BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-86/01

TIME SERIES ANALYSIS OF HOUSEHOLD HEADSHIP PROPORTIONS: 1959-1985

by

William Bell, James Bozik, Sandra McKenzie
and Holly Shulman
Statistical Research Division
Bureau of the Census
Washington, D.C. 20233

Recommended by: Nash J. Monsour

Report completed: July 2, 1986

Report issued: July 12, 1986

TABLE OF CONTENTS

1. INTRODUCTION

The Population Division of the Bureau of the Census periodically publishes projections of the number of households and families. In 1979 the Population Division issued the report "Projections of the Number of Households and Families: 1979 to 1995" (Series P-25, No. 805), and this year (1986) they released "Projections of the Number of Households and Families: 1986 to 2000" (Series P-25, No. 986). In this paper we describe the time series methodology used in the 1986 projections, evaluate the resulting and some alternative projections, and contrast this approach with the approach used in 1979.

The data available consist of 130 series, divided into 10 age groups within 13 categories. The categories, and a description of the data are found in Table A.1. Annual data from 1959 through 1985 are used.

In 1979, a regression approach was used for projecting the proportions for 1995. In some cases, the natural logarithm of the proportion ($Ln(p_t)$) was taken before performing the regression to prevent projections from going below zero. In other cases, the natural logarithm of one minus the proportion ($Ln(1-p_t)$) was taken to prevent projections from going above one. The 1995 projection was then connected to the actual 1978 value by a straight line. In this way projections for 1979 to 1994 were produced. Projections of these proportions were then combined with Census Bureau population projections to produce projections of the number of households in the various categories. Publication number 805, series P-25, contains a detailed discussion of the categories and procedure.

As noted above, in some cases the $Ln(p_t)$ transformation was used, while in others $Ln(1-p_t)$ was used. We considered use of a single transformation to prevent projected proportions from going below 0 or above 1. This is

discussed in Section 2. The logistic transformation, $Ln[(p_t)/(1-p_t)]$, was a natural one to examine; other transformations were also considered.

Section 3 discusses alternative forecasting methods, which were examined on a subset of the 130 series. These methods include ordinary and robust linear regression of the logistically transformed data on time. There was indication of the presence of outliers in some of the series.

In Section 4 ARIMA time series models are studied. Several alternative models are considered, but it is determined that insufficient historical data are available to allow effective statistical evaluation of these models. In the interest of simplicity and intuitive appeal, it is proposed to use two models--the (0,1,0) with constant and the (0,2,1)--in conjunction. Section 4 describes these models and their use.

Section 5 of the paper presents an evaluation of the alternative methods of household headship projections, using results from the 130 series. We conclude in Section 6 with a discussion of possible future avenues of research.

2. TRANSFORMATIONS

2.1 Alternative Transformations

A sample of twenty-one household headship rate series was examined in an attempt to determine what transformation, if any, of the original series of proportions was most appropriate for use in modeling and forecasting. The sample series are listed in Table A.2, and graphically displayed in Figure A.1. The transformations considered were

$$p_t = \text{original data} \quad \text{(no transformation)}$$
$$r_t = \arcsin(\sqrt{p_t})$$
$$y_t = Ln[p_t/(1-p_t)] \quad \text{(logistic)} \quad .$$

As noted earlier, the logistic transformation was considered as a natural alternative to the procedure previously used. The arcsin transformation was considered because it is effectively "in between" the logistic and no transformation, as can be seen in Figure B.1.

As far as statistical modeling of the data is concerned, the relevant question is which transformation does the most to stabilize the variance and make the data most nearly normally distributed. Stability of variance would not be expected of p_t, r_t, or y_t, since these series generally appear nonstationary. However, we might expect their first differences ∇p_t, ∇r_t, or ∇y_t , where $\nabla p_t = (p_t - p_{t-1})$, and ∇r_t and ∇y_t are similarly defined, to have stable variance. These series are still possibly autocorrelated, though the extent of this is difficult to determine due to the limited amount of data available. As a check, Table B.1 lists $r_1(\nabla y_t)$, the lag-1 sample autocorrelation coefficient of ∇y_t, for the sample of 21 series examined. It should be kept in mind that there is likely to be considerable error in these estimates since there are only 26 observations on each ∇y_t.

To assess the effects of transformations we examined time series plots, histograms, and normal probability plots of ∇p_t, ∇r_t, and ∇y_t. The time series plots and histograms were of little use, mainly because of the limited amount of data available. The normal probability plots were found to be somewhat useful for assessing distributional shape. Figure B.1 shows the arcsin and logistic transformations to be approximately linear for p_t in [.25, .75], so that they will affect distributional shape only for series with some values outside this range.

As p_t is limited to [0,1], we might expect ∇p_t to have a short upper tail in a series where the proportions are above .75, since we could not expect many large, positive ∇p_t's in such a series. One would hope the arcsin or

logistic transformation would "stretch-out" a short upper tail in ∇p_t. The behavior of the lower tail of ∇p_t is unclear in such a case. The reverse of these statements would apply to a series where the proportions are below .25. The normal probability plots were examined for each of the 21 sample series to determine whether the transformations had the desired effect. Table B.1 gives our impression of these plots.

There are significant outliers in the differenced data no matter what transformation is used. More than one positive outlier typically obscured other features of the distribution's upper tail, and analogously for negative outliers and the lower tail. Thus, it was sometimes difficult to judge if, apart from the outliers, the distribution had a short upper or lower tail.

Despite this qualification, and the fact that the range of some series was inside [.25, .75] so the choice of transformation didn't matter, the general impression was that there was some tendency for ∇p_t to have a short upper tail (except possibly for outliers) for a p_t series where the proportions are above 0.75, and a short lower tail for a p_t series where the proportions are below 0.25. The logistic seemed to correct for this better than the arcsin transformation. There was no evidence that one would be worse off using the logistic transformation, and doing so has the advantage of keeping projections within [0,1], as shown in section 2.2. Projections using the arcsin transformation will also remain within [0,1], but can be shown to eventually reach, and then reflect back off either the 0 or 1 boundary due to the periodic nature of the sine function. The fact that we are not sure of the extent of autocorrelation in the differenced series, and what effect this has on the normal probability plots, may be an important qualification in this evaluation of different transformations. However, considerably more data are required before we can remedy this.

2.2 Treatment of 0 or 1 Values With the Logistic Transformation

The approach using the logistic transformation for projections is to first project y_t, and then transform the projections to projections for p_t via the inverse of the logistic transformation :

$$p_t = \exp(y_t)/[1+\exp(y_t)] \quad .$$

The logistic is a 1-to-1 transformation of (0,1) to $(-\infty,\infty)$; thus, its inverse is a 1-to-1 transformation from $(-\infty,\infty)$ to (0,1). This guarantees that use of the logistic will force projected proportions to be in (0,1).

Minor problems arise in dealing with series containing values that are exactly 0 or 1. Two simple approaches are to treat such values as missing data, or to modify these values of 0 or 1 to values slightly above 0 or slightly below 1, respectively, before transforming. None of the twenty-one series in the sample had a value which was 0 or 1, but two of the remaining 109 series contained 0 values. For the regression models discussed later, it was simple to treat the 0 values as missing data. However, missing data create computational difficulties for time series modeling. Thus, for the time series modeling discussed in section 4, we tried replacing the 0 values by .005, which was approximately one half of the minimum of each series, ignoring the 0 values.

For the series MSIGQ14 (male secondary individuals living in guest quarters, ages 14-17), the zero values were all concentrated at the beginning of the data. For this series, using replacement values had a strong effect on the estimation procedure. The first six data points, which contained the 0 values, were omitted to avoid this situation. The result was an estimated

model that was thought to be much more reasonable. For the second series, MSIGQ18 (male secondary individuals living in guest quarters, ages 18-19), the 0 values were evenly spread throughout the series, and the replacement values seemed to cause no difficulties with the time series modeling.

3. SOME POSSIBLE MODIFICATIONS TO THE METHOD USED IN 1978

The method for projecting the proportions used in 1978 by the Population Division was basically as follows:

1. Using 15 years of data (1964-1978), the following model was fit:

$$\text{Ln}(p_t) = a_0 + a_1 t + e_t, \quad t = 1, \ldots, 15$$

p_t = proportion in year t

e_t = error term

If the estimate of $a_1 > 0$, the following alternate model was fit:

$$\text{Ln}(1-p_t) = a_0 + a_1 t + e_t$$

If a_1 is still > 0, the model with the smaller value of a_1 was used.

2. The fitted model was used to project $\text{Ln}(p_t)$ or $\text{Ln}(1-p_t)$, and thus p_t, for 1995. The 1978 p_t value was connected to the 1995 p_t projection with a straight line to obtain projections for 1979-1994.

The method above was used to obtain a series of projections called series B. Three other series, derived from series B, were also published. Census Bureau publication number 805, series P-25 discusses the different series published in 1979.

3.1 Use of the Logistic Transformation

As a first step in developing alternatives to the projection methodology used in 1978, it was natural to consider if some simple changes could be made to the 1978 method, while still retaining the basic element of linear regression on time. The first such change considered was the logistic transformation discussed in Section 2, which provides a simple means of constraining projections of the individual rates to (0,1). A revision of the 1978 method that uses the logistic transformation follows these steps:

1. logistic transformation of each series p_t to $y_t = \text{Ln}(p_t/(1-p_t))$;
2. fitting the linear model $y_t = a_0 + a_1 t + e_t$;
3. extrapolating with the fitted model to get a target value, $\hat{y}_{t*} = \hat{a}_0 + \hat{a}_1 t^*$, at some future time point t^* (e.g., t^* = the year 2000);
4. linearly interpolating between the last data point, y_n, and $\hat{y}_{t*}$ to get projections of y_t for the intervening years;
5. transforming $\hat{y}_{n+1},\ldots,\hat{y}_{t*}$ back to $\hat{p}_{n+1},\ldots,\hat{p}_{t*}$ via $\hat{p}_t = \exp(\hat{y}_t) / (1+\exp(\hat{y}_t))$, to get the projections of p_t.

Steps 4 and 5 represent another modification to the 1978 method which transformed the target value back to the p_t scale, then performed a linear interpolation to get projections of p_t for the intervening years (essentially interchanging steps 4 and 5). The drawback to linear interpolation in the p_t scale is that a linear forecast function for p_t must, if extended beyond t^*, ultimately cross the 0 or 1 boundary (unless $\hat{a}_1 = 0$). In the scheme above, the forecast function is linear in the y_t scale. This produces a curved forecast

function for p_t that, if extended beyond t^*, will be asymptotic to (i.e., approach but never reach) either the 0 or 1 boundary.

3.2 Weighted Least Squares

A second feature of the 1978 method studied was the use of weighted least squares in fitting the straight line. Weighted least squares estimates of a_0 and a_1 resulted from minimizing

$$\sum_{t=1}^{n} w_t(\mathrm{Ln}(p_t) - a_0 - a_1 t)^2$$

where $w_1,\ldots,w_n$ are the weights, and for some series $\mathrm{Ln}(p_t)$ was replaced by $\mathrm{Ln}(1-p_t)$. Ordinary least squares estimation would use $w_1 = \ldots = w_n = 1$. In 1978, the w_t's were chosen to be approximately proportional to the sample size of the Current Population Survey (CPS). In fact, the CPS sample size changed several times between 1964 and 1978 (the span of data used in 1978), but most changes were relatively minor except for an increase from 33,000 to 48,000 households in 1970. This increase of approximately 50% led to the weights $w_1 = \ldots = w_6 = 67$ (for 1964 - 1969) and $w_7 = \ldots = w_{15} = 100$ (for 1970 - 1978).

The assumptions appropriate for the use of weighted least squares are questionable for this data. The model underlying this use of weighted least squares is

$$\mathrm{Ln}(p_t) = a_0 + a_1 t + e_t, \quad \mathrm{Var}(e_t) \propto w_t^{-1} . \tag{1}$$

The assumption of $\mathrm{Var}(e_t) \propto w_t^{-1}$ comes from the usual result of variance being inversely proportional to sample size. Actually, if p_t is viewed as the observed proportion resulting from n_t i.i.d. Bernoulli trials with common

probability of success π_t, then $Var(p_t) = \pi_t(1-\pi_t)/n_t \propto \pi_t(1-\pi_t)w_t^{-1}$. Letting n_t be the sample size at time t, we are then led to the weights used if we make the additional assumptions that the differential effects of $\pi_t(1-\pi_t)$ and the effects of the logarithmic transformation can be ignored.

These last two assumptions could certainly be questioned. However, the real problem with model (1), and consequently with the weighted least squares procedure, stems from the assumption that apart from sampling error the proportions follow a straight line -- i.e., $\pi_t = a_0+a_1t$. If this were true, the graph of each series would be very nearly a straight line, with usually only slight variations about the line due to the sampling error. Few, if any, of the graphs show behavior approaching this. A more realistic model assumes that the true underlying proportions contain inherent stochastic variation u_t over time, independent of the sampling variation. This can be written $\pi_t = a_0 + a_1t + u_t$, and, ignoring the effects of the Ln transformation again, leads to the model

$$Ln(p_t) = a_0 + a_1t + u_t + e_t \qquad (2)$$

where the variance of the error term is now

$$Var(u_t + e_t) = Var(u_t) + Var(e_t).$$

Accepting the above arguments for $Var(e_t)$ being proportional to n_t^{-1}, the error variance is not approximately proportional to n_t^{-1} unless $Var(u_t)$ is small relative to $Var(e_t)$. In fact, analysis of the data suggests $Var(u_t)$ is far greater than $Var(e_t)$ in most cases. Thus, the dependence of $Var(e_t)$ on sample size can effectively be ignored. Possible exceptions occur in some of

the lowest or highest age groups for which certain series are highly volatile. Even so, the volatility in these series may well be due to a high $Var(u_t)$ producing large fluctuations in the π_t's.

Discarding model (1) in favor of model (2) removes the reason for using weighted least squares. A serious difficulty with the use of weighted or ordinary least squares on (2) is that u_t is certainly correlated over time, and most probably is a nonstationary time series. We attempt to address this issue with time series models later.

3.3 Extent of Data Used

To this point our recommended modifications to the 1978 method are to (1) use the logistic transformation, (2) do the linear interpolation in the transformed scale and then transform back, and (3) drop the use of weighted least squares. Another issue concerns how much data to use. In 1978, data from 1964 - 1978 was used, despite the availability of data back to 1959. Statistical principles generally argue for use of as much data as are available, as long as it all follows the same model. Our work has generally used all the available data, except for analyses making comparisons with the 1978 projections. Unless there is some knowledge (theoretical or empirical) suggesting that early years of data are not comparable with later years, we would be leery of discarding data. Any problems arising from use of all the data most likely stem from inappropriateness of the model used.

3.4 Treatment of Outliers

A final issue to consider in regard to modifications to the 1978 method is the possible effect of outliers on the least squares fitting and resulting projections. This suggested the possibility of replacing least squares

regression with a robust fitting scheme. This was done for the sample set of 21 series, using a robust regression routine from Interactive Data Analysis by McNeil (1977). All regressions were performed on the logistically transformed data.

Robust regression using Tukey's biweight influence function was performed three times, with c values of 2, 4, and 6. The value c allows one to adjust the amount that observations far away from the fitted line are downweighted, with more drastic downweighting the lower the value of c. McNeil recommends choosing a value of c between 4 and 10, with 10 likely to give results very close to least squares, and 4 as low as one would probably ever want to go (McNeil (1977), p. 157). The value of 2 for c was tried since c = 4 usually produced results very close to least squares. Such a low value for c was needed to achieve results different from least squares due to the strong autocorrelation in the series, as explained below.

Table C.1 gives parameter estimates for four regressions: least squares, and three robust regressions using different values of c. The parameters estimated are a_0 and a_1 in the regression equation $y_t = a_0 + a_1 t + e_t$. (Recall y_t in this case is the logistic transform Ln $(p_t/(1-p_t))$, where p_t is the proportion in question at time t.) For many series, the estimates of a_0 and a_1 for the four runs are similar. Note the value of c = 6 produced results very close to the least squares regression results, while c = 2 sometimes provided estimates different from least squares. Note specifically the series SM20, SM25, and FFH35, for which the results for c = 2 differ substantially from the others.

The series and the regression fits were also examined graphically on both the original and the tranformed (logistic) scale. Example graphs for series SM20 (Single Males, Age 20-24) are included in Figures C.1 and C.2. The

different performance for the c = 2 robust regression is apparent here. It seems c = 2 ignores too many observations.

The straight line fits are generally very poor, and since there is strong autocorrelation present, when one data point lies far from the line, the neighoring points generally do so as as well. This results in a large variance estimate even when the variance is estimated robustly. Since the cutoff point beyond which observations get zero weight depends on the product of c and the residual standard error, very low values for c are needed to compensate for the high residual standard error.

Thus, robust regression does not appear to provide significiantly better estimates of a linear relation for use in projection than least squares regression in the series that were studied. Straight lines do not fit the data well, and outliers, while a problem, are not nearly as big a problem as autocorrelation.

3.5 Conclusions

Modifications to the 1978 method that we recommend are the use of the logistic transformation, the dropping of weighted least squares, and the use of all available data in fitting models, unless there are concrete reasons (theoretical or empirical) why early data should not be used. Outliers are a problem in many of the series, but robust regression does not help. Autocorrelation in the data is a far more important problem than outliers. In fact, the strong autocorrelation in these series results in linear functions of time providing extremely poor fits to these data. Thus, our final conclusion in regard to the 1978 method is that linear regression on time is not an appropriate model for these data, and thus its use should be dropped. (Obviously, at this point, we have completely dismantled the 1978 method.) As

an alternative, in the next section we investigate the use of ARIMA time series models. The objective of these models is to explicitly model the autocorrelation structure of time series.

4. ALTERNATIVE METHODS: Time Series Models

4.1 Introduction

The general ARIMA (autoregressive-integrated-moving average) model of order (p,d,q) for y_t can be written

$$\nabla^d y_t = x_t$$

$$x_t = \theta_0 + \phi_1 x_{t-1} + \ldots + \phi_p x_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}$$

where

(i) ∇ is the differencing operator $\nabla y_t = y_t - y_{t-1}$

(ii) $\nabla^d y_t = \nabla(\nabla^{d-1} y_t)$ so, e.g. $\nabla^2 y_t = \nabla y_t - \nabla y_{t-1}$

(iii) a_t is a random error series

(iv) $\theta_0, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$ and $\sigma_a^2 = \mathrm{Var}(a_t)$ are parameters.

The approach is to difference y_t a suitable number of times d to produce a stationary series x_t (one varying about a constant mean level and whose other properties are stable over time). Then x_t is related linearly to a constant θ_o (sometimes constrained to be zero), past values of itself (the autoregressive part or 'AR'), and current and past values of the random error series (the moving average part or 'MA'). The "I" in ARIMA stands for "integrated", which is the inverse of differencing. ARIMA modeling consists of choosing the model orders (p,d,q) and fitting the model using standard

computer software to produce estimates of θ_o, $\phi_1,\ldots,\phi_p$, $\theta_1,\ldots,\theta_q$, σ_a. Forecasts can then be produced from the estimated model.

Time series observations are generally not independent of one another -- they usually exhibit a high degree of correlation over time. This makes standard statistical techniques such as regression inappropriate for time series data (often highly inappropriate). ARIMA models attempt to account for correlation in time series data with a small number of parameters (small p and q). Most time series are so strongly correlated that they require differencing once, or sometimes twice, to produce a stationary series suitable for further modeling.

ARIMA modeling is typically done with fairly long time series -- 50 or more observations for a nonseasonal series and more for a seasonal series. Time series as short as the headship rate series (27 observations from 1959-1985) present some difficult problems. It was obvious from plots of the series that virtually all the headship series would require taking $x_t = \nabla y_t = y_t - y_{t-1}$. There was a question as to whether some series might benefit from taking a second difference, $x_t = \nabla^2 y_t$. As will be seen later, this was tried via models that also introduced an MA operator, $(1 - \theta B)$, when the additional difference was included. If the second difference is not needed, the estimate of θ will be very close to 1. (For this model, θ is constrained by $|\theta| \leq 1$.) In fact, if $\hat{\theta} = 1$, one can cancel a $(1 - B)$ from both sides of the equation, which reduces the equation to that of a first difference model. So the models with $d = 2$ could default back to a model with $d = 1$, depending on estimates of the moving average parameters.

Beyond the choice of d, the usual model selection procedures are based on the fact that given models correspond to distinct patterns in the auto-correlations, $\rho_k = \text{Corr}(x_t, x_{t-k})$ $k = 1,2,3,\ldots$. If the usual sample

estimates of ρ_k are reasonably reliable, one may examine these for a pattern corresponding to a given model to select p and q. Unfortunately, the shortness of the headship series precluded any serious attempts at selection of p and q by statistical means. The approach taken was to try several models with various small (usually 0 or 1) values of p and q, examine forecasts from these models to see how they differed, and try to choose a model or models that seemed reasonable for most of the series.

Another problem with short time series is that even if a reasonable model is selected, estimates of the model parameters, and consequently forecasts calculated from the estimated model, will be unreliable. There is no way around this problem. The best course is to limit models to a very small number of parameters. While adding parameters to a model may make it more flexible and realistic, for short series these benefits are quickly offset by the increased error due to estimating more parameters.

A final problem in forecasting with short time series is that it is difficult to assess the accuracy of forecasts from a given model, or to compare the relative accuracy of forecasts from different models or procedures. Comparison of forecasts from a single origin with later observed actual values are likely to be governed by one or two random shocks after the forecast origin. This makes such comparisons of little value in deciding which model or forecasting procedure will perform well in the future. Use of a small number of different forecast origins cannot be expected to do much better, yet a small number of origins is all that would be available for short series. The accuracy of long term forecasts is especially difficult to assess -- much more so than forecasts one or two years ahead. It is not hard to see that with 27 years of data little can be said about where the series will be

15 years from now. In section 4.5, we illustrate some of the forecasting problems by calculating forecast intervals for several series.

The next section describes our efforts at empirical evaluation of forecasts from various ARIMA models. The conclusions are primarily negative -- these empirical comparisons do not offer a reliable method of selecting "best" models for each series individually, or of selecting one model to use on all series that is "best" in some sense. Thus, the choice of which model to use must be based on other considerations, such as simplicity and intuitive appeal.

While these problems have been discussed mostly in the context of ARIMA modeling, they are in fact present, in one form or another, for any forecasting procedure applied to short series. ARIMA models are neither more nor less susceptible to these difficulties than other approaches. It should also be kept in mind that just because it is difficult to use ARIMA models to adequately model the correlation structure of a short series, doing nothing about autocorrelation is not an appropriate alternative. Autocorrelation can cause serious problems for other procedures, and doing nothing about it effectively assumes autocorrelation is not present (an ARIMA(0,0,0) model). This is likely to be worse than making an attempt at modeling.

4.2 Empirical Forecasting Evaluation

Five models were studied. The notation used is that of Box and Jenkins (1970).

(0,1,0) "RANDOM WALK"

$$\nabla y_t = a_t$$

(0,1,0)-WITH-CONSTANT

$$\nabla y_t = \theta_0 + a_t$$

(2,1,0)-WITH-CONSTANT

$$\nabla y_t = \theta_0 + \phi_1(\nabla y_{t-1}) + \phi_2(\nabla y_{t-2}) + a_t$$

(4,1,0)-WITH-CONSTANT

$$\nabla y_t = \theta_0 + \phi_1(\nabla y_{t-1}) + \phi_2(\nabla y_{t-2}) + \phi_3(\nabla y_{t-3}) + \phi_4(\nabla y_{t-4}) + a_t$$

(0,2,2)

$$\nabla y_t^2 = y_t - 2y_{t-1} + y_{t-2} = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$$

All modeling was performed on the logistically transformed data. A conditional likelihood estimation procedure from the SCA time series package was used to estimate the models and produce forecasts. The conditional likelihood procedure sometimes has difficulty estimating moving average parameters with values near 1. In such instances an exact likelihood estimation is recommended. The exact likelihood estimation in SCA was therefore used for the (0,2,2) model. Occasionally this estimation procedure will not converge, in which case the model cannot be estimated in the SCA package. For the (0,2,2) model this situation occurred with eight of the 21 sample series, leaving the remaining four models to be compared. Six recent observations (1979-1984) were witheld before modeling the series and used to

compare and evaluate forecasts from the models. Plots of the forecasts were also compared with these actual observations.

Considering that the models were estimated using only 20 observations, the forecasts are likely to be unreliable. Since forecast accuracy decreases the farther one tries to forecast into the future, less emphasis should be placed on 4, 5, and 6-years-ahead forecasts. It was decided to evaluate the models based both on forecasts through 3-years-ahead and on forecasts through 6-years-ahead. For the 3-years-ahead forecast evaluation, data from 1959-1978 were used to produce forecasts for 1979-1981. Then data from 1959-1981 were used to re-estimate the model and produce forecasts for 1982-1984. Due to the inherent increase in uncertainty of the forecasts as they extend further into the future, a weighted mean absolute error and a weighted root mean square error measure were computed, with weights that decrease geometrically as the forecasts extend into the future. The measures for the 6-years-ahead forecast evaluation are defined as

$$\text{weighted MAE} = \frac{\sum_{i=1}^{6} \beta^{i-1} \left| f_{1978,i} - y_{1978+i} \right|}{\sum_{i=1}^{6} \beta^{i-1}}$$

$$\text{weighted RMSE} = \left[\frac{\sum_{i=1}^{6} \beta^{i-1} \left(f_{1978,i} - y_{1978+i} \right)^2}{\sum_{i=1}^{6} \beta^{i-1}} \right]^{1/2}$$

where $f_{j,i}$ = ith-step-ahead forecast from origin j. For the 3-step-ahead evaluations,

weighted MAE =

$$\frac{\sum_{i=1}^{3} \beta^{i-1} |f_{1978,i} - y_{1978+i}| + \sum_{i=1}^{3} \beta^{i-1} |f_{1981,i} - y_{1981+i}|}{2 \sum_{i=1}^{3} \beta^{i-1}}$$

weighted RMSE =

$$\left[\frac{\sum_{i=1}^{3} \beta^{i-1} (f_{1978,i} - y_{1978+i})^2 + \sum_{i=1}^{3} \beta^{i-1} (f_{1981,i} - y_{1981+i})^2}{2 \sum_{i=1}^{3} \beta^{i-1}}\right]^{1/2}$$

Values for β of .7 and .85 were chosen to vary the deemphasis given to longer term forecasts. These four measures were computed for all five models on all 21 series, except for the (0,2,2) model with the eight series for which the estimation did not converge. For each series, the models were ranked according to how close the forecasts produced were to the actual values as measured by the criteria above.

From example graphs of the evaluation measures for two of the series (Figures D.1 and D.2) it is clear that there is little difference in forecast evaluation between the RMSE and MAE, and between the two values of β. It should be noted here Figure D.2 (the series FPI20) is an example where the estimation did not converge for the (0,2,2) model. The number of forecast years used did in some cases influence the outcome. With the accuracy of longer forecasts in doubt, it was felt that the outcome of the 3-years-ahead analysis was more relevant. The ranks for the 13 sample series where the (0,2,2) model was estimable are displayed in Figure D.3. The rankings of all

the models, both including and excluding the (0,2,2) model, are displayed in Figure D.4. The (0,1,0) and (0,2,2) models had the most top ranks (4 each) but also had the most lowest ranks (5 and 3, respectively). No model seems obviously superior from these graphs or the tables in Figure D.4.

A nonparametric Friedman rank sum test was performed on the rankings, the result being no model significantly better than the others. The forecast evaluation measures, likewise, did not point to a clearcut choice of model to use. Rather than select different models for different series, it thus seemed desirable to choose one model for all series. The (0,1,0)-with-constant model was selected as a reasonable model that was certainly as good as any other and, in fact, had the highest individual Friedman rank sum. In addition, the (0,1,0)-with-constant model has simplicity as an appealing feature. Forecasts of this model will be a straight line from the last data point, with a slope that is the mean of the differenced data.

4.3 Robust Estimation of (0,1,0)-With-Constant

The method of estimating slope in the (0,1,0) model produces forecasts that are equivalent to connecting a line from the first observation to the last observation and extending the line for forecasting. This would seem to ignore the rest of the data. While such a forecasting procedure is appropriate if the (0,1,0) model is correct and there are no outliers in the data, an outlier at the first or last time point could cause poor forecasts. As noted in section 2.1, outliers appear to be present in many of the series. A procedure that is affected little by outliers is called "robust". Hence, use of a robust procedure to estimate the constant θ_0 in the (0,1,0)-constant model was investigated.

Kafadar (1982) discusses the formation of a "t-like" statistic that replaces the classical mean or average with a biweight estimate,

$$T_{bi} = \sum_{t=1}^{N} [x_t \cdot b(u_t)] \; / \; \sum_{t=1}^{N} b(u_t),$$

where $u_t = (x_t - T_{bi}) / (c.s_{bi})$. In our procedure, T_{bi} is an estimate of θ_o for the (0,1,0)-constant model. This is an iterative process, with $T_{bi}^{(0)} =$ median (x_t) . s_{bi} is a measure of the spread of the data that remains constant over the iterations. It is proportional to the median absolute deviation,

$$s^{(0)} = \underset{t}{\text{med}} \; | \; x_t - T_{bi}^{(0)} \; |.$$

The biweight function itself,

$$b(u) = u(1 - u^2)^2, \; u \leq 1,$$

$$= 0, \; |u| > 1$$

reduces the influence of values far away from a measure of center by downweighting them in the resulting estimate. Thus, we taper the weights given to x-values more than $(c.s_{bi})/2$ away from T_{bi}, down to zero for values more than $(c.s_{bi})$ away from T_{bi}. A graph of the biweight function can be found in figure E.1. As noted in section 3, McNeil (1977) suggests picking c to be in the range [4,10]. Results very near to use of the arithmetic mean follow for c = 10, and c = 4 usually provides great protection from outlying values.

Graphs comparing the nonrobust (0,1,0)-constant forecasts with the robust forecasts for c = 4 and 6 revealed that use of robust procedures can affect the forecasts. The robust forecasts appeared more reasonable when they

differed from the non-robust forecasts, especially in series with outliers present near the beginning or end of the data. Figure E.2 is an example of this. With c = 6, the extreme values are downweighted some, resulting in forecasts that differ from those of the nonrobust procedure. Using c = 4 produces more downweighting and forecasts that differ still further. With limited data and many series, we thought it better to strive for more protection from outliers rather than less, and so judged that choosing c = 4 was more appropriate. Thus, the modelling of all 130 series,using an (0,1,0)-constant model, was performed with the robust estimation procedure, with c = 4.

4.4 (0,2,1) Model Considered

The graphs of some series seemed to indicate a change of slope in recent years, particularly since 1978. Prime examples are the series FFH25 just considered, as well as the series MPI20 (male primary individuals, ages 20-24), FFH30 (female head of household, husband not present, ages 30-34), FFH35 (female head of household, husband not present, ages 35-40). Graphs of these series are given in Figures F.1, F.2, and F.3 respectively. This raised the possibility of using a model with a linear forecast function, where the slope is estimated in a way that gives more weight to recent differences of the data. The (0,2,1) model is such a model. Letting $x_t = \nabla y_t$, we can write the model as:

$$\nabla^2 y_t = \nabla x_t = (1 - \theta B)a_t \qquad (Ba_t = a_{t-1}).$$

This is a (0,1,1) model for the once-differenced values x_t. The (0,1,1) model produces constant forecasts of x_t that are an exponentially weighted moving average (EWMA) of the data:

$$\hat{x}_t(\ell) = (1 - \hat{\theta})[x_t + \hat{\theta}x_{t-1} + \hat{\theta}^2 x_{t-2} + \ldots]$$

Notice the weights on the observations decline exponentially and sum to one. Thus, this model produces a constant forecast of the future slope, and hence a linear forecast for y_t (which turns out to start from the last data point), with more weight given to recent observations in determining the slope.

The above formula for $\hat{x}_t(\ell)$ applies when the series is long enough, and $\hat{\theta}$ is small enough so that the weight given to x_1, $\hat{\theta}^{t-1}$, is effectively zero. For short series, or $\hat{\theta}$ values near 1, this will not be the case and the weights in $\hat{x}_t(\ell)$ will not be exactly $(1 - \hat{\theta})\hat{\theta}^j$. In fact, for $\hat{\theta} = 1$ the weights turn out to be all 1/n, so the (0,2,1) model with $\hat{\theta} = 1$ produces the same forecasts as the (0,1,0) model with a (non-robustly estimated) constant. As another way to look at this, notice if $\theta = 1$ in the (0,2,1) model we can cancel a $\nabla = 1-B$ on the left hand side with the $1 - \theta B = 1-B$ on the right to get

$$\nabla y_t = \theta_o + a_t$$

the (0,1,0) with constant model for y_t. We preferred to treat the (0,1,0)-constant model separately and not as a special case of the (0,2,1) with $\theta = 1$, so that we could use a robust estimate of θ_0. While techniques developed for robust estimation and forecasting with general ARIMA models could be used on the (0,2,1) model, these techniques are usually applied to

series longer than 27 observations, and they require considerably more effort for each series.

The (0,2,1) model was fitted to all 130 headship rate series by the exact maximum likelihood estimation procedure in the SCA time series package. As noted earlier, use of exact maximum likelihood is important in estimating models with moving average terms. For models with a $1 - \theta B$ moving average term with θ near 1, conditional likelihood tends to underestimate θ . This is crucial in our application, since many of the series probably do not require a second difference. In that case, we would expect to get $\hat{\theta} = 1$, so the model would reduce to the (0,1,0)-constant model. Unfortunately, the SCA package does not constrain $\hat{\theta}$ to the invertibility region [-1,1], with the result that for many of the series the SCA package produced an estimate of $\hat{\theta}$ greater than 1. One expects that the correct maximum likelihood estimate of θ is 1 in these cases, though we did not pursue the further work needed to verify this.

Of 130 series, 21 had values of $\hat{\theta}$ less than 1 for the (0,2,1) model. The (0,2,1) model yielded forecasts that were practically identical to the (0,1,0)-with constant non-robust forecasts in 10 of these 21 series. In some of those 10 cases the series exhibited no significant change in level over the 27 observations, so that the forecasts from both models varied little from one another. (See Figure F.4 and note the scale of the vertical axis.) In the other cases, the series did not seem to exhibit much change in slope in recent years, in which case forecasts from the (0,2,1) model differ little from forecasts from the non-robust (0,1,0)-with-constant model. In producing forecasts for the (0,2,1) model, we thus revert back to the (0,1,0)-with-constant forecasts for the 109 series for which $\hat{\theta} \geq 1$, and for the 10 series with $\hat{\theta} < 1$ where the (0,2,1) model made little difference.

In reverting back to the (0,1,0)-with-constant model, outliers may still be a problem, so robust estimation of the model was used. For the 11 remaining series, forecasts from the (0,2,1) model were used. Although they are few in number, they are among the more influential individual series with respect to their contribution to the total number of households. Hence it was felt that the total effect on the household projections due to incorporating the (0,2,1) forecasts for just the 11 series should not be ignored.

4.5 Conclusions

The modeling procedure settled on for the proportions was to use the (0,2,1) model only when (1) $\hat{\theta} < 1$, and (2) the (0,2,1) forecasts appeared somewhat different from the non-robustly estimated (0,1,0)-with constant forecasts. Otherwise, the robustly estimated (0,1,0)-with constant model was used. Requirement (1) comes from the fact that the (0,1,0)-with constant model is the special case of the (0,2,1) model with $\theta = 1$. Requirement (2) was used to protect against outliers since we could not estimate the (0,2,1) model robustly. If the (0,2,1) forecasts did not differ much from the non-robust (0,1,0)-with constant forecasts, it seemed safer to use the robustly estimated (0,1,0)-with constant model in case outliers were present. These models were used to forecast the 130 series of proportions, with the (0,2,1) model being used for 11 of these series. The resulting forecasts of the proportions were used to produce the headship projections referred to as Series A in Bureau of the Census (1986, Series P-25, No. 986).

The decision to use the (0,1,0)-with constant in conjunction with the (0,2,1) model was based largely on their simplicity and intuitive appeal. The shortness of the historical time series of proportions (27 annual observations) prevents any effective statistical discrimination between alternative

time series models. This is true both for selecting a model for any one particular series and for selecting one model to use for all series. Differencing each series at least once seems wise. Until substantially more data become available, it is difficult to say much beyond this.

5. EXAMINATION OF FORECASTS AND PROJECTIONS

5.1 Upper and Lower Bounds for Forecasts

Earlier we noted one would not have much statistical faith in forecasts for the year 2000, based on only 27 years of annual data (1959-1985). To get some idea of the precision of the forecasts from the (0,2,1) and (0,1,0)-with-constant models, upper and lower 95% confidence bounds for the forecasts were calculated using both models, for several headship rate series. These calculations proceed from the assumption that the estimated model is, in fact, the correct model for the data. Since this is not the case in practice, and since our data here are limited in length, these calculations will give only a very rough idea of the precision of the forecasts. First, we will derive the forecast error variances for the two models. We will then discuss the results of the forecast bounds for several series. The use of the two different models gave some indication of the sensitivity of the results to model choice. Finally we will address the option of a constant rate forecast, i.e.; using the 1985 value as the projection into the future.

Derivation of Forecast Error Variances

The general ARIMA (p,d,q) model,

$$\phi(B)y_t = \theta(B)a_t \qquad (1)$$

may be expressed in several ways. One way is in terms of the current and previous values of the uncorrelated random shocks a_t only:

$$\begin{aligned} y_t &= a_t + \psi_1 a_{t-1} + \psi_2 a_{t-2} + \cdots \\ &= a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \\ &= \psi(B) a_t \end{aligned} \qquad (2)$$

(Actually, this result needs to be generalized for models involving differencing, but this need not concern us here.) This form of the model may be used in calculating the variance of the forecast error, $V(\ell)$ - the variance of the difference between the actual value at some future time point $n + \ell$ (data through time point n), and its forecasted value:

$$V(\ell) = \mathrm{Var}(y_{n+\ell} - \hat{y}_{n+\ell})$$

In the form of (2) we have, since a_t is known for $t \leq n$

$$\begin{aligned} y_{n+\ell} &= a_{n+\ell} + \psi_1 a_{n+\ell-1} + \cdots + \psi_{\ell-1} a_{n+1} + \psi_\ell a_n + \cdots \\ \hat{y}_{n+\ell} &= \hat{a}_{n+\ell} + \psi_1 \hat{a}_{n+\ell-1} + \cdots + \psi_{\ell-1} \hat{a}_{n+1} + \psi_\ell a_n + \cdots \quad . \end{aligned}$$

Combining these expressions,

$$\begin{aligned} y_{n+\ell} - \hat{y}_{n+\ell} &= (a_{n+\ell} - \hat{a}_{n+\ell}) \\ &\quad + \psi_1 (a_{n+\ell-1} - \hat{a}_{n+\ell-1}) + \cdots + \psi_{\ell-1} (a_{n+1} - \hat{a}_{n+1}) \quad . \end{aligned}$$

But $\hat{a}_t = 0$ for $t > n$ so

$$V(\ell) = \mathrm{Var}(y_{n+\ell} - \hat{y}_{n+\ell}) = \mathrm{Var}(a_{n+\ell} + \psi_1 a_{n+\ell-1} + \ldots + \psi_{\ell-1} a_{n+1})$$
$$= (1 + \psi_1^2 + \psi_2^2 + \ldots + \psi_{\ell-1}^2)\sigma_a^2 \quad . \qquad (3)$$

To get estimates of the ψ's in terms of the parameters of a specific model, we apply $\phi(B)$, the general autoregressive operator, to both sides of (2),

$$\phi(B)y_t = \phi(B)\,\psi(B)a_t \qquad (4)$$

but from (1), $\phi(B)y_t = \theta(B)a_t$. Therefore, from equations (4) and (1),

$$\theta(B) = \phi(B)\,\psi(B) \qquad (5)$$

Thus, the ψ_j for (3) may be obtained by equating coefficients of B^j, $j=0,1,2, \ldots n$:

$$(1-\theta_1 B - \ldots - \theta_q B^q) = (1-\phi_1 B - \ldots - \phi_{p+d} B^{p+d})(1+\psi_1 B+\psi_2 B^2 + \ldots) \qquad (6)$$

Estimates of the ψ_j result from substituting estimates of the ϕ's and θ's in (6).

Now, the (0,2,1) model, in its general form,

$$\nabla^2 y_t = (1 - \theta_1 B)a_t$$

may be expressed in its expanded form,

$$y_t - 2y_{t-1} + y_{t-2} = a_t - \theta_1 a_{t-1} \quad .$$

From (6),

$$(1 - \theta_1 B) = (1 - 2B + B^2)(1 + \psi_1 B + \psi_2 B^2 + \ldots)$$

the ψ_j are calculated to be

$$\begin{aligned}
\psi_0 &= 1 \\
\psi_1 &= 2 - \theta_1 \\
\psi_2 &= 2(2 - \theta_1) - 1 = 3 - 2\theta_1 \\
\psi_3 &= 2\psi_2 - \psi_1 \quad\quad = 4 - 3\theta_1 \\
&\downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow \\
\psi_j &= 2\psi_{j-1} - \psi_{j-2} = 1 + (1 - \theta_1)j
\end{aligned}$$

To compute the forecast error variance in (3), an estimate of σ_a^2 was obtained from the SCA computer package.

For the (0,1,0)-with-constant model,

$$y_t - y_{t-1} = \theta_0 + a_t$$

the actual value at some future time point $n + \ell$ can be shown to be

$$y_{n+\ell} = y_n + \ell\theta_0 + \sum_{j=1}^{\ell} a_{n+j} \quad .$$

Since the best forecast of future a_{n+j}'s is zero, our forecast from the (0,1,0)-with-constant model is then

$$\hat{y}_{n+\ell} = y_n + \ell\hat{\theta}_0 \quad .$$

where $\hat{\theta}_0$ is the estimate of θ_0. Here, we calculate the forecast error variance directly:

$$V(\ell) = \mathrm{Var}(y_{n+\ell} - \hat{y}_{n+\ell}) = \mathrm{Var}(\ell(\theta_0 - \hat{\theta}_0) + \sum_{j=1}^{\ell} a_{n+j})$$

$$= \ell^2 \mathrm{Var}(\hat{\theta}_0) + \ell\sigma_a^2 \qquad (7)$$

$\hat{\theta}_0$, a function of the data through time n, is independent of $a_{n+1}, \ldots, a_{n+\ell}$. If $\hat{\theta}_0$ was calculated as the mean of the first differences, its variance would be σ_a^2/n, and the forecast error variance would be

$$V(\ell) = \sigma_a^2(\ell + \frac{\ell^2}{n}) \quad . \qquad (8)$$

The error in estimating θ_0 is not important untless ℓ is large, as $V(\ell)$ is dominated by $\sigma_a^2\ell$ for ℓ small relative to n. When θ_0 is estimated by the biweight estimate of location, as in our analysis, equation (7) still holds, with $\mathrm{Var}(\hat{\theta}_0)$ estimated by s_{bi}^2, the asymptotic variance estimate for the biweight estimate. We obtained this from a computer program by Kafadar (1982).

Forecast and Forecast Bounds Results

Figures G.1 - G.6 are graphical examples of the upper and lower bounds for three series where the two models differ (i.e. $\hat{\theta}_1 < 1$); SF25 (single females, ages 25-30), SM25 (single males, ages 25-30), and MFSP35 (married females, husband present, ages 35-44).

The difference in results may be explained in several ways. Different estimates of σ_a^2 were used in calculating the forecast bounds for the two models - s^2 for the (0 2 1) model, and s_{bi}^2 for the (0,1,0)-with-constant

model. If outliers were present in the series, s^2_{bi} should be less than s^2, contributing towards narrower bounds for the (0,1,0)-with-constant model. In addition, earlier we derived the ψ_j's for the (0,2,1) model. For the (0,1,0)-with-constant model, ψ_j is 1 for all j. If one compares the ψ_j's from both models, the (0,2,1) ψ_j's are all larger, contributing to wider bounds for the (0,2,1). Finally, the forecast error variance for the (0,1,0)-with-constant model takes into account error in estimating θ_0, while the (0,2,1) forecast error variance does not account for error in estimating θ_1. It is difficult to account for this error in estimating autoregressive or moving average parameters. This latter item would contribute towards wider bounds for the (0,1,0)-with-constant model. Unless $\hat{\theta}_1$ is very near 1 the larger ψ_j's for the (0,2,1) will have more of an effect than the $\text{Var}(\hat{\theta}_0)$ term for the (0,1,0)-with-constant. Thus, we would expect the (0,2,1) to produce the wider bounds.

In fact, in all three series considered here, the bounds for the (0,1,0)-with-constant model were narrower than those from the (0,2,1) model. This, however, does not mean that in general, the (0,1,0)-with-constant model is preferrable. Model selection should not be based on looking for narrow forecast bounds -- wider bounds may well be more correct!

Using these results, the validity of the constant series projections cannot be ruled out. They fall within the confidence bounds for all the series that we examined; however, in several series, the projections fall very near an upper or lower bound. The uncertainty of the forecast bounds themselves should be considered here; they are only <u>rough</u> estimates based on assuming knowledge of the correct model for the data. Thus, a slight change in a parameter estimate, or estimate of variance, even assuming the model is correct, could lead to the constant series falling outside the forecast

bounds, or moving further inside. We conclude by reminding the reader that we are forecasting 15 time points in the future, based only on 27 data points.

5.2 Comparison with Alternative Projections

It is of interest to compare the headship projections produced using various time series models to alternative projections produced from demographic and economic models. Census Bureau demographers compare alternative projections on the basis of total households rather than on headship rates. For the projections using time series and economic models, the headship rates are forecast and combined with population projections to obtain projections of the number of households. Census Bureau demographers formulate judgemental forecasts of household totals, implicitly incorporating their knowledge and opinions of the future course of such things as marriage and divorce rates, and their effects on household formation. As total households are used as part of the basis for comparison, it seems reasonable to directly consider the annual time series of total number of households. By fitting a time series model to this series, forecasts of total households can be produced directly, depending only on the past behavior of the total household data.

Data on total households was obtained from 1950 to 1985. The logged data was modeled and an (0,1,0)-with- constant model seemed to be reasonable. An outlier analysis indicated very large residuals for years 1980 and 1982. The outlier effects were estimated and the data and projections modified to account for these outliers. The projected total households for 1995 and 2000 from this procedure were compared to seven other projections of total households for 1995 and 2000. (See figures H.1 and H.2) Three of the projections, TCR high, TCR average, and TCR low, were obtained from an

economic model relating household formation to certain economic variables. (See Tella, Chandrasekar, and Reznek (1985).) Projections using this model require projections of the economic variables. The TCR high projections use economic projections from the Council of Economic Advisors (CEA), TCR low uses economic projections from the Housing Division of Census, and TCR average uses economic projections that are an average of the CEA and Housing Division economic projections. The projections denoted as AR010C correspond to the use of the (0,1,0)-constant model on all 130 series, and those denoted MA021 correspond to the use of the (0,2,1) model on 11 of the series and the (0,1,0) constant model on the rest. (See sections 4.2 and 4.4.) Two projections based on demographic assumptions are included, denoted as the demographic and constant projections. The constant projections are based on the assumption that the householder proportions will remain at their current 1985 level, with only changes in population structure affecting the proportions. This actually corresponds to use of a random walk without trend model, (0,1,0), for the proportions. The demographic projections are judgemental forecasts developed by Census Bureau demographers. They reflect the assumptions that recent rapid increases in the number of households formed will moderate due to an assumed leveling off of marriage and divorce rates, the completion of the passage of the baby boom cohorts past the ages where they are most likely to form their own households, and other demographic considerations.

The time series model projections based on headship rates come closest to the projections based on time series modeling of total households. As a whole, the demographic and economic model projections tend to be lower than the time series model projections, so these projections imply a deviation in the future from the past pattern of consistent increases in numbers of households. The time series direct projections of total households differ

from the others in that they do not directly account for effects of the present and future age structure of the population. In the other projections, the future age structure is taken into account explicitly in the time series and economic model projections (by projecting householder proportions and combining these with age-sex specific population projections) or implicitly in the case of the demographic judgemental projections. The impact of these age structure shifts may not be captured by the dotted projection line, thereby contributing to the projection differences.

Some differences in projections may be due to the different breakdowns of the headship rates. The time series and constant model methodologies were applied to a series of headship rates broken down by 130 age-sex-marital-householder status categories. The TCR economic models considered a breakdown of the data into only four age specific categories. The demographic judgemental projections did not explicitly involve a breakdown of the data. To investigate the effect of the different breakdowns on the headship projections, the (0,2,1) time series model was estimated for the four age categories used in the TCR models and resulting projections produced. (See Figure H.3 .) The projections of householders for the four age categories and the totals from this method (hereafter called MA021(4)) were compared to those from the TCR models, the constant projections, the MA021 model applied to the 130 series breakdown, and a middle level set of projections proposed by Census Bureau demographers. The latter three projections are included in Population Division's 1986 household projection publication. Census Bureau demographers felt the projections from the MA021 method may be too high and that assuming headship rates would remain constant was not a very likely scenario. A middle level series was created by using as projection slopes for each of the 130 series an average of the projection slopes for those series

from the MA021 and constant projections. The slope is 0 for the latter. Census demographers felt the middle level series produced results more consistent with their assumption that changes in marriage and divorce rates will slow considerably but not cease over the next 15 years. The choice of breakdown into 4 series or 130 series did have an effect. The projection level for total households for MA021(4) is lower than for MA021, in fact, it is about the same as the middle series projections. On the other hand, with regard to projections broken down by the four age specific categories, the MA021(4) projections follow the MA021 projections more closely than any of the TCR projections. It seems that both the type of model used and the type of data breakdown are important in explaining the differences in projections.

The demographic projections and the three TCR economic projections also differ from the time series model projections by incorporating outside information (other than population projections) that may relate to the formation of households. The TCR projections were formally obtained through a model relating household formation to economic variables. The demographic judgemental projections were informally obtained incorporating such outside information as expected behavior of future marriage and divorce rates, and their relation to and impact on household formations. A time series model cannot take into account any predicted change in the structure of the series beyond the last time point, or future information about related variables, unless such information can be quantified and incorporated into the model. The differences in projected total households are also due in part to the fact that outside information was not incorporated into the time series projections. If the expected behavior of the outside information and its relation to household formation are accurate, then the time series projections may prove to be too high.

6. FUTURE WORK

The work described in this paper serves as a foundation for future work. Many of the problems considered here -- what transformation to use, how to deal with the autocorrelation present in the data, treatment of outliers, dealing with short series -- must be dealt with by any approach to thehousehold projections. There are certain extensions to our work that were not pursued due to time limitations.

One obvious improvement would be to do something about outliers in the (0,2,1) model. Outliers can adversely affect both the estimate of the moving average parameter, and, if they are near the end of the series, the forecasts produced for any given value of the parameter. Unfortunately, procedures for treatment of outliers in this sort of model are only recently being developed. Martin, Samarov, and Vandaele (1983) and Bell (1983) have suggested approaches that could be tried. Computer software for the former was not available for this study. Computer software by Bell was available, but to use the procedure on all 130 individual series is a substantial investment in time and resources. Robust (outlier resistant) procedures were easy to use with the (0,1,0)-constant model, since fitting this model simply amounts to estimating the mean of the differenced data.

The primary difficulty faced in analyzing the headship rate series is the limited number of observations available on each series. Unless additional past data on each series could somehow be obtained, this suggests that to obtain more information one should consider pooling information across series. For example, if one assumed all the series followed the same model with the same parameter values, all the series could be used jointly to estimate the parameters. While such an extreme assumption is most likely inappropriate for this data, some grouping of series and use of shrinkage

estimation or Bayesian techniques might prove useful. Another approach to pooling information is to consider if a number of the series are essentially being driven by some underlying series. Perhaps there exists some aggregated series, whose movements account for much of the movement in the 130 series, which is some function (possibly linear) of the series under consideration. This can be investigated by analogues of standard multivariate analysis techniques that are suitable for time series data.

Another approach related to that mentioned above is to consider using other variables to explain the behavior of the household headship rates via some sort of model (such as a regression model). The advantages of such an approach for forecasting depend heavily on the extent to which the "other variables" are either known in advance of the headship rates (leading indicators), or can be forecast (more accurately than the headship rates directly). This approach has recently been pursued by Tella, Chandrasekar, and Reznek (1985), who review considerable additional literature on the subject. As noted above, many of the issues considered here will still be present, so that a combination of their approach with some of the ideas mentioned here may be most fruitful.

References

Bell, W. R. (1983), "A Computer Program for Detecting Outliers in Time Series," American Statistical Association 1983 Proceedings of the Business and Economic Statistics Section, 634-639.

Box, G. E. P., and Jenkins, G. M. (1970), Time Series Analysis, Forecasting and Control, San Francisco: Holden Day.

Bureau of the Census (1986), "Projections of the Number of Households and Families: 1986 to 2000," Current Population Reports, Series P-25, No. 986, Washington, D.C.: U.S. Government Printing Office.

Bureau of the Census (1979), "Projections of the Number of Households and Families: 1979 to 1995," Current Population Reports, Series P-25, No. 805, Washington, D.C.: U.S. Government Printing Office.

Kafadar, K. (1982), "A Biweight Approach to the One-Sample Problem," Journal of the American Statistical Association, 77, 416-424.

Liu, L., and Hudak, G. G. (1983), "The SCA System for Univariate-Multivariate Time Series an General Statistical Analysis," Scientific Computing Associates, Dekalb, Illinois.

Martin, R. D., Samarov A., and Vandaele, W. C. (1983), "Robust Methods for ARIMA Models," in Applied Time Series Analysis of Economic Data, Arnold Zellner (ed.), U.S. Department of Commerce, Bureau of the Census, 153-169.

McNeil, D. R. (1977), Interactive Data Analysis, New York: John Wiley and Sons, Inc.

Tella, A., Chandrasekar, K., and Reznek, A. (1985), "Developing an Economic Model of Household Rates," Census Bureau unpublished report.

APPENDICES

Table A.1 - Household Series Description

Age groups:	14-17, 18-19, 20-24, 25-29, 30-34, 35-44, 45-54, 55-64, 65-74, 75 and above		
Categories:	Single (never married)		Male Female
	Married, spouse present		Male Female
	Family Households		Married male, wife present with own household Male family householder, wife not present Female family householder, husband not present
	Nonfamily Householders		Males Females
	Secondary Individuals (Lodgers, employees, etc)		Males Females
	Group quarters (Rooming houses, convents, etc.)		Males Females

Note: The categories are not exclusive. For example, one may be included in the proportion of MARRIED MALE WIFE PRESENT WITH OWN HOUSEHOLD and MARRIED MALE WIFE PRESENT.

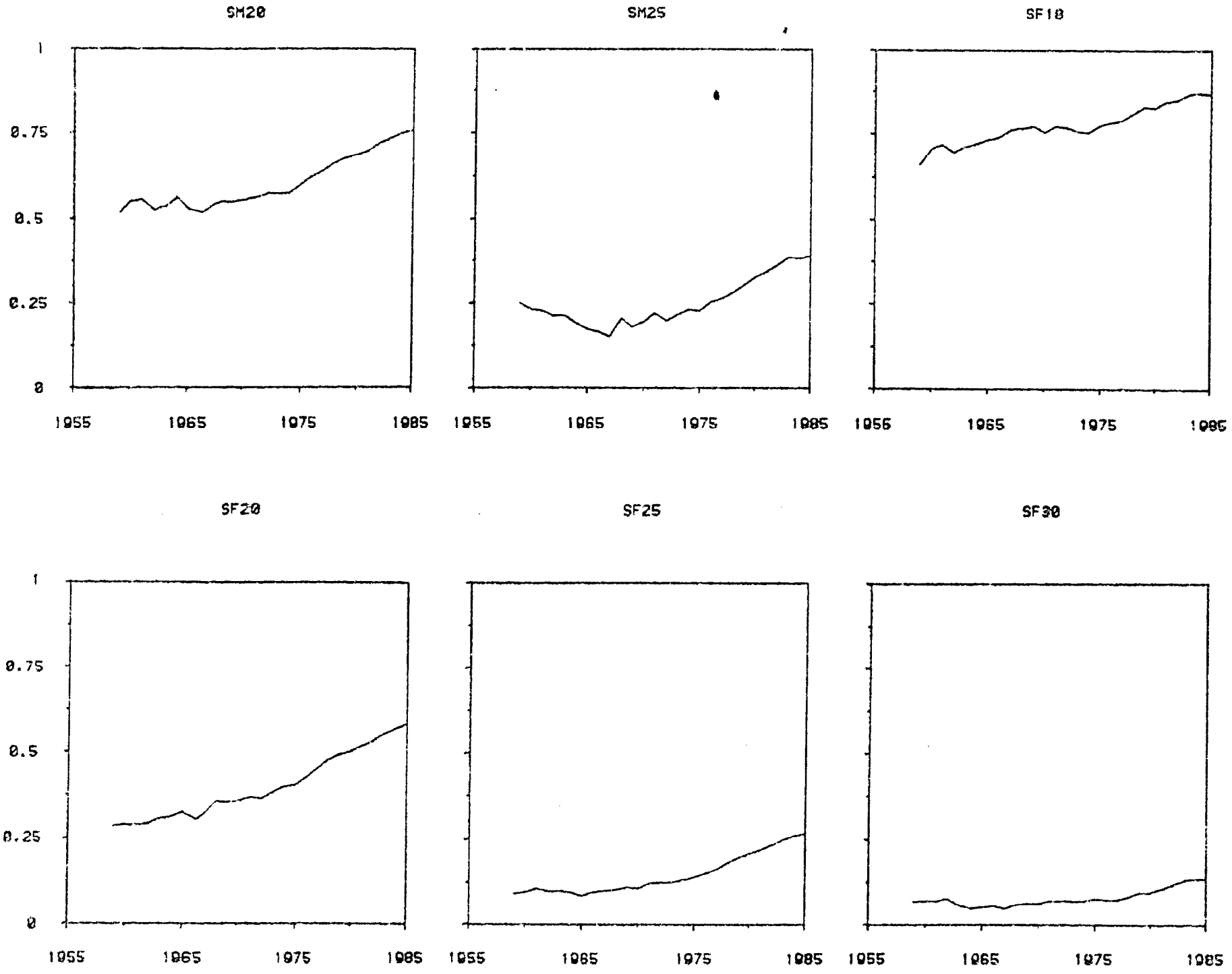
Data are proportions. The denominator depends on the respective category. For example, MARRIED MALE WIFE PRESENT is based on the number of men who have been married at any time

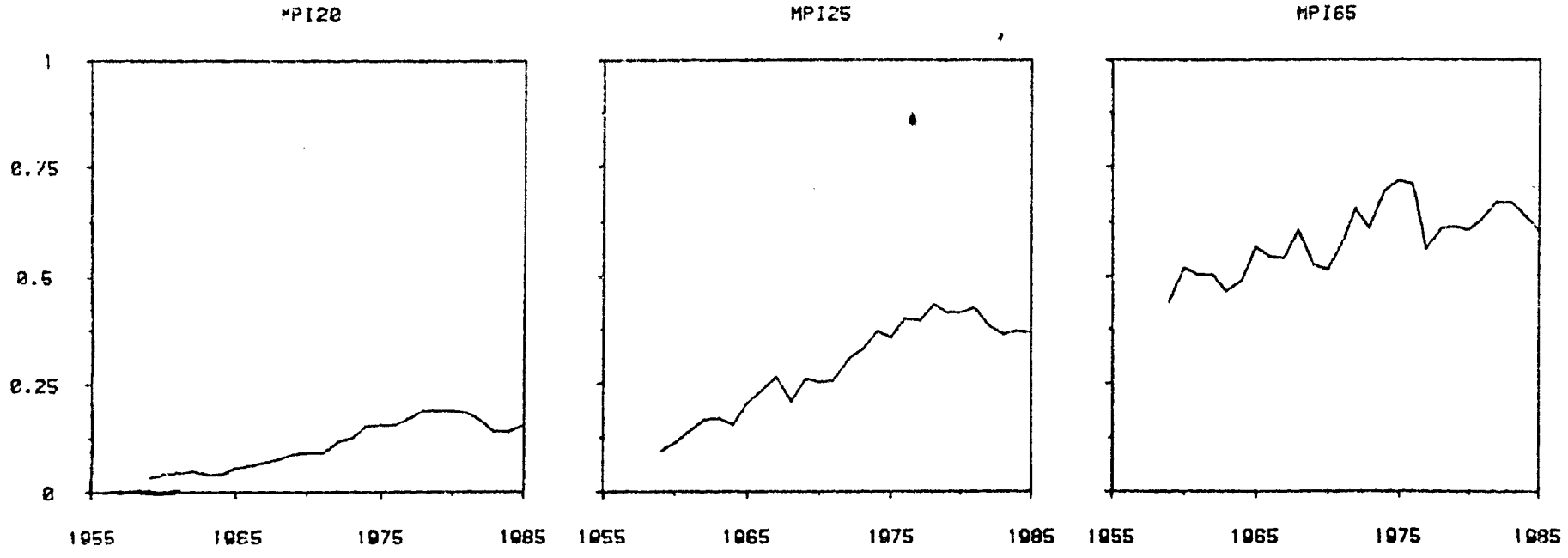
Consult Census Bureau publication 805, series P-25, "Projections of the Number of Households and Families: 1979 to 1995" for more information.

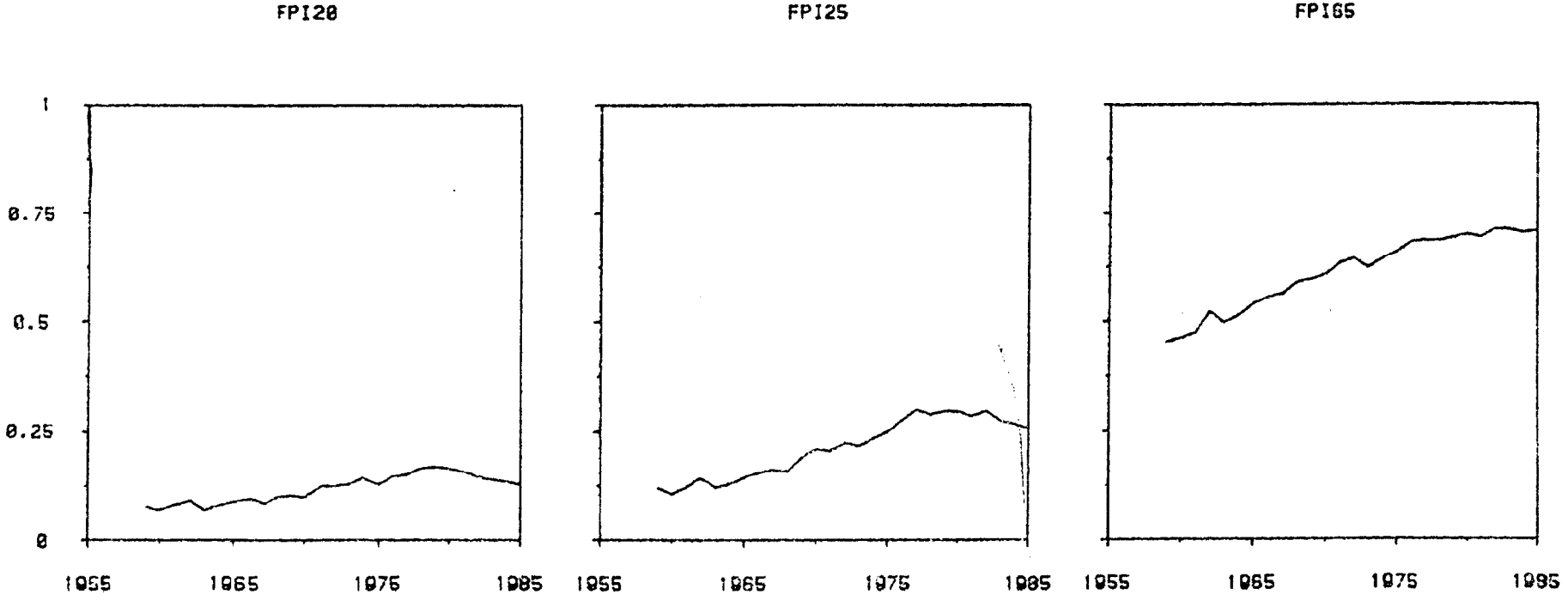
Table A.2 - Household Series Sample

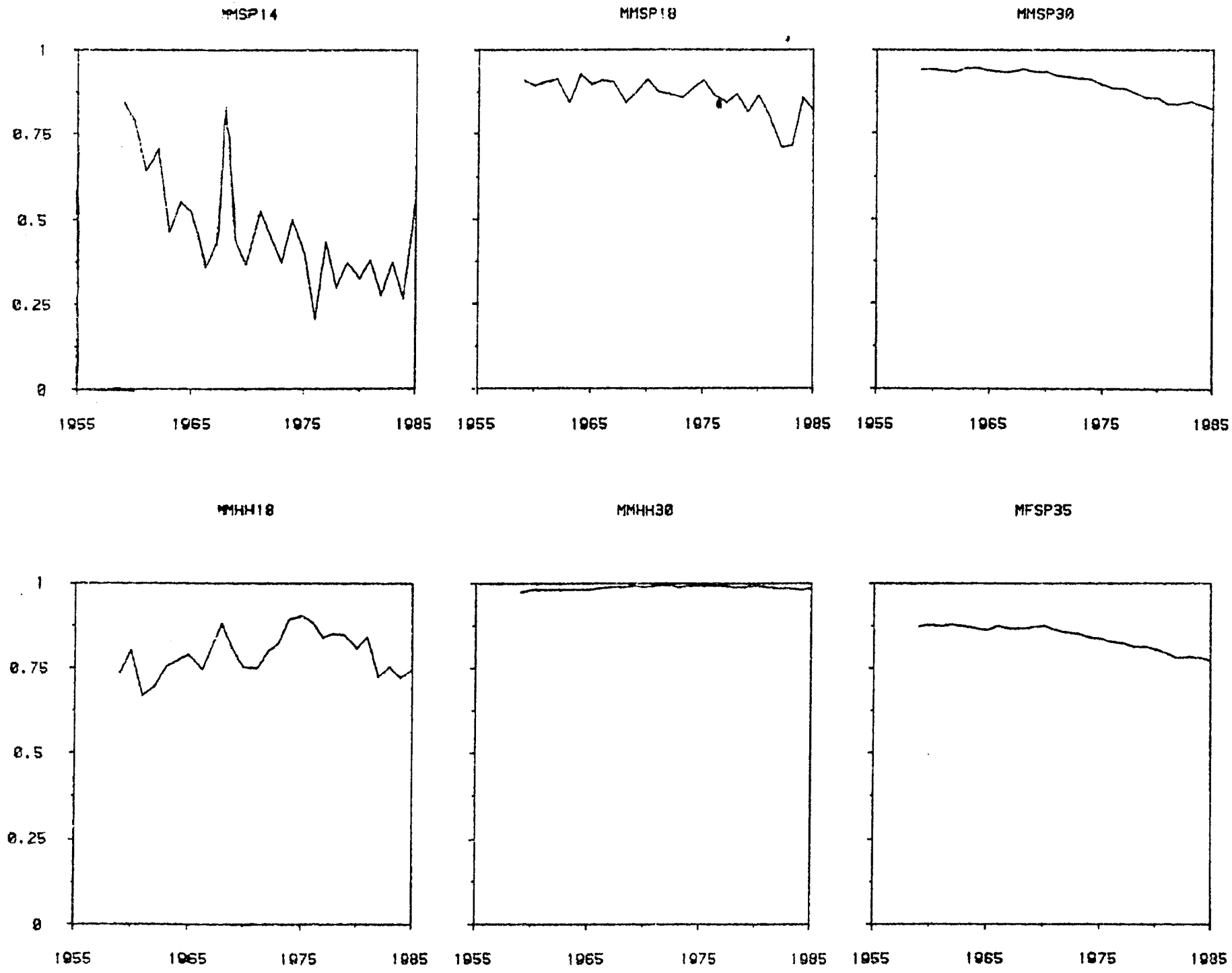
Series Name	Series Description
SM20	Single males, ages 20-24
SM25	Single males, ages 25-29
SF18	Single females, ages 18-19
SF20	Single females, ages 20-24
SF25	Single females, ages 25-29
SF30	Single females, ages 30-34
MMSP14	Married males, wife present, ages 14-17
MMSP18	Married males, wife present, ages 18-19
MMSP30	Married males, wife present, ages 30-34
MMHH18	Married males, wife present, with own household, ages 18-19
MMHH30	Married males, wife present, with own household, ages 30-34
MFSP35	Married females, husband present, ages 35-44
FFH25	Female family householder, husband not present, ages 25-29
FFH35	Female family householder, husband not present, ages 35-44
MPI20	Male nonfamily householder, ages 20-24
MPI25	Male nonfamily householder, ages 25-29
MPI30	Male nonfamily householder, ages 30-34
MPI65	Male nonfamily householder, ages 65-74
FPI20	Female nonfamily householder, ages 20-24
FPI25	Female nonfamily householder, ages 25-29
FPI65	Female nonfamily householder, ages 65-74

Figure A.1


SM20
SM25
SF10
SF20
SF25
SF30
1
0.75
0.5
0.25
0
1955
1965
1975
1985


MPI20
MPI25
MPI65
1
0.75
0.5
0.25
0
1955
1965
1975
1985


FPI20
FPI25
FPI65
1
0.75
0.5
0.25
0
1955
1965
1975
1985


MMSP14
MMSP18
MMSP30
MMHH18
MMHH30
MFSP35
1
0.75
0.5
0.25
0
1955
1965
1975
1985

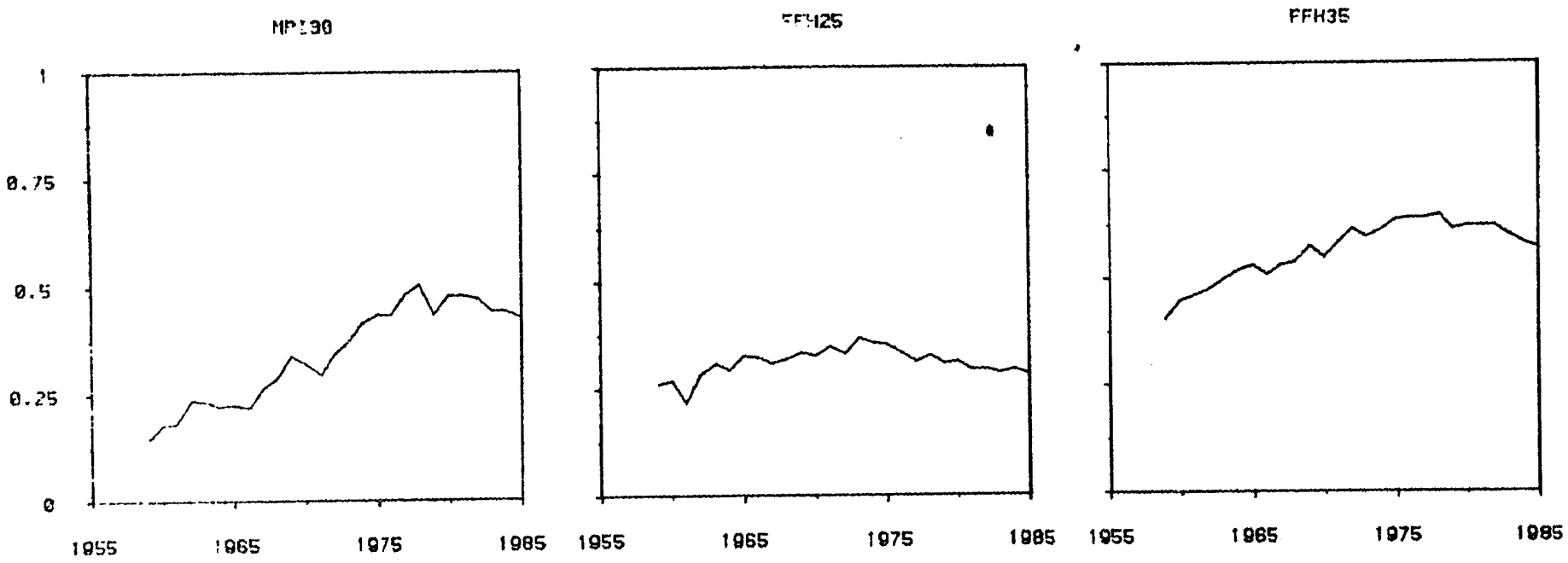

MP130
FFH25
FFH35
1
0.75
0.5
0.25
0
1955
1965
1975
1985
1955
1965
1975
1985
1955
1965
1975
1985

Table B.1

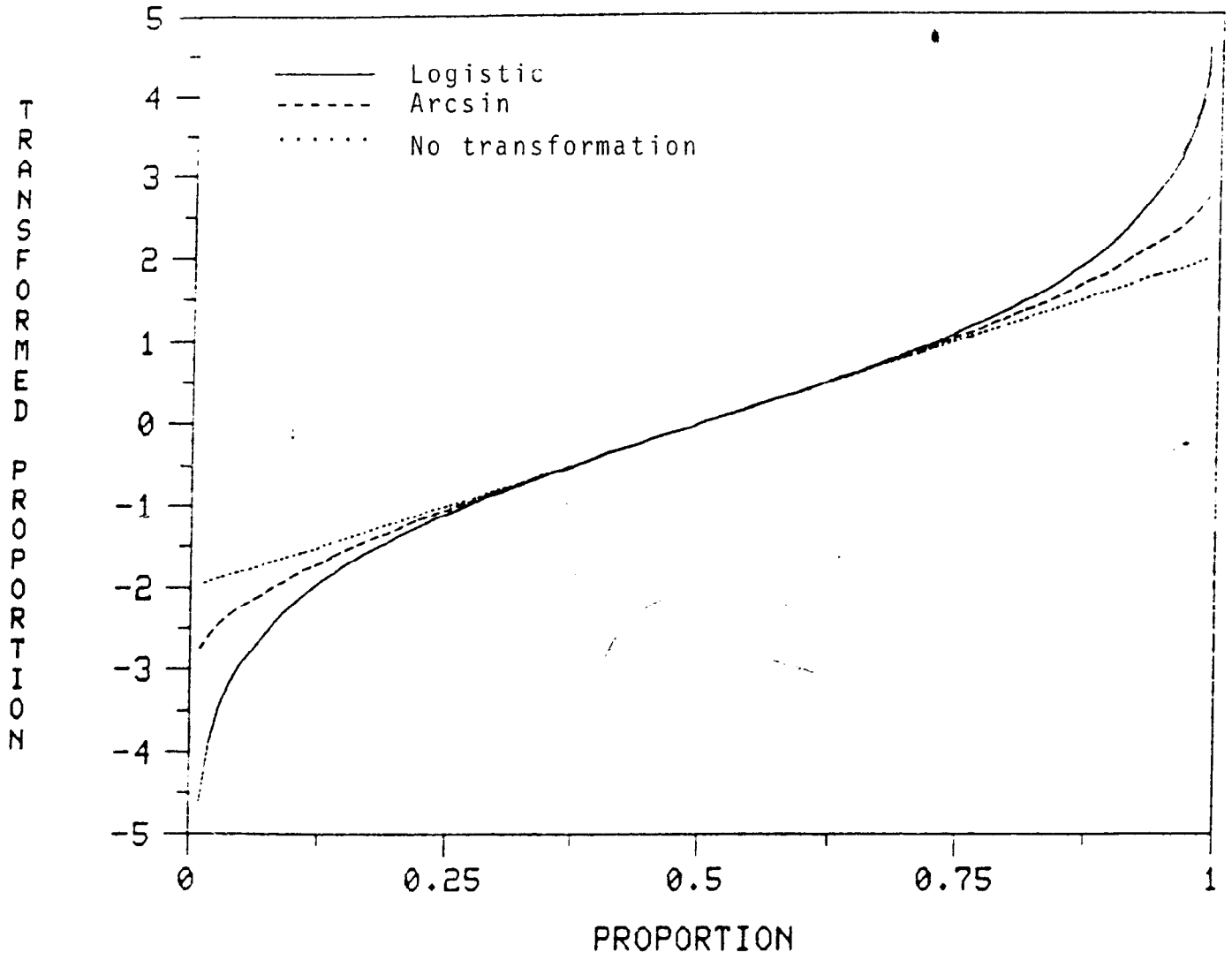
Series	Range	What Normal Probability Plots Show	$r_1(\nabla y_t)$
SM20	.1 - .75	Due to range of data, transformation does not matter. A couple outliers.	.12
SM25	.1 - .4	Original data seems to have a short logistic. One large outlier.	-.20
MMSP14	.2 - .9	Not much difference between transformations. One very large outlier in series produces two large outliers in differenced data. Series is very volatile and not that important.	-.45
MMSP18	.7 - .95	Original data might be best but series behaves very badly. Series is not that important.	-.34
MMSP30	.8 - .95	Difficult to tell - actually arcsin looks worst. There appears to be at least 3 positive outliers.	-.05
MMHH18	.6 - 1.0	Original data has a short upper tail that is well-corrected by the logistic.	-.14
MMHH30	.97 - 1.0	Original data may have a short upper tail which the logistic corrects.	-.33
MPI20	0 - .2	Original data may have a short lower tail which arcsin or logistic corrects.	.27
MPI25	.1 - .5	No apparent short lower tail in the original data. Transformation doesn't seem to matter. There are some negative outliers.	-.17
MPI30	.1 - .5	Same story as MPI25.	-.11
MPI65	.4 - .7	Transformation does not matter.	-.20
SF18	.6 - .9	Original series seems to have short upper tail that logistic seems to correct. Complicated by one large positive outlier.	-.14
SF20	.25 - .55	Transformation has little effect. One large negative outlier.	-.10

Table B.1 (continued)

Series	Range	What Normal Probability Plots Show	$r_1(\nabla y_t)$
SF25	.1 - .3	Original data may have a short upper and lower tail. Arcsin or logistic seems to help some, though logistic may produce a longish lower tail.	-.03
SF30	.05 - .15	Hard to distinguish characteristics of lower tail from presence of negative outliers.	-.05
MFSP35	.75 - .9	Original may have a short upper and lower tail. Logistic corrects short upper tail, not the short lower tail.	-.02
FFH25	.2 - .4	Original data seems to have a short lower tail corrected by logistic. One large outlier in series produces one large positive and one large negative outlier in differenced data.	-.38
FFH35	.4 - .7	Transformation has no effect. No apparent outliers - good plots.	-.08
FPI20	.05 - .2	Original data look best. Transformed data looks OK but one slight negative outlier in original is elongated by logistic.	-.37
FPI25	.1 - .3	Original seems to have short upper and lower tail corrected by logistic.	-.15
FPI65	.4 - .7	Transformation has no effect. Two negative and one positive outlier.	-.29

Figure B.1

LOGISTIC AND ARCSIN TRANSFORMATIONS



Note: The arcsin and no transformation are linearly rescaled to have the same value and slope at $p_t = .5$ as the logistic transformation.

Table C.1

Parameter Estimates Looked at in Comparing Robust versus Least Squares Regression

	C = 2 Robust		C = 4 Robust		C = 6 Robust		Least Squares	
Series	a_0	a_1	a_0	a_1	a_0	a_1	a_0	a_1
SM20	0.04	0.019	-0.12	.038	-0.11	.038	-0.10	.037
SM25	-2.17	.064	-1.76	.043	-1.64	.036	-1.61	.034
MMSP14	0.35	-.046	0.60	-.060	0.79	-.070	0.90	-.076
MMSP18	2.36	-.026	2.38	-.031	2.42	-.036	2.43	-.038
MMSP30	3.24	-.064	3.13	-.058	3.08	-.055	3.07	-.054
MMHH18	0.96	.033	1.11	.020	1.15	.018	1.18	.017
MMHH30	3.72	.052	3.82	.040	3.94	.026	3.99	.023
MPI20	-3.39	.094	-3.38	.090	-3.35	.086	-3.27	.078
MPI25	-1.99	.082	-2.02	.082	-1.95	.073	-1.94	.071
MPI30	-1.66	.071	-1.66	.072	-1.62	.068	-1.61	.069
MPI65	-0.14	.029	-0.11	.030	-0.11	.032	-0.10	.033
SF18	0.70	.043	0.70	.041	0.70	.041	0.70	.041
SF20	-1.32	.060	-1.10	.049	-1.10	.049	-1.10	.049
SF25	-2.88	.069	-2.62	.055	-2.57	.052	-2.56	.051
SF30	-3.34	.053	-3.28	.049	-2.97	.032	-2.95	.031
MFSP35	2.10	-.032	2.21	-.032	2.12	-.032	2.12	-.031
FFH25	-0.68	-.006	-0.81	.002	-0.86	.005	-0.89	.007
FFH35	-0.29	.047	-0.21	.037	-0.18	.033	-0.17	.032
FPI20	-2.66	.049	-2.62	.044	-2.59	.040	-2.58	.039
FPI25	-2.27	.069	-2.17	.061	-2.13	.056	-2.12	.054
FPI65	-0.22	.052	-0.18	.047	-0.16	.046	-0.15	.045

Figure C.1

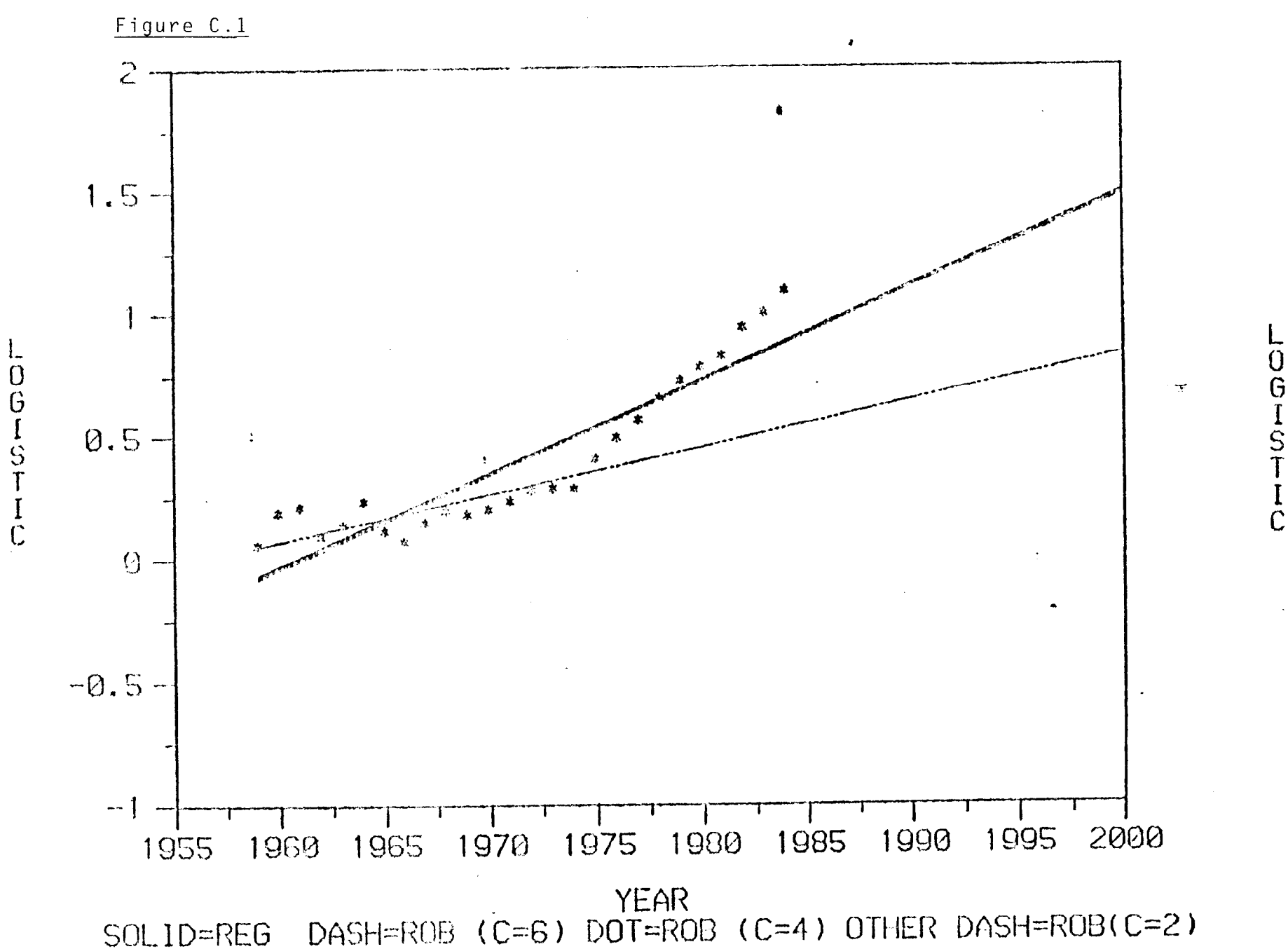

SERIES SM20
2
1.5
1
0.5
0
-0.5
-1
LOGISTIC
LOGISTIC
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
YEAR
SOLID=REG DASH=ROB (C=6) DOT=ROB (C=4) OTHER DASH=ROB(C=2)

Figure C.2

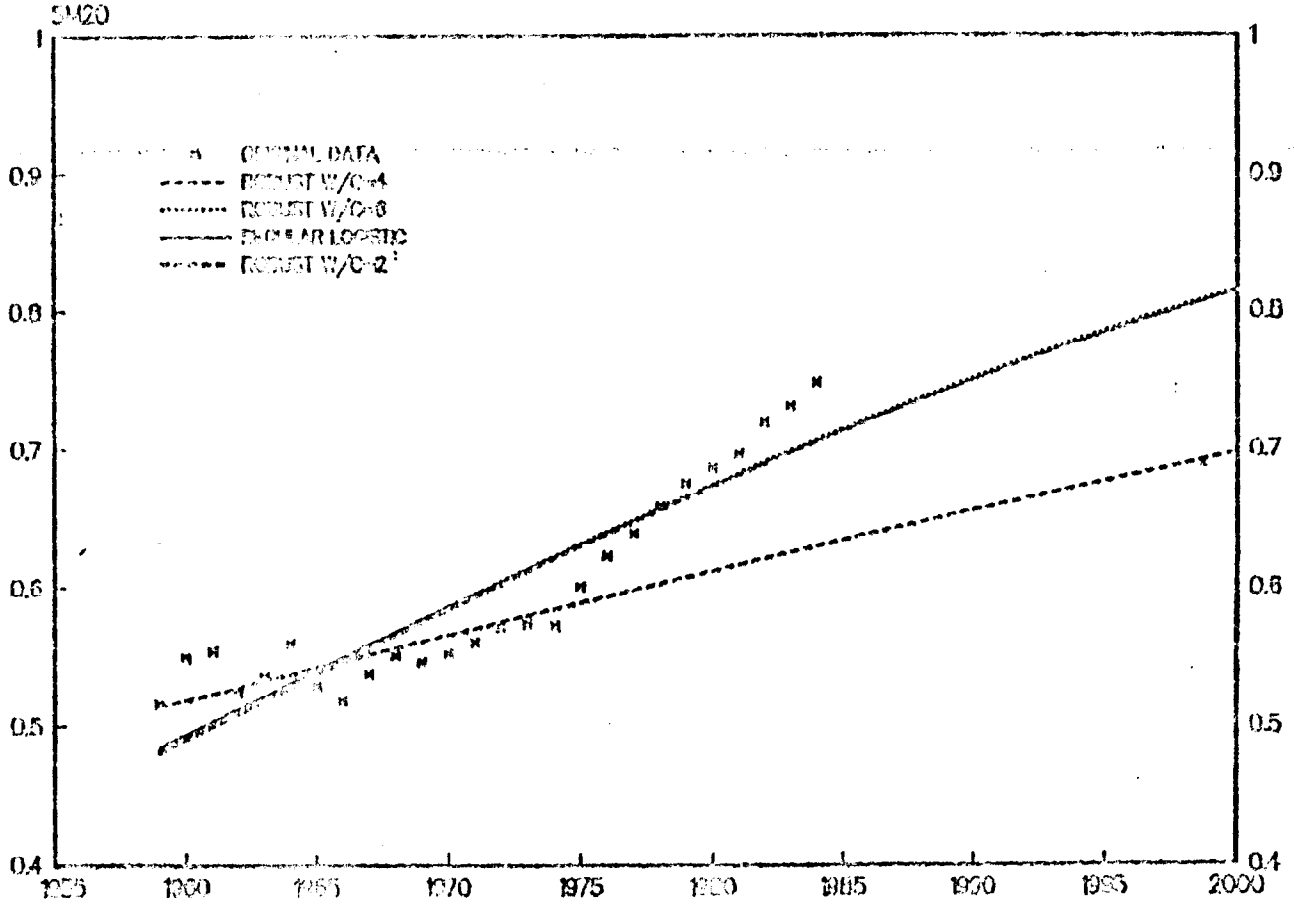


Figure D.1

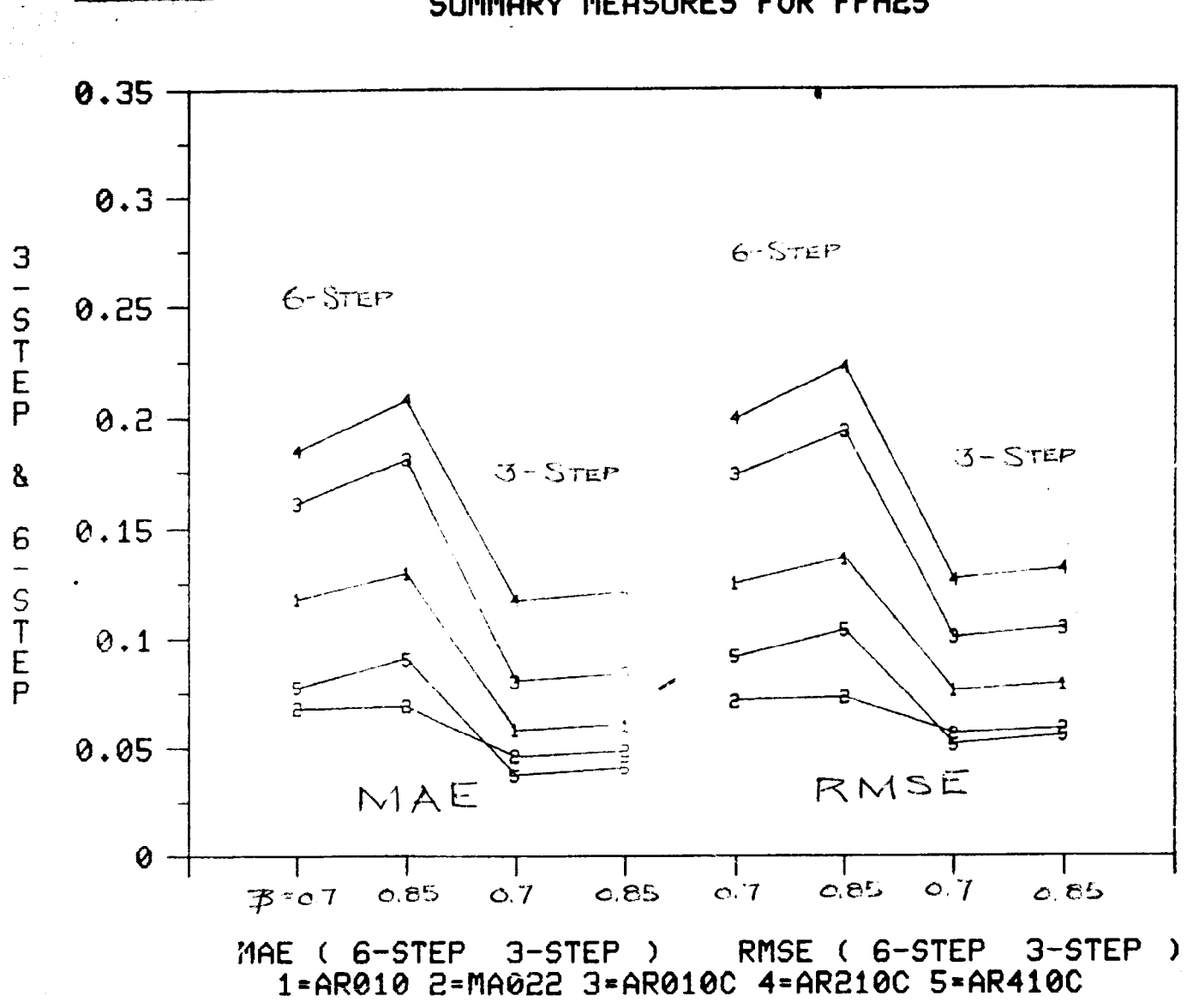

SUMMARY MEASURES FOR FFH25
3-STEP & 6-STEP
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
6-STEP
3-STEP
6-STEP
3-STEP
MAE
RMSE
β=0.7
0.85
0.7
0.85
0.7
0.85
0.7
0.85
MAE (6-STEP 3-STEP) RMSE (6-STEP 3-STEP)
1=AR010 2=MA022 3=AR010C 4=AR210C 5=AR410C

Figure D.2

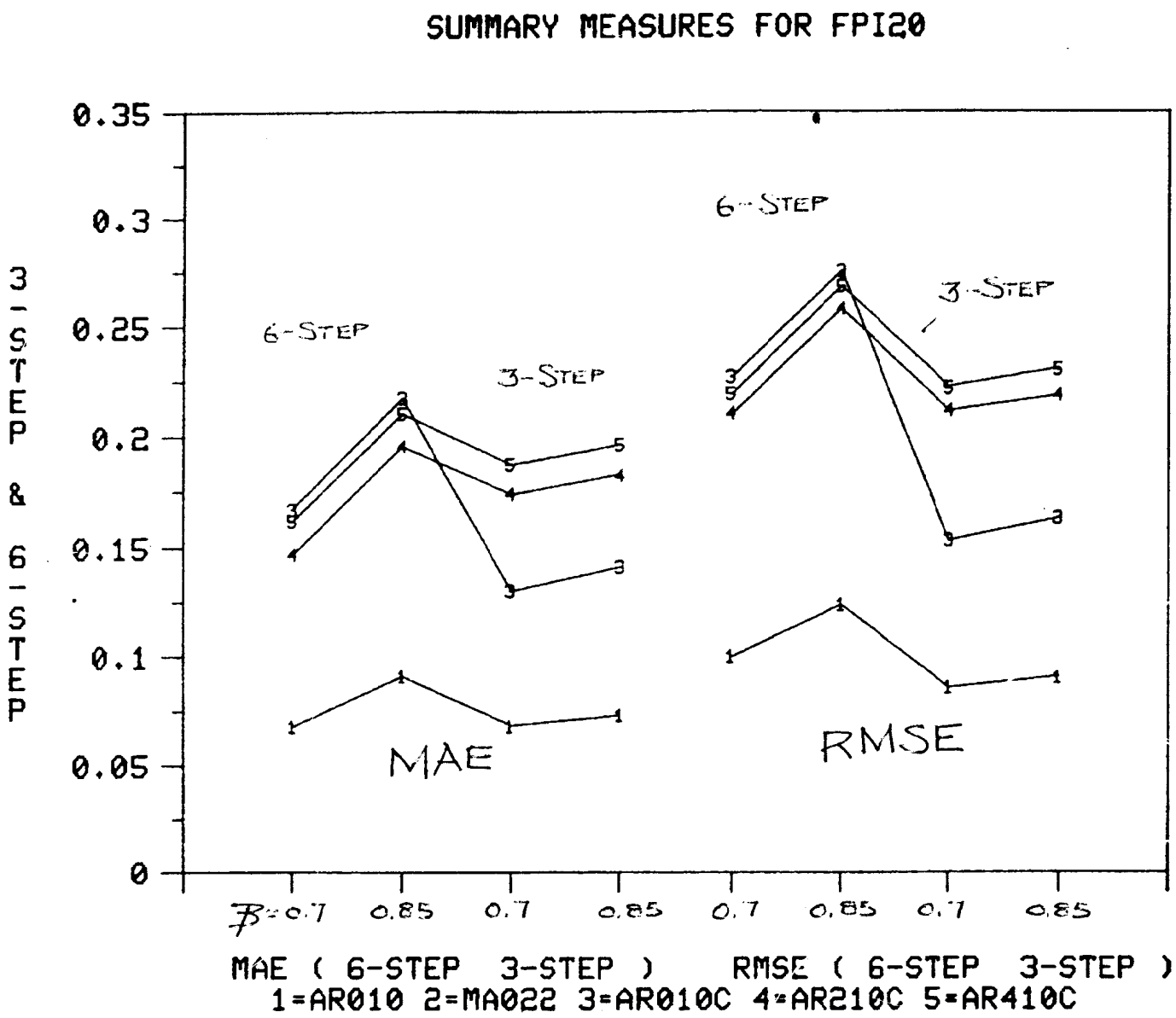

SUMMARY MEASURES FOR FPI20
3-STEP & 6-STEP
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
6-STEP
3-STEP
6-STEP
3-STEP
MAE
RMSE
B=0.7
0.85
0.7
0.85
0.7
0.85
0.7
0.85
MAE (6-STEP 3-STEP) RMSE (6-STEP 3-STEP)
1=AR010 2=MA022 3=AR010C 4=AR210C 5=AR410C

FIGURE D.3

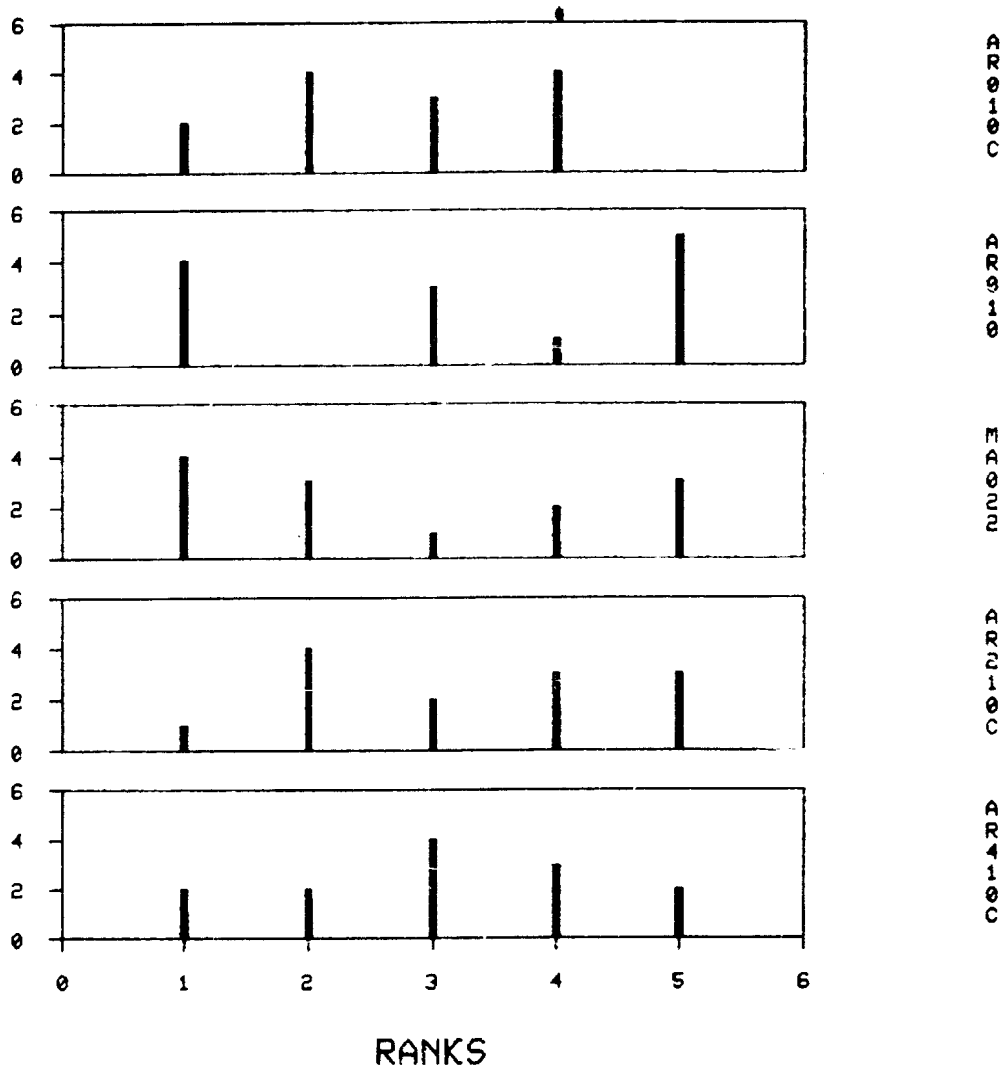


Figure D.4

Ranks

Model	1	2	3	4	5
ARCO10C	2	4	3	4	0
ARO10	4	0	3	1	5
MAO22	4	3	1	2	3
AR210C	1	4	2	3	3
AR410C	2	2	4	3	2

Note: Above ranks are for the 13 series that included the (0,2,2) model

Ranks

Model	1	2	3	4
ARO10	5	1	0	2
ARO10C	2	4	2	0
AR210C	0	1	5	2
AR410C	1	2	1	4

Note: Above ranks are for the 8 series for which we could not estimate the (0,2,2) model

Figure E.1

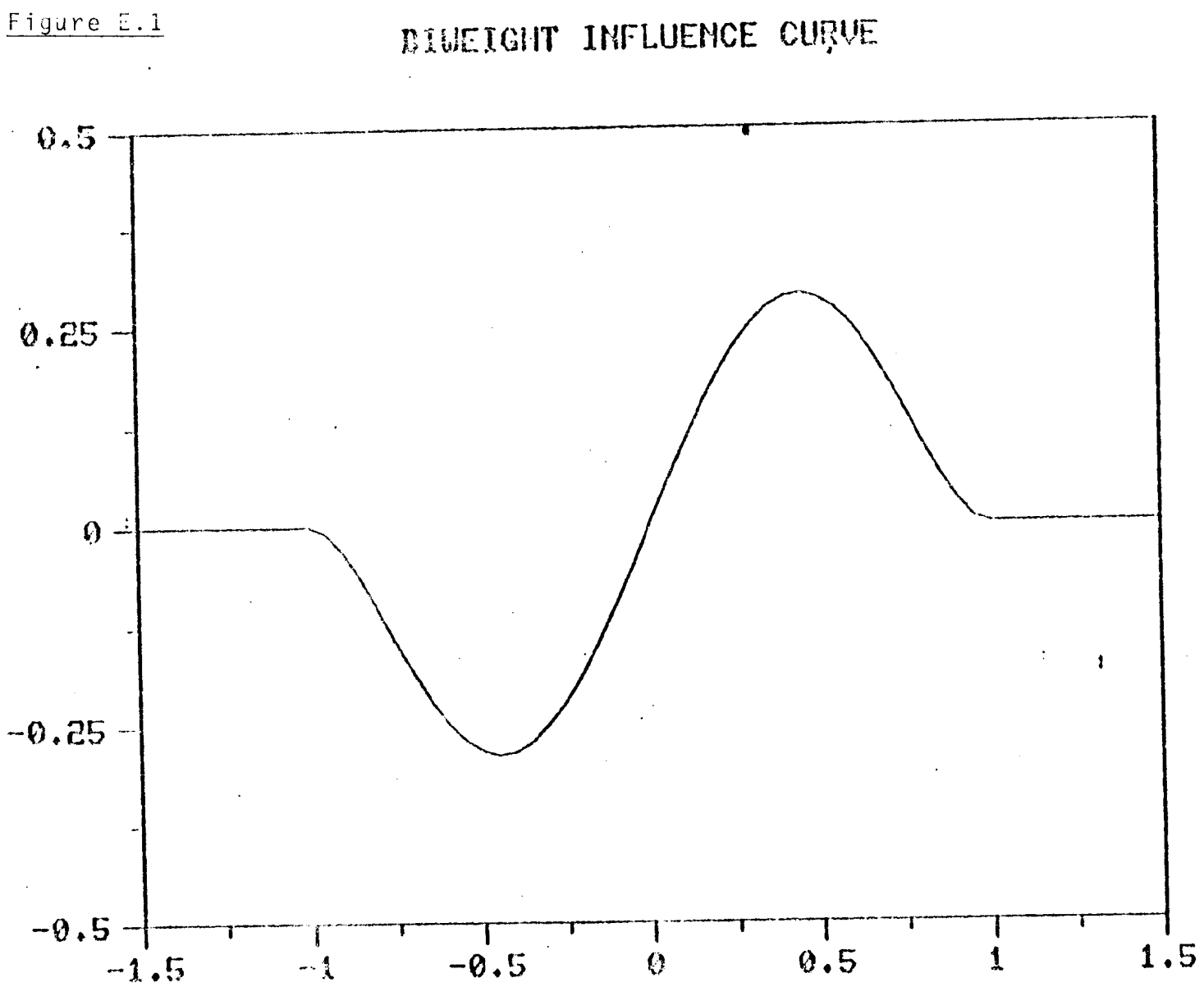


Figure E.2

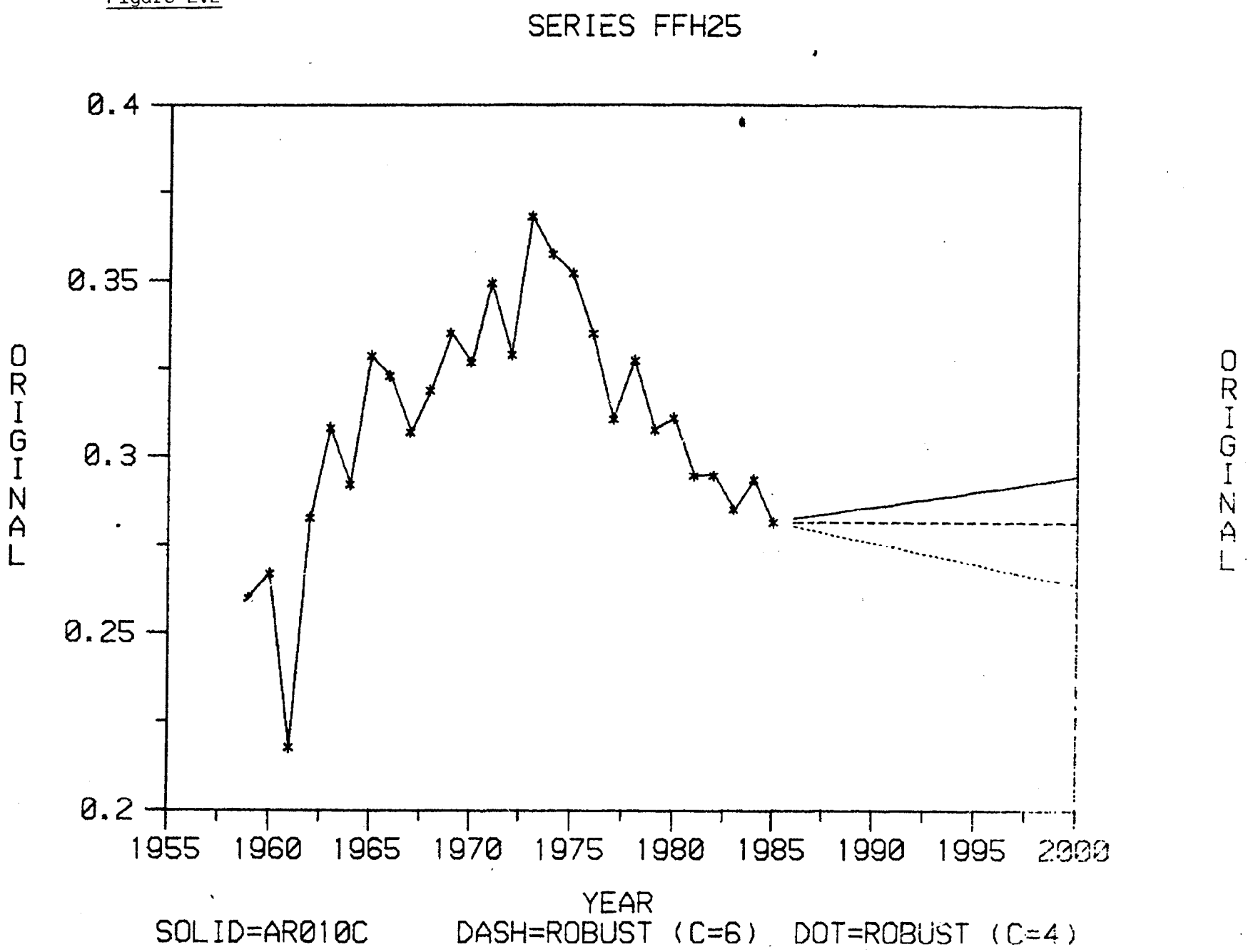

SERIES FFH25
0.4
0.35
0.3
0.25
0.2
ORIGINAL
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
YEAR
SOLID=AR010C
DASH=ROBUST (C=6)
DOT=ROBUST (C=4)

Figure F.1

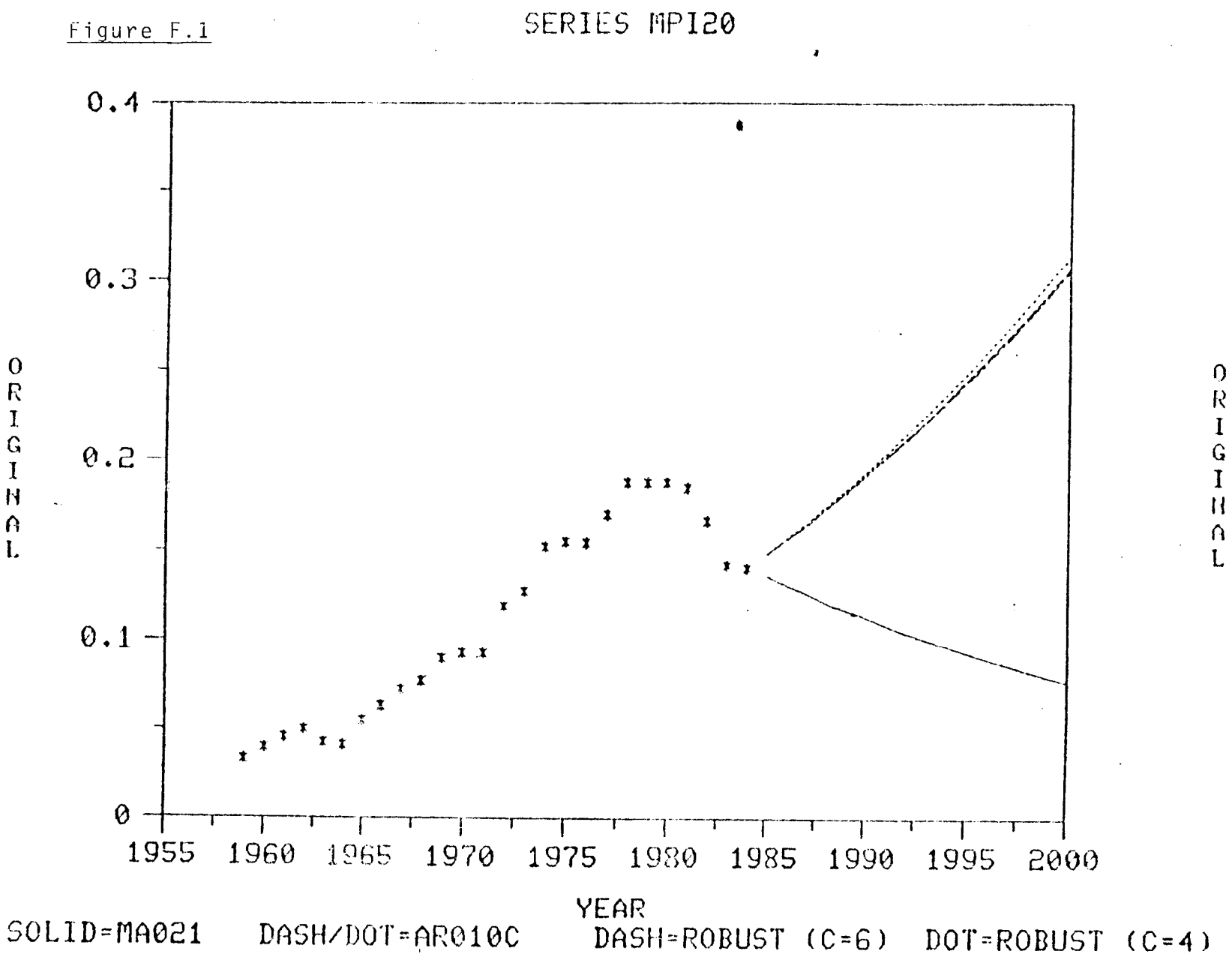

SERIES MPI20
0.4
0.3
0.2
0.1
0
ORIGINAL
ORIGINAL
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
YEAR
SOLID=MA021
DASH/DOT=AR010C
DASH=ROBUST (C=6)
DOT=ROBUST (C=4)

Figure F.2

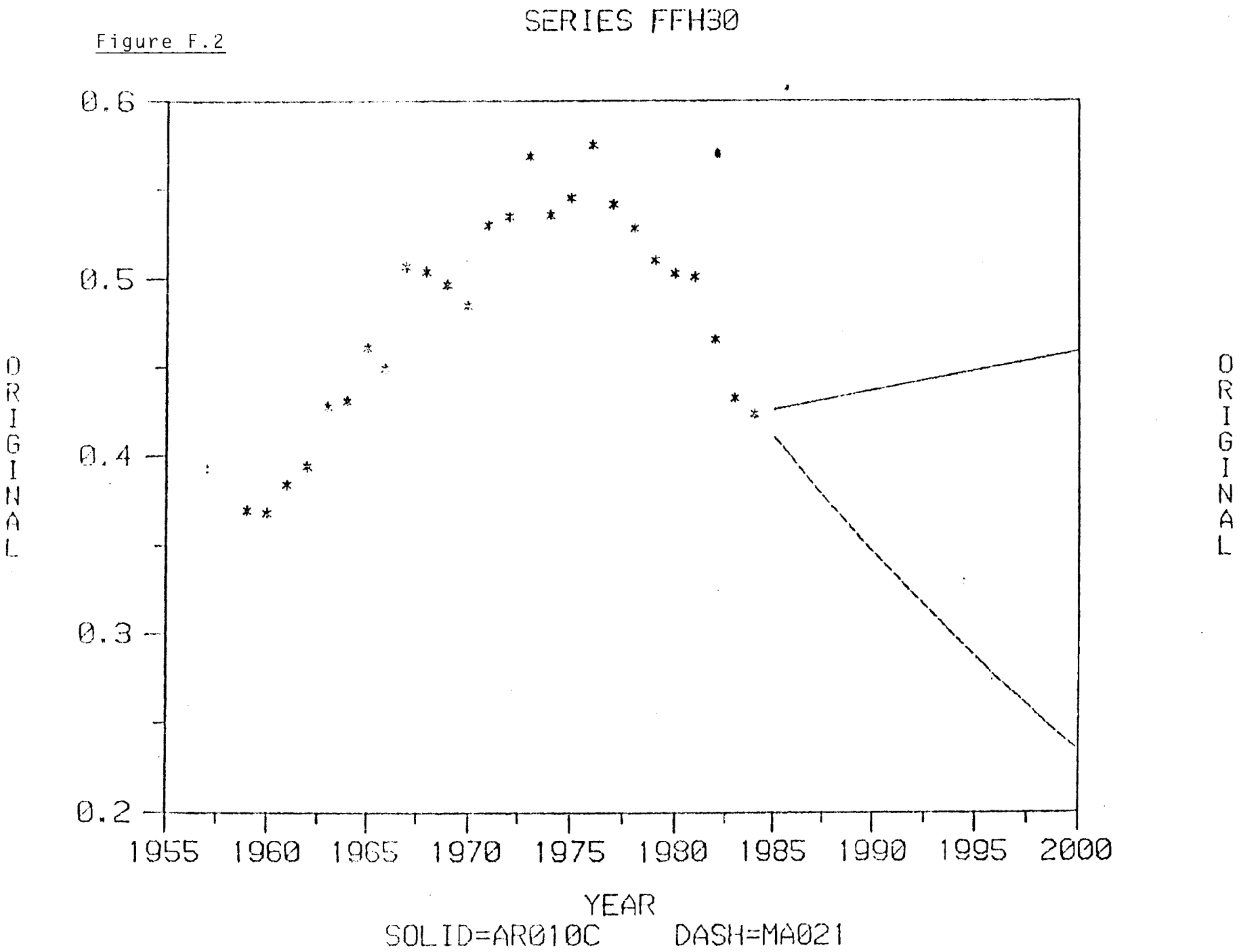

SERIES FFH30
0.6
0.5
0.4
0.3
0.2
ORIGINAL
ORIGINAL
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
YEAR
SOLID=AR010C
DASH=MA021

Figure F.3

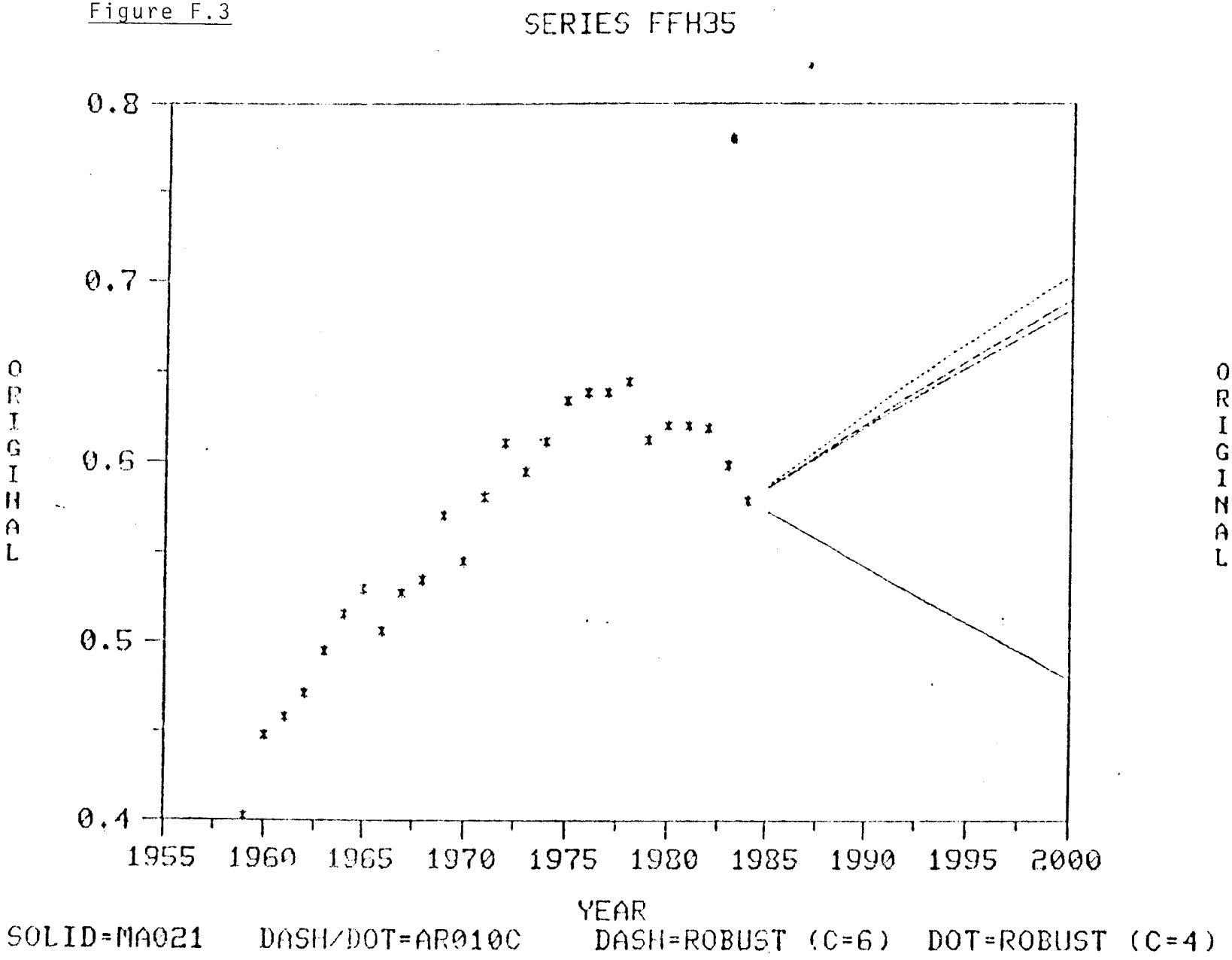

SERIES FFH35
0.8
0.7
0.6
0.5
0.4
ORIGINAL
ORIGINAL
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
YEAR
SOLID=MA021
DASH/DOT=AR010C
DASH=ROBUST (C=6)
DOT=ROBUST (C=4)

Figure F.4

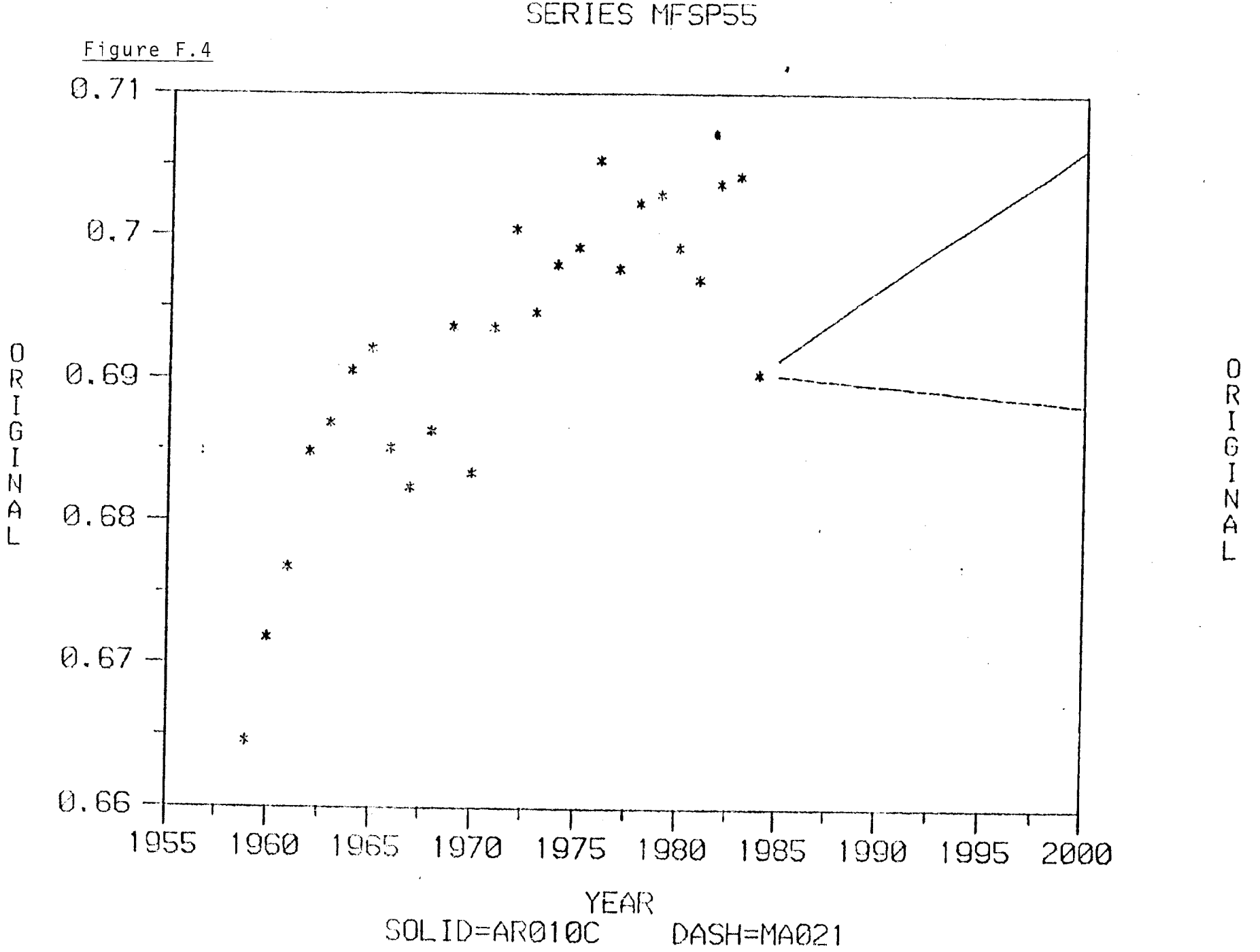

SERIES MFSP55
0.71
0.7
0.69
0.68
0.67
0.66
ORIGINAL
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
YEAR
SOLID=AR010C
DASH=MA021

Figure G.1

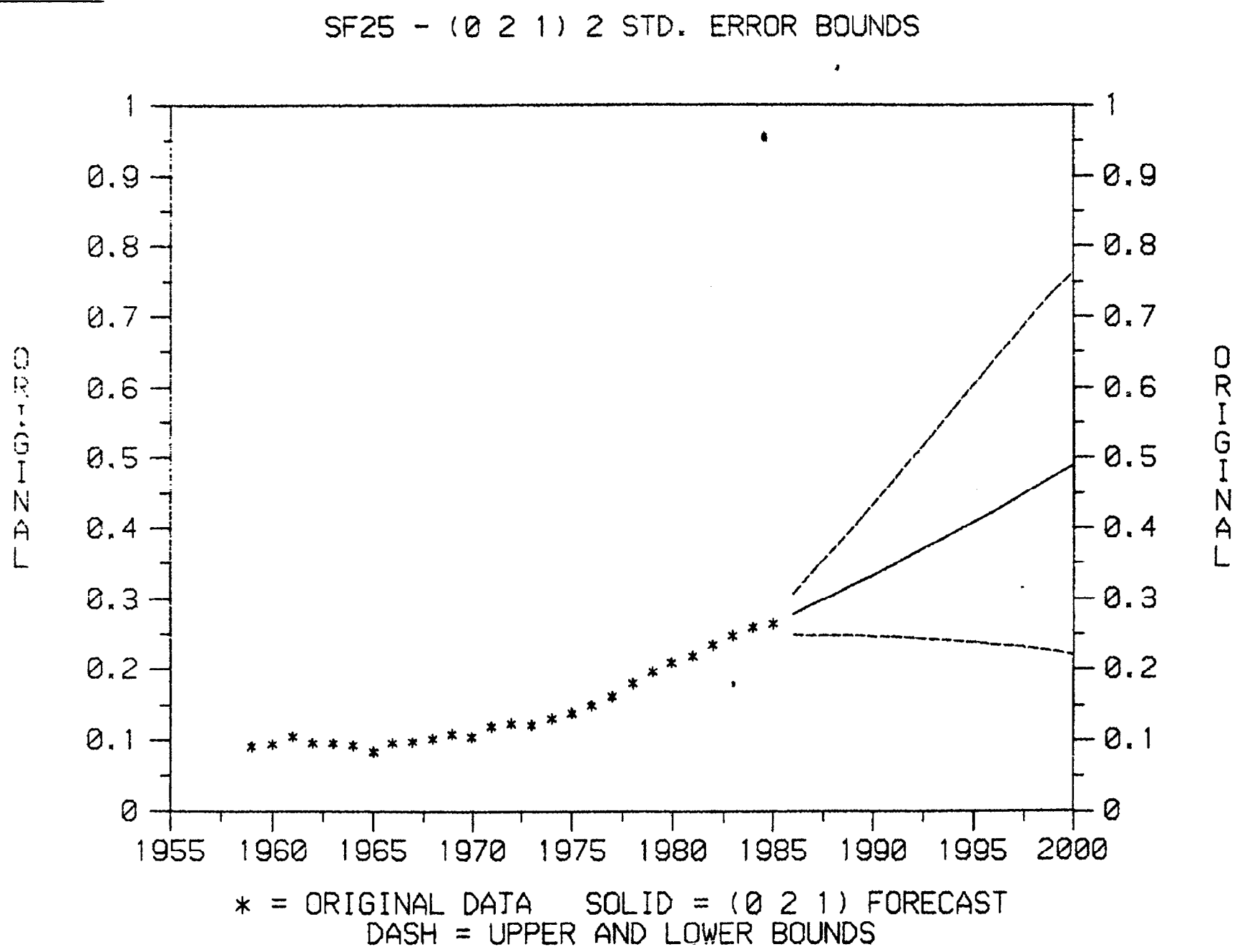

SF25 - (0 2 1) 2 STD. ERROR BOUNDS
ORIGINAL
ORIGINAL
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
* = ORIGINAL DATA SOLID = (0 2 1) FORECAST
DASH = UPPER AND LOWER BOUNDS

Figure G.2

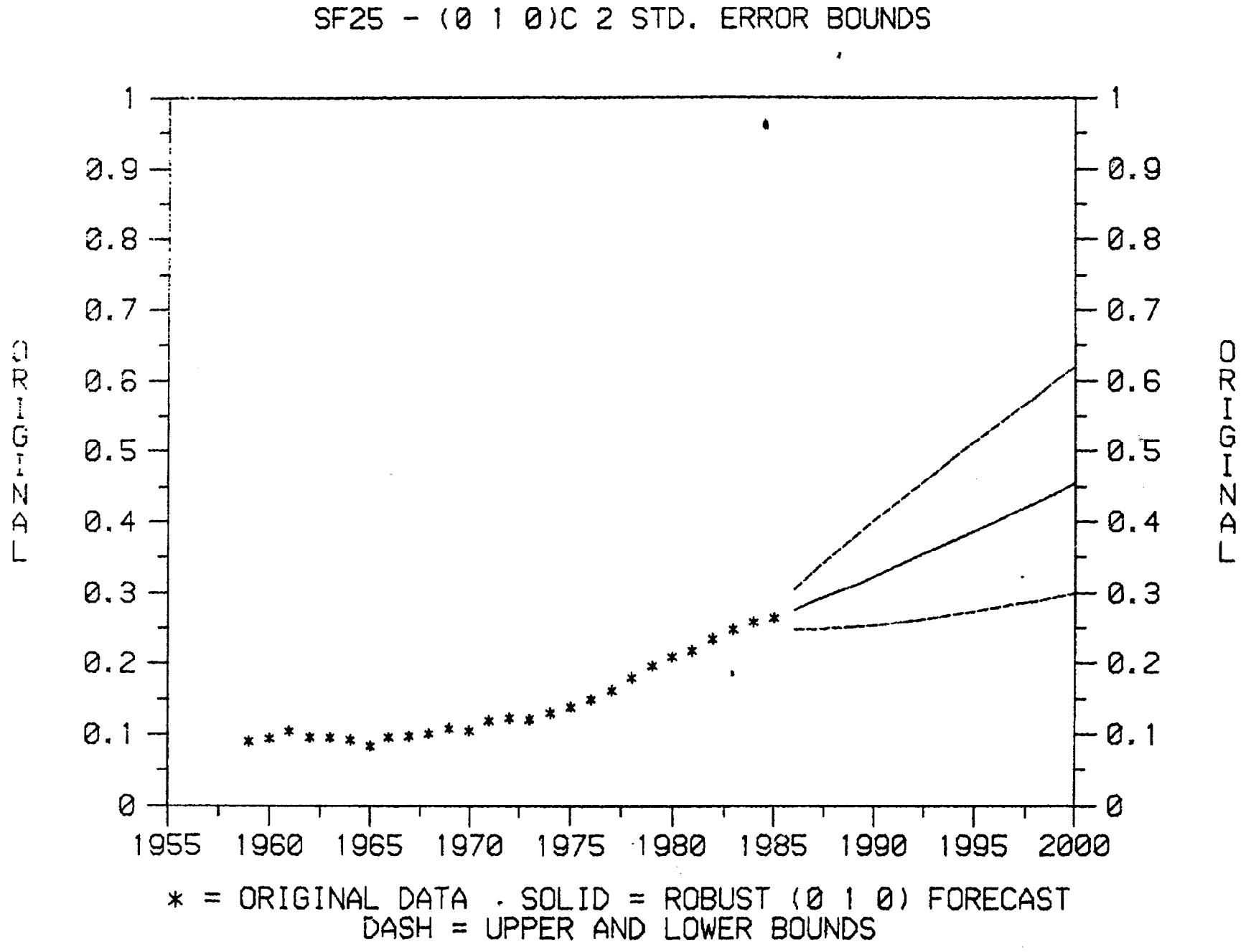

SF25 - (0 1 0)C 2 STD. ERROR BOUNDS
ORIGINAL
ORIGINAL
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
* = ORIGINAL DATA . SOLID = ROBUST (0 1 0) FORECAST
DASH = UPPER AND LOWER BOUNDS

Figure G.3

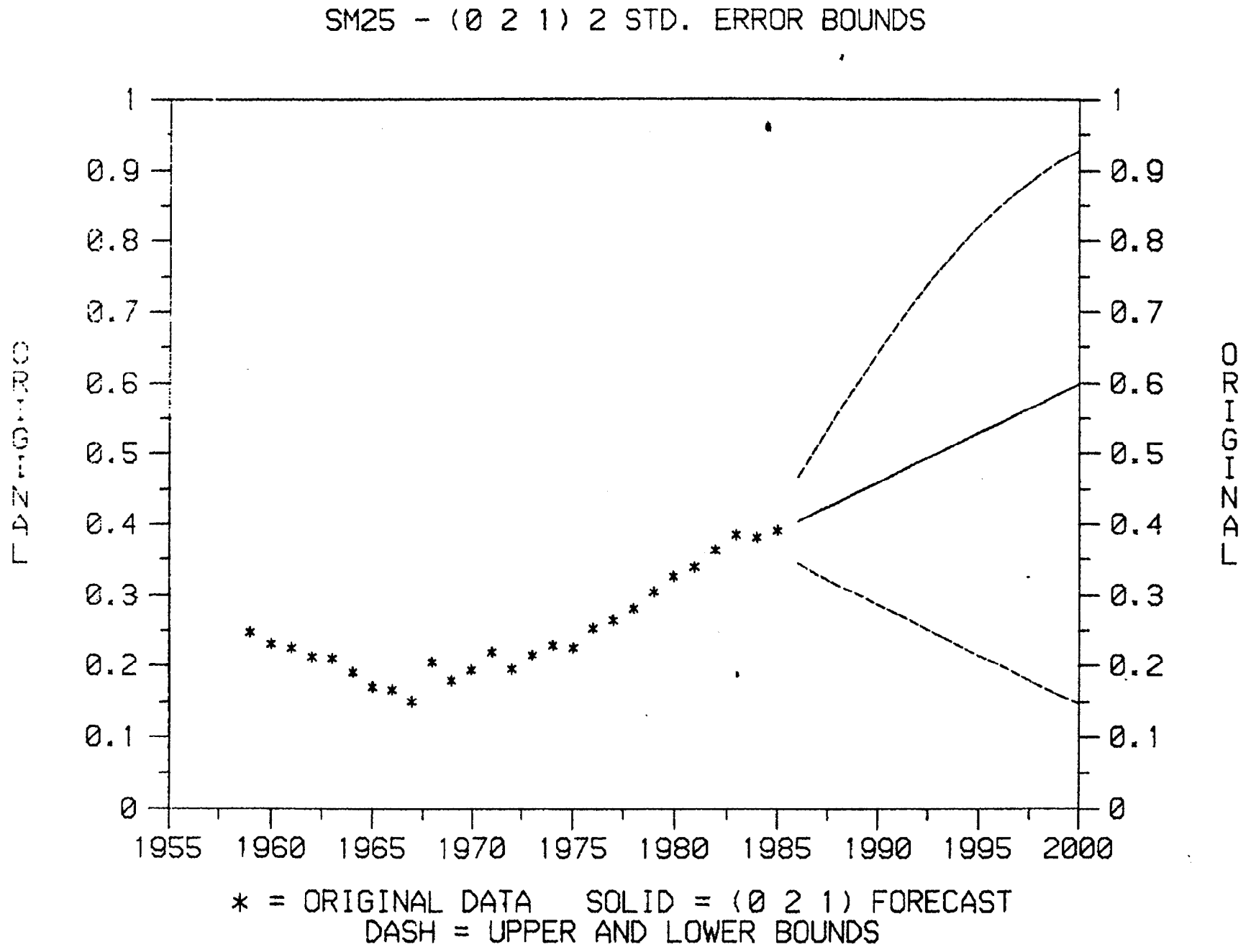

SM25 - (0 2 1) 2 STD. ERROR BOUNDS
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
ORIGINAL
* = ORIGINAL DATA SOLID = (0 2 1) FORECAST
DASH = UPPER AND LOWER BOUNDS

Figure G.4

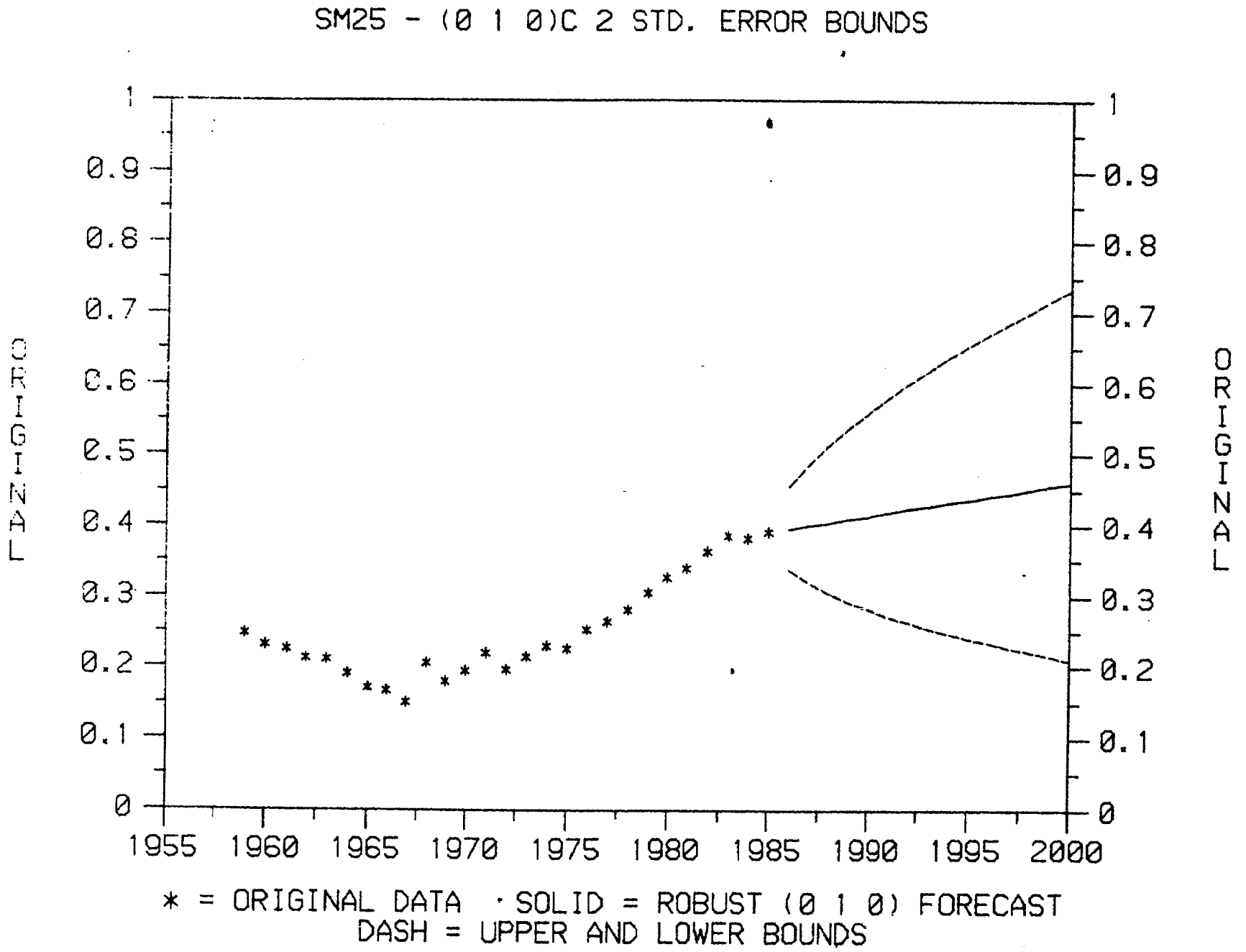

SM25 - (0 1 0)C 2 STD. ERROR BOUNDS
ORIGINAL
ORIGINAL
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
* = ORIGINAL DATA · SOLID = ROBUST (0 1 0) FORECAST
DASH = UPPER AND LOWER BOUNDS

Figure G.5

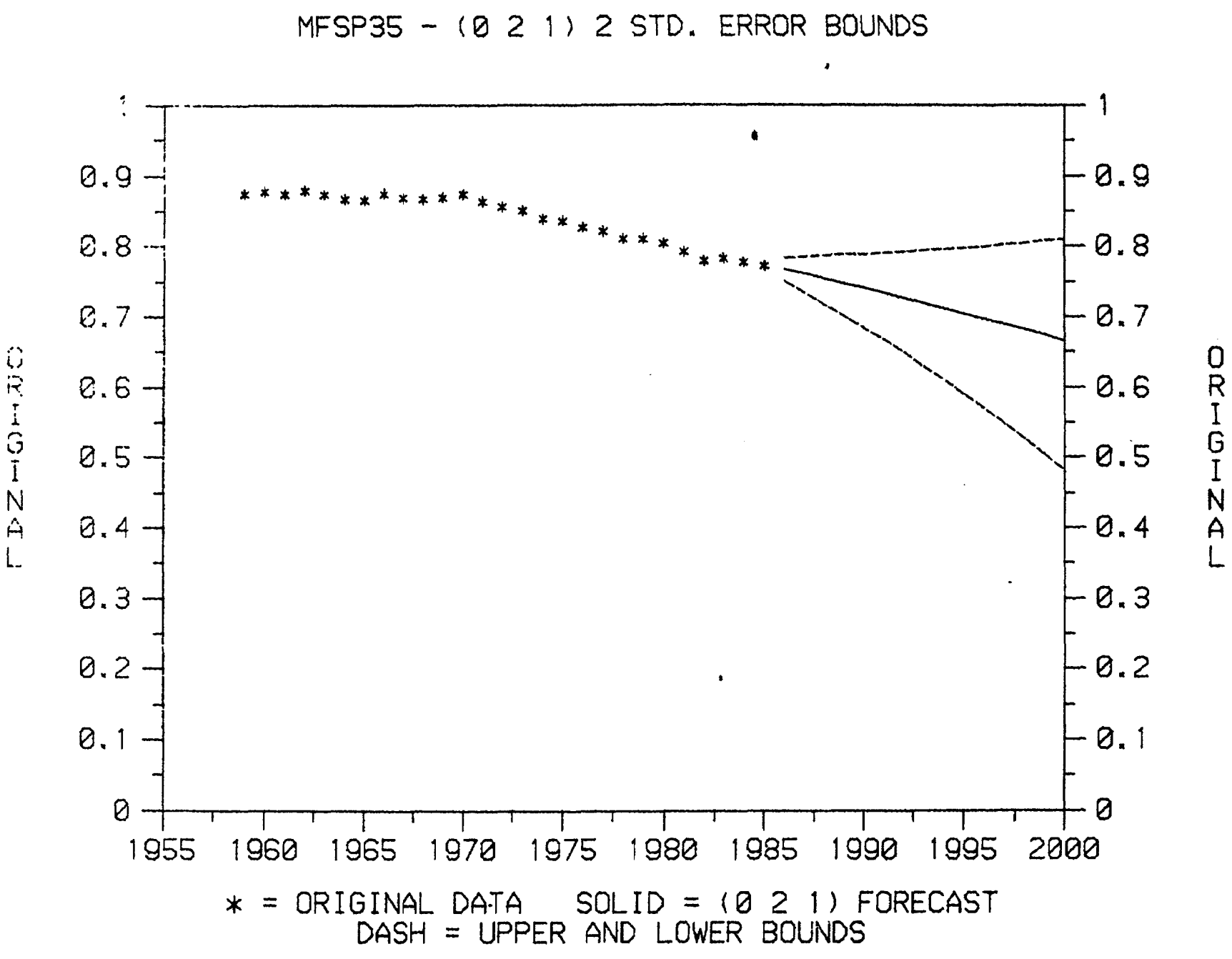

MFSP35 - (0 2 1) 2 STD. ERROR BOUNDS
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
ORIGINAL
* = ORIGINAL DATA SOLID = (0 2 1) FORECAST
DASH = UPPER AND LOWER BOUNDS

Figure G.6

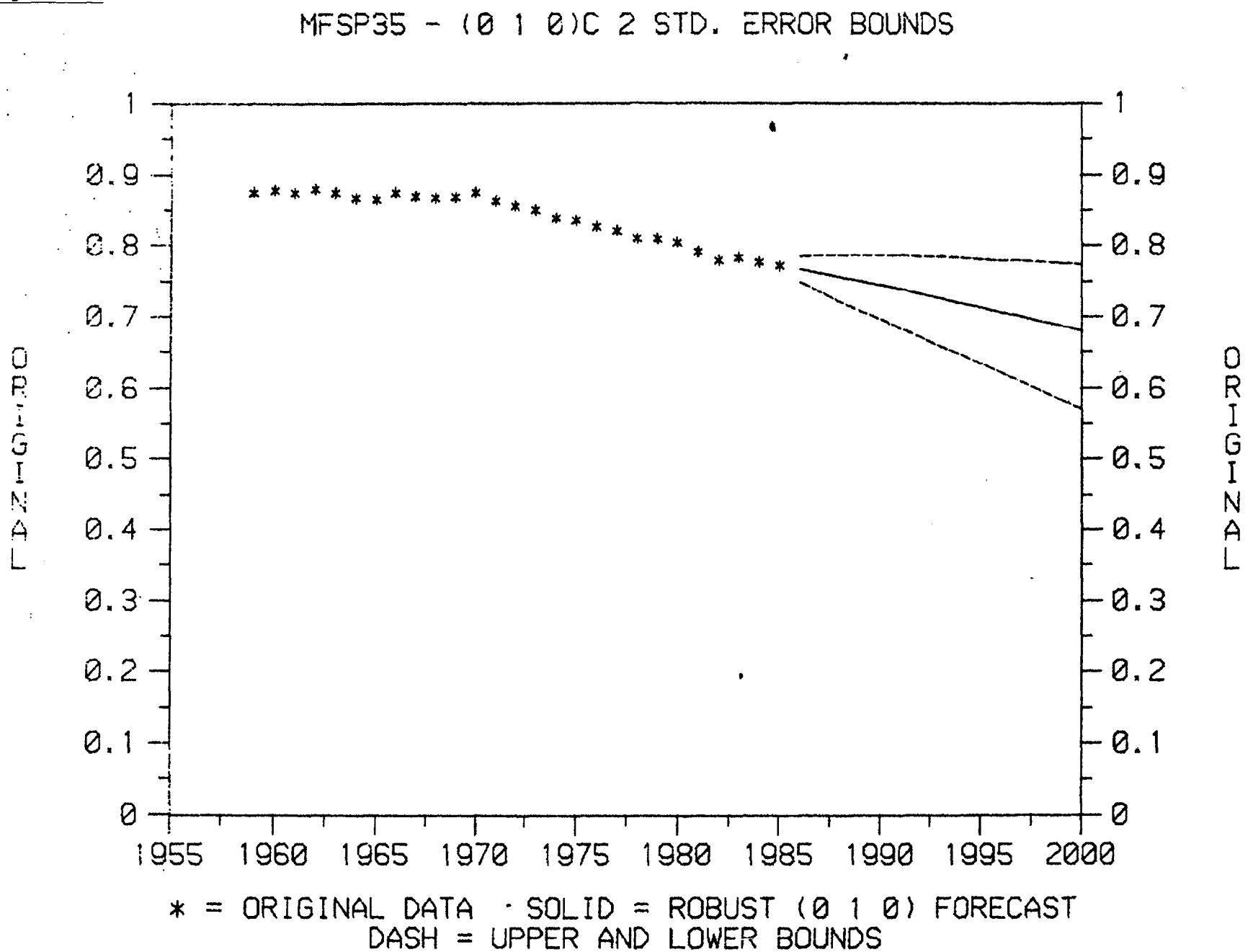

MFSP35 - (0 1 0)C 2 STD. ERROR BOUNDS
ORIGINAL
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
* = ORIGINAL DATA · SOLID = ROBUST (0 1 0) FORECAST
DASH = UPPER AND LOWER BOUNDS

Figure H.1

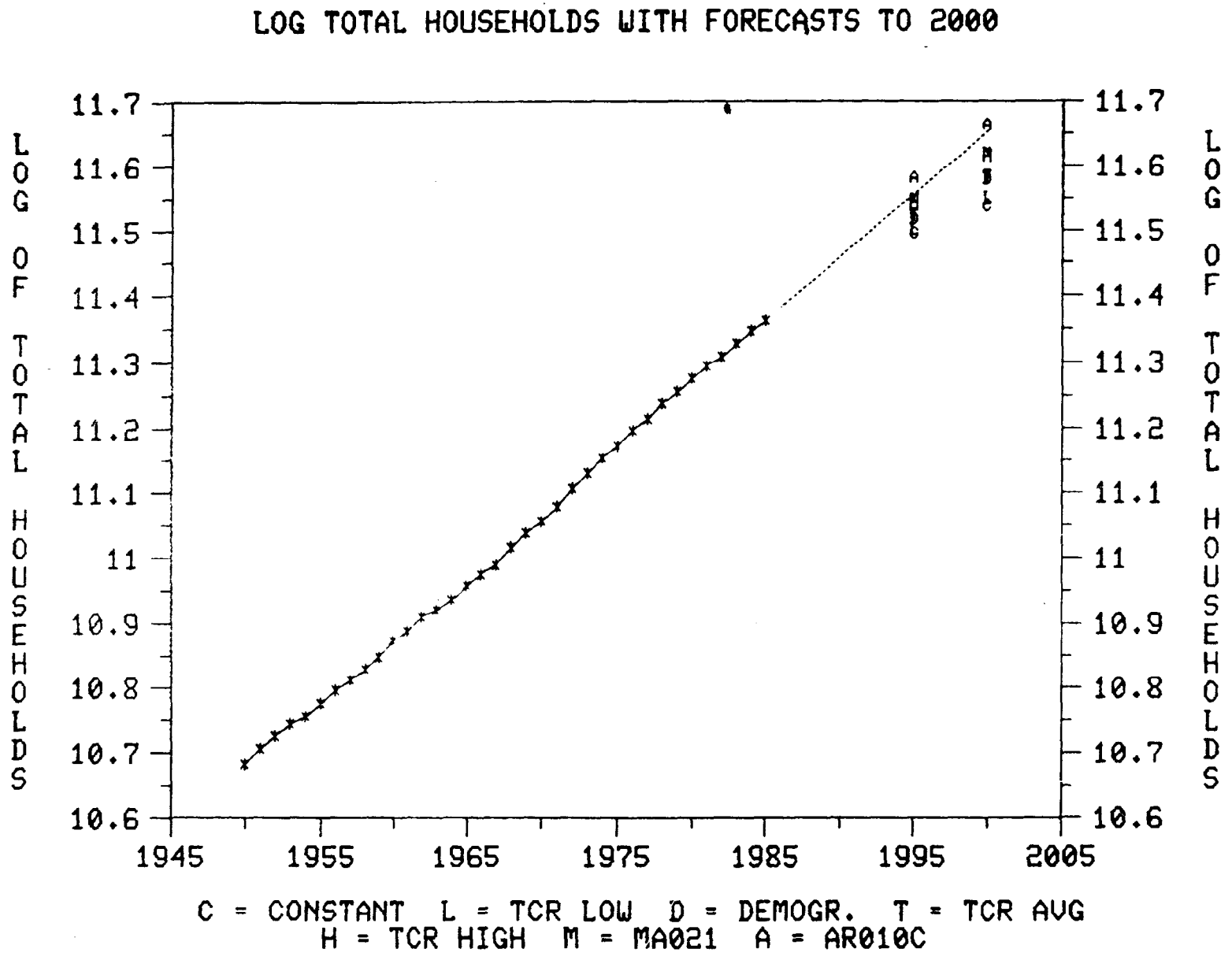

LOG TOTAL HOUSEHOLDS WITH FORECASTS TO 2000
LOG OF TOTAL HOUSEHOLDS
11.7
11.6
11.5
11.4
11.3
11.2
11.1
11
10.9
10.8
10.7
10.6
1945
1955
1965
1975
1985
1995
2005
C = CONSTANT L = TCR LOW D = DEMOGR. T = TCR AVG
H = TCR HIGH M = MA021 A = AR010C

Figure H.2

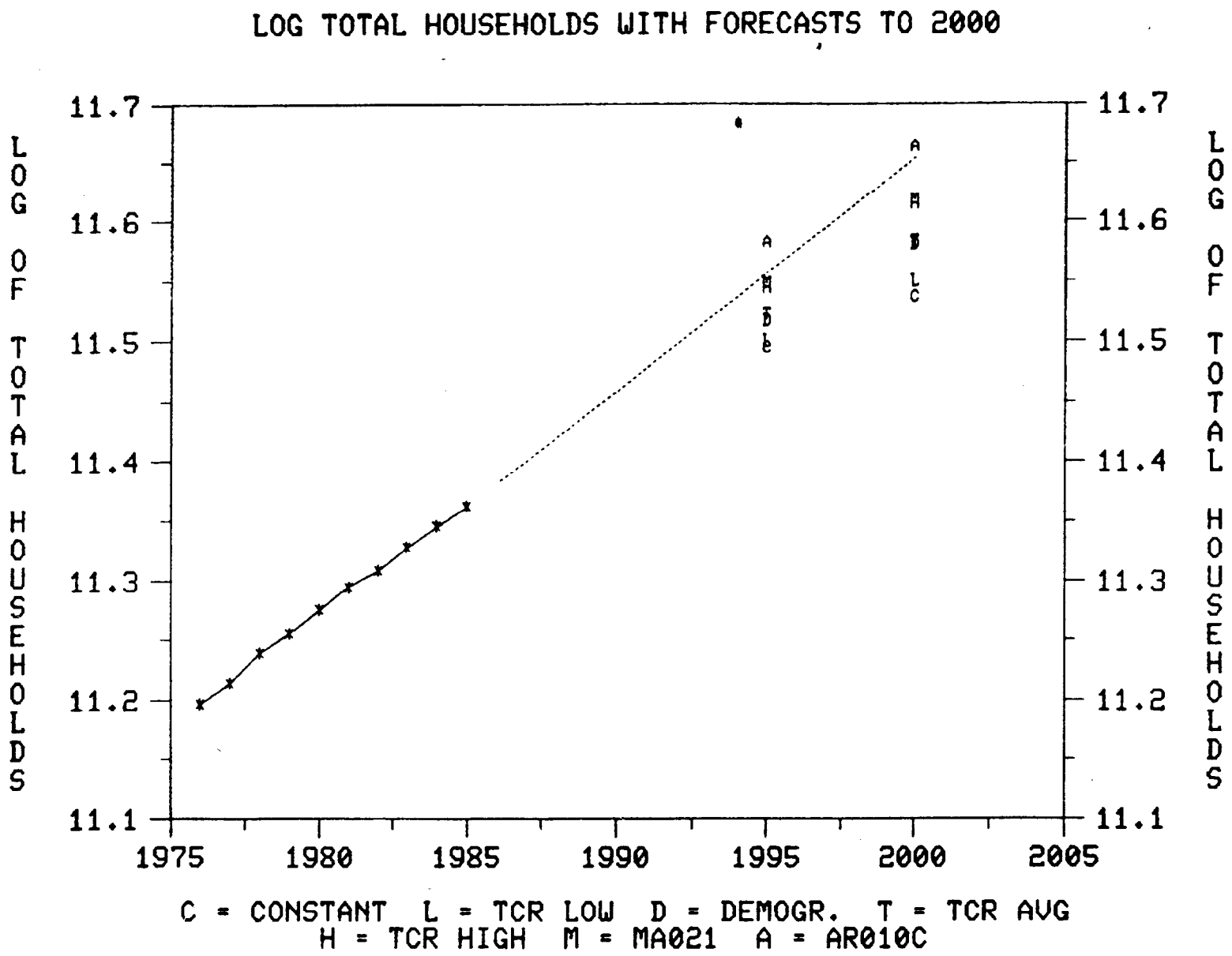

LOG TOTAL HOUSEHOLDS WITH FORECASTS TO 2000
LOG OF TOTAL HOUSEHOLDS
11.7
11.6
11.5
11.4
11.3
11.2
11.1
1975
1980
1985
1990
1995
2000
2005
C = CONSTANT L = TCR LOW D = DEMOGR. T = TCR AVG
H = TCR HIGH M = MA021 A = AR010C

Figure H.3 a

1995

Series	Total	< 25	25-34	35-54	55 +
TCR high(4)	103,235	5,710	21,428	41,063	35,034
TCR avg (4)	101,041	5,474	20,785	40,385	34,397
TCR low (4)	98,848	5,238	20,142	39,707	33,761
MA021 (130)	102,785	4,492	20,470	43,410	34,413
middle level (130)	100,308	4,316	19,927	41,951	34,115
1985 constant (130)	98,180	4,264	19,559	40,628	33,730
MA021 (4)	100,249	4,122	19,373	42,143	34,611

Figure H.3 b

2000

Series	Total	< 25	25-34	35-54	55 +
TCR high (4)	110,689	6,379	19,996	45,923	38,391
TCR avg (4)	107,262	5,998	19,116	44,794	37,354
TCR low (4)	103,835	5,616	18,237	43,665	36,317
MA021 (130)	110,217	4,882	18,924	49,511	36,900
middle level (130)	105,933	4,442	18,004	46,942	36,555
1985 constant (130)	102,440	4,299	17,455	44,703	35,983
MA021 (4)	105,497	4,103	17,412	47,290	36,691