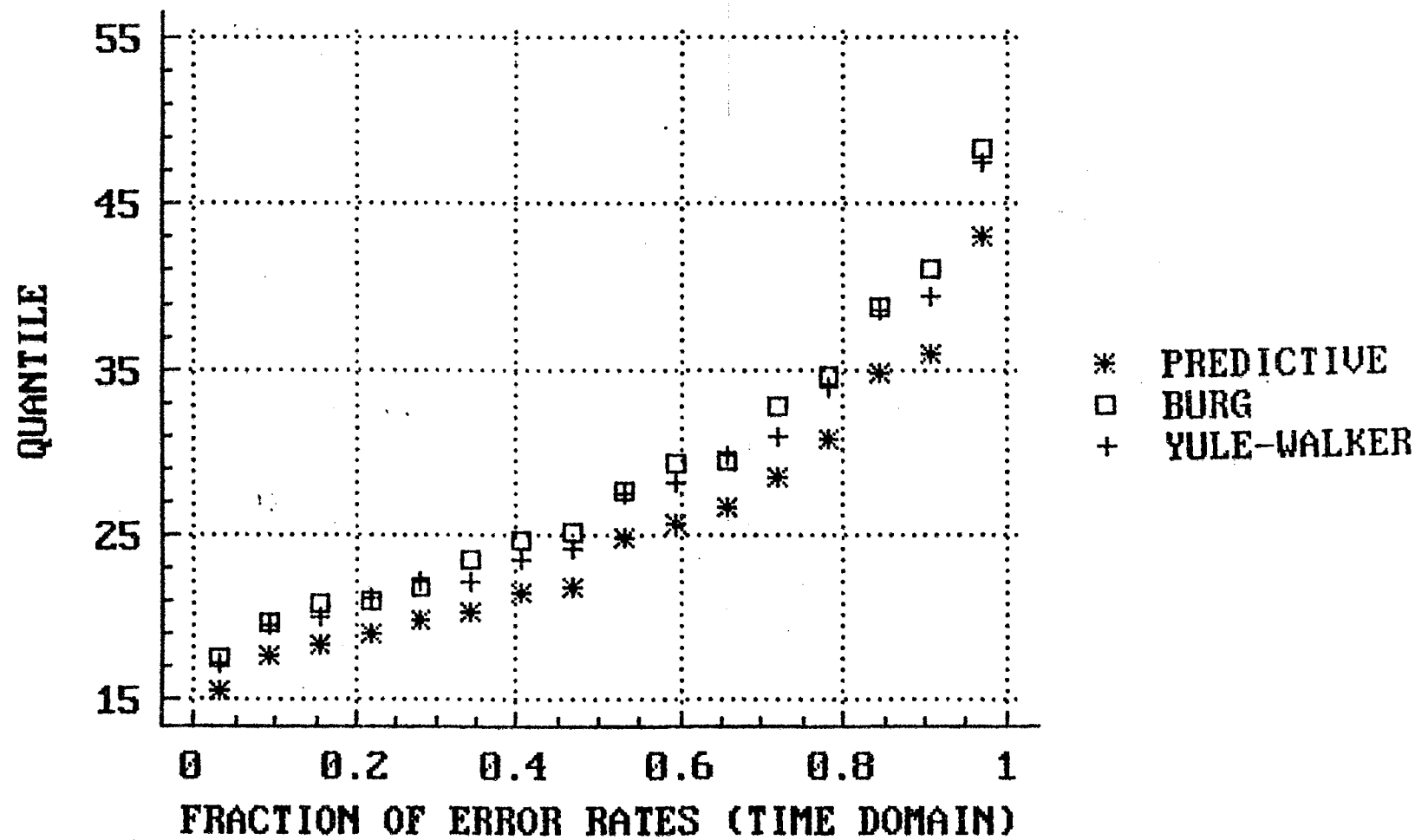


Figure 4. Quantile Plot for AR(1) vs. AR(1)

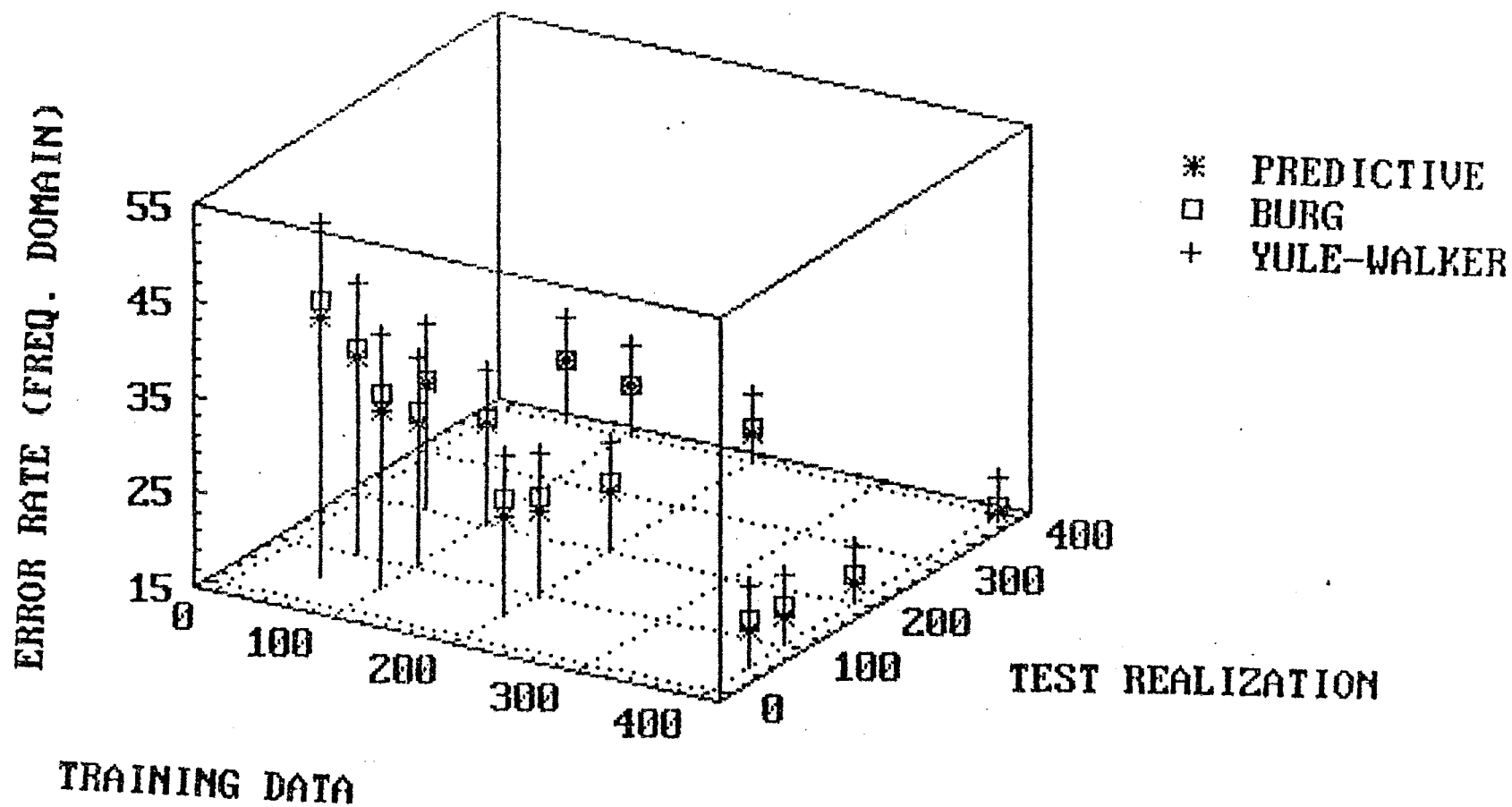


Figure 5. Error Rates and Lengths of Training Data, Test Realizations for AR(1) vs AR(1)

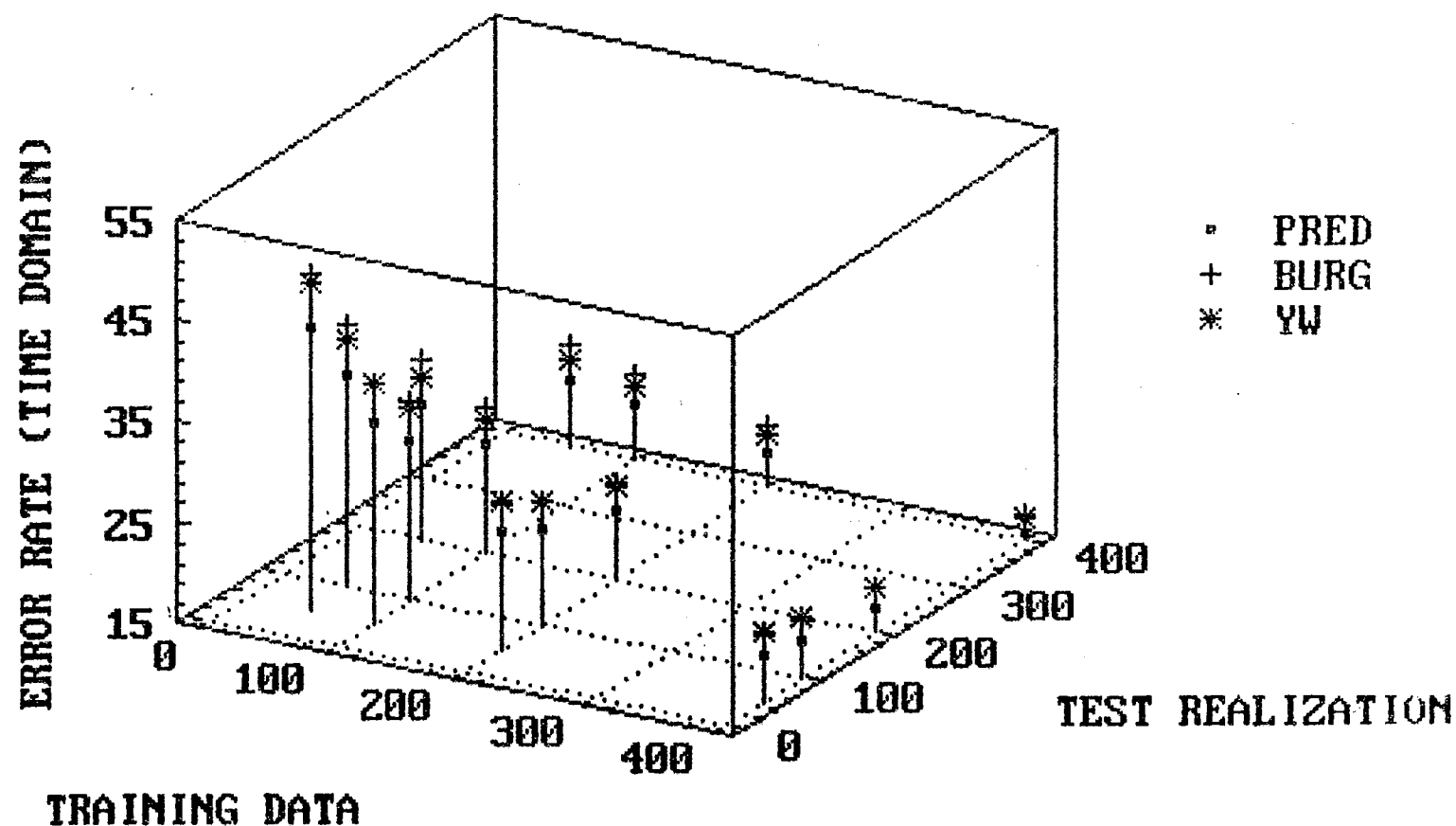


Figure 6. Error Rates and Lengths of Training Data, Test Realizations for AR(1) vs AR(1)

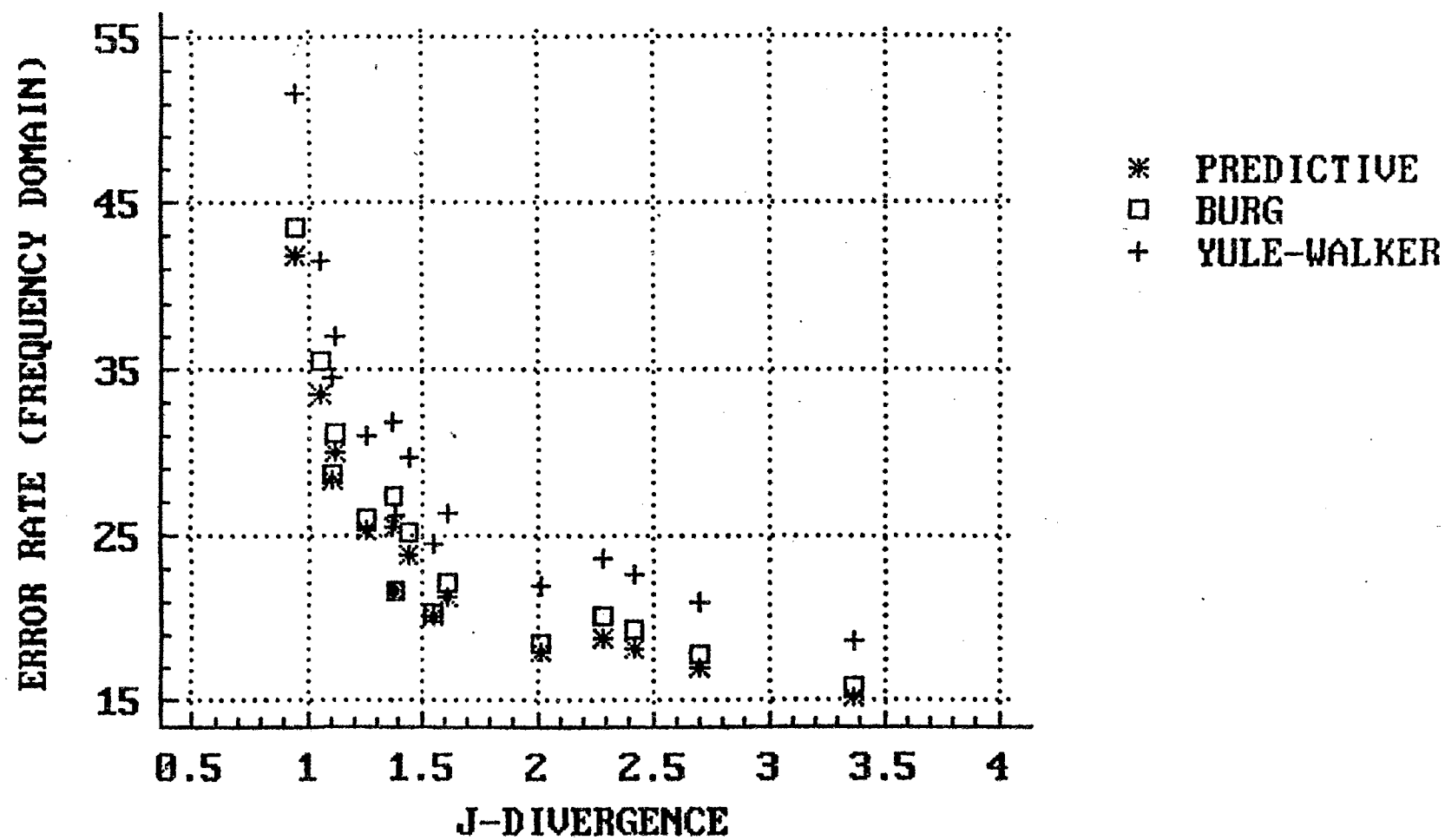


Figure 7. Error Rates and J-Divergence for AR(1) vs AR(1)

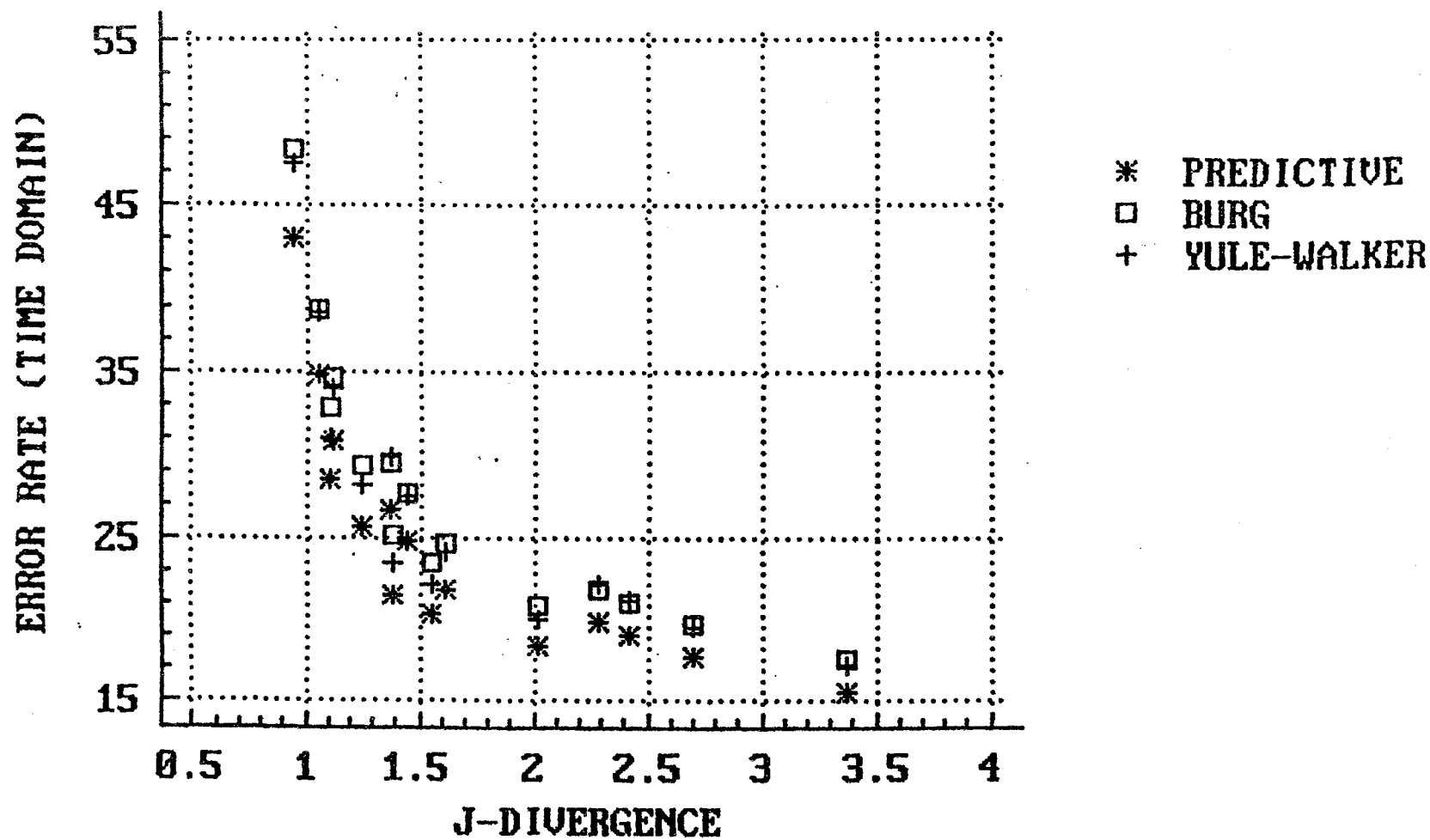


Figure 8. Error Rates and J-Divergence for AR(1) vs AR(1)

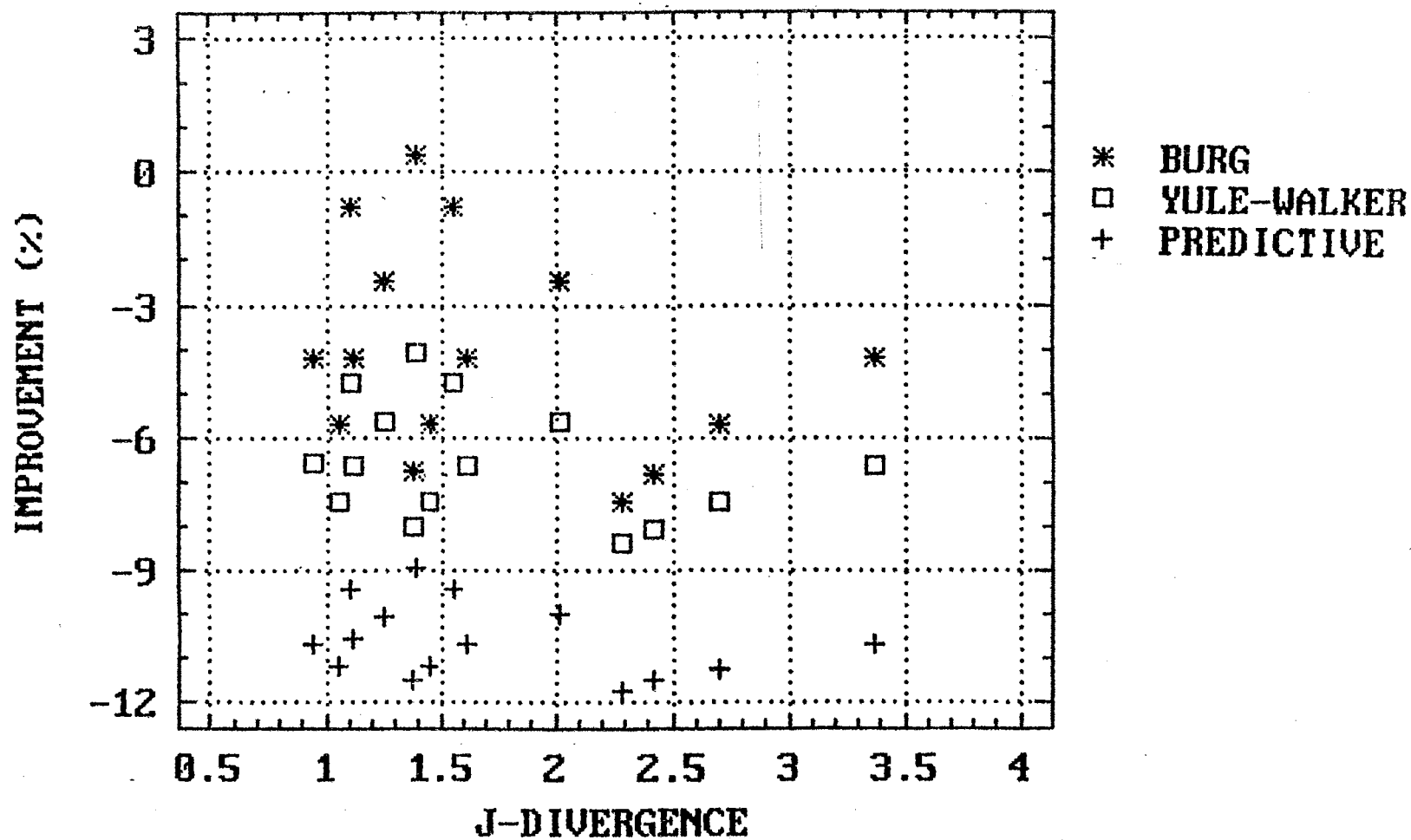


Figure 9. Improvement of Frequency over Time Domain
Methods for AR(1) vs AR(1)

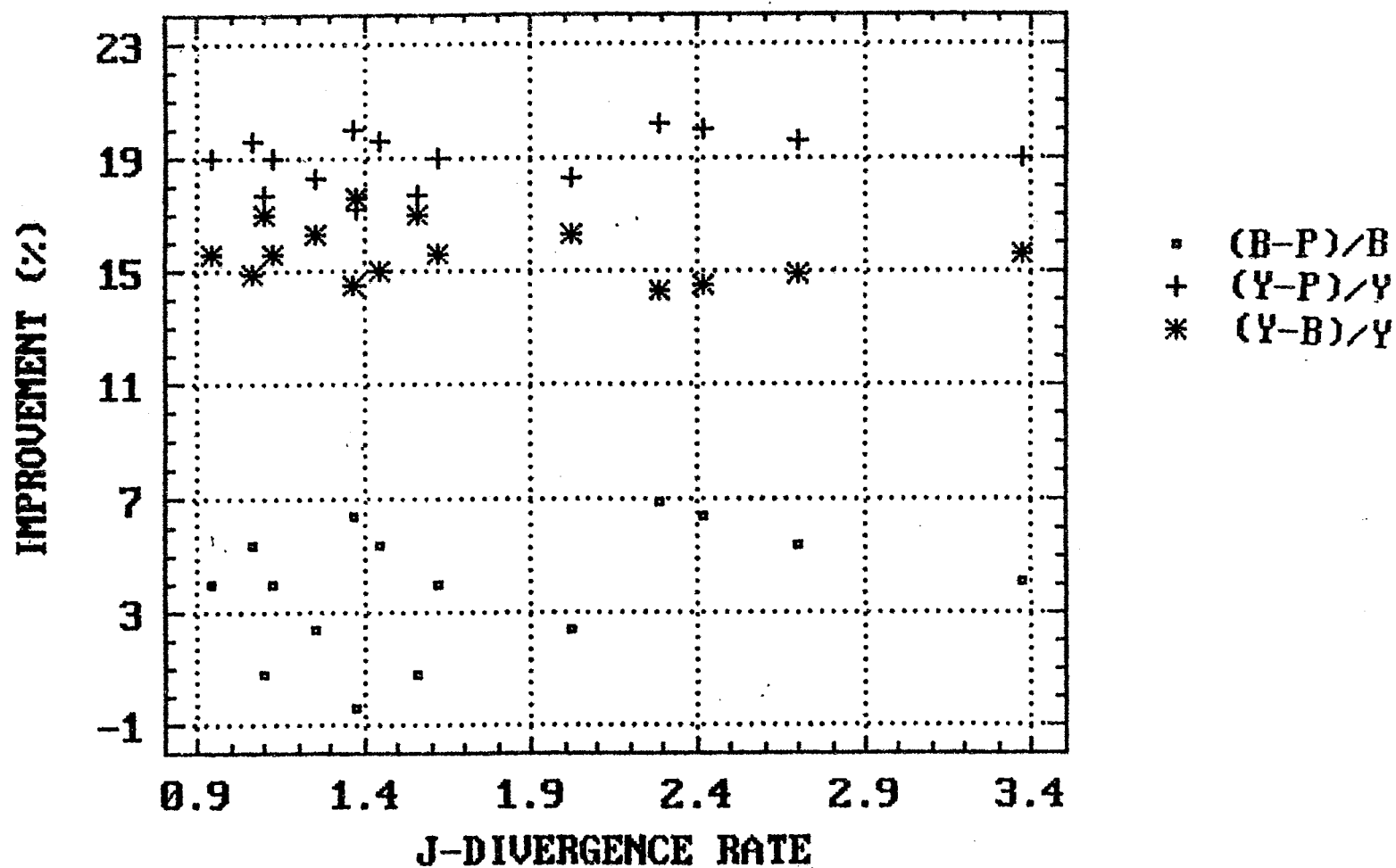


Figure 10. IMPROVEMENT RATE AND J-DIVERGENCE FOR
FOR AR(1) VS AR(1) (FREQUENCY DOMAIN)

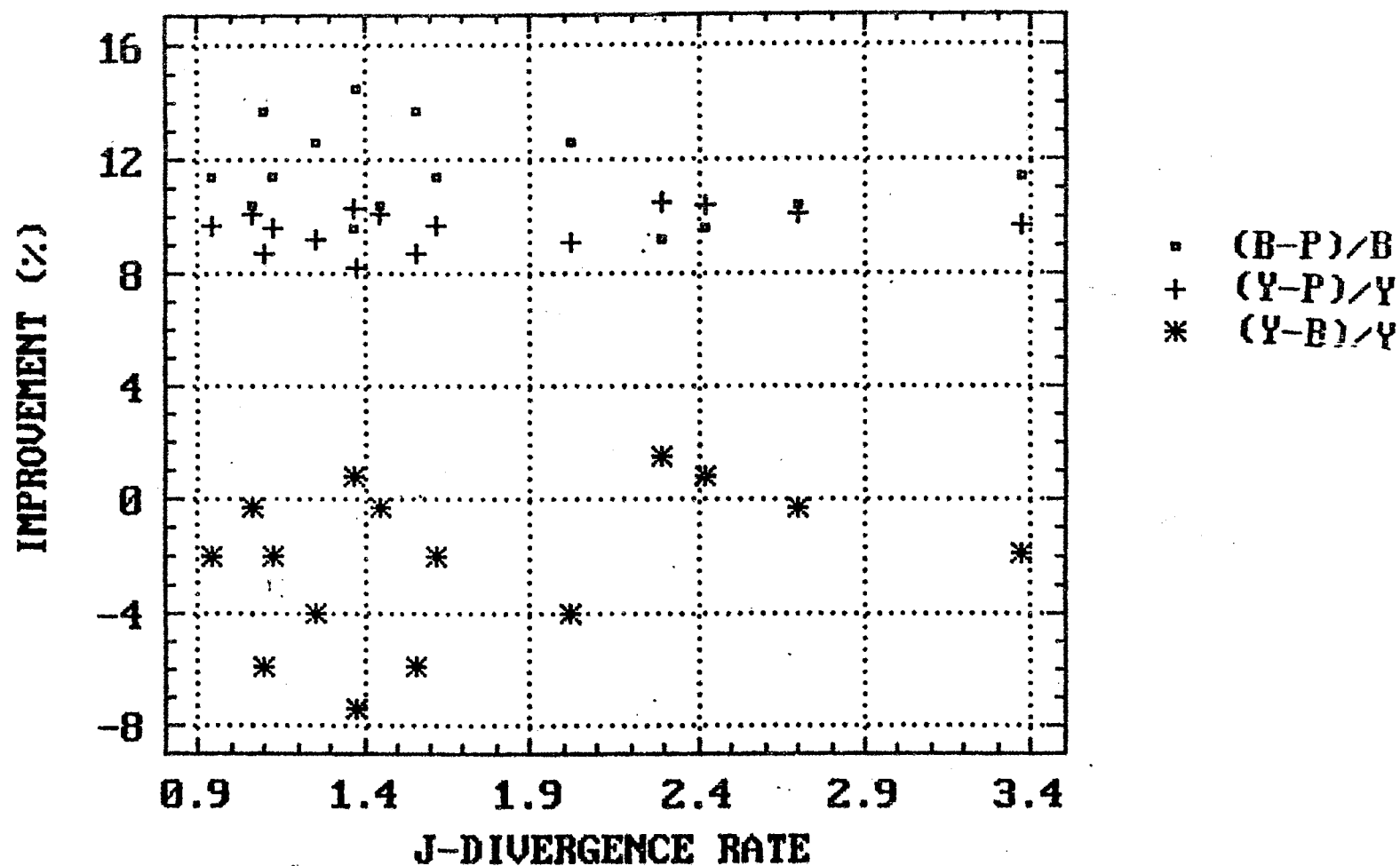


Figure 11. IMPROVEMENT RATE AND J-DIVERGENCE FOR
FOR AR(1) VS AR(1) (TIME DOMAIN)

SECTION 2

FIGURES FOR AR(2) VERSUS AR(1) CLASSIFICATION

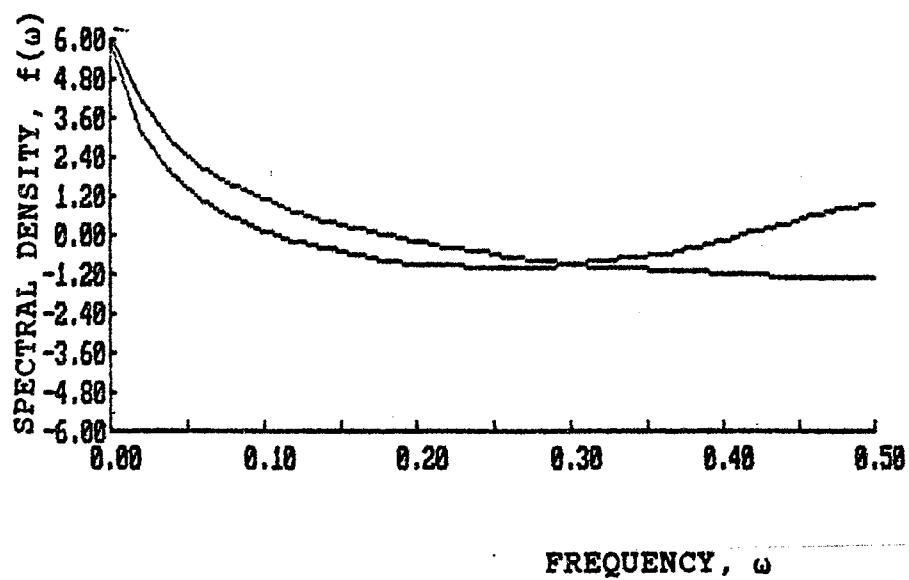


Figure 12. Spectral Plots for AR(2) and AR(1)

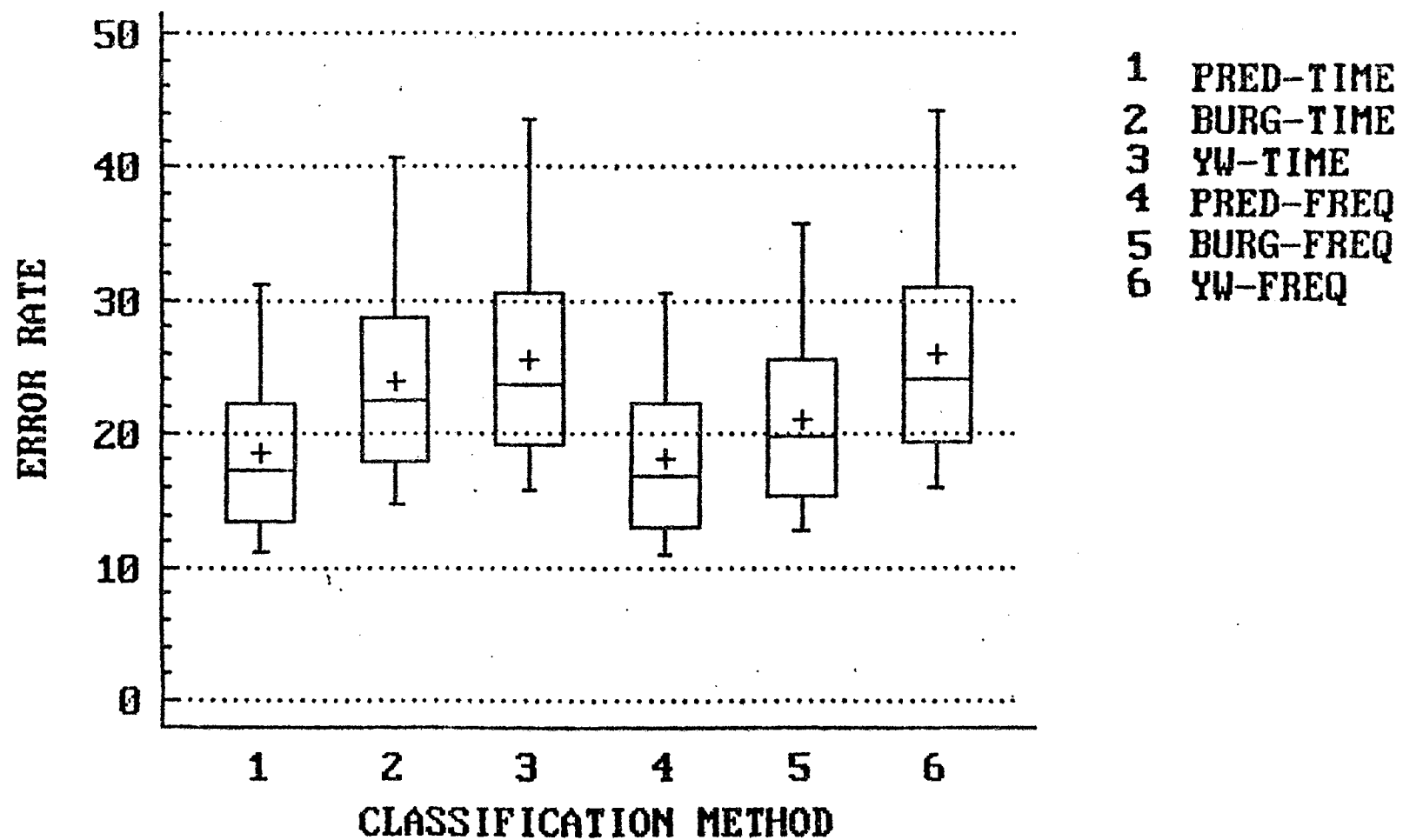


Figure 13. Box and Whisker Plot for AR(2) vs. AR(1)

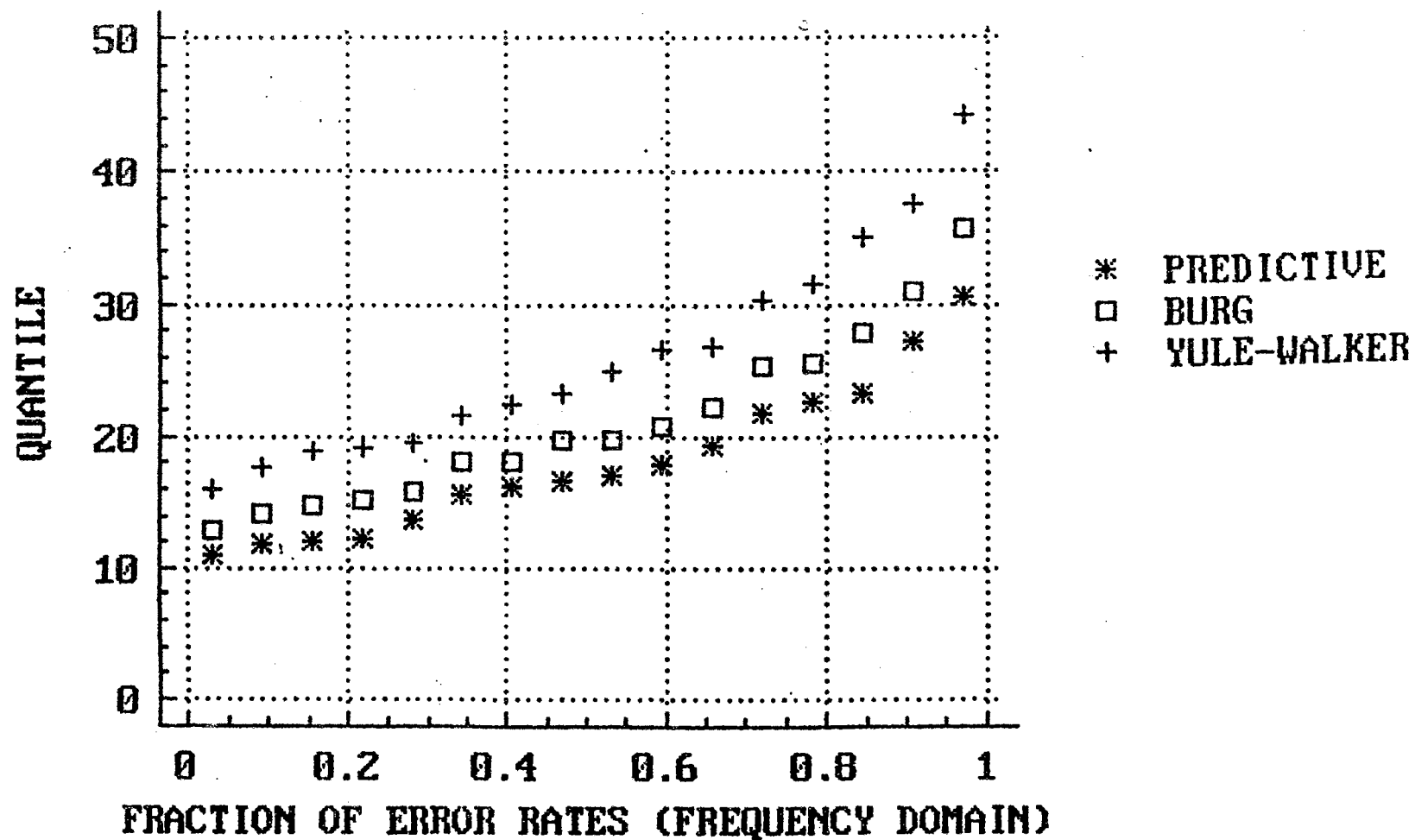


Figure 14. Quantile Plot for AR(2) vs. AR(1)

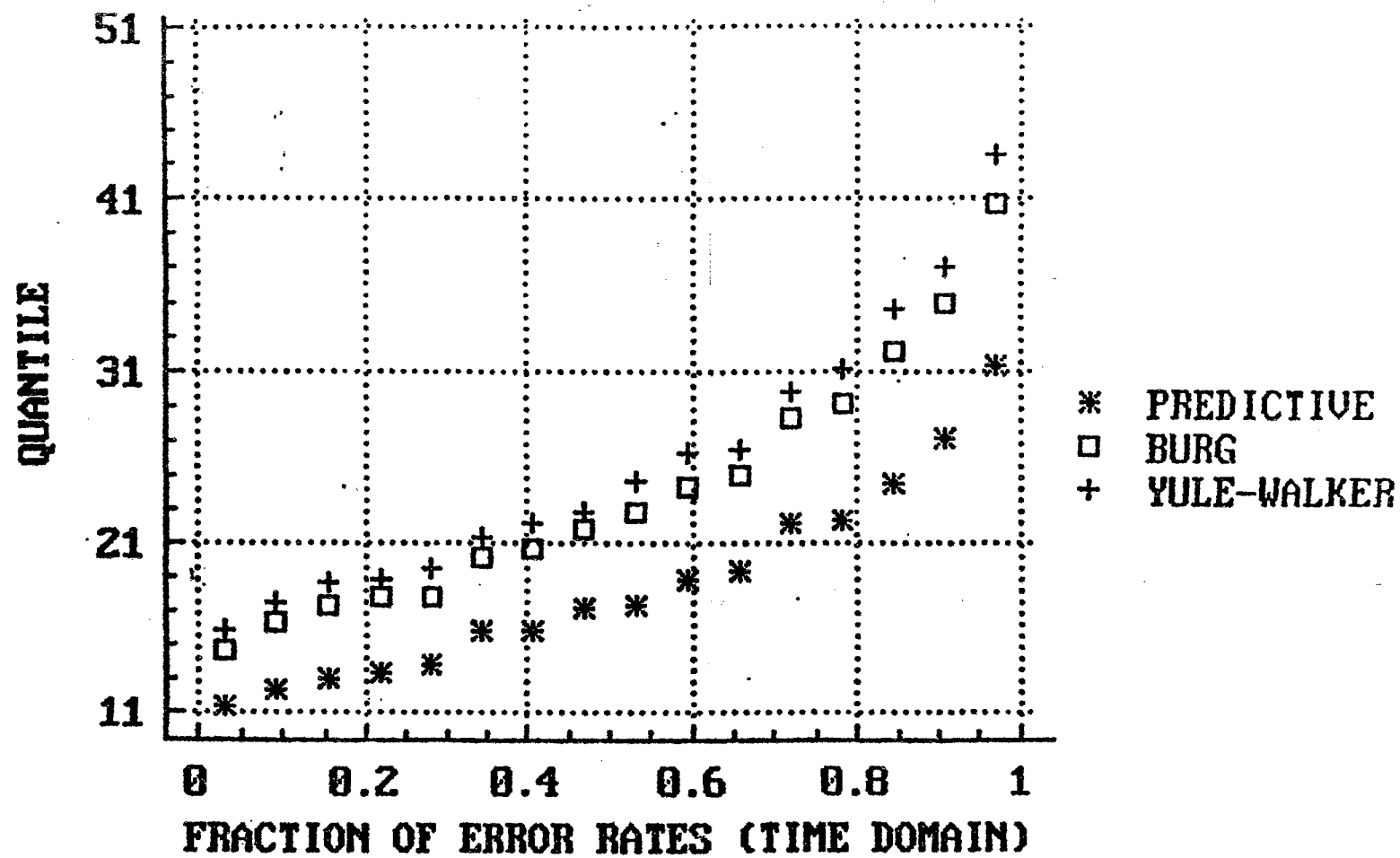


Figure 15. Quantile Plot for AR(2) vs. AR(1)

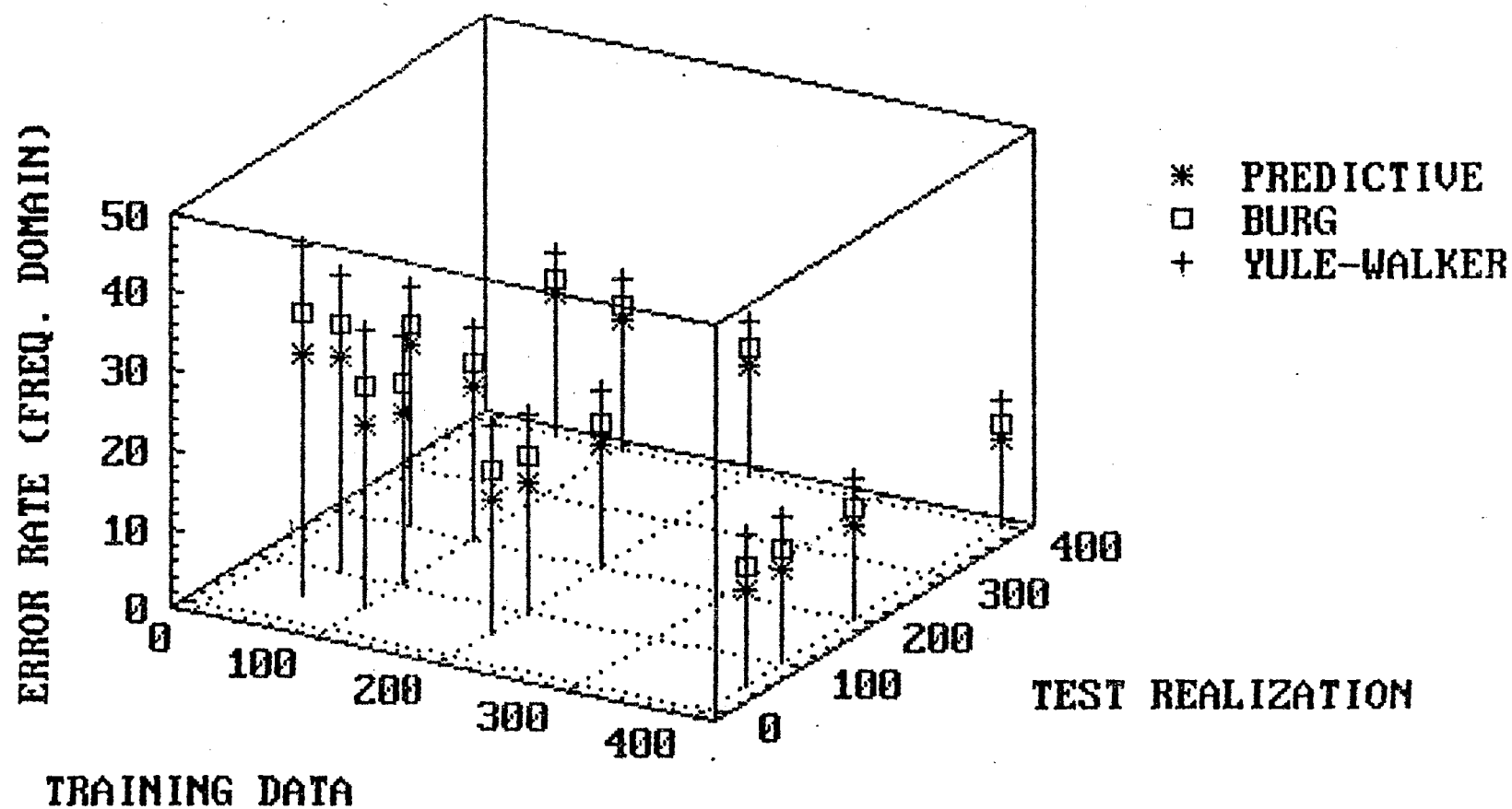


Figure 16. Error Rates and Lengths of Training Data, Test Realization for AR(2) vs AR(1)

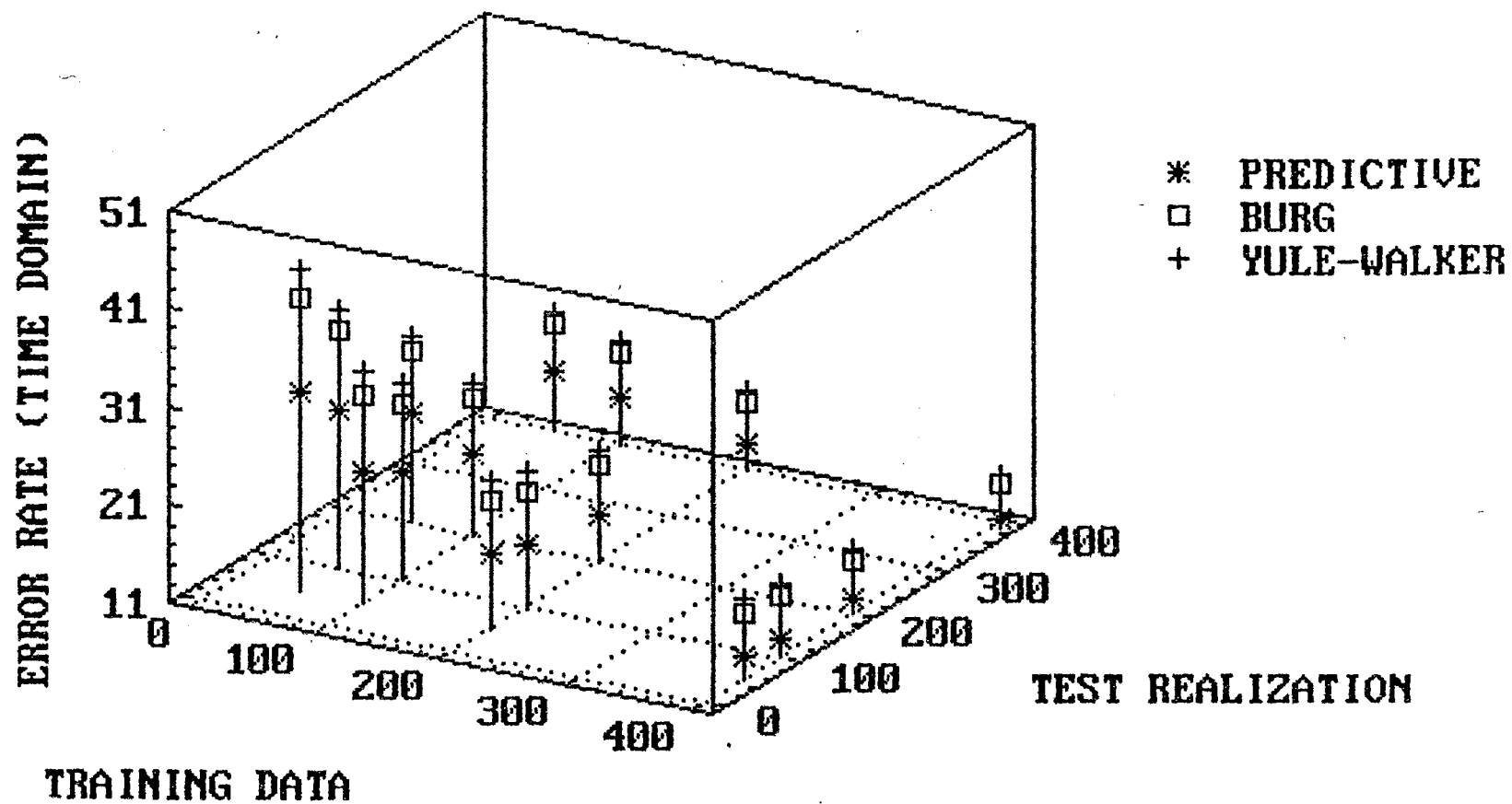


Figure 17. Error Rates and Lengths of Training Data, Test Realization for AR(2) vs AR(1)

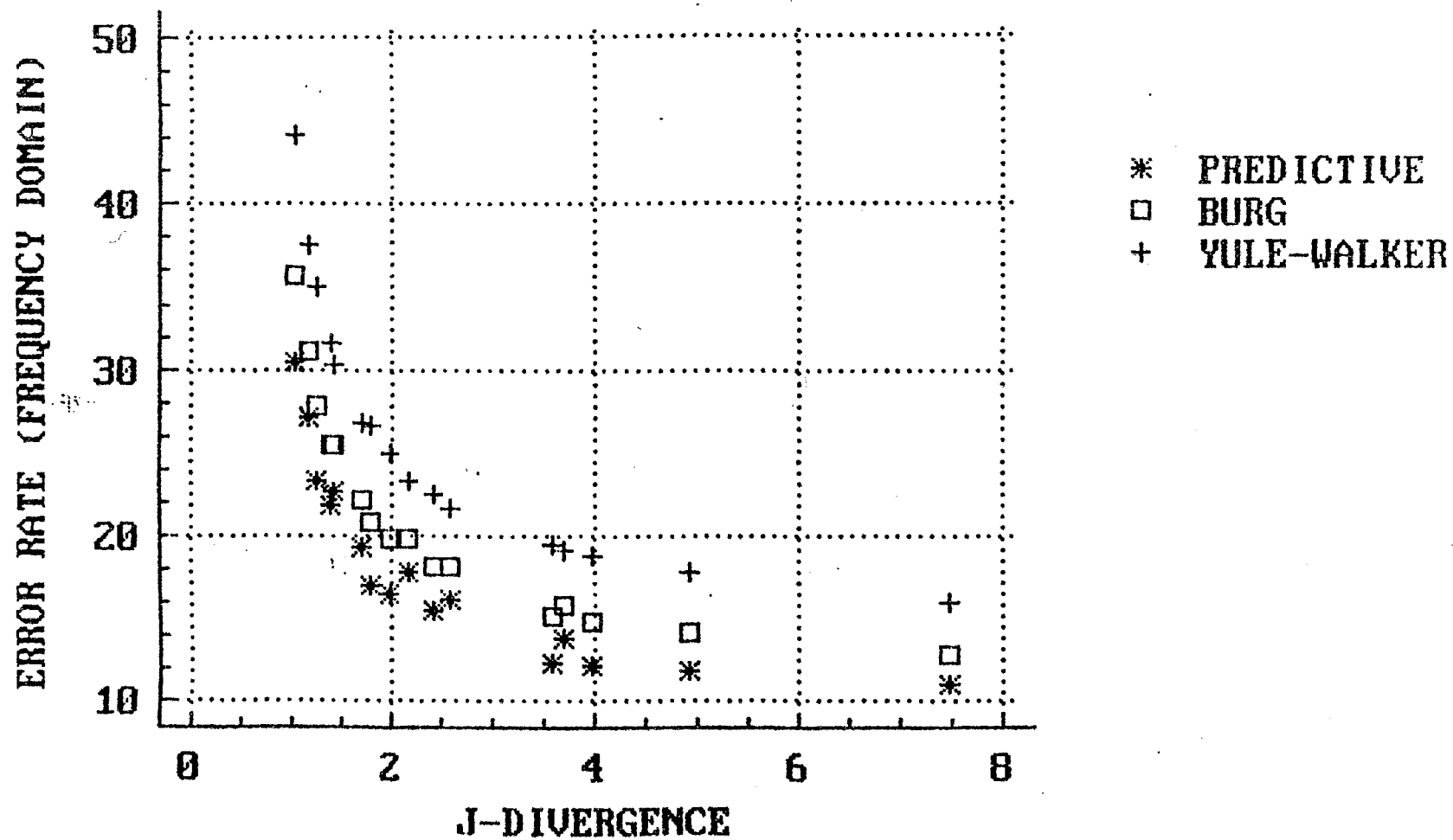


Figure 18. Error Rates and J-Divergence for AR(2) vs AR(1)

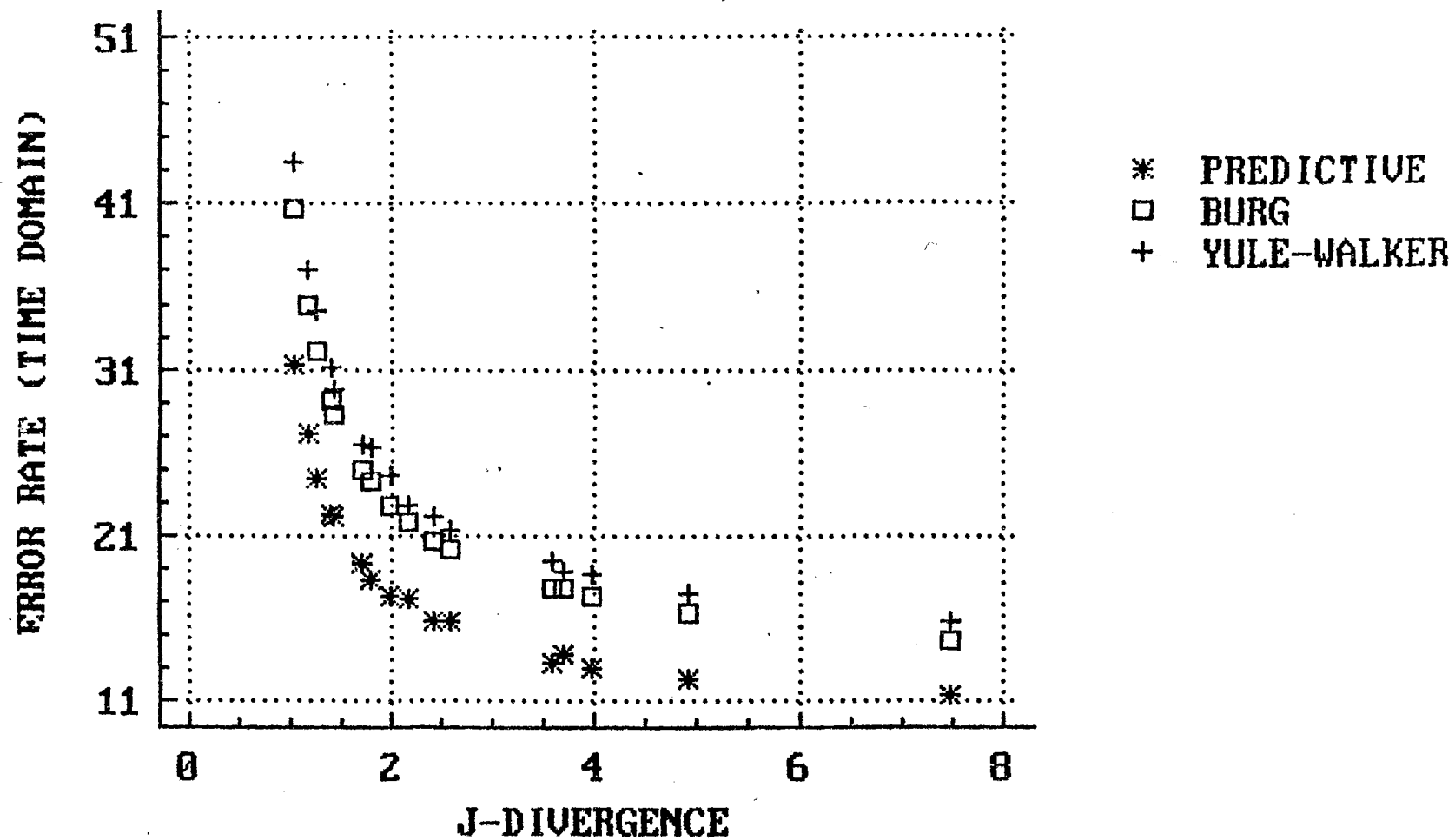


Figure 19. Error Rates and J-Divergence for AR(2) vs AR(1)

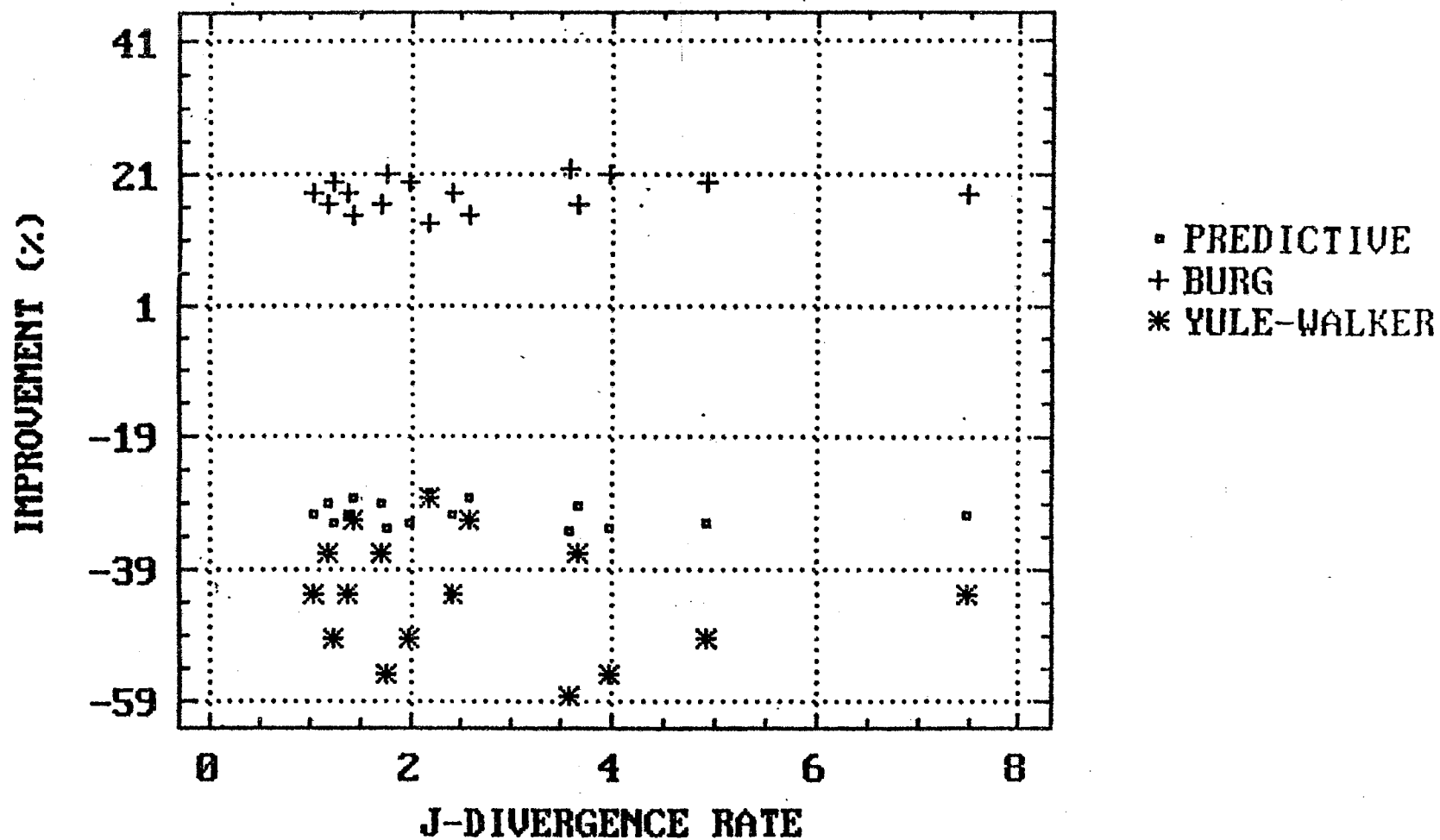


Figure 20. IMPROVEMENT OF FREQUENCY OVER TIME DOMAIN FOR AR(2) VS AR(1)

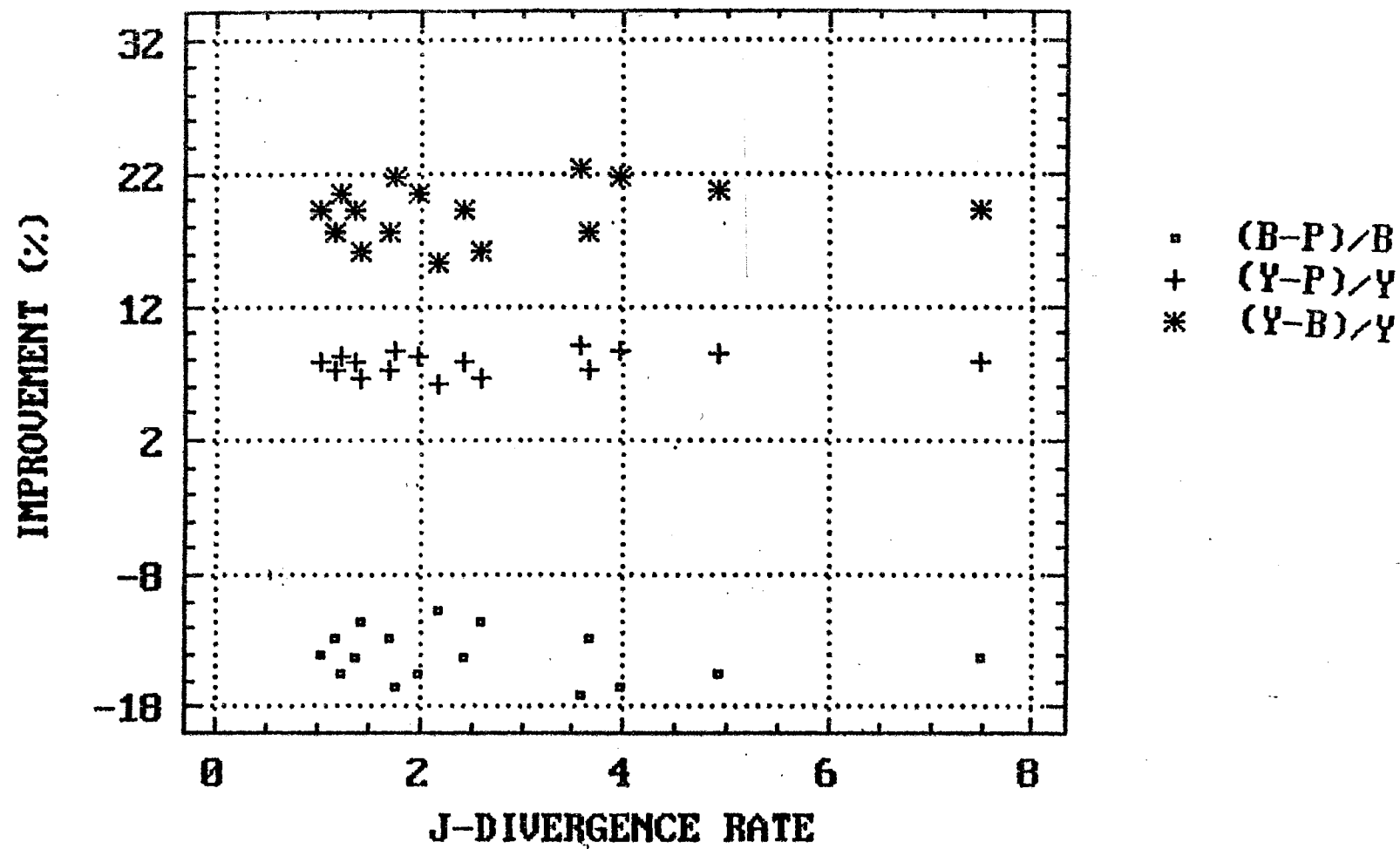


Figure 21. IMPROVEMENT RATE AND J-DIVERGENCE FOR AR(2) VS AR(1) (FREQUENCY DOMAIN)

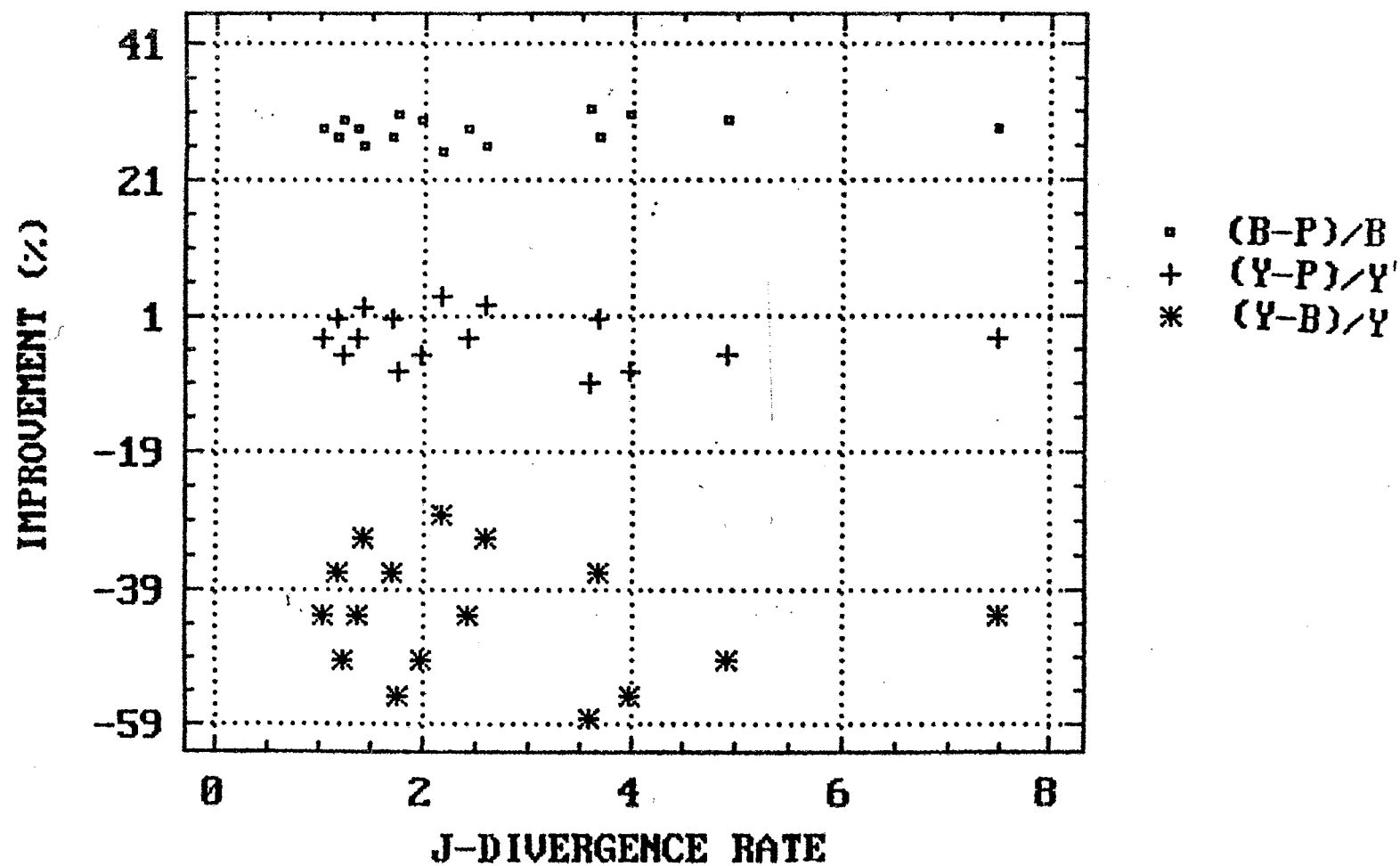


Figure 22. IMPROVEMENT RATE AND J-DIVERGENCE FOR AR(2) VS AR(1) (TIME DOMAIN)

SECTION 3

FIGURES FOR AR(2) VERSUS AR(2) CLASSIFICATION

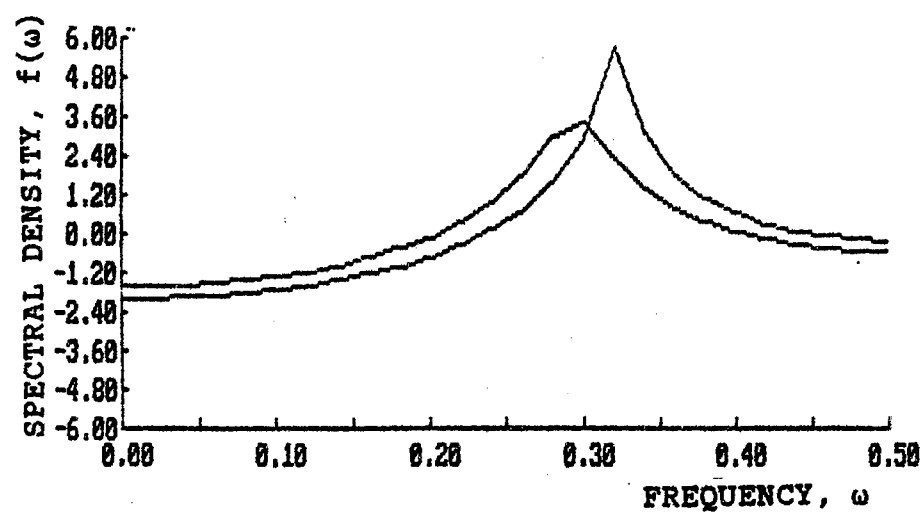


Figure 23. Spectral Plots for AR(2) and AR(2)

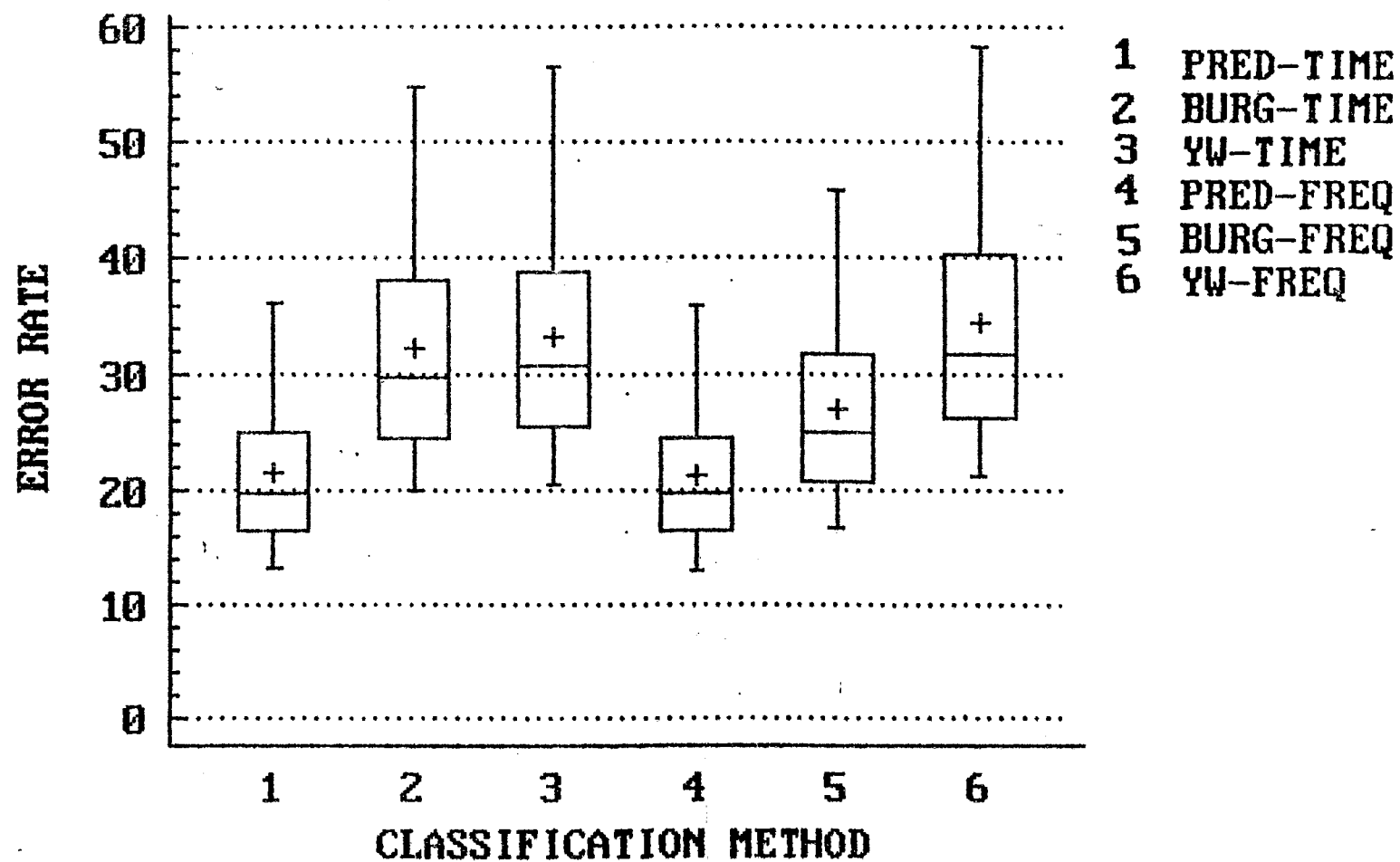


Figure 24. Box and Whisker Plot for AR(2) vs. AR(2)

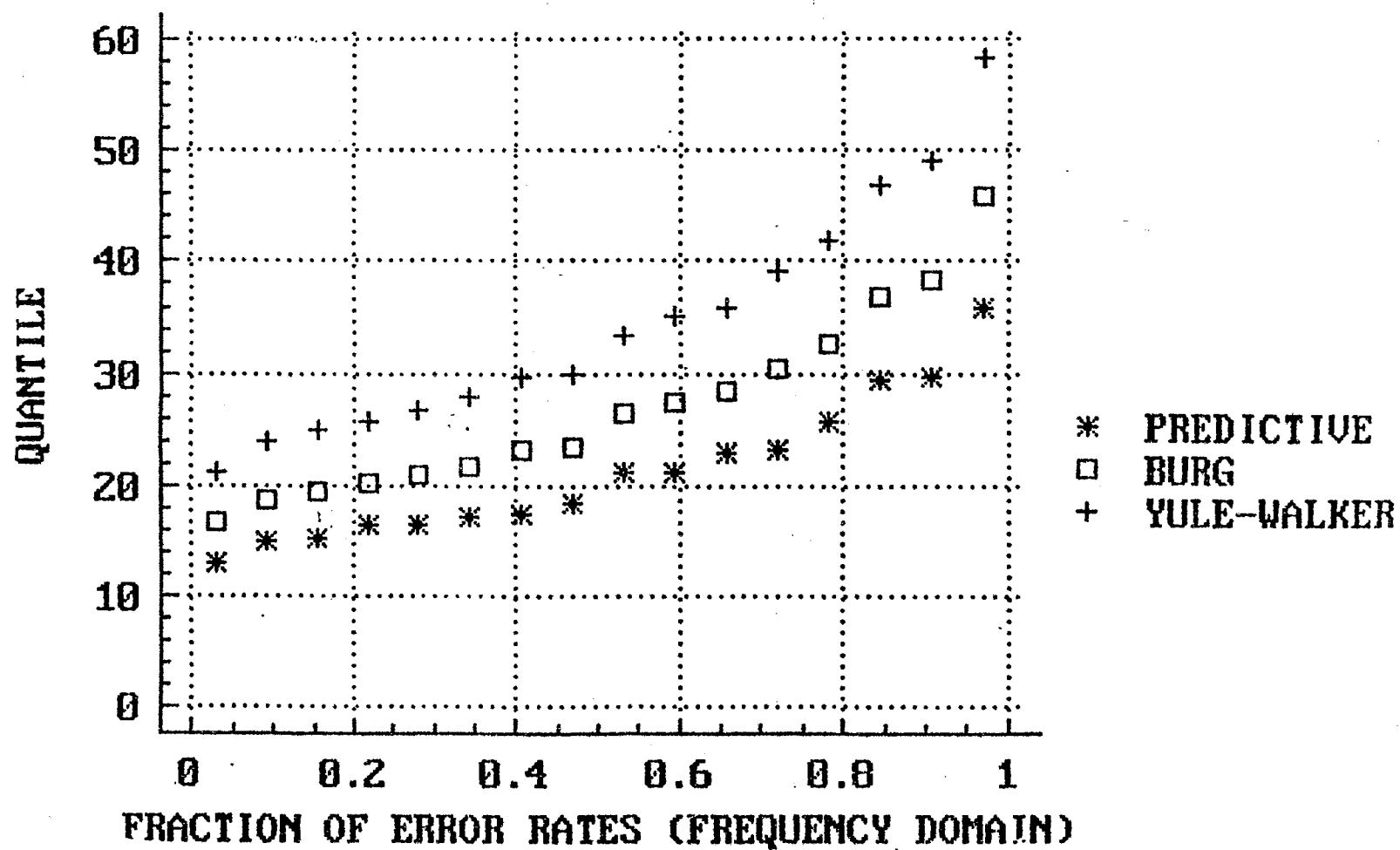


Figure 25. Quantile Plot for AR(2) vs. AR(2)

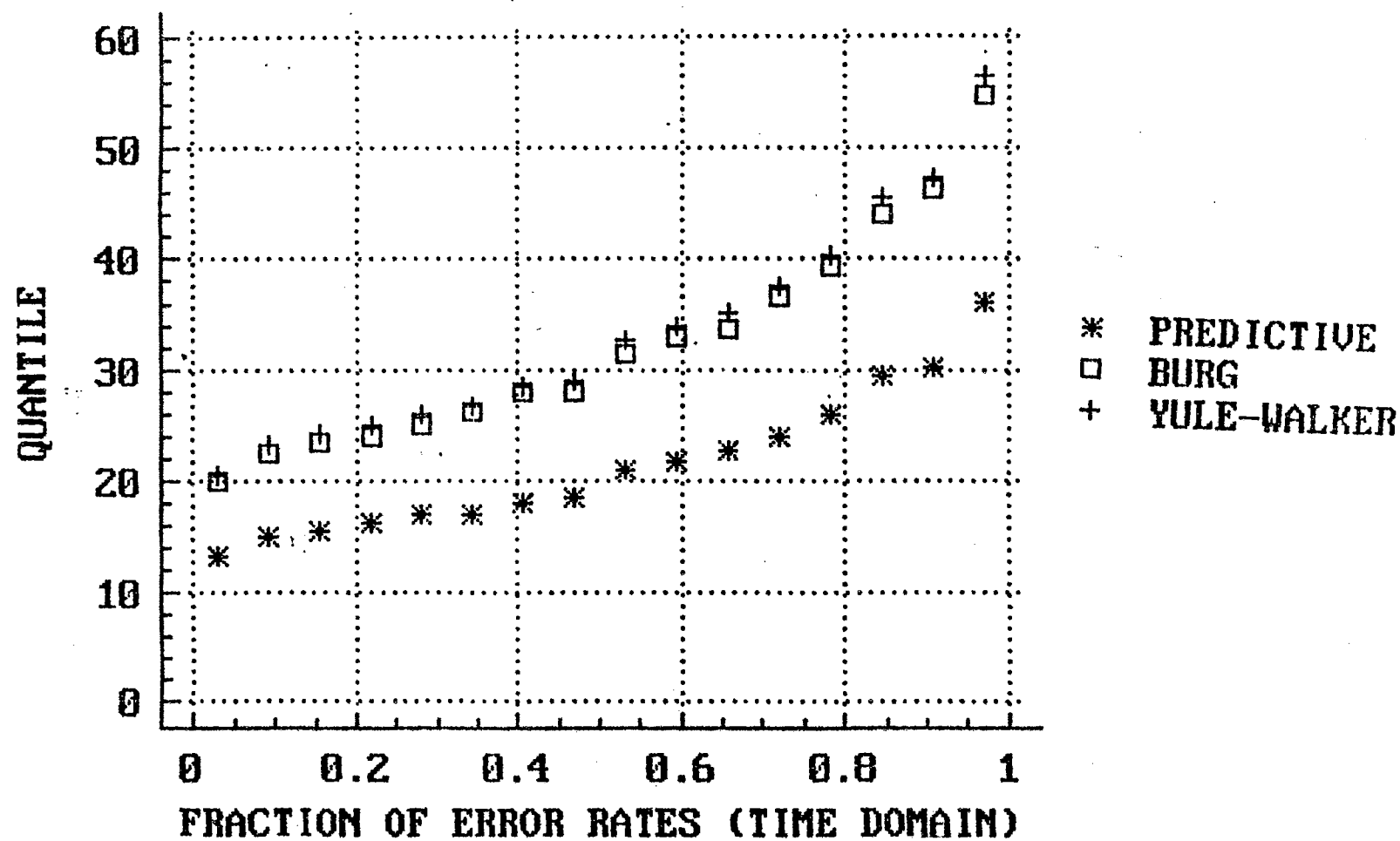


Figure 26. Quantile Plot for AR(2) vs. AR(2)

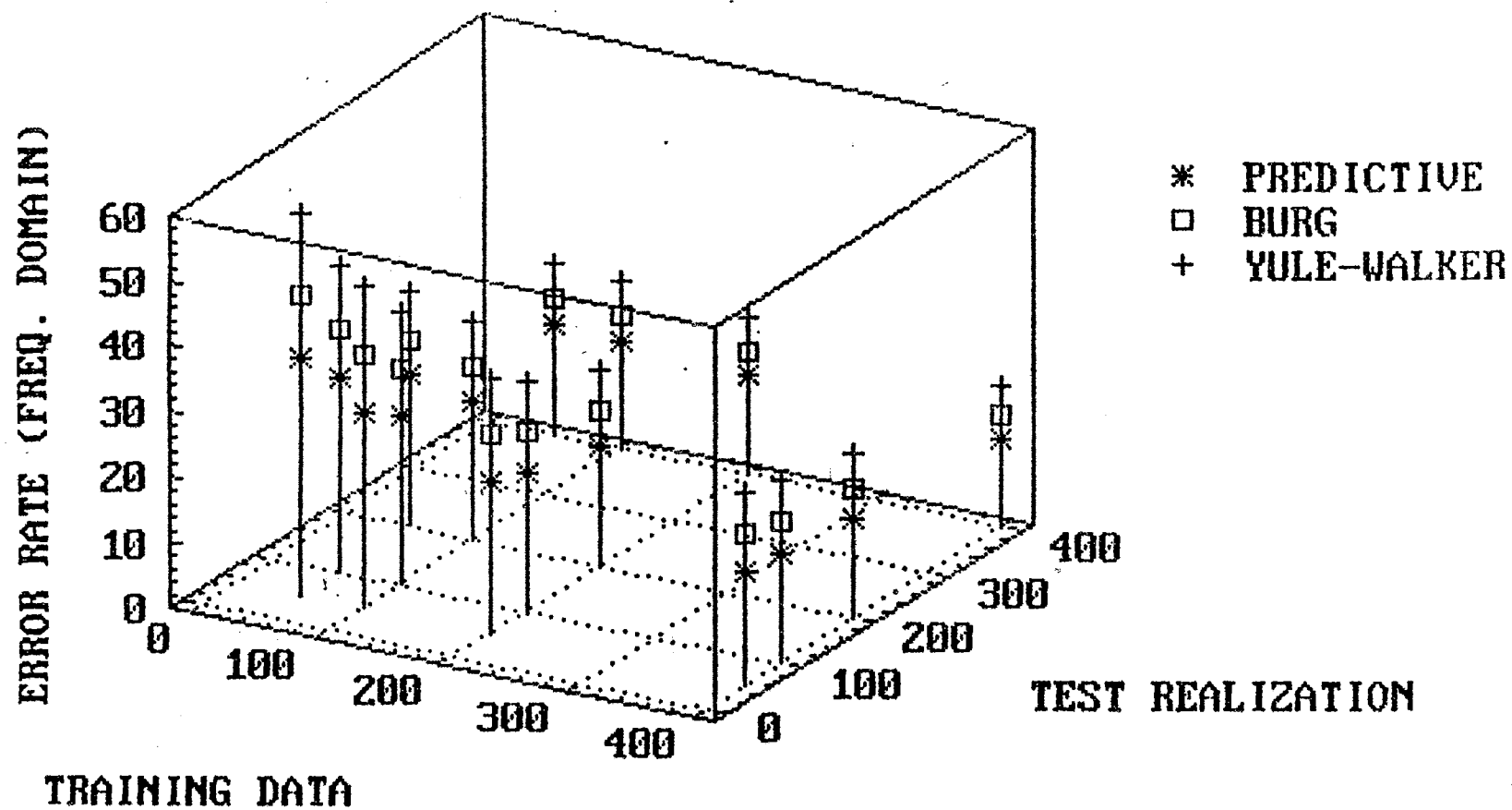
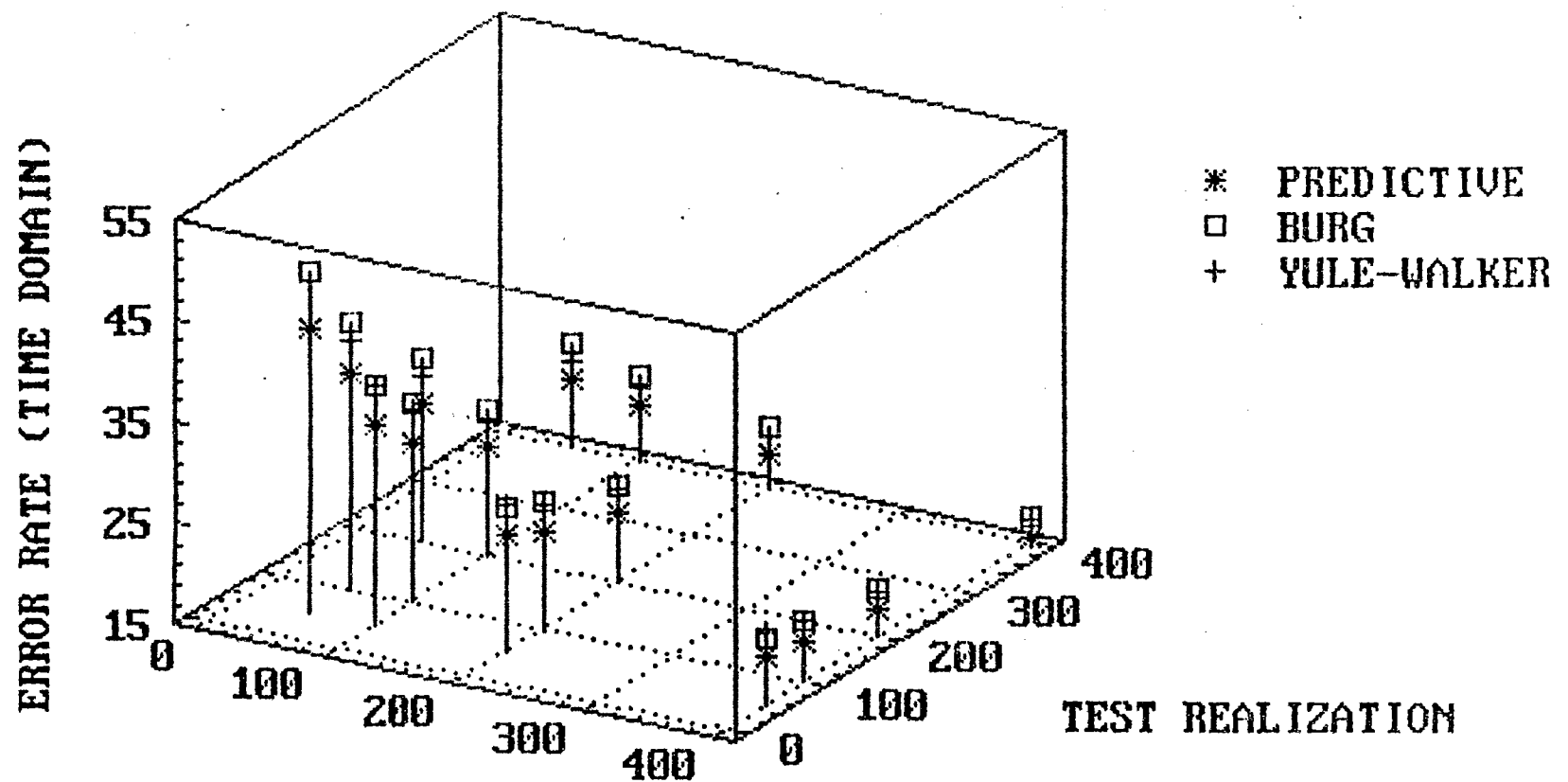


Figure 27. Error Rates and Lengths of Training Data, Test Realization for AR(2) vs AR(2)



TRAINING DATA
 Figure 28. Error Rates and Lengths of Training Data,
 Test Realization for AR(2) vs AR(2)

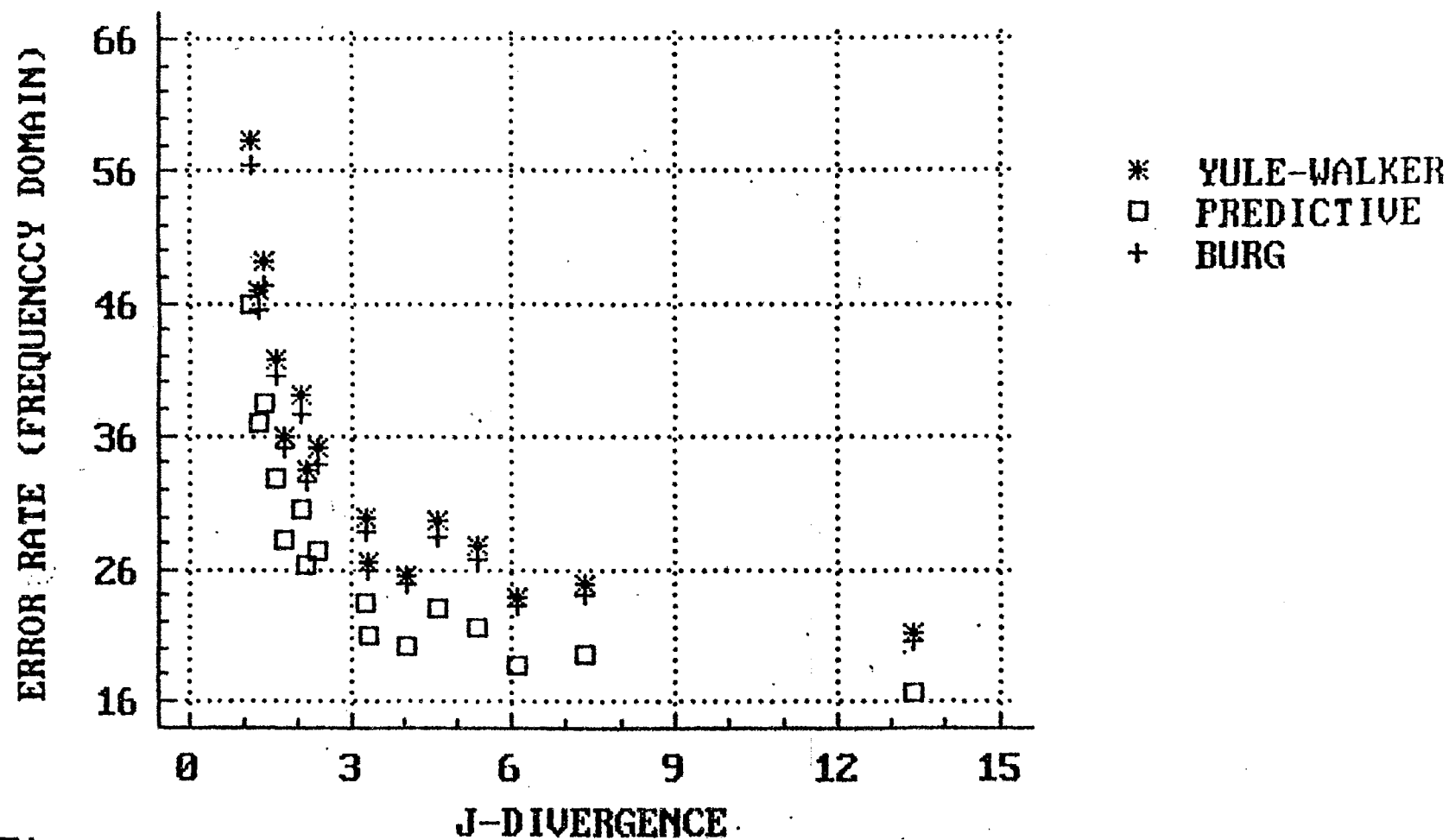


Figure 29. Error Rates and J-Divergence for AR(2) vs AR(2)

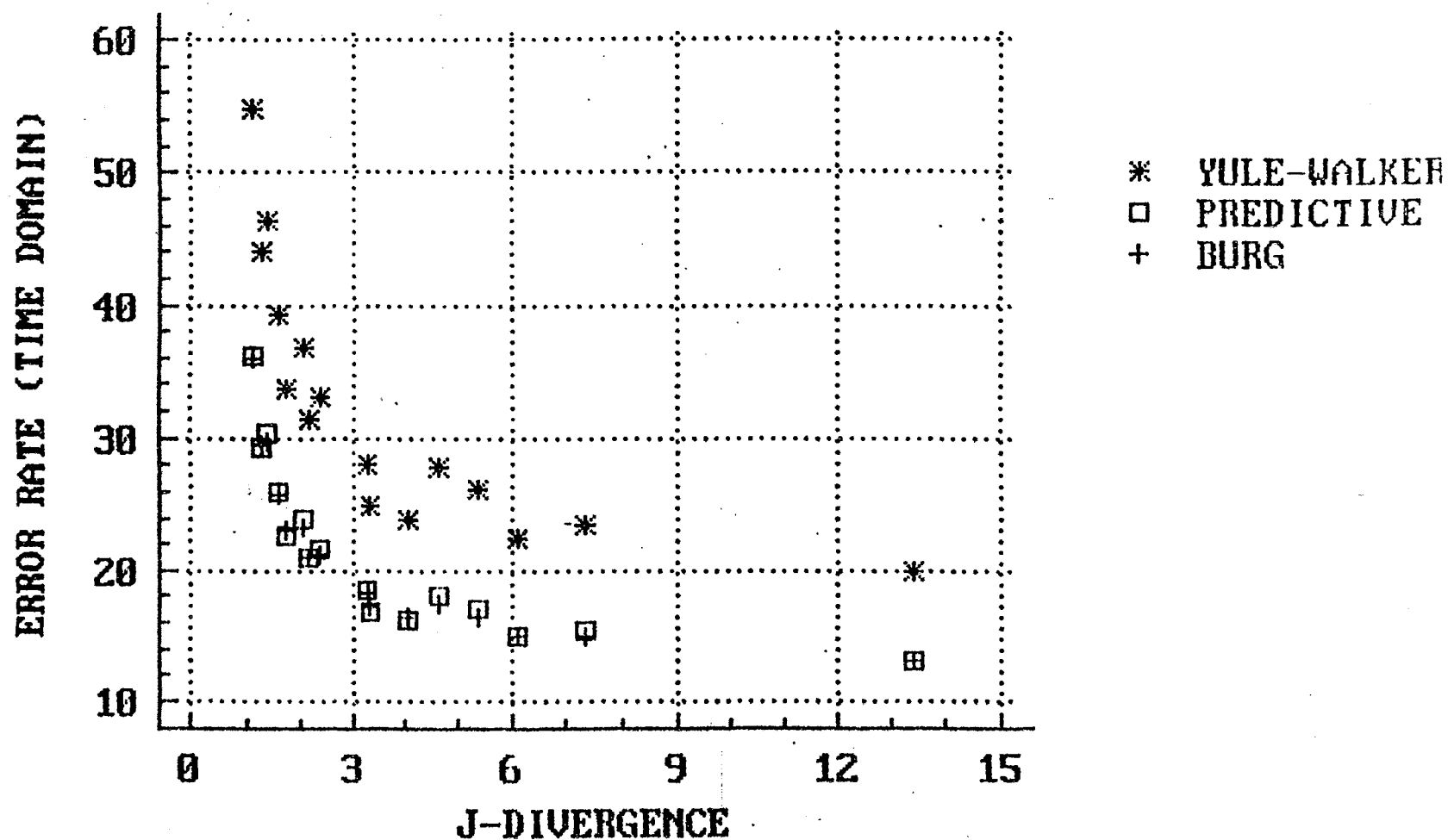


Figure 30. Error Rates and J-Divergence for AR(2) vs AR(2)

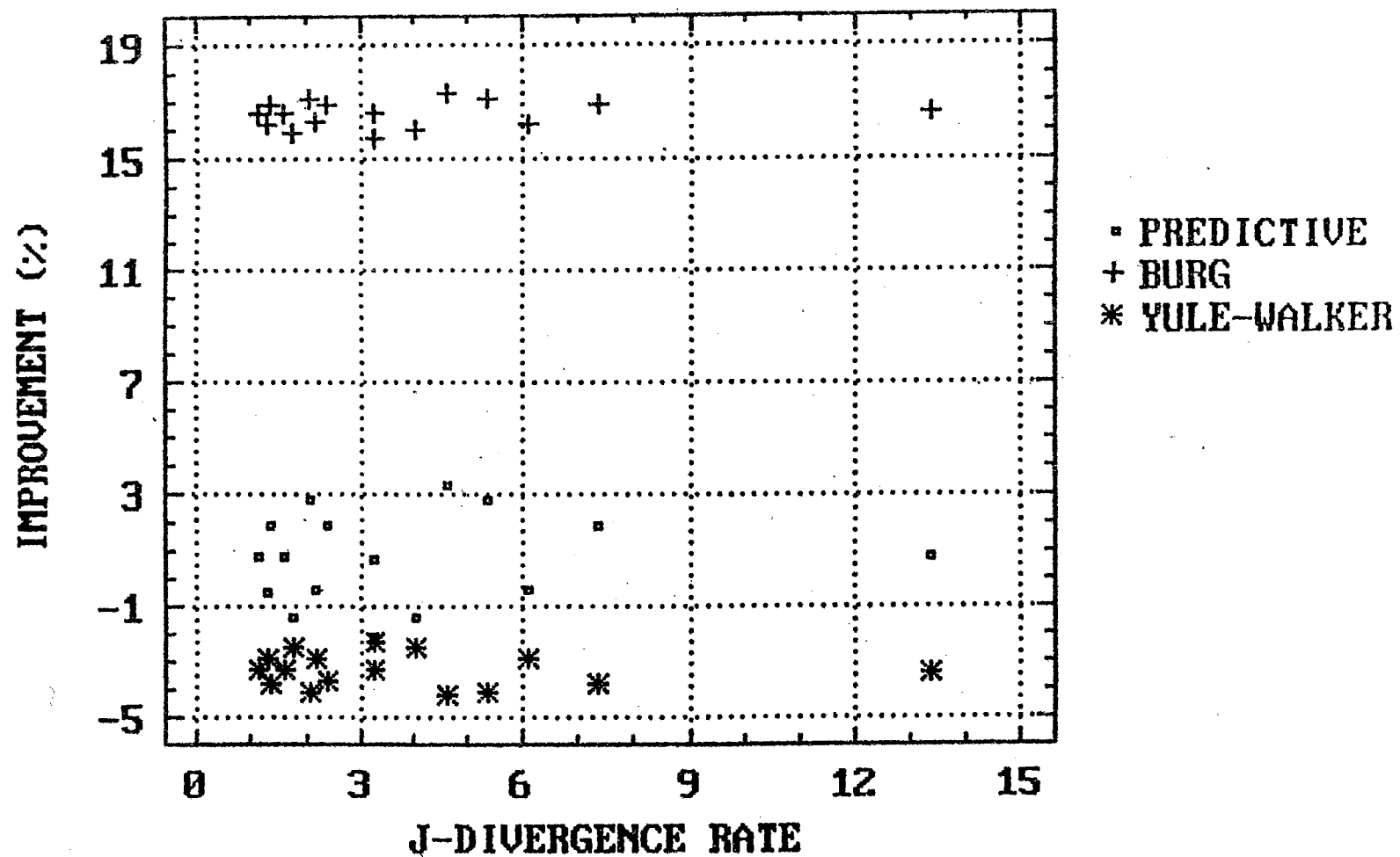


Figure 31. IMPROVEMENT OF FREQUENCY OVER TIME
DOMAIN FOR AR(2) VS AR(2)

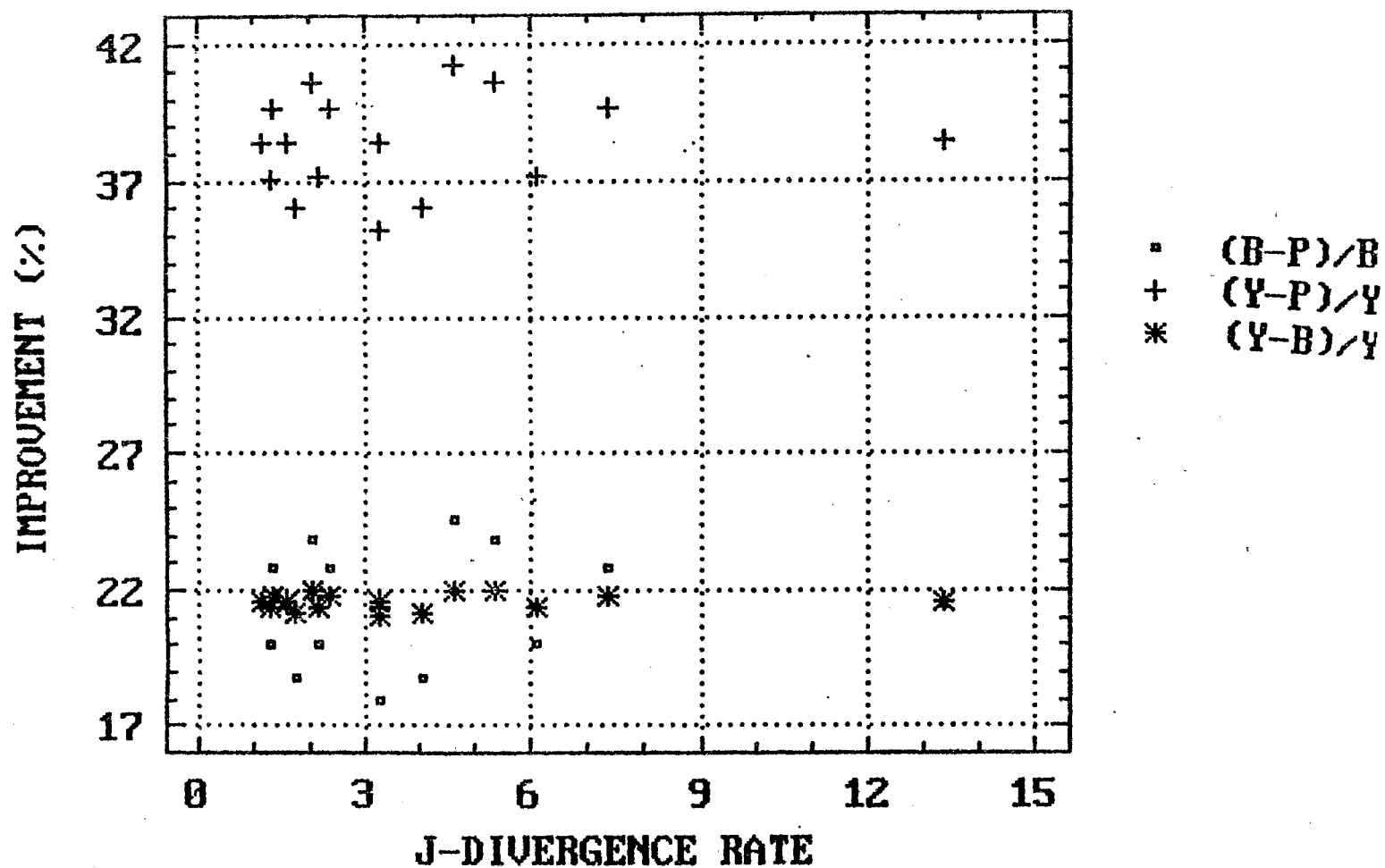


Figure 32. IMPROVEMENT RATE AND J-DIVERGENCE FOR
AR(2) VS AR(2) (FREQUENCY DOMAIN)

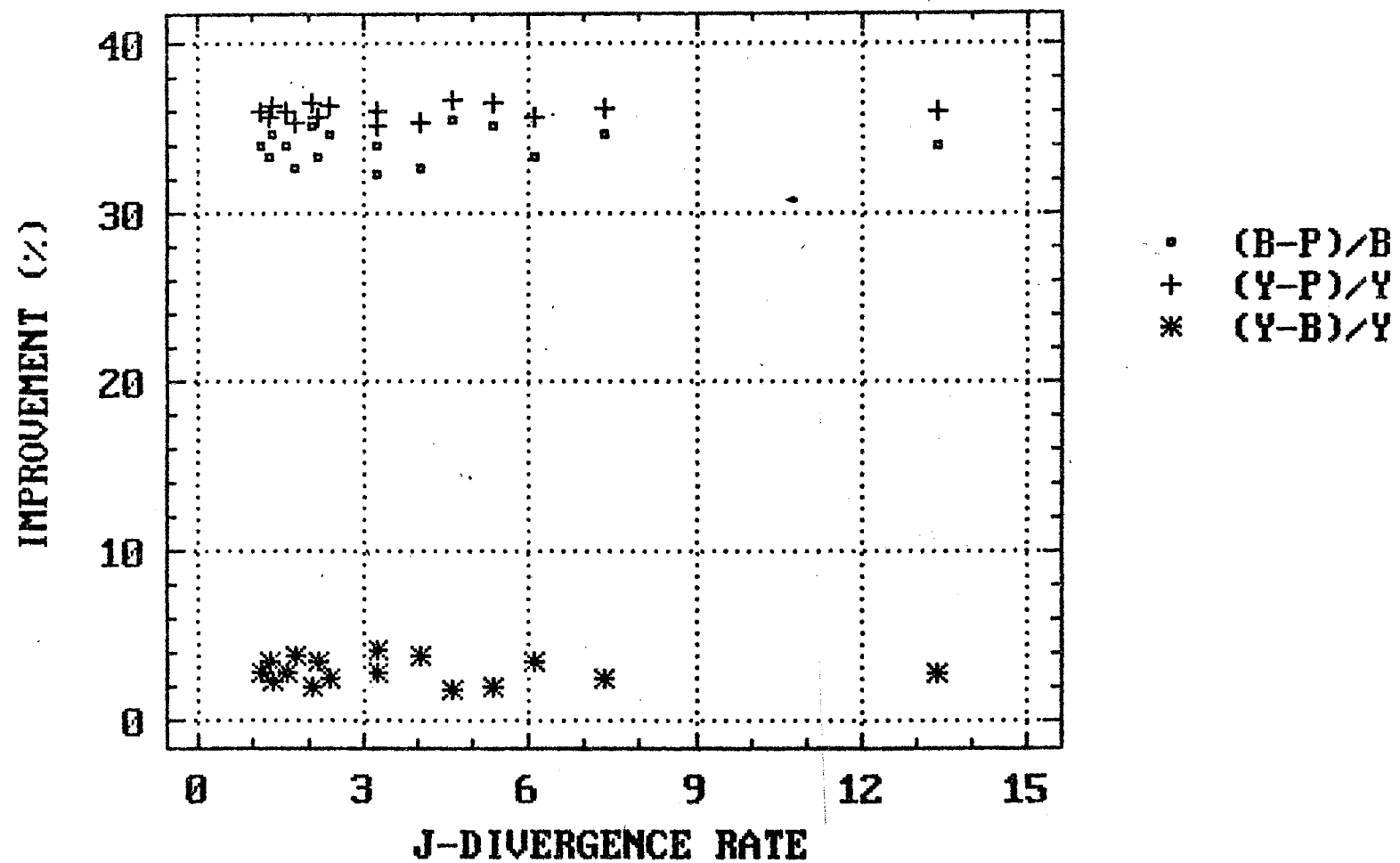


Figure 33. IMPROVEMENT RATE AND J-DIVERGENCE FOR AR(2) VS AR(2) (TIME DOMAIN)

VITA

Hon R. Tachia

Candidate for the Degree of
Doctor of Philosophy

Thesis: THE PREDICTIVE DISCRIMINATION OF AUTOREGRESSIVE
TIME SERIES WITH UNKNOWN ORDER

Major Field: Statistics

Biographical:

Personal Data: Born in Korinya, Benue State, Nigeria,
March, 15, 1953, the son of Tachia and Mbayan
Hon.

Education: Graduated from Mount Saint Michael's
Secondary School, Aliade, Nigeria, in December
1971; received Bachelor of Science Degree in
Statistics from University of Nigeria, Nsukka,
Nigeria in December, 1978; received Master of
Science Degree in Statistics from Iowa State
University in July, 1983; completed requirements
for the Doctor of Philosophy Degree at Oklahoma
State University in May, 1993.

Professional Experience: Instructor, Department of
Mathematics and Computer Science, Central State
University, Wilberforce, Ohio, September, 1987 to
June, 1989; Assistant Professor, Division of
Business and Economics, Wilberforce University,
Wilberforce, Ohio, from August 1989 to present.