Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Univariate Time Series Modelling and Forecasting using TSMARS

Doctor of Philosophy in Statistics

2006

Gerard Keogh

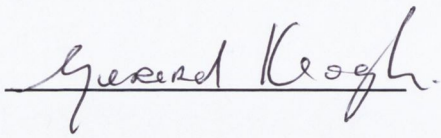
Declaration

This thesis has not been submitted as an exercise for a degree at this or any other University.

This is entirely my own work.

I agree that the Library may lend or copy the thesis on request.

Signed



Summary

This thesis studies threshold nonlinearity in time series using TSMARS, a time series extension of the Multivariate Adaptive Regression Splines (MARS) procedure of Friedman (1991a). MARS is model free and can detect and measure linear and curvilinear structure in data. In this thesis this is used to assess the degree of nonlinearity in empirical time series in official statistics published by the Central Statistics Office (CSO).

For this research Friedman's (1991a) MARS algorithm has been coded from scratch in SAS/IML. This has facilitated the study of empirical series that possess seasonality, outliers, and dependent errors. Each of these require extensions that are novel to TSMARS. These extensions are an important contribution of this thesis.

In Chapter 2 the SAS/IML TSMARS algorithm is described in detail. In simulation studies this program gives statistically equivalent results Friedman's original version (1991a).

In Chapter 3 a rigorous set of simulation studies are conducted to examine how integrated and seasonal time series should be modelled with TSMARS. We show that differencing integrated data improves the precision of the estimates. Simple seasonal models that are novel are also introduced. These distinguish so called regime dependent seasonality from simpler non-regime dependent forms. These are studied to decide whether prior seasonal adjustment is likely to affect TSMARS estimates. We show that this is not the case. Furthermore, we show that it is prior seasonal adjustment that changes the characteristics of the series and not TSMARS.

Chapter 4 provides an empirical investigation of a test-bed of twenty seasonal and non-seasonal CSO time series using four different and sophisticated TSMARS modelling variations. Novel aspects of these variations include seasonal adjustment prior to TSMARS modelling and modelling with variables lagged at 1, 2, 3, 12 and 13 past periods, to name only two. Independent predictors are also used to control for fixed effects such trading day factors. The key purpose of this modelling procedure is to look for and measure the level of nonlinearity in CSO Series. A key conclusion of this chapter is that nonlinearity is shown to be only present to a small degree. In virtue of the autoregressive nature of the modelling we stress that this conclusions is incomplete as outliers and moving average components may alter it.

In Chapter 5 TSMARS is examined from a forecasting perspective - this is novel research. Here, based on cross validation, we show that TSMARS gives consistent forecasts for simple linear and nonlinear time models. The chapter also provides one year ahead (i.e. 12-steps) forecasts for the empirical CSO data using a cross validation approach. The accuracy and precision of the forecast error is reported.

Chapter 6 implements a novel outlier treatment methodology in TSMARS. This is called the Conditional Model Outlier Treatment (CMOT) procedure. We prove that this approach ensures that the model selection mechanism in TSMARS is consistent in the presence of outliers.
Three different adjustment procedures are also set out; namely, Least Squares (LSAO), Bounded Influence (BIF) and Time Series (TSAO). The LASO method is ideal when residuals are independent. The BIF method is suitable when the residuals are independent but may deviate from normality. The TSAO

method is specific for autoregressive, threshold and additive model time series and we prove this method models the error process correctly in these cases. Simulation studies show that these treatment procedures make TSMARS consistent; that is, TSMARS is more likely to choose a correct model type in the presence of an outlier. Both the CMOT procedure and these three outlier treatment mechanisms are important contributions to TSMARS and to the subject of robust methods generally.
The outlier adjustment procedures are also run on the 'test-bed' of twenty economic time series and we find no evidence to alter the conclusions of Chapter 4 - that is, nonlinearity is only present to a small degree. Moreover, we find that the nature of the models found suggest that dependent errors may be more appropriate to model the twenty empirical CSO series.

In Chapter 7 we extend TSMARS to incorporate moving average (MA) components. This is a particularly significant development of the program as many economic time series are better modelled with MA rather than AR terms. This extension allows TSMARS to identify SETMA, ASTAR and ASTARMA models; both ASTMA and ASTARMA are novel model forms.

To gain efficiency we implement parsimonious MA estimation using conditional least squares (CLS) based on a Gauss-Newton procedure. We use two variations of the methodology; namely, Jacobi and Seidel iteration schemes. There are a number of important innovations. The Jacobi iteration de-couples regimes and estimates each separately. This Jacobi iteration is only used for finding the threshold as the residual sum of squares (RSS) in this step is not too sensitive to the method. Final estimation uses the regime dependent Seidel iteration to ensure accurate estimates. In simulation studies this new methodology is shown to be statistically sound.
Attention is then returned to TSMARS with MA component estimation of the empirical CSO series. The results show no improvement over earlier estimates. However, simpler more stable models are found showing that MA components better explain the nature and extent of nonlinearity of CSO series. This is a significant finding.

In Chapter 8 predictive intervals for TSMARS models are computed. We use two novel variations of existing parametric and nonparametric bootstrap methods. These ensure that the intervals account for explicit dependence of forecast on the last p values of a p^{th} order model of a time series. We conduct simulation studies that show the predictive intervals obtained for simple linear models are close to those obtained elsewhere. Moreover, we apply our methods to simple threshold models driven by three different forms of noise. Once again forecast intervals are shown to be accurate and consistent.

These bootstrapping methods are also used to compute predictive intervals for some of the empirical CSO Series. The results show the intervals are generally small for these data. Moreover, these intervals are in close agreement with cross validation intervals obtained in Chapter 5 when large fluctuations do not occur near the end of a time series.

Acknowledgements

I should like to acknowledge the support and assistance of a number of people. First, I should like to thank Professor John Haslett. He suggested I study TSMARS and in our discussions has shown great enthusiasm for, and given my valuable assistance doing this research. Second, I should thank my colleagues at CSO who have supported this research and with whom I have discussed many aspects of statistics. Third, I would like to thank the Director General who has generously provided me with a framework to complete this research.

Dedication

This thesis is dedicated to my darling beautiful wife Siobhan. Her love and support has sustained me throughout our years together and especially during this research.

1 Introduction

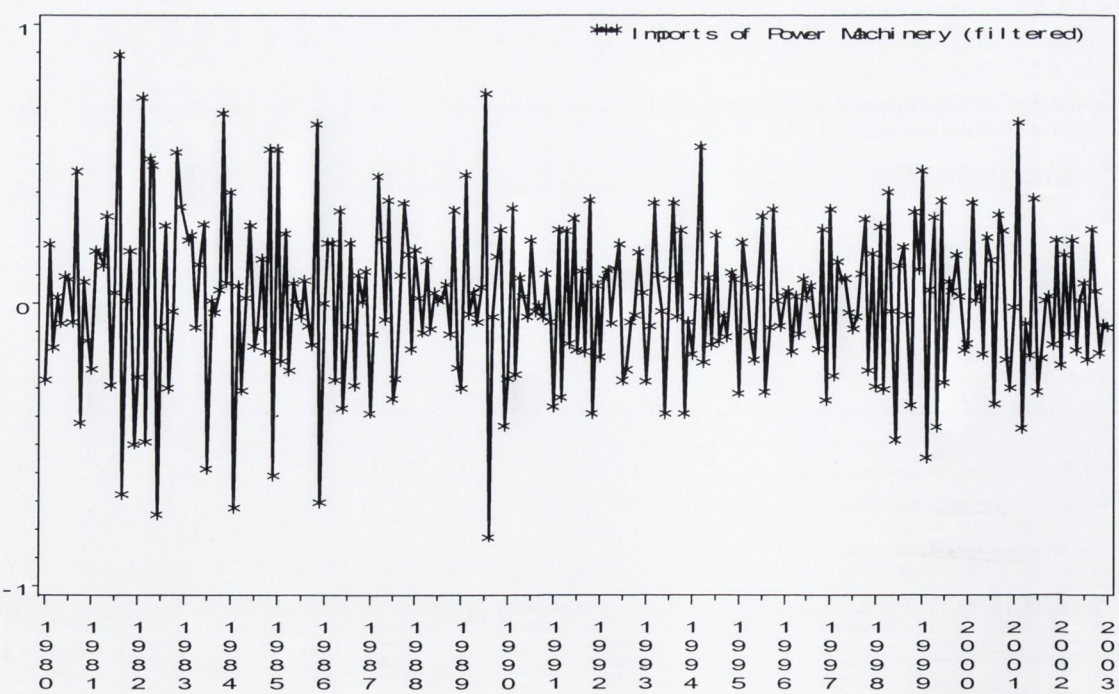
1.1 Background

This thesis studies nonlinearity in time series. The primary focus is on estimation and short term forecasting using TSMARS, a time series extension of the Multivariate Adaptive Regression Splines (MARS) procedure of Friedman (1991a). MARS is chosen because it is model free and can discern nonlinearity. In addition, the method gives a precise measure of the degree of nonlinearity, by apportioning the overall variance of the time series to linear and nonlinear components.

Of particular interest in this thesis are empirical time series in official statistics published by the Central Statistics Office (CSO). Several aspects arise in the study of these series, such as seasonality, outliers, and dependent errors. Each of these require extensions that are novel to TSMARS. These extensions constitute another important focus of this thesis. Additionally, a new version of TSMARS has been implemented in SAS/IML, a platform that will make it much more accessible to researchers in many fields.

To illustrate some of the topics of interest to this research the logged 1st differences (i.e. filtered) of the Imports of Power Machinery empirical series is plotted in Figure 1.1.1. This time series is estimated and future values forecast in later chapters. Of interest is the local or short-term behaviour of the series. This is often important in forecasting into the immediate future, typically one year ahead. The reason for this is simply that the next value is likely to depend on the last observed value of the series. Thus, for example, the Volume of Monthly Industrial Production in Ireland in October will be close to that of September, if all other economic factors remain unchanged.

Figure 1.1.1: Plot of Filtered Imports of Power Machinery Time Series



The local behaviour of the series is influenced by a number of effects. For example, in mid 1989, the large upward spike is followed by a correspondingly large downward spike. In a 1st differenced series this indicates that an additive (shock) outlier occurred in the original (i.e. undifferenced) imports series. The presence of an outlier tends to affect statistical procedures adversely giving poor estimates and forecasts. In this study handling outliers is important as empirical data are emphasised.

Another interesting feature of the series is it seems to have patches where it fluctuates more dramatically than at other times. The period between 1981 and 1984 appears more volatile than the remainder. Changes in the volatility of a time series are evidence of hetroscedasticity; that is, the variance of the series changes over time.

A less obvious, but relevant feature of the series is the frequency of positive and negative values. There are 85 positive and 192 negative values. However, the propensity of negative values is not due to a downward trend in the data. Careful scrutiny of the plot shows that a negative value tends to be followed by another negative value, while a positive value also tends to be followed by a negative value. This suggests that there may be a degree of asymmetry in the series. A tentative nonlinear statistical model for these data with error ε_t might be

$$y_t = \begin{cases} 0.1y_{t-1} + \varepsilon_t & \text{if } y_{t-1} \le 0 \\ -0.3y_{t-1} + \varepsilon_t & \text{if } y_{t-1} > 0 \end{cases} \qquad (1.1.1)$$

This model is a special case of the general class of threshold models that are important in this thesis. The model suggests that when y_{t-1} is negative, y_t will tend to stay negative and if y_{t-1} is positive, y_t will become negative. This implies that model (1.1.1) cannot be reversed; that is, y_{t-1} cannot be predicted from y_t, as the threshold depends on y_{t-1} which is unknown when time is reversed. Asymmetry caused by the presence of a threshold is central in this thesis. Moreover, it is a particular type of nonlinear phenomenon that is suited to modelling with TSMARS.

In this thesis, threshold models similar to (1.1.1) are used in Monte Carlo experiments to test extensions to TSMARS. However, the most significant and innovative aspects of the research are:

- First, the systematic treatment of seasonal threshold models; some of these models are original. In particular we show that TSMARS estimates are largely unaffected by data transformations and seasonal adjustment. These findings are applied to regular and seasonal CSO time series to ascertain the degree of nonlinearity that might be present. We find only a small nonlinear component in these series. The study of these series also covers forecasting and here we find that nonlinearity does affect forecasts for some of the series.
- Second, as TSMARS is based on least squares estimation it is not robust against outliers. A set of three methods, varying in complexity and efficiency, to treat outliers within TSMARS is developed and tested. The key development here is that the outlier treatment methods are built into the core MARS procedure in an efficient way that ensure the model selection criteriion is robust. In simulation studies the efficacy of this approach is adumbrated. In addition the extended TSMARS procedure is applied

to the CSO series and here we find that outliers do not alter our conclusion regarding te extent of nonlinearity present.

- Third, a systematic treatment of threshold models that possess dependent errors is created and tested; here, once again, some original models are defined. This is a particularly significant development of the TSMARS procedure in that parsimonous threshold moving average type models are incorporated in to the TSMARS framework in a consistent and efficient manner. This extension is also used to test CSO series for nonlinear moving average components. The particularly appealing finding of this research is that the resulting TSMARS models tend to be simpler than their counterparts obtained using the original version of TSMARS.
- Fourth, two novel adaptations of bootstrap methods for obtaining stantard errors for TSMARS forecasts are also given. In simulation studies these are show to produce reasonable forecast error intervals that are in line with earlier research. These techniques are then used to generate forecast intervals for the CSO series that are shown to be close to cross validation intervals obtained in an earlier chapter.

In the remainder of this chapter, a review of relevant nonlinear time series models and their properties is given. In section 1.3 nonlinear seasonal models are discussed, some of which are new. Section 1.4 describes nonparametric methods, highlighting the family of methods to which MARS belongs. Finally, section 1.5 sets out the main research themes; that is, estimation and forecasting of (potentially) nonlinear time series. The section closes with a review of the chapters and contributions in the remainder of the thesis.

1.2 Nonlinear Time Series Models and their Properties

Typically the analysis of nonlinear time series models is more difficult than linear models with additive Gaussian innovations. One reason for this is that lagged innovations as well as lagged values of the variable y_t, may be combined so that it is impossible to express y_t as a linear combination of innovations alone. Moreover, the innovations may be non-Gaussian and give rise to nonlinear effects even where the variables and innovation are combined linearly in a statistical model. In this section the statistical and dynamic aspects of a number of relevant nonlinear time series models are discussed.

1.2.1 Preamble

The general form of the nonlinear time series model is

$$y_t = f\left(y_{t-1},\ldots,y_{t-p},\varepsilon_t \ldots \varepsilon_{t-q}\right) \tag{1.2.0}$$

This is a rather intractable stochastic difference equation where the independent innovation ε_t and its lagged values are implicit. A somewhat simpler form specified by, for example, Tsay (2000) assumes the innovation is additive giving the model

$$y_t = f\left(y_{t-1},\ldots,y_{t-p},\varepsilon_{t-1} \ldots \varepsilon_{t-q}\right)+\varepsilon_t \tag{1.2.1}$$

It is clear that this definition encompasses the linear models of the previous section but also allows for models based on much more general functional forms.

An immediate difficulty with model (1.2.1) is that it gives no clue as to the degree or form of nonlinear model that may be useful. Moreover, there are no obvious measurable quantities, such as the lag autocovariance function of the linear model, that are known to fully summarise the underlying dynamics. This is in stark contrast to nonlinear dynamical systems theory where basic tools such as the correlation dimension and Lyapounov exponents are important invariants (Tong 1990 Chapter 2). These can be used to decide whether data is generated from a nonlinear system. The usefulness of these invariants however is limited when the model is stochastic (Tong 1990 Chapter 4)). Indeed it is this lack of precise invariants for the nonlinear stochastic model (1.2.1) that, in part, makes it difficult to distinguish a linear model from a nonlinear one for a given time series realisation. The implications of this are explored further in the remainder of this section and in the Appendix.

1.2.2 The univariate stochastic linear model and the MA representation

A linear model for a given time series variable y_t, observed at a time point $t = 1 \ldots n,$ is assumed to be driven by an unobserved error process ε_t called the innovation. In addition, it is presumed that the innovation is white noise (WN), that is

$$\begin{aligned} E(\varepsilon_t) &= 0 && \forall\, t \\ E(\varepsilon_t^2) &= \sigma_\varepsilon^2 && \forall\, t \\ E(\varepsilon_s \varepsilon_t) &= 0 && \forall\, s, t \text{ and } s \neq t \end{aligned}$$

where $E(\cdot)$ is the expectation function. This definition specifies that the errors are uncorrelated. Linear models driven by uncorrelated errors are defined as weakly stationary linear models. If it is further assumed that the errors are independent and identically distributed (i.i.d.), the error process is distinguished as strict WN. Therefore, normally distributed errors are strict WN random variables. A linear model driven by i.i.d. errors is known as a strictly stationary linear model.

A time series y_t is said to be causal, if it is caused by a WN process up to time t. This means that y_t can be represented as an infinite MA process. Harvey (1993, Chapter 8) defines a linear time series model according to the invertible MA representation

$$y_t = \sum_{j=0}^{t-1} \psi_{j,t} \varepsilon_{t-j} + \lambda_{0,t} \tag{1.2.2}$$

where $\psi_{j,t}$ are non-stochastic weights, $\lambda_{0,t}$ is a (fixed or random) initial condition and the innovations ε_t is a sequence of independent random variables with mean 0. Using the MA representation (1.2.2) the statistical properties of y_t can be deduced.

Without doubt the most familiar case of model (1.2.2) occurs when the innovations are Normal. In this instance the statistical properties are fully described by second order moments and the predictive distributions are also Normal. This model (see Tong 1990 Chapter 1) is generally called a Linear Gaussian Random Process (LGRP).

The invertible MA representation can be obtained by repeatedly substituting for lagged values. This technique can also be used to derive the (generalised) MA representation of nonlinear models. Consider the following stochastic perturbation of the AR(1) model

$$y_t = \phi\, \varepsilon_{t-1} y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$

This model is a special case of the general class of bilinear models. Back substituting for the lagged value y_{t-1} gives the equation

$$y_t = \phi\, \varepsilon_{t-1}\left(\phi\, \varepsilon_{t-2} y_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t = \phi^2 \varepsilon_{t-1}\varepsilon_{t-2} y_{t-2} + \phi\, \varepsilon_{t-2}^2 + \varepsilon_t$$

and by repeating this y_t can be written as

$$y_t = \phi^k\, y_{t-k} \prod_{i=1}^{k} \varepsilon_{t-k} + \sum_{i=0}^{k-1} \phi^i\, \varepsilon_{t-i} \prod_{j=1}^{i} \varepsilon_{t-j}$$

Assuming $|\phi| < 1$ the deterministic component y_{t-k} is negligible and therefore

$$y_t = \sum_{i=0}^{k-1} \phi^i\, \varepsilon_{t-i} \prod_{j=1}^{i} \varepsilon_{t-j}$$

The mean of this process is clearly zero while the variance, obtained by squaring the MA representation is $V(y_t) = \sigma^2 + \phi^2 E(\varepsilon_{t-1}^4)$. The variance of the process then depends on the kurtosis of the Normal distribution.

This example shows how repeated substitution can be used to derive the mean and variance of a simple nonlinear time series model. While it has proved effective in this case, more complex nonlinear models cannot be expressed so readily in MA form. Consequently their statistical properties are not easily derived.

1.2.3 Limitations of the univariate stochastic linear model

In the last subsection the definition of a stationary stochastic linear time series model was given in terms of second order moments of the innovations. Stationarity is important because it specifies that the model is stochastically stable. Therefore any predicted value of the model will remain bounded; that is, $\left|E\left(y_t^k \middle| Y_{t-1}\right)\right| < \infty\ \forall\, t$ and $k \geq 1$, where Y_{t-1} is the set of all lagged values.

Beyond the LGRP defined in the previous subsection, the next level of complexity in model (1.2.2) is the Linear non-Gaussian Random Process (LnGRP), see Tong (1990 Chapter 1). Certain aspects of these are examined in detail in Rosenblatt (2000). A simple example of such a model is the AR(1) model

$$y_t = \phi_1 y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim U[-½, ½] \tag{1.2.3}$$

where $U[-½, ½]$ is the Uniform distribution over the range –½ to ½ and has mean 0 and variance 1/12. This model is strictly stationary and linear in the state variable y_t. As a consequence the first and second order moments are straightforward to compute. These however are not sufficient to characterise the full stochastic behaviour of this process which depend on the upper and lower limits. It is also worth

mentioning that this model can be accurately estimated by least squares and produces residuals that are (asymptotically) uniformly distributed. That is regression

$$E(y_t|y_{t-1}) = \phi_1 y_{t-1}$$

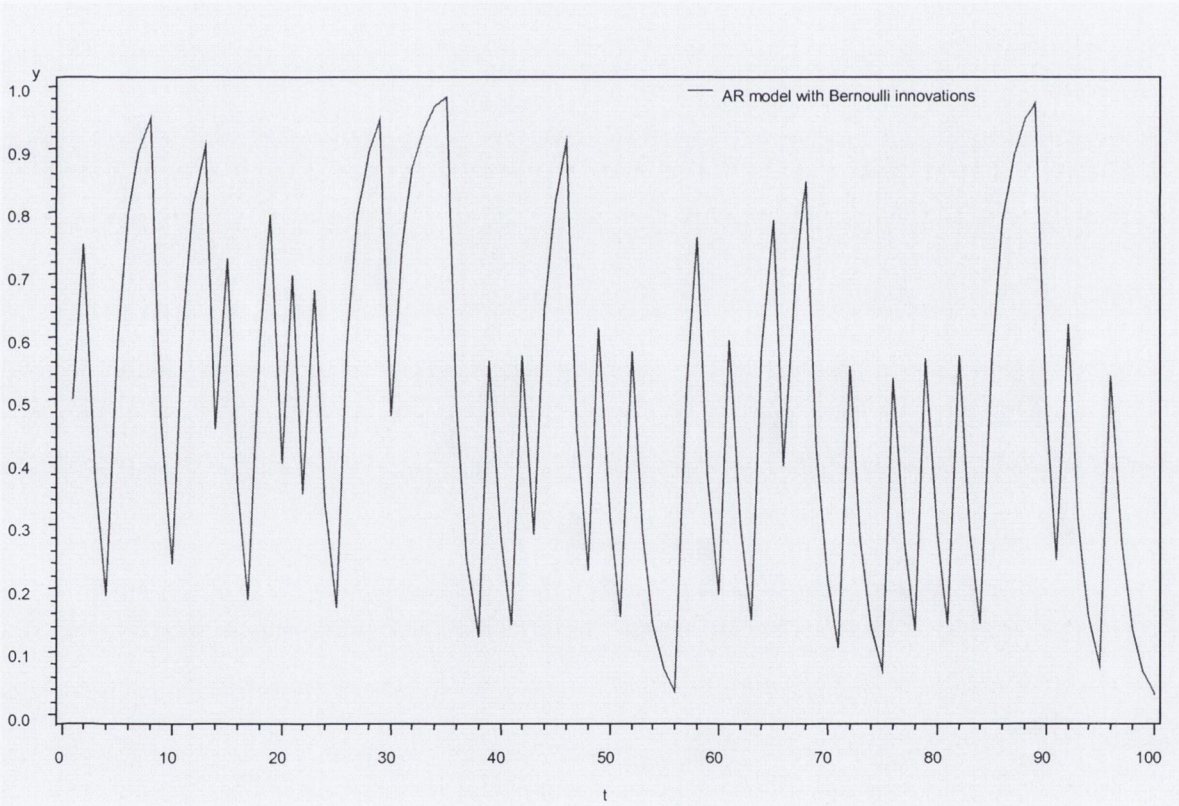
gives estimated uniform residuals $\hat{\varepsilon}_t = y_t - \hat{\phi}_1 y_{t-1}$. This suggests that least squares based estimation methods, such as TSMARS, may prove useful for models driven by non-Normal disturbances.

The uniform AR(1) model (1.2.3) has the property that the innovations are independent and the series autocorrelated. However, even in simple nonlinear processes, such as the bilinear model

$$y_t = \phi_1 \varepsilon_{t-1} \varepsilon_{t-2} + \varepsilon_t$$

with innovation $\varepsilon_t \sim i.i.d[0, \sigma^2]$, (i.e. strict WN random variables with mean 0 and variance σ^2) the series itself tuns out to be a WN process (see Harvey 1993, Chapter 8). Therefore the absence of autocorrelation only means that a linear model is inappropriate. However, there may be a nonlinear model that is appropriate. In terms of prediction, the linear least squares predictor remains the best linear predictor, but there may be a nonlinear predictor that is better.

Figure 1.2.1: Simulated time series plot from AR model (1.2.4)



Worse still, the distinction between linearity and nonlinearity gets blurred for certain LnGRP. Even simple LnGRP with independent innovations can turn out to be irreversible. That is the backward regression $E(y_{t-1}|y_t)$ model is not the reverse of the forward regression model $E(y_t|y_{t-1})$. Tong (1990, Chapter 1) illustrates nicely the type of unexpected effect that can occur using the AR(1) model

$$y_t = \frac{1}{2} y_{t-1} + \varepsilon_t \qquad \varepsilon_t = \begin{Bmatrix} 1/2 \text{ with probability } 1/2 \\ 0 \text{ with probability } 1/2 \end{Bmatrix} \tag{1.2.4}$$

and ε_t is independent of $y_s, s < t$. The stationary distribution of y_t is then uniformly distributed on [0,1]. Figure 1.2.1 shows a plot of a time series simulated from this model. This time series displays a saw-tooth pattern with often longer curved upward movements followed by short sharp downward movements. Thus, a model that is linear in the state variable turns out to be irreversible, that is asymmetric and therefore nonlinear, in time.

This example shows that even though a model can be formally represented as an MA model (1.2.2) the process itself is not symmetric. Therefore the representation (1.2.2) would appear to be too weak to specify a linear model. In this thesis a linear model is defined according to (1.2.2) with the additional requirement that the marginal distributions defined by forward and backward regressions are reversible. In general only processes with Gaussian innovations are examined in this thesis.

1.2.4 Some well known univariate stochastic nonlinear time series models

Within the framework of model (1.2.0) some simple models that are nonlinear in the parameters can be readily postulated. In finance for example, volatility is often studied using a multiplicative noise models known as an ARCH (autoregressive conditional hetroscedastic) or GARCH (Generalised ARCH) model, see Harvey (1993, Chapter 8) or Fan & Yao (2003). The simplest form of (G)ARCH model directly relevant to this thesis is ARCH(1) model of lag order 1, given by

$$y_t = \sigma_t \varepsilon_t \qquad \sigma_t^2 = \alpha + \beta y_{t-1}^2 \quad \alpha > 0, \beta \geq 0 \tag{1.2.5}$$

and if, in addition $\varepsilon_t \sim N(0,1)$ the model is conditionally Gaussian. The key fact however about this model is that the evolving variance is related to the previous times series value in a nonlinear way. That is, large time series values tend to be followed by large values while small values tend to be followed by small ones. As a consequence the distribution tends to have heavier tails than the Normal. It can be readily shown using the Law of Iterated Expectations (see Chapter 8, Harvey 1993) that the unconditional variance of the process is given by

$$Var(y_t) = E(y_t^2) = \beta/(1-\alpha)$$

and since the process is a Martingale Difference it is WN but not strict WN. The ACF of the squared observations also has the nice property that it follows an AR(1) process (see Harvey 1993, Chapter 8).

Closely related to the ARCH model is the Bilinear model. A simple example of this model can be obtained by squaring the ARCH(1) model above giving

$$y_t^2 = \alpha \varepsilon_t^2 + \beta y_{t-1}^2 \varepsilon_t^2$$

and putting $u_t = y_t^2, \eta_t = \alpha \varepsilon_t^2$ and $\phi = (\beta/\alpha)$ gives the bilinear form

$$u_t = \phi u_{t-1} \eta_t + \eta_t$$

In this case it is interesting to note that the multiplicative ARCH model based on the general form (1.2.0) can be expressed as an additive error model of the form (1.2.1).

Bilinear models (see Tong 1990) involve sums of products of lagged variables of y_t and the innovations. In general the simplest bilinear model (which may be termed 1st order) is given by

$$y_t = (\phi_0 + \phi_1 \varepsilon_t) y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim i.i.d.(0, \sigma^2) \tag{1.2.6}$$

These models are studied in Tong (1990) among other places. As in subsection 1.2.2, using repeated substitution of past values, model (1.2.6) can be expressed as a sum of products of the lagged innovations. This can be used to establish that the model is stationary if $\alpha^2 + \beta^2 \sigma^2 < 1$. It is worth mentioning that the relationship between ARCH and Bilinear models is used to characterise neatly the modelling results of a simulated ARCH(1) model studied in Chapter 2.

The method of repeated substitution in model (1.2.6) suggests the possibility of models that involve sums and products of the innovations only. These, quite naturally, are called nonlinear MA models. Among the simplest is

$$y_t = \varepsilon_t + \beta \varepsilon_{t-1} \varepsilon_{t-2} \qquad \varepsilon_t \sim i.i.d.(0, \sigma^2) \tag{1.2.7}$$

An important contribution of this thesis is the development of a procedure to estimate models of this type within the TSMARS framework.

1.2.5 Univariate stochastic nonlinear AR and MA time series models based on a threshold

Among all types of nonlinear time series model, the study of the Nonlinear Autoregressive (NLAR) model has generated greatest interest. The NLAR(p) model of order p has the general form

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) + \varepsilon_t \tag{1.2.8}$$

and obviously the linear AR(p) model driven with Gaussian noise is the simplest.

Particularly relevant to the research described in this thesis, is subclass of model (1.2.8) known as the self-exciting threshold autoregressive (SETAR) model (Tong 1990). A SETAR model is in fact a number of distinct linear AR models that depend on a threshold lag parameter. This parameter effectively separates each AR model into two or more distinct regimes. The simplest SETAR model is the 2-regime SETAR(2,1,1) model

$$y_t = \begin{cases} \phi_1 y_{t-1} + \varepsilon_t & \text{if } y_{t-d} \geq r \\ \phi_2 y_{t-1} + \varepsilon_t & \text{if } y_{t-d} < r \end{cases} \tag{1.2.9}$$

In this model the two AR(1) models are distinguished by the value of the threshold r and lag value d. Model (1.2.9) is a special case of the general l-regime SETAR$(l, p_1, p_2, \ldots p_l)$ model, the notation indicating l, the number of regimes and p_j the AR order in each regime, conditional on $r_{j-1} \leq y_{t-d} < r_j$ with $r_j \in \Re$.

Using indicator functions $I(\bullet)$ to distinguish each regime, model (1.2.9) can straightforwardly be rewritten as follows

$$y_t = (\phi_1 I[y_{t-d} \leq r] + \phi_2 I[y_{t-1} > r]) y_{t-1} + \varepsilon_t \tag{1.2.10}$$

with obvious generalisations to 3 or more regimes and higher AR orders. Thus for example the general l -regime model with mean μ and variance σ_j in each regime can be compactly written as

$$y_t = \mu + \sum_{j=1}^{l} \left(\Phi_j^p(B) + \sigma_j \varepsilon_t\right) I[r_{j-1} \le y_{t-d} < r_j]\, y_t \qquad (1.2.11)$$

with $\Phi_j^p(B)$ being the AR polynomial of arbitrary order p in regime j. This expression follows straightforwardly from (1.2.10). The general l -regime model SETAR Model (1.2.11) specifies that the threshold variables y_{t-d} are simple. Clearly, the threshold variables can have more complex functional forms. In particular the threshold variable may be the k^{th} differenced value $y_{t-d} - y_{t-d-k}$ - this type of threshold we distinguish as complex.

This model, with Gaussian errors, has been widely used in the literature in diverse areas, including economics (Tiao & Tsay 1994), environmental sciences (Melard & Roy 1988), finance (Li & Lam 1995), population dynamics (Stenseth et. al. 1999), as well as the original application to the Canadian Lynx data described in Tong (1990). The popularity of this model lies partially in its simplicity as it partitions the data according to the state variable. This can preserve stationarity. In contrast, change-point models where the regime changes with time result in non-stationary processes (Fan & Yao 2003, Chapter 4). In addition, there is no literature on SETAR models driven with non-Gaussian errors.

Repeated substitution of past values y_{t-1}, y_{t-2} etc. fails for SETAR models due to the presence of thresholds and so these models cannot be expressed as MA processes. As a consequence, the properties of even the simple SETAR model (1.2.9) are still a matter of study. The general SETAR model is even less well understood though it has been shown to be strictly stationary if (a) $\sigma_1 = \ldots = \sigma_l$ and (b) either $\max_{j=1}^{l} \sum_{k=1}^{p} |\phi_{j,k}| < 1$ or $\sum_{k=1}^{p} \max_{j=1}^{l} |\phi_{j,k}| < 1$, (see Chapter 4 Fan & Yao 2003).

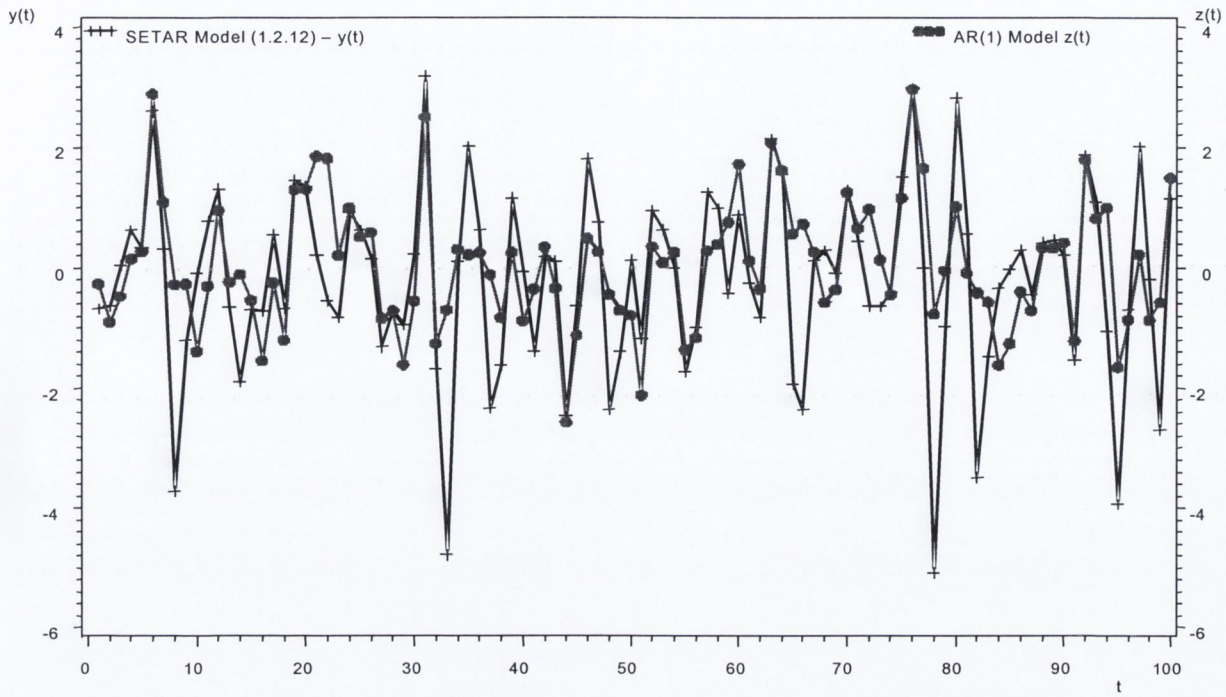
where $\sigma_1, \ldots, \sigma_l$ is the innovation standard deviation in each regime and $\phi_{j,k}$ is the AR(k) parameter in regime j.

These conditions are sufficient but not necessary. So, as noted in Fan & Yao (2003), among other places, it remains a challenge to prove the necessary conditions for the simple 2-regime model (1.2.9) to be ergodic (see Appendix). For example, it is immediately clear that model (1.2.9) with parameters $\phi_1 = 0.5, \phi_2 = -0.5, r = 0$ and $\varepsilon_t \sim N(0,1)$ is strictly stationary. However, for the SETAR(2,2,2) model (see Tong 1990, Chapter 7)

$$y_t = \begin{cases} 0.25y_{t-1} - 0.4y_{t-2} + \varepsilon_t & \text{if } y_{t-2} \le 1 \\ 0.5y_{t-1} - 1.2y_{t-2} + \varepsilon_t & \text{if } y_{t-2} > 1 \end{cases} \qquad (1.2.12)$$

with $\varepsilon_t \sim N(0,1)$, stationarity cannot be assumed. Time series realisations from the model however do appear to be stationary. An important implication of this issue relevant to this thesis is that simulation studies are restricted to simple strictly stationary SETAR models. These are used to address a number of basic research questions and the knowledge is used to inform modelling choices for empirical data.

Figure 1.2.2: Plot of SETAR(2,2,2) model and AR(1) model

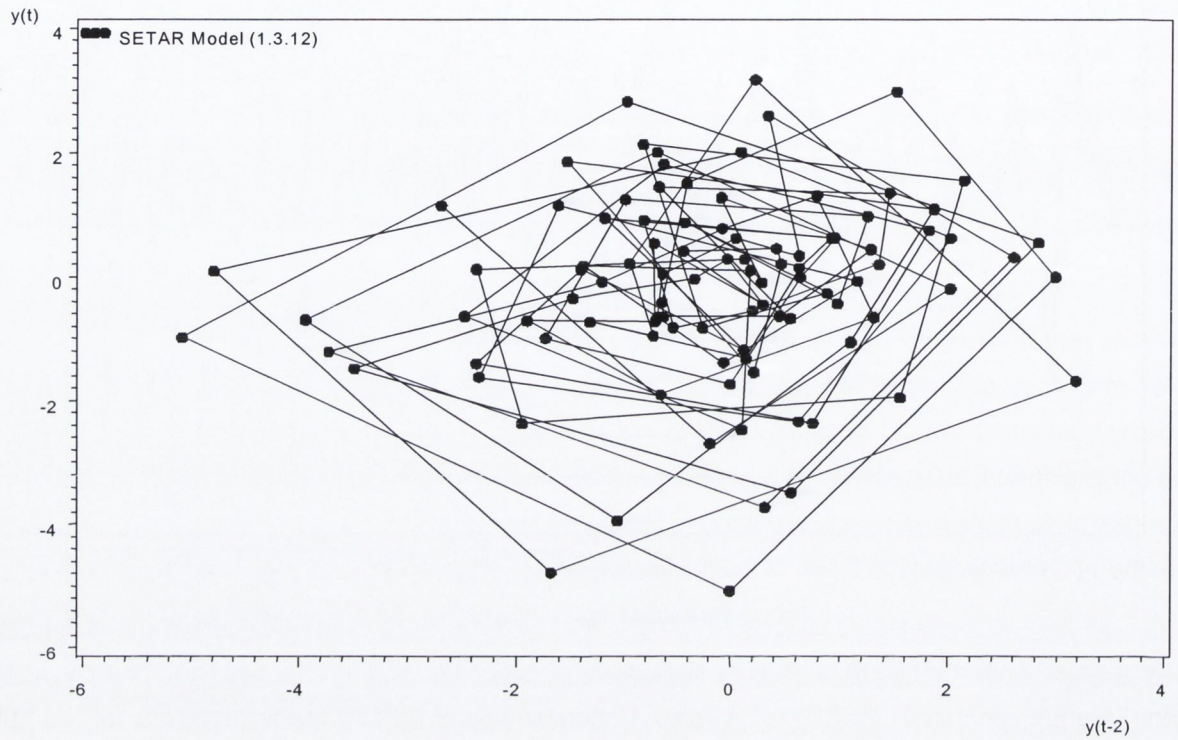


Another second important implication is analysis tools appropriate for linear models (e.g. the correlation function) are not very useful for SETAR models. Understanding of SETAR models tends to rely on data-analytic techniques and non-parametric methods. However, perhaps the simplest and most useful technique for understanding state variable relationships is scatter plot analysis. For model (1.2.12) a time series plot is shown in Figure 1.2.2. Also shown for comparison is an AR(1) model realisation with parameter $\phi_1 = 0.447$. This is an 'equivalent' AR(1) model obtained by taking the square root of the area spanned by the 2-dimensional vectors specified in each regime (i.e. [0.25,-0.4],[0.5,-1.2]). It is clear from Figure 1.2.2 that where the SETAR model data values go above the threshold value (= 1), the series tends to jump to a negative value two steps later, oscillates and then settles down again to something like an an 'equivalent' AR(1) model. As an aside, it is worth noting that stationarity for SETAR models may be easier to show by inferring it from an 'equivalent' linear model.

Displayed in Figure 1.2.3 is the scatter plot of y_t against y_{t-2}. The important feature of this scatter plot is that there is a small 'hole', that is absence of scatter plot points in the centre at (0, 0). Indeed any realisation from model (1.2.12) is likely to possess a hole in the scatterplot that uses the threshold variable as the predictor (i.e. x-axis). The existence of a hole in the plot indicates that the underlying skeleton (i.e. the associated deterministic system, see Tong 1990) may possess a limit cycle. The system therefore is often used to model predator prey type situations such as the Canadian Lynx data (Tong 1990). Here, populations increase where the food source is abundant and then fall away rapidly as the expanded population quickly runs out of that food. In addition the data cannot be reflected along the line

$y_t = y_{t-2}$ and this is indicative of asymmetry. The data generating process is therefore not reversible and so is not a linear Gaussian process.

Figure 1.2.3: Scatter Plot of SETAR(2,2,2) Model (1.2.12) data



1.2.6 Generalisations

A number of models are associated with or are generalisations of the SETAR model that are also important in this thesis. The first is the Additive Model studied in Tsay (1993). The form of this model is

$$y_t = f_1(y_{t-1}) + f_2(y_{t-2}) \ldots + f_p(y_{t-p}) + \varepsilon_t \qquad (1.2.13)$$

That is, a nonlinear function acting on each lagged variable separately. The model can be estimated via the Generalised Additive Modelling (GAM) technique of Hastie and Tibshirani (1990). Tsay (1993) uses this approach to model, among others, data arising from the additive sine model given in Chapter 2 of this thesis.

The second and perhaps most obvious generalisation of the SETAR model (1.2.11) is the SETMA model. The simplest example is the SETMA(2,1,1) model (see Tong 1990) given by

$$y_t = \begin{cases} \theta_{1,1} + \theta_{1,2}\,\varepsilon_{1,t-1} + \varepsilon_{1,t} & \text{if } y_{t-1} \le r \\ \theta_{2,1} + \theta_{2,2}\,\varepsilon_{2,t-1} + \varepsilon_{2,t} & \text{if } y_{t-1} > r \end{cases} \qquad (1.2.14)$$

This model is a special case of the 2-regime SETMA(2;q,q) model of order q studied by De Gooijer (1998). The model has the obvious generalisations to three or more regimes with varying orders in each and can be written in a compact form similar to (1.2.11). Note also that in Chapter 7 mention is made of a related model (not specified elsewhere in the literature) and referred to in this thesis as the Innovation Excited Threshold Moving Average model. Here the threshold is placed on the lagged innovation ε_{t-d}

and not on the lagged series value y_{t-d}. This model is not studied here but it could be estimated by generalising the approach in Chapter 7 even further.

When the functions in the additive model (1.2.13) are thresholds, e.g. $f_1(y_{t-1}) = \phi_1 y_{t-1} I[y_{t-1} < r]$ it is immediately clear that certain SETAR models fall within the class of additive models. SETAR models are also within the class of Adaptive Spline Threshold Autoregressive (ASTAR) models defined by Lewis & Stevens (1991). In effect the ASTAR model is a generalisation of the SETAR model wherein each regime, the AR model is replaced by a model that is bilinear (or possibly trilinear) in the lagged values of y_t. The full model formula is given in Chapter 2, equation (2.4.1). This model is the basis of much of the empirical studies undertaken in this thesis in Chapters 4 and 5. The model is estimated using the MARS algorithm of Friedman (1991). Analysis of the ASTAR model from a theoretical perspective has not been carried out due to its complexity.

The final generalisation of (1.2.11) is the ASTMA and ASTARMA models. These models incorporate MA components into the adaptive spline model framework. These models have not appeared in the literature. They are set out and investigated in Chapter 7, equations (7.2.10) and (7.2.11) respectively.

Related to the SETAR model is the Functional-Coefficient Autoregressive (FAR) Model of Chen & Tsay (1993). Indicator functions are replaced by smooth transition coefficient functions $f_i(y_{t-d})$ giving the form

$$y_t = f_1(y_{t-d}) y_{t-1} + f_2(y_{t-d}) y_{t-2} \ldots + f_p(y_{t-d}) y_{t-p} + \varepsilon_t \qquad (1.2.15)$$

Chen & Tsay (1993) demonstrated that this model is ergodic if the transition functions are bounded and all roots of the associated characteristic function lie inside the unit circle.

There are several other nonlinear models appearing the literature related to the threshold model. In particular the Smooth Transition AR (STAR) model is closely related to the SETAR and FAR models. These smooth out the threshold discontinuity using a sigmoid or logistic function. deBruin (2002), among other, investigates the properties of these models and uses them to study asymmetry in economic cycles. Other models that use the threshold principle are Markov switching models. These are also popular models for studying asymmetry in economic cycles, in for example GNP. The theory and methodology of these models is set out in detail in Kim & Nelson (1999).

Finally, general nonparametric models are also used to study nonlinear data. The approach uses kernel smoothing or locally weighted regression methods to estimate the conditional mean and/or variance of a time series (see Fan & Yao 2003). A justification for the usefulness of nonparametric models for time series data comes from the whitening by windowing principle (Hart 1996). This states that where there is strong mixing the process will have short memory and in this case smoothing methods suitable for independent data can be deployed for dependent processes. This principle is adopted in a simple outlier treatment method in Chapter 4.

1.2.7 Summary

The statistical and dynamic aspects of a number of nonlinear time series models that occur in the literature have been set out. In particular the focus has been on models that are relevant to this thesis. It was shown

that even simple linear non-Gaussian process give rise to asymmetry. On this basis a definition of the linear model has been given which in fact is quite strict. Weaker definitions (as referred to above, Harvey 1993) are based on the general MA representation (see Tong 1990, section 4.10) but problems remain with these, especially when the innovation distribution is not smooth.

The properties of some well-known simple nonlinear models were examined. The connection between ARCH and Bilinear models was emphasised and studied using the back substitution method. This method works for linear models. However, for more complex models this is not fruitful (see Appendix).

Threshold models and their generalisations were also set out. These models tend to be suitable in situations where there is asymmetry. Thus predator-prey situations are often a worthwhile application area. It was also emphasised that analysis remains to be done to fully understand these models. Generalisations of threshold models included the SETMA model. In this thesis (see Chapter 7) a novel extension of TSMARS is developed to estimate these models and the more general ASTARMA model. This extension is the first automatic nonparametric procedure that treats dependent errors in a systematic methodology.

A major gap has been left in the models discussed so far as no seasonal models have been set out. The seasonal ARIMA (SARIMA) is set out in the Appendix. SARIMA models are important in this thesis as they are used to benchmark the performace of other nonlinear seasonal models. A second linear seasonal model called the PAR model is set out in the next section. Seasonality however has been largely ignored in the context of nonlinear models except for a few specific instances to be referred to in Chapters 3 and 4. An important step forward here is that seasonal data is a vital element of this research on nonlinear time series.

1.3 Skeletons and Frames

The noise free case of the general nonlinear time series model (1.2.0) is referred to as the skeleton (Tong 1990). This is obtained directly from the (1.2.0) by simply setting $\varepsilon_t = 0, \forall t$, giving

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) \tag{1.3.1}$$

This equation is the deterministic version of the NLAR(p) model (1.2.8). A skeleton plot of this equation is a time series with the noise stripped off.

In this thesis the novel concept of a frame in introduced. A frame is a plot of a piecewise linear skeleton function $f(\bullet)$ over the arranged sequence of n data values $\{x : y_{\min} = x_1, x_2 = x_1 + c, x_3 = x_1 + 2c, \ldots$ $x_n = x_1 + (n-1)c = y_{\max}\}$ where c is a constant (i.e. step-length) and $y_{\min}, y_{\max}$ are the minimum and maximum of the time series y_t. With this definition when $y_{t-1} = x$ then $y_{t-2} = x - c$ and $y_{t-3} = x - 2c$ etc. Thus the $p-$dimensional piecewise linear skeleton $f(y_{t-1}, \ldots, y_{t-p})$ can be plotted as $f(x, x-c \ldots, x-(p-1)c)$ which is a function of x alone.

Somewhat more generally a product-frame can defined as a frame composed of piecewise products of degree 1 in the predictor variables (e.g. xy but not xy^2). A frame therefore is made up of linear piecewise additive components while the product-frame allows piecewise curvilinear components. Based on this definition, a product-frame is the skeleton of the ASTAR model.

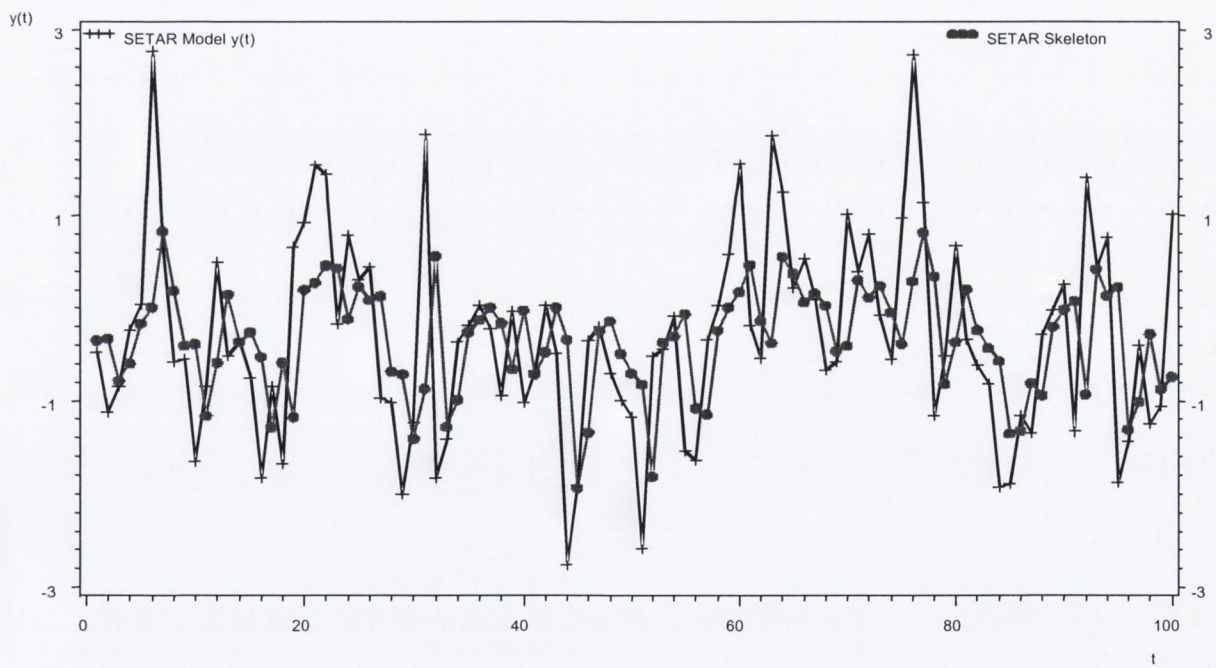
Frames are useful visualisation tools for SETAR and more general ASTAR model approximations produced by TSMARS. Take the simple case of the SETAR(2,1,1) model driven by normally distributed noise $\varepsilon_t = N(0, 1/4)$

Model	Skeleton
$y_t = \begin{matrix} 0.7y_{t-1} + \varepsilon_t & if\ y_{t-1} \le 0 \\ 0.3y_{t-1} + \varepsilon_t & if\ y_{t-1} > 0 \end{matrix}$	$y_t = \begin{matrix} 0.7y_{t-1} & if\ y_{t-1} \le 0 \\ 0.3y_{t-1} & if\ y_{t-1} > 0 \end{matrix}$

(1.3.2)

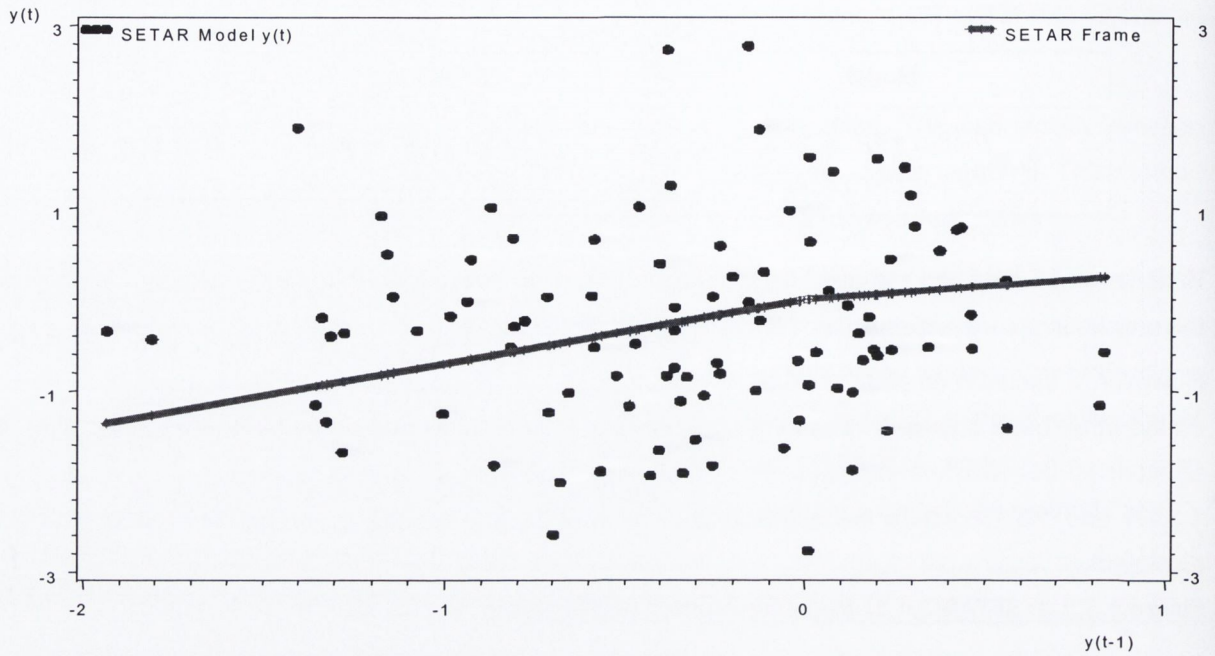
Variations of this model are used for simulation purposes throughout this thesis. Clearly the skeleton of this model is also the frame (i.e. the piecewise linear representation in the ordered values of $x = y_{t-1}$). However, a skeleton emphasises time evolution based on (1.3.1) with the noise stripped off. The frame on the other hand shows the state variable representation (1.3.2) and emphasises the (piecewise) linear relationship between the variables. The contrast between these two representations is given in Figures 1.3.1.1 and 1.3.1.2 where the skeleton plot, as well as the frame plot for SETAR model (1.3.4), are displayed.

Figure 1.3.1.1: SETAR(2,1,1) Model Realisation and Skeleton



The skeleton plot given in Figure 1.3.1.1 is useful as it shows how the noise affects the model. However, as a time series plot the model structure is hard to discern. In contrast, the frame representation in Figure 1.3.1.2 clearly shows the model structure in terms of the state variables based on the ordered sequence y_{t-1}. Of course here the frame coincides exactly with the underlying SETAR function. In particular the slope of the AR model in each regime is emphasised; this is the value of the corresponding regime parameter value. Moreover, this representation suggests that each regime of the SETAR model is independent, at least to an initial approximation. This regime independent approximation is used in later chapters as a starting point for developing TSMARS for more complex time series data.

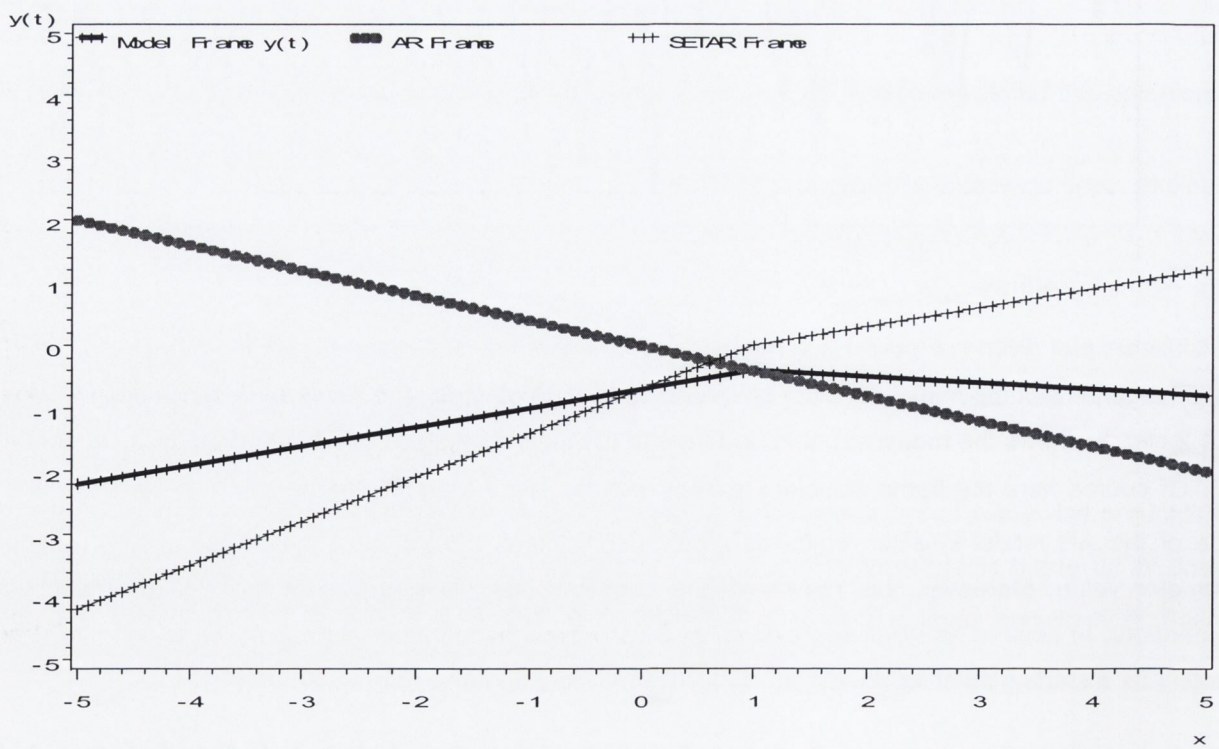
Figure 1.3.1.2: SETAR(2,1,1) Model Realisation Scatterplot and Frame



An important advantage of the frame is the fact that it is based on an ordered sequence of the state variables. Thus, for example, if $y_{t-1} = x$ is chosen as the predictor variable, then $y_{t-2} = x - c$ is obtained by a simple translation one step to the right. Consider the i.i.d. noise driven SETAR(2,2,2) model

$$y_t = -0.4y_{t-1} + \begin{matrix} 0.7y_{t-2} + \varepsilon_t & if\ y_{t-2} \le 0 \\ 0.3y_{t-2} + \varepsilon_t & if\ y_{t-2} > 0 \end{matrix} \qquad (1.3.3)$$

Figure 1.3.1.3: SETAR(2,1,1) Model Realisation Scatterplot and Frame



Here the frame for y_{t-1} is a line with slope –0.4 while the frame for y_{t-2} is identical to that displayed in Figure 1.3.1.2 with predictor y_{t-2}. However, because the state variables are ordered, the frame for y_{t-2} can be overlaid on that for y_{t-1} but lagging by one step. The frame for model (1.3.5) is obtained by the simple addition of the frame for y_{t-1} and the lagged frame for y_{t-2}. These three frames are displayed in Figure 1.3.1.3, where the underlying function for model (1.3.5) is now displayed as a function of one variable, namely $x = y_{t-1}$, rather than a plane in the two-dimensional space of y_{t-1} and y_{t-2}.

The frame for model (1.3.5) is now clearly identifiable as a simple piecewise linear function of the variable $x = y_{t-1}$. However, explicit knowledge of the dependence of the threshold on y_{t-2} is now lost, as the representation in Figure 1.3.1.3 suggests the appropriate time series model for $y(t)$ is

$$y_t = \begin{array}{ll} 0.3y_{t-1} - 0.7 + \varepsilon_t & if\ y_{t-1} \le 1 \\ -0.1y_{t-1} - 0.3 + \varepsilon_t & if\ y_{t-1} > 1 \end{array} \tag{1.3.4}$$

This representation is striking as it implies that this model is 'equivalent' to model (1.3.3). Moreover, this suggests that it might be easier to prove a SETAR model is stationary by showing it is equivalent to a model that is already known to be stationary.

1.4 Some Nonlinear Seasonal Models

1.4.1 Seasonality and Seasonal Models

Seasonal data are important in many application areas and in economic applications they are central. A simple key that can be used to unlock seasonality is the 2-way layout as given in Table 1.4.1.1 for Irish Total Imports in £m.

Table 1.4.1.1: Irish Total Imports £m

	Month											
Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
1996	2498	2385	2567	2483	2391	2267	2248	1945	2286	2553	2523	2333
1997	2398	2398	2747	2898	2698	2662	2671	2372	2831	3147	3111	2932
1998	3118	3193	3569	3283	3347	3307	3266	2821	3259	3480	3659	3413

Clearly this layout suggest a simple 2-way fixed effects additive ANOVA model as the basis for analysis of a seasonal time series y_t

$$y_t = \mu + \tau_i + \delta_j + \varepsilon_t \tag{1.4.1}$$

where μ, τ_i and δ_j represent the overall, annual and seasonal effects respectively. Denoting the number of season by s and using seasonal indicator functions, often called seasonal dummies $D_{j,t}$ $= \begin{cases} 1 & \text{tmod(s)} + 1 = \text{j} \\ 0 & \text{otherwise} \end{cases}$ $(j = 1 \ldots s)$, equation (1.4.1) can be rewritten as

$$y_t = \mu + \tau_i + \sum_{j=1}^{s} \delta_j D_{j,t} + \varepsilon_t \tag{1.4.2}$$

This model and the notion of splitting the data into an annual component and seasonal components can be used as a basis to build seasonal models.

1.4.2 Models based on Periodic Autoregression

The first generalisation of (1.4.2) is the Periodic Autoregression (PAR) model of Franses (1996). In this model the constant parameters δ_j are replaced with lag order s AR polynomials, and B is the usual back shift operator, giving the PAR model

$$y_t = \mu + \tau_i + \sum_{j=1}^{s} \delta_j B_s D_{j,t} + \varepsilon_t \tag{1.4.3}$$

It is clear that, in the PAR model, the within period effects are modelled by an AR polynomial; this is usually of low order. A straightforward nonlinear generalisation of this model is the threshold form called the periodic ASTAR model of Lewis & Ray (1997, 2002). In this case the linear AR terms are replaced by (usually low order) SETAR terms. The periodic ASTAR is used in Chapter 4 to model seasonal economic test data.

Clearly, the seasonal component in (1.4.1) need not be linear or even piecewise linear but may have a more general functional form. Further, there is also the possibility of allowing MA components into the seasonal part of model (1.4.1). In this case (1.4.3) would become the PARMA model

$$y_t = \mu + \tau_i + \sum_{j=1}^{s} \delta_j B_s D_{j,t} + \varepsilon_t + \sum_{j=1}^{s} \delta_j B_s D_{j,t} \varepsilon_t \tag{1.4.4}$$

This model has not been specified or studied in the literature. In principle this model and its ASTARMA generalisation could be estimated using the methods of Chapter 7 though this avenue of research is not pursued in this thesis.

1.4.3 Models based on Standard Seasonal Lags

An alternative approach to the PAR method for treating seasonality is to generalise the linear SARIMA (seasonal ARIMA, see Appendix) model. Take for example the simple linear seasonal AR model (referred to as Model 1 in Chapter 3)

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-s} + \varepsilon_t \tag{1.4.5}$$

Nonlinear versions of this model can be straightforwardly conceived. For example the additive model (1.2.12) can be readily applied to give an additive seasonal model

$$y_t = f_1(y_{t-1}) + f_s(y_{t-s}) + \varepsilon_t \tag{1.4.6}$$

When the nonlinear functions in (1.4.6) are thresholds the resulting models are SETAR. In Chapter 3 the SETAR(2,1,1) model is augmented by the seasonal fluctuation $\rho_2 g(t,s)$, giving the model

$$y_t = \mu + \begin{cases} \rho_{11} y_{t-1} & \text{if } y_{t-1} \geq 0 \\ \rho_{12} y_{t-1} & \text{if } y_{t-1} < 0 \end{cases} + \rho_2 g(t,s) + \varepsilon_t \tag{1.4.7}$$

Here, it is obvious that the threshold is placed on the regular component and so the seasonal fluctuation is independent of the threshold. When, in contrast the seasonal fluctuation depends on the threshold we get the model

$$y_t = \mu + \begin{matrix} \rho_{11}\, y_{t-1} \\ \rho_{21}\, y_{t-1} \end{matrix} + \begin{matrix} \rho_{12}\, g_1(t,s) \\ \rho_{22}\, g_2(t,s) \end{matrix} + \varepsilon_t \qquad \begin{matrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{matrix} \tag{1.4.8}$$

This model is called the Regime Dependent Seasonal SETAR model. It is introduced and studied in simulations in Chapter 3. Clearly, generalisations to more complex threshold structures are possible and models of this seasonal ASTAR flavour are estimated for test data in Chapter 3.

Models (1.4.7) and (1.4.8) are based solely on autoregressions and quite naturally MA components can be included. This type of model occurs in Chapter 7, where some of the modelled test data result in Regime Dependent Seasonal ASTARMA models.

1.4.4 Comments

Where seasonal data arise in applications there are numerous instances of the linear SARIMA (or SRAIMA+ type) methodologies having been used with success. In contrast the area of nonlinear seasonal models is largely ignored except for a few specific cases mentioned in Chapter 3. In particular very little analysis has been done on seasonal nonlinear models other than through simulation studies.

In this section a number of seasonal models have been set out. Some of these appear in the literature while others are given here for the first time. These models are used to study the effects of seasonality in a nonlinear context through extensive simulation studies in this thesis.

1.5 Nonparametric Smoothing Methods

It is assumed in this thesis that little or no knowledge about a time series is available, other than obvious factors such as integration, trading effects or seasonality. In this situation, finding the true or even best parametric model is infeasible as the population of models is infinite. The alternative is to adopt a nonparametric (i.e. model free) method. Here, of course, there are many methods to choose from. In the remainder of this section the main families of smoothing methods are outlined with the focus on conditional mean smoothers. From these TSMARS is chosen for study in this thesis and the reasons for this choice are also given.

1.5.1 Linear smoothers

Consider again the NLAR(p) model (1.2.8). Linear smoothers approximate the nonlinear function $f(\bullet)$ with a flexible class of (simpler) functions based on deterministic AR terms $y_{t-1},\ldots,y_{t-p}$. Approximating the function in this way is generally known as scatterplot smoothing. This emphasises overall functional form at the expense of a potentially complete understanding of the stochastic evolution.

More formally, let $\mathbf{t} = (t_1,\ldots,t_n)^T$ and $\mathbf{y} = (y_1,\ldots,y_n)^T$ be the set of time series observations. A linear scatterplot smoother of $\mathbf{y}$ against $\mathbf{t}$ at each time point t_0 is given by the linear form (see Buja, Hastie & Tibshirani 1989):

$$S(t_0|\mathbf{t},\mathbf{y}) = \hat{y}_t = \sum_{i=1}^{n} s_i(t_0,\mathbf{t}) y_i \tag{1.5.1}$$

for some set of weights $s_i(t_0,\mathbf{t})$. In matrix form (1.4.1) can be written as

$$\hat{\mathbf{y}} = S\,\mathbf{y} \tag{1.5.2}$$

where the smoother matrix S depends on $\mathbf{t}$, as well as the particular smoother, but not on $\mathbf{y}$.

Examples

1. Moving Average, Local Regression and Kernel Smoothers

   The most familiar form of smoother is the moving average. This group also includes the slightly more general local regression smoother. These produce an estimate at t_0 by averaging the response values in a neighbourhood around t_0.

   A related method is the popular kernel smoothing. Here the smoothing matrix has elements $s_{ij} = c_i\, d(t_i, t_j; \lambda)$, where d is an inverse distance measure, c_i is chosen so that the rows sum to unity and λ is the window size (i.e. smoothing parameter). When, for example, $d(\bullet)$ based on the Gaussian distribution the smoother is the well-known Nadaraya-Watson nonparametric regression estimator (see Pena et. al., 2000).

2. Regression Spline

   At predefined points, known as knots, a regression spline partitions the predictor space into distinct intervals. A polynomial (e.g. a constant indicator $I(t-t_0) = \begin{cases} 1 & t >= t_0 \\ 0 & otherwise \end{cases}$, with knot t_0), called a basis function, is then fitted to the data in each interval by least squares. The smoother matrix therefore is made up of blocks of so-called Hat matrices, one for each interval.

   The knots are a set of predefined points in the predictor space. When the independent variable (time) is the predictor the regression spline is simply a change-point regression. However, if lagged values of y_t are used to define the predictor space, then the regression spline is a state variable change-point regression; that is, the approximation is in fact the parametric SETAR model (1.2.11).

3. Smoothing Spline

   Another popular choice of smoother is the cubic smoothing spline. Here the objective is to trade off accuracy of (least squares) fit against smoothness (see Hastie, Tibshirani & Friedman 2001); that is, to find an approximation $\hat{y}_t$ to y_t that minimises:

$$\sum_{i=1}^{n} \{y_i - \hat{y}_i\}^2 + \lambda \int [\hat{y}''(\tau)]^2 \, d\tau$$

   where λ is a tuning parameter. The solution to this Lagrange variational problem is a natural cubic spline with knots at each distinct predictor value and a smoother matrix given by the generalised ridge regression (see Hastie, Tibshirani & Friedman 2001)

$$S = N(N^T N + \lambda \Omega_N)^{-1} N^T$$

   where N and Ω_N are constructed from some appropriate set of basis functions such as B-splines.

   Other Popular Smoothers

The Locally Weighted Running-line Smoother (LOWESS) of Cleveland (1979) combines the strict local nature of running lines and the smooth weights of kernel smoothers.

Wavelet smoothing uses a complete orthonormal basis and then shrinks the coefficients by selection toward a sparse representation. Wavelets are useful where it is necessary to represent either smooth, piecewise constant or bumpy, that is fractal type functions in an efficient way (see Hastie, Tibshirani & Friedman 2001).

1.5.2 Higher dimensional smoothing based on Additive models

In p-dimensions the linear least squares estimate is likely to provide a poor approximation to a more general surface. A convenient nonparametric smoothing solution involves applying a multivariate Gaussian kernel. This generally tends to give poor results because the distribution of available data values gets more sparse as the dimension of the predictor space increases. This is known as the *curse of dimensionality* (Bellman 1961). As a consequence fewer points are available in a local neighbourhood for use in computing the local smoothed approximation.

Backfitting Algorithm

A more pragmatic alternative than direct multivariate smoothing is to generalise the linear model. A key feature of the linear model is that it is additive in the predictor effects. The Additive Model (1.2.13) of Buja, Hastie & Tibshirani (1989), and Hastie and Tibshirani (1990) builds on this concept for the (model free) component functions $f_i(\bullet)$ in (1.2.13). As mentioned above, the model can be estimated via the Generalised Additive Modelling (GAM) technique of Hastie and Tibshirani (1990).

As mentioned breifly in subsection 1.2.6, an efficient procedure for fitting a GAM is backfitting algorithm of Buja, Hastie & Tibshirani (1989). This cycles through each predictor in turn applying a linear smoother, based on cubic smoothing splines to represent the component functions $f_i(\bullet)$. The resulting program that fits Additive Models in this way is called BRUTO (Hastie and Tibshirani 1990).

It is also worth mentioning that the benchmark X11 (Shiskin et. al. 1967) seasonal adjustment program uses a forerunner of the GAM technique to separate seasonal and regular cycles in empirical series (see Hastie and Tibshirani 1990).

Regression Spline Algorithm

In the Additive Model (1.2.13), an alternative to smoothing splines is to use regression splines to approximate the component functions. The regression spline adopted is a simple piecewise linear function of the form

$$b(x) = (x - x_0)_+ = \begin{cases} x - x_0 & x >= x_0 \\ 0 & otherwise \end{cases} \qquad (1.5.3)$$

The notation $(x - x_0)_+$ is shorthand to represent a spline that is supported to the right of the knot point x_0. The Additive Model (1.2.13) is then approximated by linear splines $b(y_j) = (y_j - y_{jm})_+$ with knot points y_{jm} according to the formula

$$y_t = \sum_{j=1}^{p} \sum_{m=1}^{M_j} \beta_{jm} \left(y_j - y_{jm} \right)_+ \qquad (1.5.4)$$

The parameters β_{jm} can then be found by regressing onto the space spanned by these linear splines.

Automatic Fitting Algorithms

In practice there are two major drawbacks to both the backfitting and regression spline methods. These are that the number of basis functions and the knot positions have to be chosen at the outset. An alternative is to adaptively fit basis functions, so that, when added together they give a flexible and parsimonious Additive Model fit.

In Friedman and Silverman (1989) an Additive Model fit is built up by looking at each variable and data point/knot on that variable, in turn, to form a piecewise linear spline basis function. This is temporarily added to the model by regression and the lack-of-fit computed. When all variables and knots have been examined, the basis function that gives the greatest improvement in lack-of-fit is added to the model. The cycle is repeated until there is no improvement in the overall lack-of-fit. This procedure is adaptive; that is, it automatically chooses the knot points, parameters and linear spline functions that make up the Additive Model approximation. The procedure forms the basis of the TURBO program of Friedman and Silverman (1989).

1.5.3 The rationale for TSMARS

MARS and therefore TSMARS, is basically a generalisation of the TURBO program. It is an adaptive procedure that automatically chooses the knot points, parameters and linear spline functions. The key difference is that MARS allows Cartesian tensor-products of piecewise linear splines as basis functions in its model approximation. It can therefore represent more complex functional forms, involving products, than a straightforward additive model.

MARS is suited to situations where the data may be a mixture of smooth and non-smooth forms. It can automatically adapt to different forms in different regions of the predictor space. For example, the MARS model (see Chapter 2 for a definition) includes piecewise curvilinear functions such as

$$f(x,y) = \begin{cases} xy & x >= 0 \\ x & otherwise \end{cases}$$

This function of course cannot be written as an Additive Model as it involves a product. A MARS model therefore is more flexible than an Additive Model which represents curved regions by approximating planes.

As stated at the beginning of this chapter, modelling and forecasting empirical time series is of particular interest. Furthermore no a priori assumptions are made about the data other than additive Gaussian error. The key task is to decide whether or not these time series data are nonlinear. MARS is suited to this because it is model free and can discern both linear and curvilinear structure in the data. Moreover, MARS can compute the contribution to the overall variance of each linear or curvilinear component. Thus, the method not alone provides evidence for deviations from linearity but also give a precise measure of it.

These advantages, combined with the fact that MARS is adaptive and can be efficiently implemented, are among the main reasons it has been chosen for this research.

Forecasting is also relevant. Little or no actual forecasting results have been reported for TSMARS and no information is available on future prediction intervals. Forecasting, as stated at the outset, is an important focus of this thesis. Here this gap in the literature is addressed.

The version of TSMARS used in this thesis has been programmed from scratch in SAS/IML. The justification for doing so is due to two factors. First, Friedman's (1991a) original core code is not available; only an executable version of the program can be obtained. This means that any adaptation can only be done outside the core code. This severely limits experimentation and improvement. A version is however available in the R-language (Hastie 1996) but was rejected for this research as the code is complex and hard to decipher. Moreover, it is not proven in the sense defined in Chapter 2.

The second reason for creating the SAS/IML version was that it facilitated a complete understanding of the MARS algorithm. This meant that any adaptations could be built into the code and tested thoroughly. The development of this version also has the advantage that it makes MARS available, for the first time, to the wider statistical community using SAS. This is an important contribution of this research.

As stated in Section 1.1, three novel methods of outlier treatment are included in TSMARS and these are fully described in Chapter 6. This represents one of the key extensions of TSMARS developed in this thesis. The reason it is necessary to incorporate outlier treatment into TSMARS is because the estimation methodology of MARS is based on least squares. Friedman (1991a), as a consequence, remarks that MARS and therefore TSMARS, is not robust. This is a severe drawback when studies based are on empirical data that may often possess outliers.

The reason outlier treatment is not available in MARS is due to the extra computation involved. Any extension can dramatically increase computation time rendering the procedure useless in all but the simplest of problems. In Chapter 6 efficiency is maintained by adjusting for outliers only on the most recently accepted best fitting model. This adjusted model is then used to find the next best fitting model. This proves efficient as it adds little extra computational burden on MARS. It is referred to as the Current Model Outlier Treatment (CMOT) methodology. It is novel to MARS and moreover, the concept of building in outlier treatment in this way would appear to be novel to any stepwise or stagewise fitting procedure.

Finally, it is worth mentioning some developments related to MARS that have appeared. Friedman (1991b) included logical/categorical predictors into the MARS model and called the resulting function estimating method, an Adaptive Spline Network (see Friedman 1991c). Friedman (1993) followed these innovations by introducing Fast MARS. It reduced computation time at the expense of some accuracy. Another variation of MARS, called PolyMARS, has also been introduced by Stone et. al. (1997) for classification problems. This platform is not studied in this thesis.

1.6 Research Themes

The primary purpose of this research is to see if there is evidence for nonlinearity in CSO time series. This question is explored through the two main themes of estimation and forecasting. A key element of

the approach is that the size of the nonlinear part of an estimated time series model is quantified. In addition the form of the nonlinearity is identified. The estimated model is assessed in terms of within-sample predicted fit according to

$$\hat{y}_t = E(y_t | y_{t-1}, \ldots, y_{t-p}, \varepsilon_{t-1} \ldots \varepsilon_{t-q}) \tag{1.6.1}$$

and out-of-sample forecasting at h-steps ahead according to

$$\hat{y}_{T+h|T} = E(y_{T+h} | y_{T-1}, \ldots, y_{T-p}, \varepsilon_{T-1} \ldots \varepsilon_{T-q}) \tag{1.6.2}$$

To address the primary research question a number of issues have to be resolved such as the modelling method, estimation procedure and data specific issues. These will be outlined in more detail in the remainder of the section, along with a description of the main contributions of this thesis.

1.6.1 Nonlinear time series modelling forecasting

With nonlinear data there may be no obvious choice of model; even where there is, model estimation is not as straightforward as in the linear case. For example, estimation of the general form (1.2.0) with Gaussian innovations has not been attempted in the literature. However, specific forms such as the bilinear model (1.2.6) and Nonlinear MA (NLMA) model (1.2.7) can be estimated by either Conditional Least Squares (CLS), Maximum Likelihood (ML) or the Kalman Filter, see Tong (1990) and Harvey (1993) respectively.

In tandem with the interest in stationarity, mixing and invariant properties, most of the estimation activity has focussed on the NLAR(p) model (1.2.8) driven by Gaussian disturbances. Here once again specific forms have tended to dominate. The family of ARCH and GARCH models are widely used in finance and can be estimated by CLS, ML or the Kalman Filter (see Harvey 1993). In this case, as with the bilinear model and NLMA the estimator is best in the least squares sense though it is not necessarily the best.

The threshold models of subsection 1.2.4 are within the class of NLAR models. These can be estimated by OLS once the position and lag order of the threshold variable is determined. Tsay's test for a threshold uses the idea of ordered regression to find the threshold; that is, values of the time series are assigned to each regime based on the threshold lag variable y_{t-d}, and then y_t is regressed on its lagged values separately on each regime. This is repeated for each value t = 1 to n, of threshold lag variable y_{t-d}. At some point t the BIC will be a minimum; this is the desired threshold value. Once the threshold is determined, the parameters in each regime can be estimated by OLS. This methodology is used in Tong (1990) to study the Canadian Lynx data and also the Sunspot Series.

In general parametric modelling of a possibly nonlinear time series is too inflexible. The reason for this is phenomenological. Unless there is good reason to believe in a particular model based on empirical observation, such as the ARCH models in finance or threshold models for predator prey systems, there is potentially an infinite range of possible nonlinear models to choose from. One way around this difficulty is to adopt a nonparametric modelling technique.

Nonparametric time series analysis is a broad field. It does however rely on a few core methodologies. These are kernel estimation, locally weighted regression and wavelet smoothing as well as state-domain

smoothing techniques, such as GAM models (1.2.12), and FAR models (1.2.15) and Neural Network (i.e. STAR model) approaches. All of these approaches have a good pedigree and methods for estimation are described in various places including Pena (2000) and Fan & Yao (2003).

Another alternative to estimating (1.2.8) nonparametrically is described in Krantz and Schreiber (2004). This approach is a dynamical systems based method and uses reconstruction (delay) vectors. Basically the method, called local projective noise reduction, consists of fitting a series of locally linear models via principal components to delay vectors that are close (in some sense). The closeness measure then becomes a bandwidth parameter of the method that must be course grained enough to distinguish the noise. Krantz and Schreiber (2004) suggest the method is reliable for estimating certain invariants such as Lyapounov exponents.

In this thesis the nonparametric time series modelling route is followed. The method used throughout is TSMARS. This method was first adopted for time series by Lewis & Stevens (1991) and has been used, since then, by Lewis & Ray (1997) among others. The method of estimation is based on least squares and is fully described in Chapter 2. Essentially it generalises the notion of ordered regression to an exhaustive search across all lagged variables and allows for product terms in its approximation. In this way it builds a flexible approximation to the underlying nonlinear function in (1.2.8) using Cartesian products of linear splines. Thus in 3-dimensions for example, MARS can be thought of as building a 'frame' using locally linear planes that approximate the surface given by the nonlinear function.

Estimating a time series model is important but forecasting unknown future values is perhaps more important. Forecasting with linear models is well understood; in this case the forecasts are minimum mean square error (MMSE) forecasts. The forecast at h-steps ahead from time point T is given by the out-of-sample estimate (1.6.2). This formula is of course very similar to the within-sample one (1.6.1) except for the explicit dependence of the future values on the last p, q observed values.

For the linear model the distribution of the forecast errors is readily available from the model (see Wei 1990). For the linear AR(1) model the forecast value at h-steps ahead is easily seen to be $\hat{y}_{T+h|T} = \phi_1^h y_T$. Using this, and the MA representation of the series, the standard error of the forecast can be computed (see Wei 1990). However, as stated earlier forecasts from a linear model will in general not be best possible when the underlying process in nonlinear. Putting aside the complex area of non-Gaussian processes, the NLMA model admits a better nonlinear forecast than its linearised counterpart (see Harvey 1993, Chapter 8). Worse still, even within the Gaussian framework, the 3-step ahead forecast cannot be explicitly computed for the SETAR(2,1,1) model (1.2.11). This is because the forecast is conditional on the unknown future value y_{T+2} compared against the threshold. Methods for approximating the forecast error in this case are available based on the Chapman-Kolmogorov relation (see Tong 1990 and de Bruin 2002 among others). These methods tend to be somewhat limited in their usefulness as the algebraic manipulations can be difficult. When greater flexibility is required the smoothing approach of Fan & Yao (2003) or the bootstrapping technique is generally preferred to obtain confidence bands for forecasts. Very little empirical study however has been done to ascertain the power of these techniques on real rather than simulated time series.

1.6.2 Chapter and contributions review

In Chapter 2 the MARS algorithm is described in detail. Based on this description, the first contribution of this research is the development of a new version of the MARS program written in SAS/IML. This program is assessed against some of the results given for Friedman's original version (1991a) and is shown to give statistically equivalent results. The program is then adapted for time series and the resulting TSMARS program is then benchmarked against identical studies in the literature.

The type of empirical data that arises in practice is often nonstationary and seasonal. TSMARS has been rarely used for this type of data. It is a flexible nonparametric state-variable smoother and so should not be influenced by nonstationary components. However, a nonlinear log transformation and/or differencing may affect the quality of the TSMARS fitted model. It is therefore natural to ask to whether or not data should be transformed prior to TSMARS modelling. This question is explored in detail in Chapter 3 through simulation studies. The issue of seasonal data is also considered; specifically, whether or not it is better to seasonally adjust data prior to TSMARS modelling. Once again this topic is addressed through detailed simulations using newly defined nonlinear seasonal models. For both stationarity and seasonal adjustment, this is the first rigorous study that has taken place for models arising from TSMARS.

Chapter 4 is an empirical investigation of a test-bed of CSO time series. Here four different TSMARS modelling implementations are put into practice to model seasonal data. While some aspects of the implementations adopted have already appeared in the literature (see Chapter 3 for details) a significant portion are new. These include seasonal adjustment prior to TSMARS modelling and modelling with variables lagged at 1, 2, 3, 12 and 13 past periods, to name only two. Independent predictors are also used to control for fixed effects such trading day factors. The modelling procedure in all four implementations is sophisticated. It is designed to ascertain the nature and extent of the nonlinearity in CSO data. This, important contribution, is the first time a study on this scale and level of detail using TSAMRS has been reported for time series data.

In Chapter 5 the usefulness of TSMARS is examined from a forecasting perspective. This is the first time an in-depth study of this nature has been conducted in the literature. Initially, forecasting at h = 1 to 5 steps ahead on the linear and nonlinear time models studied in Chapter 2 is conducted. The purpose of this is to show that TSMARS gives consistent forecasts and can therefore be used for further research. The second half of Chapter 5 provides one year ahead (i.e. 12-steps) plug-in forecasting for the empirical data. One year ahead forecasting is the most important in short-term analysis and is conducted for each implementation for the five years 1998 through 2002. The approach is to retain the relevant year, say 1998, as a cross-validation set and use it to compute actual forecast errors. The accuracy and precision of these is reported. Once again this is a new contribution to both empirical forecasting and the understanding of TSMARS as a forecasting tool.

The linear SARIMA+ model has two important ingredients that are not part of TSMARS. These are a systematic way of treating outliers and built-in MA components. Even for non-ordered data, MARS itself is not robust; in particular, for official time series, the absence of outlier handling is a major drawback. Some work on outlier adjustment for TSMARS has been conducted (see Chapter 6). However, the approaches assume only model coefficient parameters are influenced by the existence of an outlier. Threshold value

estimates are neglected. This problem is overcome in Chapter 6 where outlier adjustment is efficiently incorporated into the threshold and coefficients estimation procedure. In addition, three adjustment procedures are made available. A fast method suitable for independent data, a robust bounded influence procedure and a time series based method that is accurate for SETAR models. Each of these three options is tested in simulation studies. They are then tested on the empirical data to see if the accuracy of the estimates obtained in Chapter 4 are influenced by outliers. The addition of efficient outlier adjustment to the threshold and parameter selection mechanism in TSMARS, is an important new contribution to the estimation procedure in TSMARS for both independent and dependent data.

A major flaw in TSMARS is the lack of any means to incorporate MA components. Many economic time series are better modelled with MA terms rather than AR terms. Trying to fit MA terms to the residual of a TSMARS fit is not ideal as MA components may have been partially smoothed in the original TSMARS fit. Adding MA component estimation to TSMARS is set out in Chapter 7. This is a significant step forward in the development of TSMARS. It is implemented using conditional least squares (CLS) based on a Gauss-Newton procedure. Two variations of the methodology adopted; these are Jacobi and Seidel iteration schemes. There are a number of important innovations. First, the Jacobi iteration assumes that each regime, of say a SETMA model, is independent. Accordingly the regimes can be de-coupled and estimated separately. An efficient algorithm is outlined for this purpose. This Jacobi iteration is however only used in the inner (costly) loop of TSMARS for finding the threshold as the residual sum of squares (RSS) in this step is not too sensitive to the method.

For the final estimation of the model the more costly Seidel iteration is used. In this case the assumption of regime independence is dropped. This iteration is required to ensure accurate basis functions are computed. This step is required because perturbations of the basis function based on the independence assumption causes instability. This results in poor TSMARS estimates.

This enhanced TSMARS procedure is used in Chapter 7 for MA component estimation in simulations of linear and threshold MA time series. It is also used to study the empirical data to identify whether or not nonlinear MA components are evident in these data.

Little or no space in the literature devoted to forecasting with TSMARS. In Chapter 8 the issue of forecasting for TSMARS is taken beyond the cross validation approach of Chapter 5. Essentially equation (1.5.2) describes the average value of the underlying predictive distribution. In the nonlinear context where asymmetry may occur, the predictive distribution often deviates from normality. As a consequence, it is often better to obtain quantiles rather than standard deviations to characterise the predictive errors. This topic is explored in Chapter 8 using bootstrapping techniques designed to take account of the explicit dependence of the forecast on the last available value. These techniques are studied using simulated models and are then applied to a small but relevant set of official time series. The bootstrapping methods and the results described are both new contributions to the study of time series forecasting using TSMARS.

2 MARS and Time Series Estimation using MARS

The research described in this thesis is built upon the smoothing program MARS (Friedman 1991a). The first section of this chapter describes the salient features of MARS. It also gives any amendments made (e.g. special parameters) to develop the SAS/IML version used to produce results reported here. Recall, from Chapter 1, a SAS/IML version was written because Friedman's original code is available and it facilitated a more complete understanding of the method. Furthermore, the version in R-code was complex and as yet it appears to be proven in the time series literature. From this point on, where the term MARS (and TSMARS) is used it will be taken to mean the SAS/IML version.

The remaining sections of this chapter are devoted to 'proving' MARS through simulation studies. The objective is to show that MARS produces an approximation that is statistically equivalent to Friedman's original version. In order to prove MARS, data is simulated from a statistical model and MARS is applied. In a proportion of simulations, the type of MARS model that results is different to the original model from which the data were simulated. These MARS models are termed 'incorrect models'. Following Lewis and Stevens (1991), results in this thesis are quoted only for correct models found in simulation studies. Incorrect models occur firstly, because the sample size used in the simulation may be small and secondly, because the statistical properties of the lack-of-fit criterion used in MARS are unknown.

In some of the simulations conducted in this chapter it is possible to distinguish the subset of correct models. In these cases MARS is deemed to be performing correctly, if at least two-thirds of models are 'correct' and this fraction increases with the sample size. This figure is chosen because a good statistical test of evidence for SETAR effect, generally has power of at least two-thirds (see Pena et. al. 2000, Chapter 10). In addition, this lower bound reflects the fact that Friedman's original code produced a correct model, in at least two-thirds of simulation studies on selected models. These studies, among others, are repeated here with the SAS/IML version of MARS.

Having fitted the model, parameter estimates and associated predicted values are compared with their true values in terms of accuracy, precision and consistency. MARS will be judged to be performing properly if estimated parameters, based on correctly estimated models, are within two standard errors of their true values. Matching this criterion ensures both the form and the accuracy, precision and consistency, of the MARS approximation is statistically equivalent to Friedman's original version.

2.1 The MARS Algorithm

MARS is a flexible nonparametric smoothing program developed by J. Friedman for data modelled as independent cases. The algorithms are described in Friedman (1991a) and subsequent enhancements to the functionality are described in Friedman (1991b, c).

MARS seeks to approximate the underlying functional relationship between a single response variable *y* and a vector of *p* predictor variables $\mathbf{x} = (x_1, \ldots x_p)$ through the nonlinear regression equation

$$y = f(x_1, \ldots x_p) + \varepsilon \tag{2.1.1}$$

over some domain $D \subset \mathbf{R}^p$ that contains n samples of data, namely $\{y_i, \mathbf{x}_i\}_{i=1}^n$ and having i.i.d. error ε. Specifically, the regression function $f(\mathbf{x})$ is modelled as a linear combination of M+1 basis functions

$$\hat{f} = \beta_0 + \sum_{m=1}^{M} \beta_m \, b_m(\mathbf{x}) \tag{2.1.2}$$

Each basis function $b_m(\mathbf{x})$ is a tensor product of K_m linear spline functions over the predictor variables and takes the form

$$b_m(\mathbf{x}) = \prod_{k=1}^{K_m} \left[s_{k,m} \left(x_{v(k,m)} - t_{\xi(k,m)} \right) \right]_+$$

The quantity K_m is the interaction degree of the spline functions and is normally limited to 3. Each spline function $s_{k,m}\left(x_{v(k,m)} - t_{\xi(k,m)}\right)$ is in turn composed of 3 elements. These are the sign denoted by $s_{k,m}$, a single predictor variable $x_{v(k,m)}$ and knot point $t_{\xi(k,m)}$ that splits the domain of $x_{v(k,m)}$ into two sub-regions according to

$$b_{k,m}\left(x_{v(k,m)} \middle| s_{k,m}, t_{\xi(k,m)}\right) = \left[s_{k,m} \left(x_{v(k,m)} - t_{\xi(k,m)} \right) \right]_+$$

with the "+" subscript denoting the positive part of the argument, i.e. $[z]_+ = \begin{cases} z, & \text{if } z > 0 \\ 0, & \text{otherwise} \end{cases}$.

If $x_{v(k,m)}$ takes on a set of unordered categorical values $\{c_1, \ldots, c_K\}$, that is a set where no distance relation exists between the values, then

$$b_{k,m}\left(x_{v(k,m)} \middle| s_{k,m}, t_{\xi(k,m)}\right) = H(x_{v(k,m)} \in A)$$

where the indicator function H takes the value 1 if $x_{v(k,m)}$ is in any unique (permutation) subset of the categorical values and 0 otherwise.

The key to understanding MARS lies in the matrices $v(k,m)$, $\xi(k,m)$ and $s(k,m)$. Each basis function $b_m(\mathbf{x})$ is associated with a corresponding column in each matrix and these record the list of predictor variables, their signs and knots in v, ξ and s respectively. The iterative solution for the knot $t_{\xi(k,m)}$, variable $x_{v(k,m)}$, sign $s(k,m)$ and interaction K_m defining each basis function, the model size M and regression parameters β_m involves MARS starting with a basis function set that consists of the single constant basis function

$$b_0(\mathbf{x}) = 1.$$

A new basis function $b_{M+1}(\mathbf{x})$ is then added to the existing set $\{b_0, \ldots b_M\}$ by conducting, in a forward stepwise manner, a nested exhaustive search looping over:

- the existing set of basis functions - that is columns of v, and
  - all predictor variables not already in the selected column of v, and
    - all values of that predictor variable not already used as knots in ξ.

This process generates a sequence of combinations of basis function, new variable and knot (i.e. split point on that variable) that make up a potential new basis function. Each of these is temporarily added in turn to the existing set of basis functions. The resulting set is then used to compute the weighted least

squares approximation $\hat{f}$ (i.e. regression spline approximation) and modified Generalised Cross Validation (GCV) lack-of-fit score

$$GCV(M) = \frac{ASR}{\left[1 - \frac{(d+1)M}{n}\right]^2} \tag{2.1.3}$$

Here the Average Squared Residual is ASR = $\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_i\right]^2$, d is a user-specified constant. The GCV Criterion was originally conceived by Craven and Wahba (1979) and has the following form

$$GCV(M) = \frac{ASR}{\left[1 - \frac{C(M)}{n}\right]^2} \tag{2.1.4}$$

Here the function $C(M)$ is a complexity cost penalty included to account for the increased variance associated with increasing model complexity (i.e. the number of basis functions M). The GCV is therefore intended to trade-off some accuracy for smoothness in the MARS approximation $\hat{f}$. In a regression situation with an MxN basis function matrix B, where only the model coefficients are being estimated the complexity cost is (see Friedman 1991a)

$$C(M) = trace\left(B\left(B^T B\right)^{-1} B^T\right) + 1 \tag{2.1.5}$$

In MARS this is equal to the number of linearly independent basis function, that is $M+1$. Using (2.1.5) in 2.1.4) gives the original GCV proposed by Craven and Wabha (1979).

Mars however makes extensive use of the data values to construct a basis function set. Thus the basis function parameters, namely the matrices of variables, knots and signs $v(k,m),\ s(k,m),\ \xi(k,m)$ respectively, are not independent of the response values. This reduces the bias of the model but increases the variance since additional parameters are being adjusted to help fit the data at hand. This means that (2.1.5) no longer reflects the effective number of parameters being estimated in the nonlinear MARS model. To compensate for this Friedman (1991a) further penalises (2.1.5) according to

$$\ddot{C}(M) = C(M) + d\,M = (d+1)\,M \tag{2.1.6}$$

where d as mentioned above is a user specified smoothing constant, usually set at 3, that penalises (i.e. charges) for models having a larger number of basis functions. This charge is proportionately reduced when a potential new basis function is linear or additive by a factor of 1/3 and 2/3 respectively. This gives a slightly greater preference to linear and additive functional forms over curvilinear forms. Friedman (1991a) also notes that in simulation studies MARS in not sensitive to the value of the smoothing constant and works well with the effective but crude choice $d = 3$. We have adopted this choice throughout this thesis.

In MARS as the forward stepwise search proceeds, the potential basis function giving the smallest GCV, that is, that best adapts to the shape of the underlying function is chosen as the candidate basis function to enter the existing set. On completion the candidate enters the existing set of basis functions and a new column is added to the matrices v, ξ and s. This new column records all existing variables, knots and

signs of the "parent" basis function as well as the new variable, knot and sign chosen in the search. The parent is simply the basis function chosen from the existing set in outer loop of the search. Where a candidate basis function is a combination of both categorical as well as continuous functions it simply replaces the parent basis function already in the existing set. This procedure continues until there is no reduction in the GCV score, or a pre-defined number of basis functions are reached, or the ASR is sufficiently small.

Following the forward stepwise search a backward iterative search is applied. This selectively deletes basis functions one-at-a-time if the associated GCV is reduced, thereby producing a "good" set of basis functions for approximating the underlying function. When this procedure is complete, the ANOVA decomposition (see below), is used to replace the linear spline basis function set with a set of cubic spline basis functions (see Friedman 1991a).

The MARS algorithm outlined above has been coded as a SAS macro with the core forward/backward stepwise searches written in SAS/IML (Interactive Matrix Language). A key component affecting the efficiency of the algorithm is the computation of the ASR from the weighted least squares approximation $\hat{f}$. The approximation is computed via (weighted) orthogonalisation using the Modified Gram-Schmidt procedure (see Golub & Van Loan 1996); this is numerically more stable than the standard version of Gram-Schmidt. This provides a convenient and relatively cheap method for computing the improvement in the ASR, since as each additional basis function is added to the existing set, the improvement is simply $-\beta_{M+2}^2 \left\| b_{M+2} \right\|^2 / n$, as the residual from the existing set is also orthogonal to b_{M+2}.

Linear dependence within the basis function set can occur and this is controlled using a tolerance criterion on the set of orthogonalised basis functions. Setting this tolerance is crucial to the quality of the approximation, since too fine a value can introduce noise while too large a value may exclude important features. A default a value $2\mathrm{X}10^{-4}$ has been chosen based on experience, though a larger value is sometimes required especially where the data are known to possess a significantly linear component.

Some additional features of this SAS implementation of MARS include the standardisation of the response and predictors to improve computational stability; these are re-coded to their input values on exit. The maximum number of allowable basis functions, M is set to $\max(5n^{1/4} + p, n/4, 50)$ while the interaction degree of the linear splines K_m, and the number of predictor variables p, are chosen *a priori*. Also, every third order statistic is chosen as a potential knot point in the predictor space. This avoids excess noise occurring in the solution and reduces the overall computation time by a factor of 3.

2.2 The ANOVA Decomposition

The result of applying the MARS algorithm is an approximation based on a model of the form

$$\hat{y}_t = \hat{f}_t = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} \left[s_{k,m} \left(x_{v(k,m)} - t_{\xi(k,m)} \right) \right]_+ \qquad (2.2.1)$$

As noted in Friedman (1991a), this constructive representation, that is, based on forward stepwise selection and backward deletion of basis functions, provides little insight into the nature of the approximation. However, by simply rearranging terms, (2.2.1) can be cast into form that reveals a lot about the predictive relationship between the response and co-variates. The idea is to collect together all

linear basis functions, then second order tensor product basis functions, then all third order terms in (2.2.1) and so on. An example will best illustrate the idea. Take the following MARS approximation based on the co-variates x_1, x_2, and x_3

$$\hat{f} = 0.05 + 0.1x_1 - 0.5x_3 - 0.18(x_1 - 0.35)_+ + 0.3x_1(x_2 - 0.5)_- + 0.2x_2x_3 - 0.1x_1(x_2 - 0.5)_-(x_3 - 0.3)_+$$

This MARS model can be rewritten as an additive combination of the form

$$\hat{f} = \begin{array}{ll} f_0^* + & = 0.05 + \\ f_1^*(x_i) + & = 0.1x_1 - 0.5x_3 - 0.18(x_1 - 0.35)_+ + \\ f_2^*(x_i, x_j) + & = 0.3x_1(x_2 - 0.5)_- + 0.2x_2x_3 - \\ f_3^*(x_i, x_j, x_k) + & = 0.1x_1(x_2 - 0.5)_-(x_3 - 0.3)_+ \end{array}$$

This is a sum over all basis functions that involve a single variable, of two variables only and involving three variables only. Similarly, the full MARS model (2.1.3) can be recast into the form

$$\begin{aligned} \hat{f} &= f_0^* + f_1^*(x_i) + f_2^*(x_i, x_j) + f_3^*(x_i, x_j, x_k) + \ldots \\ &= \beta_0 + \sum_{\substack{m=1 \\ K_m=1}}^{M_i} \beta_m b_m(x_i) + \sum_{\substack{m=1 \\ K_m=2}}^{M_{i,j}} \beta_m b_m(x_i, x_j) + \sum_{\substack{m=1 \\ K_m=3}}^{M_{i,j,k}} \beta_m b_m(x_i, x_j, x_k) + \ldots \end{aligned} \tag{2.2.2}$$

Owing to the similarity of this expression with the decomposition of the analysis of variance for contingency tables, Friedman (1991a) refers to (2.2.2) as the ANOVA decomposition of the MARS model (2.2.1).

In this thesis a small variation on (2.2.2) is adopted. The single variable function $f_1^*(x_i)$ is further split into a purely linear part and the remainder. Thus, in the example above

$$f_1^*(x_i) = \begin{array}{ll} f_{1,L}^*(x_i) + & = 0.1x_1 - 0.5x_3 - \\ f_{1,T}^*(x_i) & 0.18(x_1 - 0.35)_+ \end{array}$$

where L denotes the linear part and T represents the remainder, made up purely of splines based on a knot. In this thesis, functions of the form $f_{1,T}^*(x_i)$ will be called threshold functions while functions in (2.2.1) that involve products will be called curvilinear. The set of threshold and curvilinear functions is collectively called nonlinear functions.

The most useful aspect of the ANOVA decomposition is that overall variance (i.e. total sum of squares) can be formally written as

$$\begin{aligned} V(y) &= V(\hat{f}) + V(\text{residual}) \\ &= V(f_0^*) + V(f_{1,L}^*(x_i)) + V(f_{1,T}^*(x_i)) + V(f_2^*(x_i, x_j)) + V(f_3^*(x_i, x_j, x_k)) + \ldots + V(\text{residual}) \\ &= V(\text{constant}) + V(\text{linear}) + V(\text{nonlinear}) + V(\text{residual}) \end{aligned} \tag{2.2.3}$$

where $V(\bullet)$ is the variance (note, that this formal decomposition follows from the fact that the basis functions in MARS are orthogonal). This partition of the overall variance is crucial, as it provides a measure of the extent of nonlinearity in a given set of data. In the results reported in later chapters, the extent of nonlinearity will be expressed as a percentage of the overall variance according to

$$\%\,\text{Nonlinearity} = \frac{V(\text{nonlinear})}{V(y)} \times 100\%$$

A fundamental aspect of this research is, evidence for nonlinearity in a dataset is obtained using MARS and moreover, a precise measure of that nonlinearity is provided through the ANOVA decomposition. Using the ANOVA decomposition to measure nonlinearity in this way is a novel element of this research.

2.3 *Simulation Studies using MARS*

As pointed out by Friedman (1991a), a reasonable concern with MARS is that it might find considerable spurious structure in data where the *signal-to-noise* ratio is small. In particular, this means random noise in the data, is picked up by MARS as false structure, thereby giving a misleading indication of the association between the predictor and response variables.

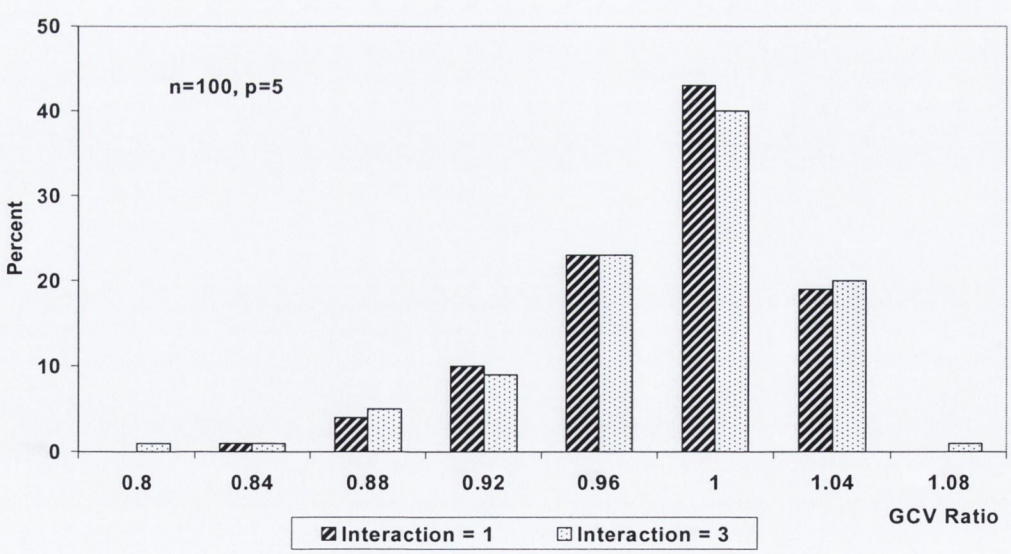
2.3.1 Random Noise Simulation Study

A straightforward way to decide whether the MARS approximation suffers in situations where the signal-to-noise ratio is small is to apply it to random data. Here we follow Friedman (1991a) and generate 100 data sets y, for which the response values in each case are drawn from the Standard Normal Distribution (i.e. N(0,1)) and five predictors (quite independent of y) are drawn from a uniform distribution on [0,1].

For each simulation run the ratio of the GCV of the final model to that of the mean (i.e. $\hat{f} = \bar{y}$) is computed; in this case the ratio should be 1. The distribution of GCVs is displayed in Figure 2.3.1.1.

In contrast, if MARS finds a spurious signal then the resulting GCV ratio will be a lot less than 1. In this case, the averages for both simulations with 5 independent predictors, 100 observations and maximum interaction K_m set to 1 and 3 respectively, is about 0.98. This is not a noticeable fall in the GCV ratio and is similar to that obtained by Friedman (1991a). Therefore MARS does not find structure in noise where there is none. MARS is therefore performing properly in this simulation.

Figure 2.3.1.1: Random Data Simulation on 100 Data sets



2.3.2 Additive Function Simulation

Clearly, it is equally important that MARS should find the true structure where it exists. So, for example, when the data are additive MARS should not introduce spurious interactions. Friedman (1991a) demonstrates that this is the case by conducting a simulation study based on the nonlinear function

$$f(\mathbf{x}) = 0.1e^{4x_1} + \frac{4}{1+e^{-20(x_2-0.5)}} + 3x_3 + 2x_4 + x_5 + 0 \bullet \sum_{i=6}^{10} x_i \qquad (2.3.1)$$

Note: the coefficients of all variables x_6 to x_{10} are zero.

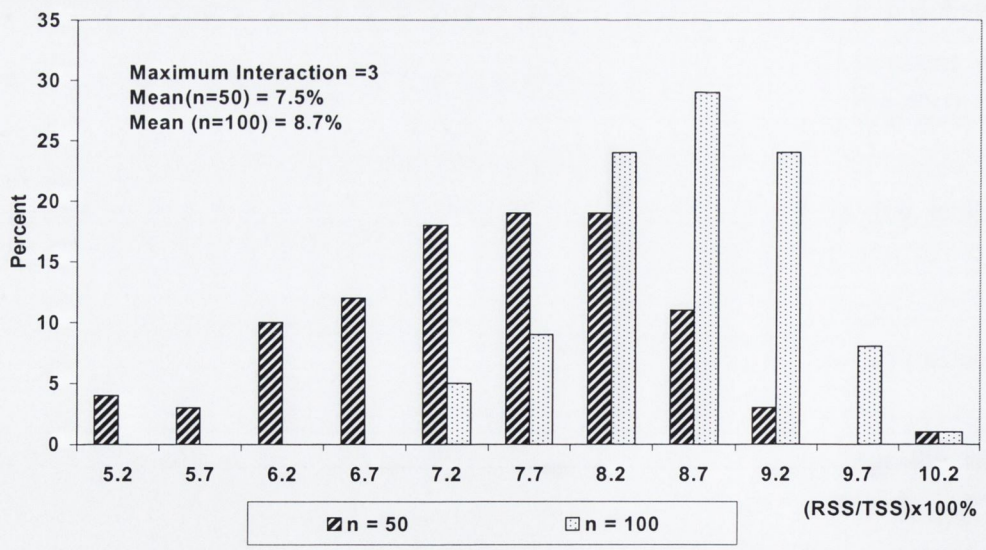
A similar simulation study to Friedman's is repeated here for n = 50 and 100 observations. In each case (p=10) covariate vectors were randomly drawn from the uniform distribution on [0,1]. The corresponding response values were assigned according to

$$y_i = f(\mathbf{x}_i) + \varepsilon_i \qquad 1 \le i \le n$$

with the noise term ε_i randomly generated from the Standard Normal distribution. Here the signal to noise ratio is 3.28 and the true underlying function accounts for 91% (a large proportion) of the variance (see Friedman 1991a).

Figure 2.3.2.1 displays the histograms obtained from the 100 simulation runs for the ratio of the residual sum of squares (RSS) to the total sum of square (TSS) by 100%. In both cases the fit is very good with the average value for these ratios (inset in the top left of the figure) being 7.5% and 8.7% respectively. This means that most of the noise (9% of the TSS) has been filtered out leaving virtually a perfect signal – that is, none of noise is left is the MARS estimate of the underlying function. MARS is performing correctly and properly as the accuracy and precision of fitted models is optimal in this simulation.

Figure 2.3.2.1: Histogram of results for Additive Data Simulation



The histograms above provide a good indication of the accuracy of the MARS estimate. The histogram in n=50 case shows the RSS is smaller than the n=100 case. Their means are 7.5% and 8.7% respectively. The n=50 case is therefore a slightly poorer fit, as the error accounts for 9% of the overall variance. Thus, in both cases MARS gives a very good approximation to the data and moreover, this approximation improves with increasing sample size.

However, this is not the whole story, as the approximations may involve second order or higher level interactions between the variables in a basis function. The table below is based on the ANOVA decomposition, equation (2.2.2). It shows that the approximation mainly comprises linear and additive basis functions. That is, the approximation is wholly made up of about 6.5 linear/threshold functions and these account for about 50% of the MARS model variance - the mean accounts for the rest. Also, the fact that no basis function has interactions would indicate the constraints, on basis function selection in MARS is somewhat different to those in Friedman (1991a) and appears to result in more conservative models. This is to be expected, since as stated earlier, this version of MARS gives slightly more preference to the selection of linear and threshold basis than does Friedman's original.

Number of observations	Function type	Average Number of Basis functions	Percentage of Approximation
50	Mean	1	48
	Linear/ Additive	6.6	52
100	Mean	1	54
	Linear/ Additive	6.4	46

2.4 Nonlinear Time Series MARS (TSMARS) Modelling

MARS, as proposed by Friedman (1991a) and outlined above, is a nonparametric approach to nonlinear regression modelling using adaptive regression spline basis functions fit by least squares. It was conceived and designed for situations where the response and covariate predictor variables are independent, with the addition of weighting to handle heteroscedasticity. MARS can be readily adapted to handle time dependent data by using lagged values as predictors.

Consider the univariate linear AR(p) time series model. It arises directly from the linear multivariate regression model by simply letting the predictor variables for the t^{th} value in the time series $\{y_t\}$ be the lagged values $y_{t-1}, y_{t-2}, \cdots, y_{t-p}$. Similarly, with MARS, by letting the predictor variables be the p lagged values of the times series we get Time Series MARS (TSMARS). This approximates a (nonlinear) time series $\{y_t\}$, with M+1 spline basis functions of the lagged predictors defined by knots (i.e. partition points) $y^*_{\xi(k,m)} \in y_{t-1}, y_{t-2}, \ldots, y_{t-p}$, and takes the form

$$\hat{y}_t = \hat{f}_t = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} \left[s_{k,m} \left(y_{v(k,m)} - y^*_{\xi(k,m)} \right) \right]_+ \qquad (2.4.1)$$

The resulting nonlinear time series model defined by this approximation is known as the Adaptive Spline Threshold Autoregressive (ASTAR) model of Lewis & Stevens (1991). This model, as pointed out by Lewis & Stevens (1991), admits the l -regime SETAR$(l, p_1, p_2, \ldots p_l)$ model of Tong (1990) as a special case. This being obtained by simply restricting the interaction degree of the spline basis functions to K_m =1. Thus, for example, the SETAR(2,1,1) model with i.i.d. noise

$$y_t = \mu + \begin{Bmatrix} \beta_1 y_{t-1} & if\ y_{t-1} \geq 0 \\ \beta_2 y_{t-1} & if\ y_{t-1} < 0 \end{Bmatrix} + \varepsilon_t \qquad (2.4.2)$$

can be rewritten in ASTAR model form as follows

$$y_t = \mu + \beta_1[(+1)(y_{t-1})]_+ + \beta_2[(-1)(y_{t-1})]_+ + \varepsilon_t$$

with the sign "-1" inverting the orientation of the predictor y_{t-1} in the sub-region to the left of the knot, that is threshold point situated at zero. TSMARS, when applied to data arising from say a SETAR(2,1,1) process, can be expected to recover the skeleton (see section 1.3) of the model by adaptively choosing the single lagged predictor, its threshold point and the parameters μ, β_1 and β_2. As noted in Lewis & Stevens (1991), the advantage of this approach over Tong's (1990) methodology for SETAR modelling, is that TSMARS admits continuous models with possibly more than 1 threshold. By contrast, Tong's approach admits discontinuous models that are usually limited to a single threshold, due to the difficulties associated with the threshold selection process.

2.5 Model Estimation Simulation Studies using TSMARS

TSMARS is designed to find autoregressive linear and/or nonlinear structure in a univariate time series, by detecting the presence or otherwise of a threshold (i.e. knot) in the set of lagged predictors. So, by analogy with the simulations studied above for independent data, it is important to test whether this SAS/IML implementation of TSMARS will find the frame when the data are simulated from a linear or nonlinear AR process. To decide this a set of simulations on univariate times series models that have been used in the literature are studied. In all but one case study the first 100 generated sample values of each simulated series are discarded to allow for "burn in".

The simulation results of this chapter are displayed as frames with more detail given in tables later. The displayed frames are in fact the average values of all frames generated by each simulation run. This average frame is computed pointwise for each data value in range of data values in the predictor space (e.g. y_{t-1} running from –2 to 1 in Model (1.3.4)). In addition, this pointwise computation of the frame allows the pointwise calculation of the standard error of the frame at each data value. The two standard error limits are also displayed with the average frame (which is referred to simply as the frame in the relevant figures). We begin with a simulation study of an AR(1) model.

2.5.1 Simulation of a linear AR(1) model

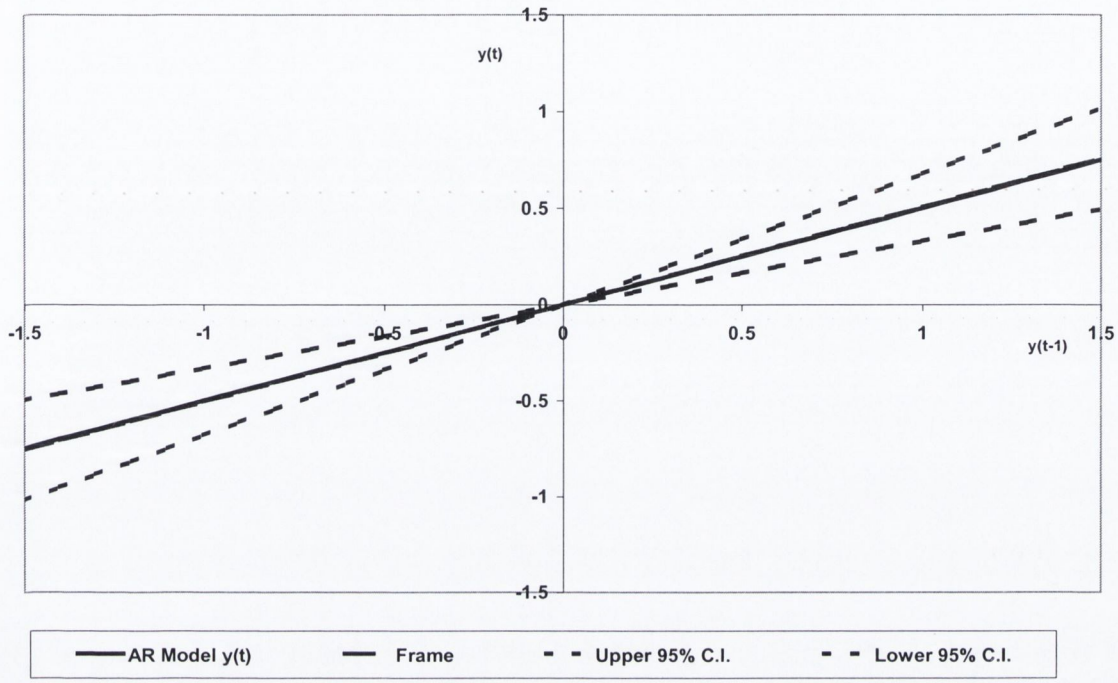
This simulation study examines the ability of TSMARS to identify a simple linear AR(1) model with known coefficients. The example is taken from Lewis & Stevens (1991) and the data are generated from a stationary AR(1) model with autoregressive parameter ρ, $(|\rho| < 1)$, driven by normally distributed noise $\varepsilon_t = N(0, \sigma^2)$

$$y_t = \mu + \rho\, y_{t-1} + \varepsilon_t \qquad t = 1, 2, \ldots, n \qquad (2.5.1)$$

Several simulations were performed for $\rho = 0.5, 0.7$ and 0.9 and with $\mu = 0$ and $\sigma^2 = 1$. Figure 2.5.1.1 shows the frame and Table 2.5.1.1 shows the simulation results for $\rho = 0.5$; results for $\rho = 0.7$ and 0.9

are similar to those reported here for $\rho = 0.5$. Note, in this case the frame is simply the straight line with slope 0.5 equal to the AR parameter value.

Figure 2.5.1.1: AR(1) Model Frame Simulation for ρ = 0.5



In each simulation 100 data sets were generated. TSMARS was then called with response y_t, one lagged predictor y_{t-1} and maximum interaction degree set to 1. Simulation experiments were repeated for two sample sizes of n = 100 and 250. These experiments were repeated again allowing 3 lagged predictors $y_{t-1}, y_{t-2}, y_{t-3}$ and maximum interaction degree of 3. All simulations were conducted with the smoothing parameter set to 3

The frame displayed in Figure 2.5.1.1 is virtually exact, as the line showing the slope of the AR Model parameter is overlaid by the frame. The upper and lower confidence lines are also virtually exact as their slopes agree almost exactly with their true asymptotic values.

Table 2.5.1.1: AR(1) Model Simulation Results for ρ = 0.5

						$Std.Err.(\hat{\rho})$	
No of lagged predictors	Maximum Interaction Degree	Basis function Tolerance	n	Number of AR(1) Models found	$\hat{\rho}$	Actual	True
1	1	$1.5X10^{-2}$	100	99	0.505	0.081	0.087
			250	100	0.505	0.058	0.055
3	3	$1.5X10^{-1}$	100	95	0.503	0.084	0.087
			250	100	0.505	0.058	0.055

Displayed in Table 2.5.1.1 are the number of times an AR(1) model is correctly identified from the 100 simulation data sets. Also given is the average value of the estimated parameter $\hat{\rho}$ and its "Actual"

standard error computed from the correctly identified models. For comparison, the true (asymptotic) standard error for AR(1) data with n observations is also given.

Where possible in this thesis, results are reported for correct models. A correct model is defined as a TSMARS model whose form is identical to the model the data are simulated from. Correct models are identified by comparing the actual form of the variable, knot and sign matrices $v(k,m)$, $\xi(k,m)$ and $s(k,m)$ in the TSMARS model to their true form as defined by the time series model. So, in this particular simulation study TSMARS models found are correct if

$$\hat{y}_t = \hat{\rho}\, y_{t-1}$$

and the form of matrices are $v(k,m) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, $\xi(k,m) = \begin{bmatrix} 0 & 0 \\ 0 & y_{\min} \end{bmatrix}$, and $s(k,m) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$.

That is, correct models involve one linear function in the lagged variable y_{t-1} and a knot at the minimum value of y_t. All other models are defined as incorrect models. If, as on line 1 of Table 2.5.1.1, TSMARS returns one case of the constant mean value model, this TSMARS model is incorrect. Including parameter estimates from incorrect models will bias the average values reported in the results. In particular, including the incorrect model will result in a parameter estimate that is smaller than reported value $\hat{\rho} = 0.505$.

Looking at Table 2.5.1.1, when the number of lagged predictors is 1 and the basis function tolerance is $1.5X10^{-2}$, TSMARS correctly identified 99 and 100 models, for simulations with 100 and 250 observations respectively. The corresponding parameter values are close to their true values and, as expected, tend to it with increasing n. There standard errors follow a similar pattern. When the number of lagged predictors equals 3, the only difference is that a higher tolerance is required to get broadly similar results. Comparing the results in Table 2.5.1.1 with those given in Lewis & Stevens (1991), virtually no difference is observed other than a slight upward bias in the parameter values in Table 2.5.1.1. The results in figure 2.5.1.1 and Table 2.5.1.1 show that the TSMARS models these data correctly and properly.

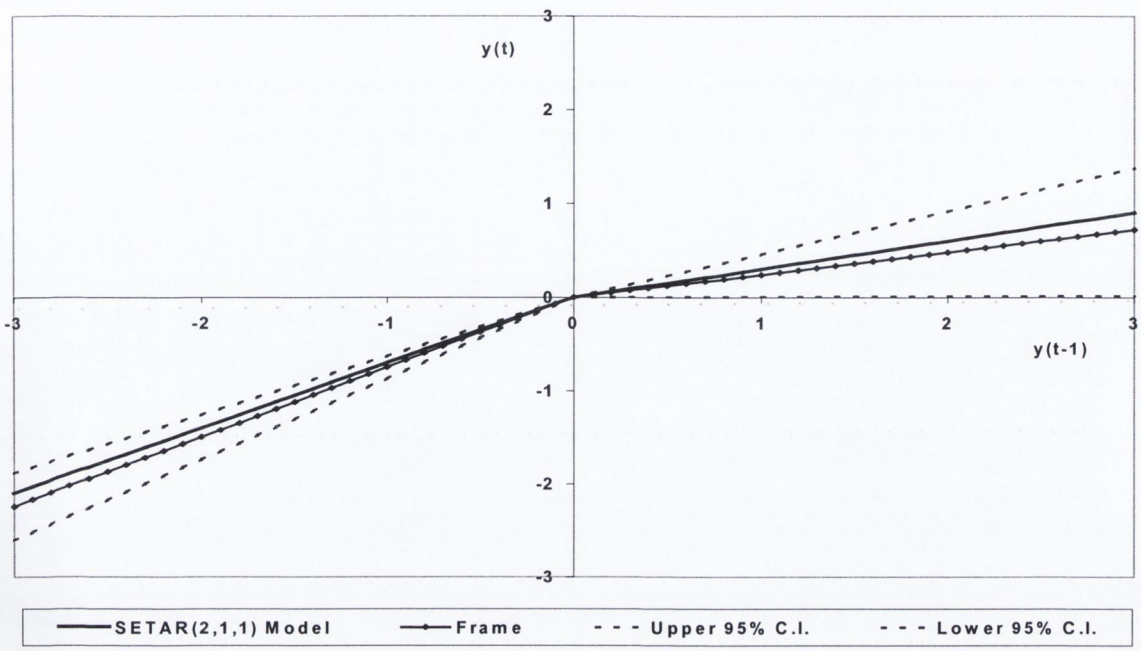
2.5.2 Simulation of a SETAR(2,1,1) model

The second simulation study is once again taken from Lewis & Stevens (1991) and repeated here for comparison. The simulation involves generating 100 data sets each consisting of n = 250, 500 and 750 samples respectively, from the model

$$y_t = \begin{cases} \rho_1 y_{t-1} + \varepsilon_t & \text{if } y_{t-1} \le 0 \\ \rho_2 y_{t-1} + \varepsilon_t & \text{if } y_{t-1} > 0 \end{cases} \qquad (2.5.2)$$

driven by normally distributed noise $\varepsilon_t = N(0, 1/4)$ with parameters values $\rho_1 = 0.7$ and $\rho_2 = 0.3$. TSMARS was applied to each data set with the basis function tolerance set at $2X10^{-3}$ and smoothing parameter set at 3. The frame for the case n = 500 is displayed in Figure 2.5.2.1 while further results are shown in Table 2.5.2.1 where once again the number of correctly identified SETAR(2,1,1) models is given, the average values of the parameters ρ_1 and ρ_2, labelled Est., along with their standard errors labelled Std. Err., for the correct models. With the SETAR model the threshold must also be estimated. This in fact is the knot value identified by TSMARS. The average knot value and its standard error is also given for the correctly identified models which in this case involve two piecewise linear functions in the lagged variable y_{t-1}.

Figure 2.5.2.1: SETAR(2,1,1) Model Frame Simulation for $\rho_1 = 0.7$, $\rho_2 = 0.3$ and threshold = 0.0



It is clear from Figure 2.5.2.1 that the frame slope is very close to the actual slope parameters in each regime of the SETAR(2,1,1) model. More importantly both of these line plots lie well within the upper and lower confidence bands obtained for the simulated frames.

In Table 2.4.2.1 more detail is given. The estimates of ρ_1 and ρ_2 and their standard errors improve with increasing sample size. This behaviour is in line with the results of Lewis & Stevens (1991). A similar trend is also observed with the number of correctly identified models. It is worth remarking that Lewis & Stevens (1991) do not report results for sample sizes under 500, where in this case, the performance of MARS falls below the proof criterion. The threshold value is not reported in Lewis & Stevens (1991).

Table 2.5.2.1: SETAR(2,1,1) Model Simulation Results for $\rho_1 = 0.7$, $\rho_2 = 0.3$ and threshold = 0.0

				$\hat{\rho}_1 = 0.7$		$\hat{\rho}_2 = 0.3$		threshold (knot)	
No of lagged predictors	Maximum Interaction Degree	n	Correct SETAR(1) Models found	Est.	Std. Err. $\hat{\rho}_1$	Est.	Std. Err. $\hat{\rho}_2$	Est.	Std. Err.
1	1	250	48	0.781	0.093	0.169	0.131	-0.010	0.088
		500	71	0.748	0.060	0.239	0.110	-0.019	0.088
		750	84	0.740	0.049	0.310	0.088	-0.064	0.079
3	3	250	46	0.778	0.102	0.160	0.141	-0.009	0.103
		500	66	0.749	0.063	0.238	0.110	-0.017	0.091
		750	81	0.741	0.048	0.320	0.070	-0.062	0.045

However, here it is estimated accurately with a slight negative bias and small standard error. These results show that TSMARS is performing both correctly and properly on sample sizes of 500 and above. This is in line with results obtained by Lewis & Stevens (1991) using Friedman's original MARS code.

2.5.3 Simulation of a EXPAR(1) model

In this simulation the ability of TSMARS to estimate an EXPAR(1) model is examined. The model is

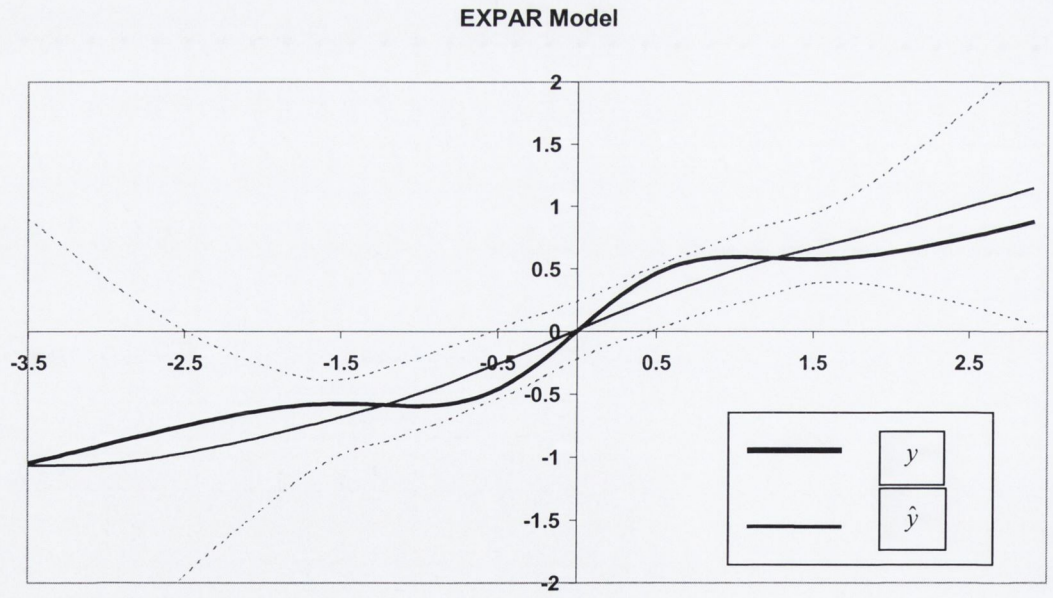
$$y_t = (0.3 + 0.8\, e^{-y_{t-1}^2})\, y_{t-1} + \varepsilon_t \qquad (2.5.3)$$

where ε_t is generated from the Standard Normal distribution. The model is taken from An & Tong (1991) and they noted that it has also been studied elsewhere in the literature. For the purposes of this simulation 50 data sets of length n = 100 were generated from the EXPAR(1) model and TSMARS called with 3 predictors $y_{t-1}, y_{t-2}, y_{t-3}$ and maximum interaction degree of 1. On completion of each TSMARS call, the resulting TSMARS model was used to estimate the frame. This is computed from on a set of ordered x values that cover the range of data values of y_t. In this case the x-range is –3.5 to 3 and predictors are x_1=x-0.1, x_2=x-0.2 and x_3=x-0.3. This process generated 50 sample frames over the range of x values that were summarised to give an average frame response value $\hat{y}(x)$ and its standard error. This frame is plotted in Figure 2.5.3.1 along with its 2 standard error bounds (dotted in red) and the true function

$$y(x) = (0.3 + 0.8\, e^{-x_1^2})\, x_1.$$

From this simulation study TSMARS has over smoothed these data. On the left of the plot the estimate is slightly concave so TSMARS appears to be introducing a slight bend to pick up the kink in the function.

Figure 2.5.3.1: Frame of EXPAR(1) Model Simulation Results



On the right of the plot a similar behaviour is also observed. On these simulations TSMARS tended, in most cases, to pick either an AR(1) model or SETAR(2,1,1) model with equal frequency and this accounts for the shape of the frame. It is clear however that the frame does pick up the general form of the

underlying function and this stays within the upper and lower standard error bands. TSMARS is once again performing properly on these data though with slightly less accuracy than in the AR and SETAR models simulated earlier. It was also observed on a number of calls that the forward stepwise strategy did pick up other basis functions. These however were automatically deleted because they contributed less than 1% to the total sum of squares. Finally, given that the function has smooth transition kinks it might be expected that a STAR(1) model could be a more appropriate to these data

2.5.4 Simulation of a Nonlinear Additive Sine model

This simulation examines the ability of TSMARS to model the nonlinear periodic time series

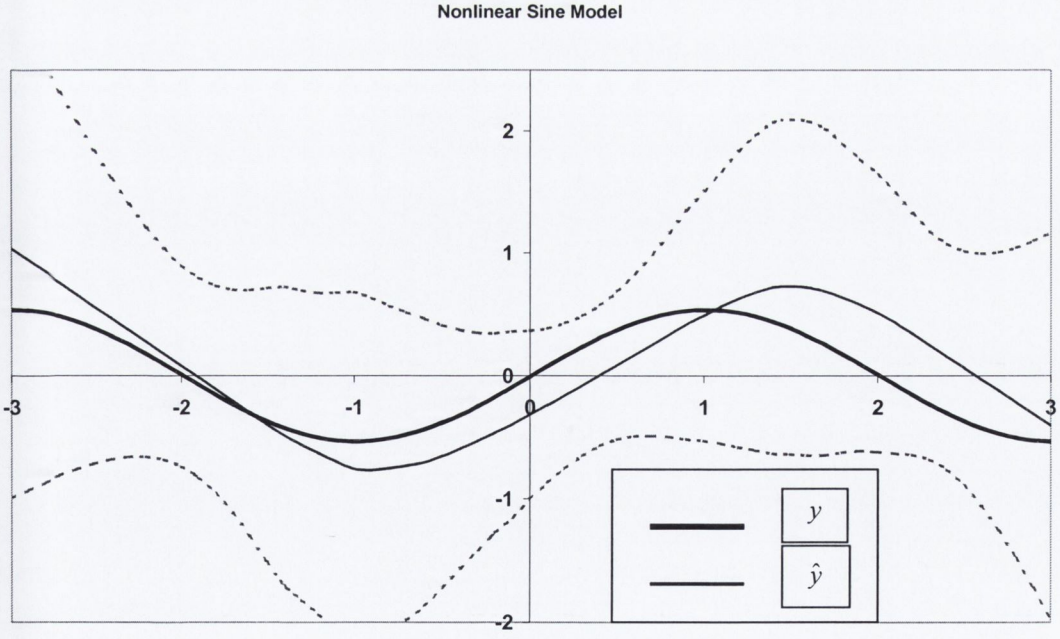
$$y_t = 1.5\,Sin\left(\frac{\pi\, y_{t-}\, 2}{2}\right) - 1.0\,Sin\left(\frac{\pi\, y_{t-3}}{2}\right) + \varepsilon_t \qquad (2.5.4)$$

where ε_t is Standard Normal noise.

This time series was studied by Chen & Tsay (1993a) using the backfitting programs Alternating Conditional Expectation (ACE) of Brieman & Friedman (1985) and the Generalised Backfiting (BRUTO) of Hastie & Tibshirani (1990). They found that both of these programs worked well on time series generated from this model.

In this study 50 data sets, each of length n = 200 were generated and TSMARS called with 3 predictors $y_{t-1}, y_{t-2}, y_{t-3}$ and maximum interaction 1. The resulting TSMARS model was then used to generate the frame over a set of x-values that correspond roughly to the range of values of y_t. The average frame response value $\hat{y}(x)$ and its standard error (dotted in red) are displayed in Figure 2.5.4.1.

Figure 2.5.4.1: Frame of Nonlinear Sine Model Simulation Results



The plot shows the true function as a thick (black) line with the estimated frame as a thin (blue) line. Overall TSMARS generates a frame that follows the true function reasonably well, as the true function is well within the standard error bounds. However, the period of the frame is slightly out of phase. This is

mainly due to the fact that TSMARS, in a number calls chose a model that only contained the lag 3 predictor. The average value of the ratio of GCV for the fitted model to the constant model was reasonable at 0.75 with a standard error of 0.16. This indicates that TSMARS smoothed the noise acceptably while not over-fitting these data. It is clear from the plot that the shape, accuracy and precision of the frame in Figure 2.5.4.1, indicates that TSMARS performs properly on these data.

2.5.5 Simulation of an ARCH model

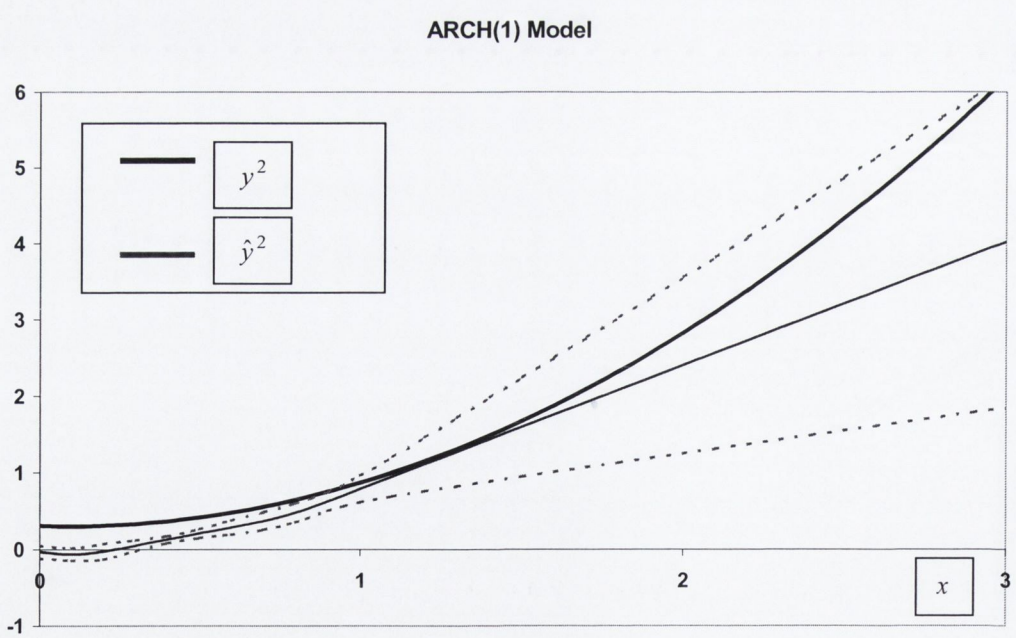
This study examines how well TSMARS can handle changes in variance and has not been considered elsewhere in the literature. The simulations are based on the ARCH(1) model

$$y_t = \sigma_t \varepsilon_t \quad \text{with} \quad \sigma_t^2 = \alpha + \beta y_{t-1}^2. \tag{2.5.5}$$

Specifically, the ARCH parameters are α = 0.3 and β = 0.7 with the noise ε_t generated from the Normal distribution with mean 0 and standard deviation 2/3. Once again 50 data sets were generated according to this model each having a sample size of n = 200. TSMARS is called for each data set with response y_t^2 and a single lagged value y_{t-1}^2. The resulting TSMARS model is then used, as before, to predict the bilinear frame associated with the squares of these ARCH data.

This bilinear modelling approach is preferred over direct ARCH modelling, as the underlying function in the ARCH model is a constant. A frame based on a constant is not particularly informative. In contrast, bilinear modelling captures the changing variance as a curve in the squares of the data values. The TSMARS approximation to this curve provides an informative frame that is more useful to visually judge the quality of this approximation on accepted models.

Figure 2.5.5.1: Bilinear Frame associated with ARCH(1) Model Simulation



In this study, an acceptable model is identified, by making a second call to TSMARS with weighting applied. On this second call, data simulated from (2.5.5) are modelled directly with response y_t and

predictor y_{t-1}. Weights are the Winsorized inverses of the predicted value obtained from modelling y_t^2 in the first call (i.e. $1/\hat{y}_t$). These capture the fact that the variance σ_t depends on t. Now, if the outcome of the first call is correct, weights will be accurate. Therefore, TSMARS should return the constant mean model on this weighted second call. This defines a correct model. Any simulation that does not return a constant mean model on the second call is incorrect and the associated bilinear model is also rejected. When this two-call process is applied to the 50 simulated data sets, it resulted in 31 acceptable bilinear model estimates. The frame plotted in Figure 2.5.5.1 is based on this set of acceptable modes.

The average bilinear frame response value $\hat{y}^2(x)$ and its standard error (dotted in red) are displayed in Figure 2.5.5.1 for the 31 accepted models. Examining the plots it appears that TSMARS does quite well where the x-values are of moderate size but the estimates however do not bend sufficiently at the end points. For the larger values of x this may mean that TSMARS will underestimate in regions where there is a large change in variance in ARCH data. With small x-values TSMARS once again underestimates and would appear to distinguish small variance in the ARCH model simply as noise. Of course this is not a drawback but in some cases a slightly negative value is returned. With the methodology used here this could result in no ARCH model being defined by TSMARS in regions where the variance is small. The simulated frames here show that TSMARS is once again performing correctly and properly. However, the estimates may need some attention where the variance is very large or very small. Methods to deal with this type of situation through moving average components are introduced in a later chapter.

2.5.6 Simulation of a Markov model

The last simulation study in our sequence is taken from Lai & Wong (2001) who used Friedman's (1991a) TSMARS program to look at one-step-ahead forecasts arising from a Markov chain model. Here we repeat their exercise but instead look at the ability of TSMARS to fit the data arising from their Markov chain. The chain $\{y_t\}$ has state space [0,1] and transition p.d.f.

$$p(y|x) = \begin{cases} e^{1-x} & \text{if } 0 \le y \le x \\ e^{1-x} - e^{y-x} & \text{if } y \le x \le 1 \end{cases} \tag{2.5.6}$$

To obtain the next value in the chain it is necessary to evaluate the c.d.f. $F(y_t|y_{t-1})$. We draw u from the uniform distribution on [0,1] and solve the nonlinear equation

$$F(y_t|y_{t-1}) = u.$$

The resulting Markov chain $\{y_t\}$ has a nonlinear regression function

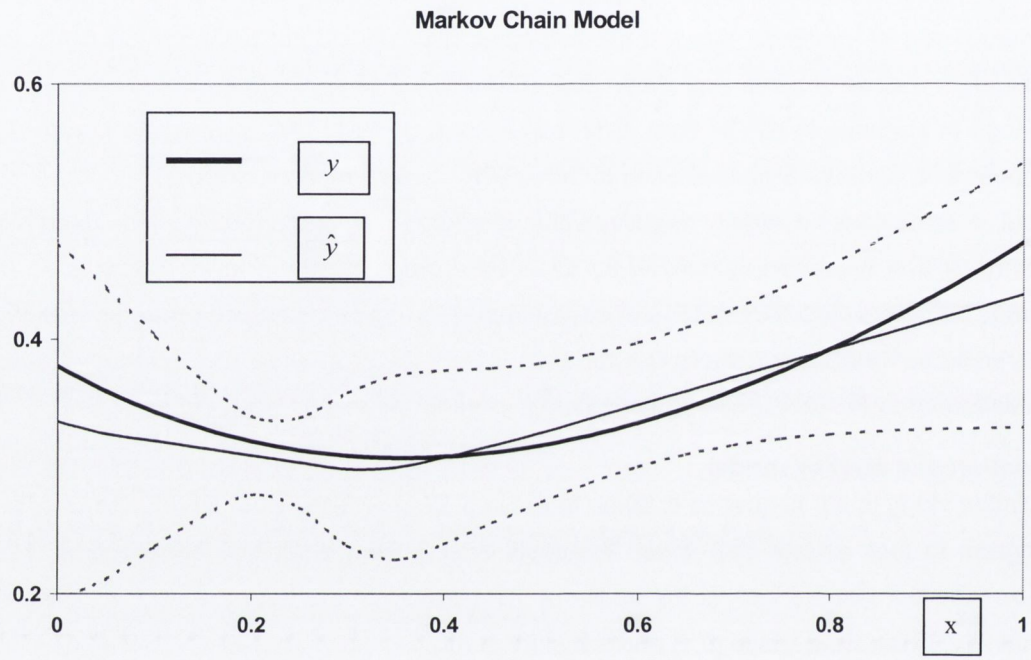
$$E(y_t|y_{t-1}) = y_{t-1} - 1 + \exp(1 - y_{t-1})/2$$

and the residuals are not i.i.d. and do not have normal marginal distributions.

The simulation study involves generating 50 data sets, each containing 300 sample values according to this Markov chain model. TSMARS was called for each data set with one lagged predictor y_{t-1}. Of the 50 calls to TSMARS model 25 were rejected as they only comprised the mean value. The remaining 25 TSMARS models were then used to generate an average frame response value $\hat{y}(x)$. Both, this and its standard error (dotted in red) are displayed in Figure 2.5.6.1 as well as the true regression function $E(y_t|y_{t-1})$.

Examining the plots in Figure 2.5.6.1 it is clear that TSMARS captures the underlying shape of the regression function quite well. This is particularly pleasing given the data set is of moderate size and also the residuals are not normally distributed. That said, at both ends of the x-value range the fit is a little poorer then elsewhere but still under 10% in error. Thus, while TSMARS performs slightly below standard in terms of correct models, it is once again performing properly. This 10% error may account for the poor forecasting ability of TAMARS when compared to the Stochastic Neural Network approach of Lai & Wong (2001).

Figure 2.5.6.1: Frame of Markov Chain Model Simulation



2.6 Concluding Remarks

The MARS algorithm and the details associated with coding it in SAS have been outlined. The ANOVA decomposition was also described. This decomposition is particularly important to the research reported in this thesis, as it provides a measure of the extent of nonlinearity in a given set of data.

With the program in place the remainder of the chapter focussed on 'proving' MARS. Basically, the objective was to ensure MARS and TSMARS would give statistically sound estimates in a variety of situations. Estimates were regarded as statistically sound if they were equivalent to those reported in the literature for Friedman's original code.

The results of the simulations showed that MARS did not find structure where there was only noise and the reverse was also true. Simulations on time series data with TSMARS also gave correct estimates that were accurate and precise. The novel concept of a frame was introduced. This visualisation tool facilitated easy comparison of the original function and the simulated TSMARS model. Based on the frames and other tabulated simulation results TSMARS is judged to perform correctly and properly. That is, MARS is proven as it gives sound estimates that are statistically equivalent to Friedman's original code.

3 Data Transformations and Seasonality

This chapter assesses whether a data transformation or seasonal adjustment is necessary prior to modelling a time series with TSMARS. TSMARS is a flexible model free smoother. Therefore, it is reasonable to infer that predicted values obtained by transforming a time series, followed by modelling and then reversing the transformation, should be close to predicted values obtained by modelling without any transformation. In this chapter the validity of this claim is checked for certain types of data transformation and separately for seasonal adjustment. In the context of TSMARS this claim has not been addressed in the literature. In this chapter it is tackled using a series of novel simulation studies. In later chapters the inferences made will ensure that studies of empirical data will be based on sound modelling principles.

Essentially, TSMARS is checked as follows. Data is simulated from a basic model, such as the SETAR(2,1,1) model. To these data, either a log or exponential transformation is applied or seasonal effect added. This data is then modelled with TSMARS to get the approximation series $\hat{y}_t$. The sequence of transformations are then applied in reverse to this approximation. The true model is fit to the resulting series giving a set of implied parameter estimates. TSMARS will be judged to be unaffected by a transformation or seasonal adjustment, if these implied parameters are within two implied standard errors of their true values.

Seasonal adjustment may also influence the nonlinear characteristics of a time series. As a consequence, poor values of implied parameters may be the result of seasonal adjustment rather than TSMARS modelling. To identify whether seasonal adjustment is responsible, Tsay's threshold F-test (see Appendix, Relevant Statistical Tests) is applied. If this test is negative for a series that possessed a threshold prior to seasonal adjustment, then poor implied parameter values are not due to TSMARS modelling. In this event, TSMARS will also be judged to be unaffected by seasonal adjustment, even though the reported implied parameter values appear poor.

3.1 TSMARS and Data Transformations

Many economic time series are differenced or natural log transformed prior to modelling. These transformations are usually applied to make a series stationary in level and variance respectively. In this section, a number of simulation studies are conducted to see whether it is necessary, to transform data prior to TSMARS modelling. This question is relevant because the quality of the TSMARS approximation could suffer, in both level and variance, by the use of a transformation when its use is inappropriate.

Simulation studies are conducted based on AR and SETAR models respectively. The Box-Cox transformation (in modified form) for number of parameter values ($0 \leq \lambda \leq 1$) combined with ($\Delta = 1$) or without ($\Delta = 0$) 1st difference operator is adopted. Thus a time series y_t, simulated from an AR (or SETAR) model is transformed according to

$$x_{\lambda,t} = \begin{cases} y_t^{\lambda} & \lambda > 0 \\ y_t^{\lambda} \mathrm{g}_e(y_t) & \lambda = 0 \\ \log_e(y_t) & \end{cases}$$

and where this is combined with the 1st differe operator the resulting transformed series is

$$\Delta x_{\lambda,t} = \begin{cases} y_t^{\lambda} - y_{t-1}^{\lambda} & \lambda > 0 \\ \log_e(y_t) - \log_e(y_{t-1}) & \lambda = 0 \end{cases}$$

This last transformation of course corresponds to the period-on-period growth rate of the original series.

On each simulation, once the transformed series is generated it is modelled with TSMARS giving the estimated series $\hat{x}_{\lambda,t}$ $(\Delta\hat{x}_{\lambda,t})$. This is then back transformed giving the TSMARS estimate of y_t, namely $\hat{y}_t$. We then act as though $\hat{y}_t$ is the observed series and fit an AR (or SETAR) model to the series to get parameter estimates. These are called implied parameter estimates, to reflect the fact that they are the values implied from fitting a model of the correct form (e.g. AR(2)) to $\hat{y}_t$. If the series $\hat{y}_t$ is good approximation to the original series, the implied parameter estimates can be expected to be close to their true values. The accuracy of the implied parameters is reported for 100 simulated time series. Their quality reflects the impact of the transformation process on the TSMARS approximation.

In addition to examining the effect of a transformation on a series simulated from AR or SETAR model, the effect of the transformations on data that are transformed with the exponential function is examined. That is, $\exp(y_t)$ is taken as the original series and the effect of the proposed transformation(s) on the TSMARS approximation to this series is also assessed.

3.1.1 Simulations based on an AR model

The time series model adopted for this set of simulations is based on the AR(2) model

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \varepsilon_t \qquad t = 1, 2, \ldots, n \tag{3.1.1}$$

where ε_t is $N(0, \sigma^2)$ white noise. Two sets of parameters are chosen, both having $\mu = 0$ and $\rho_2 = -0.5$ and $\rho_1 = 1$ or 1.5 respectively. These resulting models are borderline stationary and integrated IAR(1) models respectively.

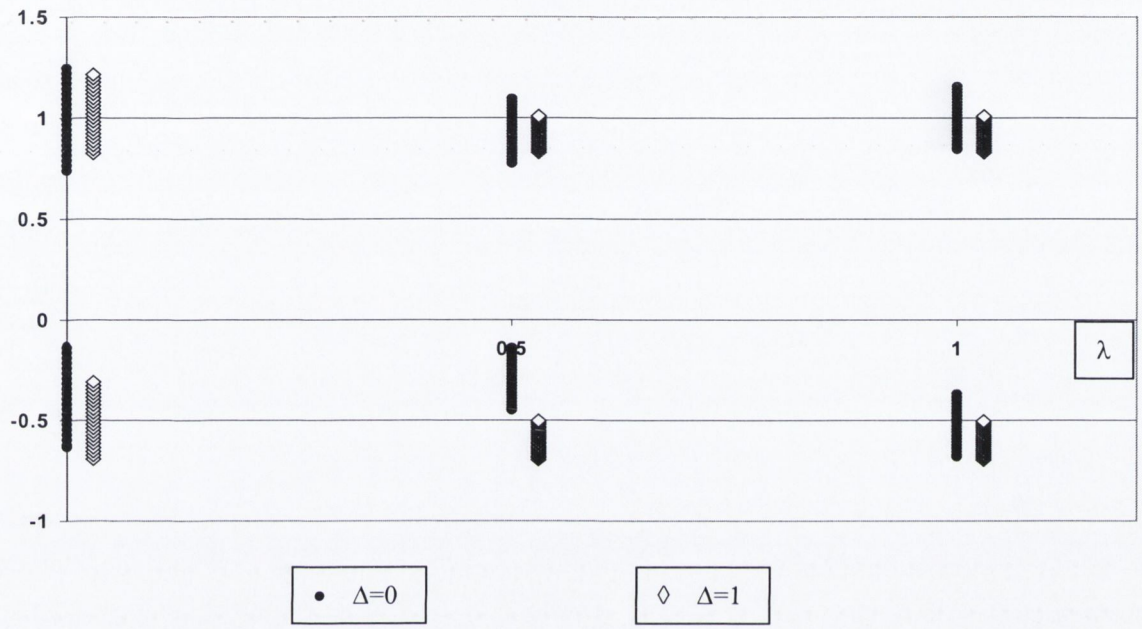
In this study 100 time series, each having n = 100 values are simulated from the model. Each time series is adjusted by adding an appropriate constant, namely $\mu = 1 - \min(y_t)$ to ensure the values all remain positive. The test procedure outlined above is applied to each series for a combination of Box-Cox parameter values and/or 1st difference operator and the implied parameter estimates obtained. For ease of comparison, the mean and two standard error (S.E.) limits, for the $\rho_1 = 1$ and $\rho_2 = -0.5$ case are graphically displayed in Figure 3.1.1.1. The bimodality that arises in the plot for each value of λ reflects this fact. The detailed figures as well as the figures for the $\rho_1 = 1.5$ case are given in the Table Appendix, see Table 3.1.1.1.

In Figure 3.1.1.1 the implied parameter estimates obtained without the Box-Cox transformation (i.e. λ = 1) and with the 1st difference operator are captioned with a diamond symbol. As the Box-Cox parameter λ is decreased no fall off in the accuracy of the implied parameter estimates is discernible, though their precision is slightly affected. Thus, the TSMARS approximation is unaffected by a smooth transformation of these data.

In contrast, when the 1st difference operator is not applied (i.e. Δ = 0) the accuracy and precision of the results (captioned with a thick dot) is influenced, but only to a small degree (for example, in the λ = 0.5

case $\rho_2 = -0.5$ is biased). The effect is least when λ = 1, that is no Box-Cox transformation is applied and raw data is modelled. These observations suggest that differencing can marginally improve estimates on borderline stationary data but inappropriate use of a Box-Cox transformation can lead to slightly poorer estimates. This inference is also borne out by similar results obtained for the IAR(1) simulations given in Table 3.1.1.1.

Figure 3.1.1.1: AR(1) Model Simulation Results $\rho_1 = 1$



Simulations were also carried out on exponential data with the results given in the Table Appendix, Table 3.1.1.2. Once again, where λ = 0 (i.e. the log transformation) and the 1st difference operator used the implied estimates are accurate and precise. However, as λ is increased the accuracy and precision of the estimates suffers. In particular, where the 1st difference operator is not applied (Δ = 0), the accuracy and precision of the estimates is poor. Based on these simulations, the use of appropriate data transformations prior to modelling with TSMARS is worthwhile for borderline stationary and integrated data.

3.1.2 Simulations based a SETAR model

The time series model adopted for this set of simulations is based on the SETAR model. Once again two variations are used, both driven by white noise $\varepsilon_t \sim N(0, 1/4)$. The first model variation is

$$y_t = \begin{cases} \mu + \rho_{11}\, y_{t-1} + \rho_{12}\, y_{t-2} + \varepsilon_t & \text{if } y_{t-2} \geq 0 \\ \mu + \rho_{21}\, y_{t-1} + \rho_{22}\, y_{t-2} + \varepsilon_t & \text{if } y_{t-2} < 0 \end{cases} \qquad t = 1, 2, \ldots, n \tag{3.1.2}$$

having parameters $\mu = 0, \rho_{11} = \rho_{21} = 1, \rho_{12} = -0.7$ and $\rho_{21} = -0.3$. Therefore, this model satisfies the necessary conditions (see subsection 1.2.5) for it to be borderline stationary; since $\rho_{11} = 1.0$ in both regimes.

The second model variation is

$$y_t = \begin{cases} \mu + \rho_{11}\, y_{t-1} + \rho_{12}\, y_{t-2} + \varepsilon_t & \text{if } y_{t-1} - y_{t-2} \geq 0 \\ \mu + \rho_{21}\, y_{t-1} + \rho_{22}\, y_{t-2} + \varepsilon_t & \text{if } y_{t-2} - y_{t-2} < 0 \end{cases} \qquad t = 1, 2, \ldots, n \qquad (3.1.3)$$

having parameters $\mu = 0, \rho_{11} = 1.7, \rho_{21} = 1.3$ and both $\rho_{12} = -0.7$ and $\rho_{22} = -0.3$ as in (3.1.2). Taking 1st differences of this model gives the stationary SETAR(2,1,1) model

$$z_t = \begin{cases} \mu + \rho_{12}\, z_{t-2} + \varepsilon_t & \text{if } z_{t-1} \geq 0 \\ \mu + \rho_{22}\, z_{t-2} + \varepsilon_t & \text{if } z_{t-2} < 0 \end{cases} \quad \text{with} \quad z_t = y_t - y_{t-1}$$

Thus model (3.1.3) with the chosen parameters is an integrated ISETAR(2,1,1) model.

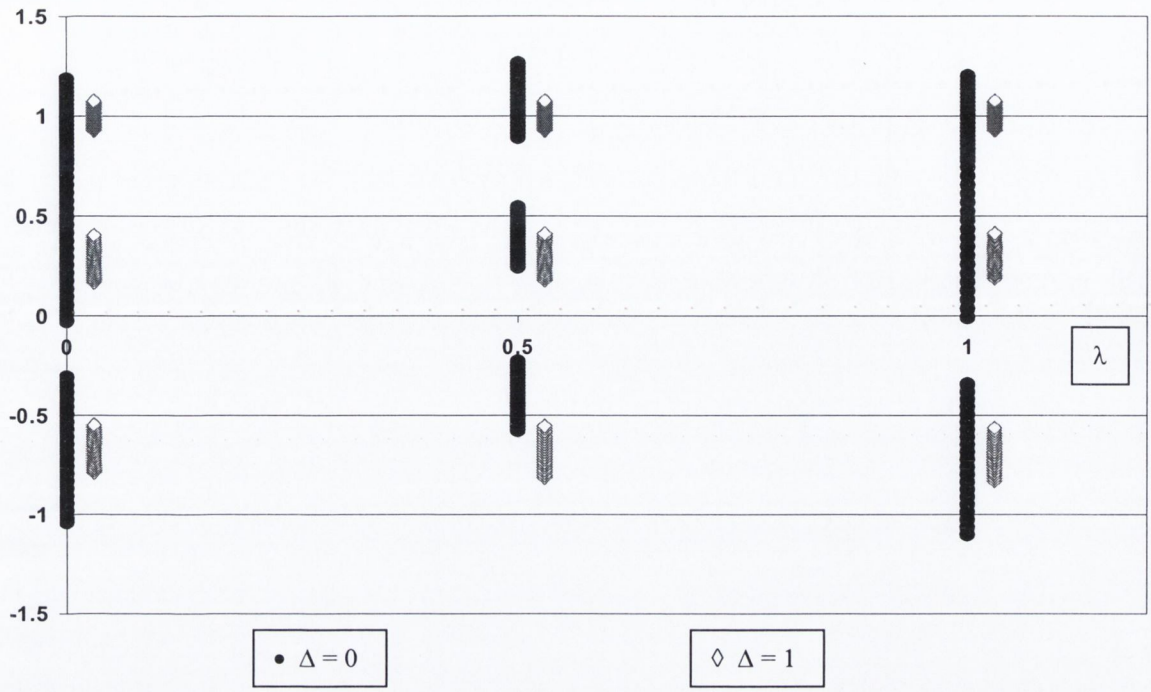
As in last subsection, 100 time series of length n = 500 are simulated according to the appropriate model and adjusted by adding an appropriate constant; namely, $\mu = \min(y_t) + 1$ to ensure the all values remain positive. The test procedure outlined earlier is applied to each series for a combination of Box-Cox parameter values and/or 1st difference operator and the implied parameter estimates obtained. For ease of comparison, the mean and two standard error (S.E.) limits, for the borderline stationary model are graphically displayed in Figure 3.1.1.1 (note, for graphical clarity, the parameter estimates of –0.3 are displayed as 0.3, since otherwise they would be hidden by the –0.5 values). . The trimodality that arises in the plot for each value of λ reflects this fact. The detailed figures, as well as the figures for ISETAR model (3.1.3) are given in the Table Appendix, see Table 3.1.2.1.

In Figure 3.1.2.1 the data are borderline stationary and therefore modelling may benefit from differencing the simulated data. This case, where the 1st difference operator applied without the Box-Cox transformation (i.e. λ = 1), is captioned with a diamond and shown on the extreme right of Figure 3.1.2.1. The (diamond) line segments associated with each parameter show that these implied parameter estimates are accurate and have narrow precision bands (compared to their undifferenced counterparts). In addition, as the Box-Cox parameter λ is decreased no fall off in accuracy or precision is discernible. This pattern is repeated for the integrated ISETAR(2,1,1) data (see Table 3.1.2.1). Thus, as with the AR data, when threshold data should be differenced, the TSMARS approximation of the differenced series is unaffected.

On the other hand when the 1st difference operator is not applied the accuracy and precision of the results (captioned with a dot) is influenced. The line segments made up of dots on the extreme right, show the implied parameter estimates obtained when TSMARS is used to model the raw simulated data. These estimates are less accurate and have larger S.E. bands than their differenced counterparts. This pattern is repeated for λ = 0.5 and 0. These results are similar to those obtained for the AR data.

However, it will be noticed in Table 3.1.2.1 that the results for the integrated model are better than those obtained for the borderline model. Valid models were not distinguished for models based on the raw (i.e. undifferenced) data. In this instance, it is clear that the Box-Cox transformation has no effect. Therefore TSMARS always finds a good approximation to this integrated data. From the results in Table 3.1.2.1, it appears that differencing does not affect the accuracy or precision of estimates on integrated data and using a Box-Cox transformation may well be futile.

Figure 3.1.2.1: SETAR(2,1,1) Model Simulation Results



Simulations were also carried out on exponential data with the results given in the Table Appendix, Table 3.1.2.2. In this case the empirical size of the implied parameter estimates is obtained when $\lambda = 0$ (i.e. the log transformation) and the 1st difference operator is used. The relevant implied parameter estimates are accurate and precise. It is clear that for increased values of λ the accuracy and precision of the estimates does suffer in the integrated data case. Looking at the left side of the table, that is where the 1st difference operator is not applied ($\Delta = 0$), the accuracy and precision of the estimates, is poorer for integrated data. The results observed here are similar to those obtained for the AR model where the data are borderline stationary. The similarity is not observed for integrated data.

Based on the results, it is clear from this study that TSMARS estimates are not influenced by a data transformation.

3.1.3 Summary

The simulation studies conducted on borderline stationary and integrated data were designed to check if TSMARS estimates were influenced by a data transformation. It was shown that differencing does not greatly improve the accuracy of estimates but can improve their precision on borderline stationary data. This conclusion did not hold up for integrated data though differencing did not improve the estimates. Thus TSMARS will in general 'adapt' to produce regression models that account for the level of an integrated time series without the explicit need to difference. The variance of the estimate can suffer, as the standard deviation of the parameter estimates obtained without differencing, was sometimes two or more times that obtained with differencing. Therefore, these results infer that it is best to difference time series that appear to be integrated prior to modelling with TSMARS. This will ensure precision is maintained.

Using a Box-Cox transformation does not appear to affect TSMARS estimates but the log transformation can help when modelling growth data to stabilise the variance. This means that economic time series in particular should be 'log' tested prior to modelling with TSMARS.

3.2 Seasonal Adjustment and TSMARS

Benchmark procedures such as X11 (Shiskin 1967) or Tramo-Seats (Gomez & Maravall 1996) are routinely used to seasonally adjust economic time series. The seasonally adjusted series or 'signal' is generally regarded as the true representation of the underlying process that is not disturbed by seasonal fluctuations. Short-term effects such as year-on-year or period-on-period growth rates, computed from the signal, reflect changes in the underlying process and are often more useful than their counterparts computed on the observed data.

In general there are two approaches to time series analysis for seasonal data. Either, the observed time series can be seasonally adjusted and a model for the signal estimated, or a full model incorporating seasonal effects can estimated directly from the observed data. TSMARS can be applied based on either approach and the quality of the resulting fit may be dependent upon the approach adopted. In this section, prior seasonal adjustment followed by TSMARS modelling is compared to direct modelling, to decide which approach gives better estimates. To assess this question a number of simulation studies are conducted on simple seasonal time series models.

Each simulation study conducted is based on an AR(1) or SETAR(2,1,1) model. To this basic model an s-period seasonal fluctuation $\rho_2\, g(t,s)$ is added of the form $\rho_2\, y_{t-s}$, or $\rho_2 \begin{cases} 1 & \text{tmod(s)} = 2 \\ 0 & \text{otherwise} \end{cases}$, or $\rho_2\, \text{Sin}(2\pi t / s)$.

Thus, for example in the AR(1) case 3 types of model are simulated, namely

Model 1: $y_t = \rho_1\, y_{t-1} + \rho_2\, y_{t-s} + \varepsilon_t$

Model 2: $y_t = \rho_1\, y_{t-1} + \rho_2 \begin{cases} 1 & \text{tmod(s)} = 2 \\ 0 & \text{otherwise} \end{cases} + \varepsilon_t$

Model 3: $y_t = \rho_1\, y_{t-1} + \rho_2\, \text{Sin}(2\pi t / s) + \varepsilon_t$

with ε_t distributed N(0,1). Note that Model 1 is the parsimonious seasonal autoregressive model AR(1)$(1)_s$. Model 2 defines seasonality that depends explicitly on the 2nd season and Model 3 allows the seasonality to vary smoothly over time.

Based on the chosen model a quarterly (s=4) time series y_t of length n+56 is simulated and a constant added to the series $\mu = 1 - \min(y_t)$ to ensure the values stay positive. This series is checked for significant seasonality at the 1% level by regressing it on seasonal dummies $D_{i,t} = \begin{cases} 1 & \text{tmod(s)} + 1 = \text{i} \\ 0 & \text{otherwise} \end{cases}$ $(i = 1 \ldots s)$ according to the model

$$y_t = \delta_1\, D_{1,t} + \delta_2\, D_{2,t} + \delta_3\, D_{3,t} + \delta_4\, D_{4,t} + u_t \qquad (3.2.1)$$

(u_t is white noise) and computing the F-statistic based on the corrected regression sum of squares. This is used to test whether all the seasonal means $\delta_i = 0$. Only those simulated series showing significant seasonality are used for this assessment.

The seasonal adjustment operator adopted is the quarterly 57-term linear symmetric moving average approximation to X11, of Laroque (1977), given by the formula

$$S_{28} = c_0 + \sum_{i=1}^{28} c_i \left(B^i + F^i \right)$$

where B and F are the backward and forward shift operators respectively. The weights c_i used in the approximation are reproduced in Table 4.1 of Franses (1996). This approximation is preferred to X11 for seasonal adjustment tests because it is linear and not data dependent. In contrast X11 adapts to the characteristics of a time series and so cannot be relied upon to make an identical transformation to each simulated series. Resulting conclusions may therefore be confounded.

Each time series is simulated with n+56 values. This is to ensure that seasonal adjustment of the middle n values (i.e. excluding the first 28 and last 28 values) is not spoiled due end-effects. In these simulation studies, only the middle n values of the original simulated series are used for modelling and comparison of results.

In order to assess the relevance of seasonal adjustment to each seasonal time series the following two methods are adopted:

1. Direct: TSMARS is called using the n middle values of a simulated time series that possesses significant seasonality with predictors $y_{t-1} \ldots, y_{t-s}$ and $g(t,s)$ appropriate to the model type (i.e. 1-3). The gives an approximation $\hat{y}_t$. We act as though this is the observed series and fit the original AR(1) or SETAR(2,1,1) with s-period seasonal fluctuation model to get implied parameter estimates.
2. Seasonal Adjustment: The alternative involves generating the seasonally adjusted series $x_t = S_{28}(y_t)$ and the seasonal series s_t for the middle n simulated time series values. The seasonally adjusted series is then modelled using TSMARS with lagged predictors $x_{t-1,\ldots,}x_{t-2s}$ giving an estimated series $\hat{x}_t$. To this we add back the seasonal series giving the approximation $\hat{y}_t = \hat{x}_t + s_t$ and once again assume this to be the observed series. The original AR(1) or SETAR(2,1,1) model with s-period seasonal fluctuation is then fitted to this series and implied parameter estimates obtained.

The assumption underlying this assessment of both methods is that implied parameter estimates should be close to the true model parameter values and the degree of closeness indicating which method is best.

It is important to draw attention to the fact that the moving average seasonal adjustment method affects the underlying dynamics of a time series. This is due to y_t and ε_t being independent, but their seasonally adjusted values are not (see Hylleberg 1992, p36). As a consequence, the ACF of the seasonally adjusted series tends to die away slowly. To attempt to capture this behaviour the number of lagged predictors used to model the seasonally adjusted series x_t is increased to 2s and the threshold parameter (used to control basis function dependence) in decreased to $1.5X10^{-8}$.

3.2.1 Simulation based on a seasonal AR(1) model

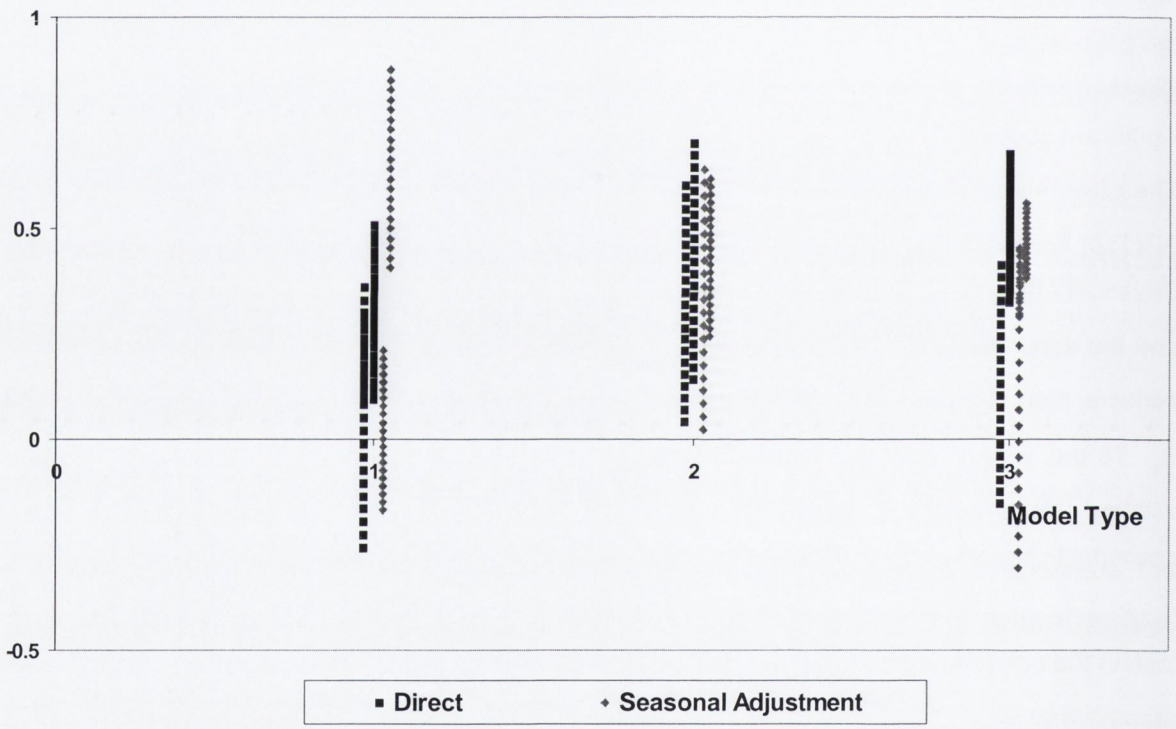
The time series model adopted for this set of simulations is based on the AR(1) model augmented by the seasonal fluctuation $\rho_2\, g(t,s)$, giving the model

$$y_t = \mu + \rho_1 y_{t-1} + \rho_2 g(t,s) + \varepsilon_t \qquad t = 1, 2, \ldots, n \tag{3.2.2}$$

where ε_t is $N(0, \sigma^2)$ white noise. The parameters are $\mu = 0,\ \sigma = 1$ with the remaining parameters based on the chosen model (1-3), that is, the seasonal fluctuation $g(t,s)$. The AR(1) parameter ρ_1 in each model is constant. For Model 1, ρ_1 = 0.15 while in Model 2 ρ_1 = 0.4 and in Model 3 ρ_1 = 0.2. For each of these values the seasonal parameter ρ_2 takes values ¼, ½ and ¾ respectively. This set of nine simulation runs, each consisting of 100 simulated series, is repeated for both the Direct and Seasonal Adjustment methods, giving a 18 different sets of results for 18 separate simulation runs.

Each simulation consist of generating a sufficient number of time series values of length n = 100 + 56 according to each model and testing each for seasonality. This is repeated until 100 series are generated showing significant seasonality. Each of these series is modelled either directly or by prior seasonal adjustment as outlined above. For the worst case, namely ρ_2 = ¼, the two standard error (S.E.) bands for

Figure 3.2.1.1: Implied Parameter Estimates for AR(1) Model with Seasonality (worst case shown)



each parameter estimate are shown in Figure 3.2.1.1 across all three model types. Note, for visual clarity, the dot and diamond lines are placed slightly to the left and right respectively of their Model Type value while the mean is at the centre of the band. The results for the Direct Method are shown with a thick line of dots while those for Seasonal Adjustment are captioned with a line of diamonds. In each case two lines are shown, each represents the standard error band of that parameter estimate, ρ_2=0.25 while ρ_1=0.15, 0.4 and 0.2 for each model type respectively. The bimodality that arises in the plot for each model type reflects this fact. More results are given in Table 3.2.1.1 (see Table Appendix) which also gives the Mean

(over the 100 simulations) Residual Root Mean Square Error (RMSE) computed from y_t and the relevant approximation $\hat{y}_t$.

The results from Model 1, where the seasonality is stochastic, show the direct method, that is modelling the simulated series using TSMARS, gives implied parameter estimates that are closest to the true values. Both methods produce estimates of the AR(1) parameter ρ_1 with a slight bias but standard deviations are similar. It is striking that the implied seasonal parameter estimate (i.e. ρ_2) is significantly biased for the Seasonal Adjustment method (top left-hand diamond column) for Model 1. In Model 2 the seasonality is categorical and induced via a level shift in season 2. Once again the estimate is biased. For Model 3, which displays periodic seasonality this pattern is repeated.

Table 3.2.1.1 (see Table Appendix) shows the results from all simulations. Both the accuracy and precision of the estimates is shown to improve as the seasonal parameter (i.e. ρ_2) is increased. Thus, as the seasonality is more readily distinguished both methods perform equally well. The Mean RMSE results for all 3 models is always just under 1.0 for the Direct Method while it is about 0.8 for the Seasonal Adjustment method. Given the innovation standard deviation is σ = 1, it would appear that the Seasonal Adjustment method tends to over fit the simulated data.

The conclusion from these seasonal AR(1) based simulations is that the presence of seasonality has not influenceed the quality of the TSMARS approximation.

3.2.2 Simulation based on a seasonal SETAR(2,1,1) model

The time series model adopted for this set of simulations is based on the SETAR(2,1,1) model augmented by the seasonal fluctuation $\rho_2\, g(t,s)$, giving the model

$$y_t = \mu + \begin{cases} \rho_{11}\, y_{t-1} \\ \rho_{12}\, y_{t-1} \end{cases} + \rho_2\, g(t,s) + \varepsilon_t \quad \begin{matrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{matrix} \qquad t = 1, 2, \ldots, n \tag{3.2.3}$$

where ε_t is $N(0, \sigma^2)$ white noise. The parameters are $\mu = 0$, $\sigma = ¼$ while the remaining parameters are based on the chosen model (1-3), that is, on the seasonal fluctuation $g(t,s)$.

In these simulations 100 series of length n = 500 + 56, showing significant seasonality are generated. Each series is modelled either directly or by prior seasonal adjustment. The two standard error (S.E.) band, with the central mean value, of the resulting implied parameter estimates for Model 1 only, are reported in Figure 3.2.2.1, with results for Models 2 and 3 in Table 3.2.2.1 (see Table Appendix). Thus in Figure 3.2.2.1 three lines are shown, each represents the standard error band of that parameter estimate. The trimodality that arises in the plot for the regular parameters $\rho_{11} = 0.2, \rho_{12} = 0.1,$ and each respective value of the seasonal parameter $\rho_2 = 0.25, 0.5$ and 0.75 reflects this fact.

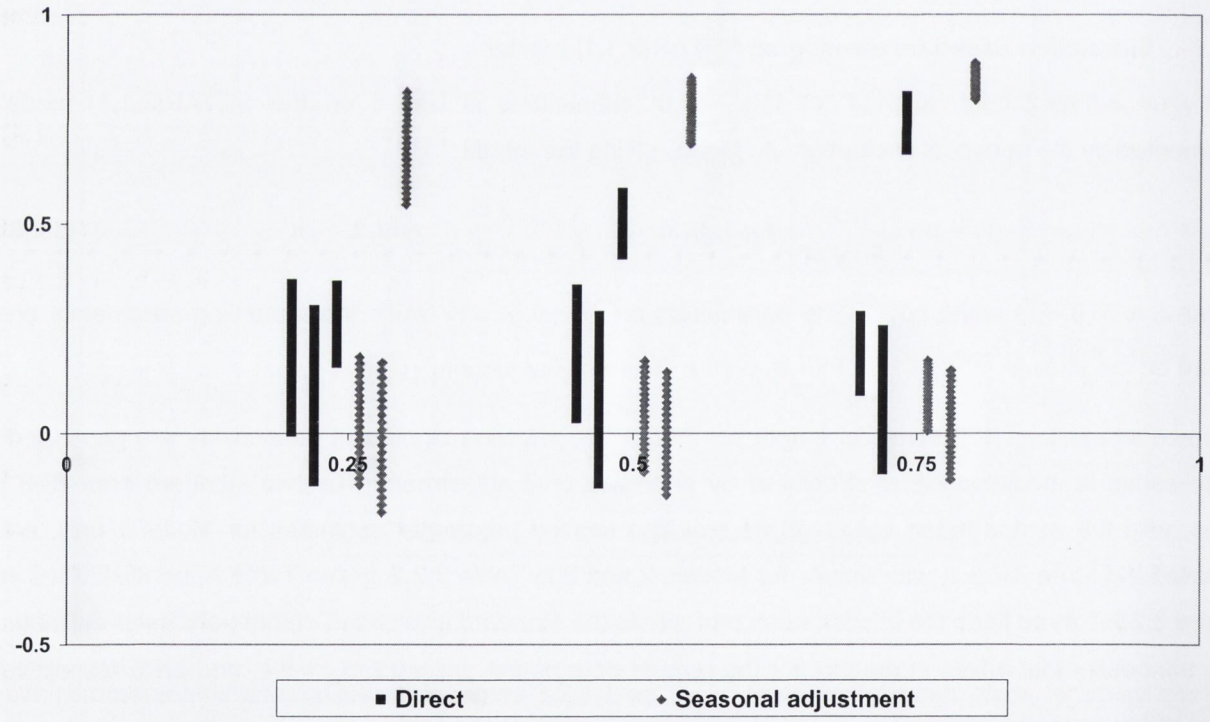
The distribution of the implied parameter estimates displayed in Figure 3.2.2.1 for Model 1 show that the Direct Method is much better than the Seasonal Adjustment Method. The Seasonal Adjustment Method gives estimates that are particularly biased for the seasonal parameter ρ_2 (c.f. the three columns captioned with a diamond running across the top of the figure). This bias decreases as ρ_2 increases and

the seasonality gets stronger; this is demonstrated by the Seasonal Adjustment (diamond) and Direct (dot) columns getting closer as ρ_2 increases.

The results displayed for Model 1 are the worst set obtained. Results from Models 2 and 3, where the seasonality is deterministic, are generally good (see Table 3.2.2.1). However, for Model 2, the Seasonal Adjustment method gave Residual RMSE values that are poor, when compared to the standard deviation of the noise σ = ¼. This suggests a potential dilemma; that is, either (a) the estimates may be poor as a consequence of seasonal adjustment, or (b) seasonal adjustment is working correctly and TSMARS modelling of these data has caused problems.

To answer this dilemma, Table 3.2.2.1(a) reproduces from Table 3.2.2.1, implied parameter values obtained using the Seasonal Adjustment method on Model 2 with ρ_{12} = 0.75. These are contrasted with implied parameter values obtained from the seasonal adjusted series, that is $x_t = S_{28}(y_t)$. These values are obtained by regressing x_t against left and right threshold basis functions (threshold = 0) of x_{t-1} and an indicator series that has value 1 in season 2 and is 0 otherwise. Also, provided are the actual empirical estimates (that is, those based on the seasonal SETAR(2,1,1) regression for y_t) of the parameters.

Figure 3.2.2.1: SETAR(2,1,1) Model 1 with Seasonality Simulation Results (worst case)



It is clear from Table 3.2.2.1(a) that seasonal adjustment is effective since the seasonal parameter ρ_2 = 0, for x_t, as expected. However, the associated SETAR(2,1,1) parameter ρ_{11} is accurate but has a very large standard deviation. This induces the poor precision in the corresponding estimate (0.304) from Table 3.2.2.1. Since the empirical estimate of this parameter is correct (0.70, 0.116), it must be inferred

that seasonal adjustment has altered some of the characteristics of the simulated time series x_t. Thus the fault in the modelling is not due to TSMARS but to prior seasonal adjustment which has altered the underlying nonlinear characteristics of the regular SETAR(2,1,1) data.

Table 3.2.2.1(a): Seasonal Adjustment effects on Model 2 Parameters

	ρ_{11} =0.7		ρ_{12} = 0.3		ρ_2 = 0.75	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Seasonal Adjustment (Table 3.2.2.1)	0.51	0.304	0.33	0.027	0.73	0.018
Seasonal Adjusted Series x_t	0.74	0.600	0.45	0.045	0.00	0.001
Empirical Estimates (y_t)	0.70	0.116	0.30	0.020	0.75	0.014

These results and associated observations support the view that the Direct Method gives the most reliable TSMARS estimates when the data are seasonal. Also, since the Direct Method worked relatively well on all three models we can say that there is no preference for a particular model type.

3.2.3 Simulation based on regime dependent seasonal SETAR(2,1,1) models

The seasonal SETAR(2,1,1) model (3.2.3) studied above assumes that seasonality is constant across the two separate regimes of the process. This assumption can be relaxed allowing the seasonality to be different in different regimes. The resulting regime dependent seasonal SETAR(2,1,1) model has different seasonal fluctuations $g_1(t,s)$ and $g_2(t,s)$, in respective regimes and takes the form

$$y_t = \mu + \begin{matrix}\rho_{11}\, y_{t-1} \\ \rho_{21}\, y_{t-1}\end{matrix} + \begin{matrix}\rho_{12}\, g_1(t,s) \\ \rho_{22}\, g_2(t,s)\end{matrix} + \varepsilon_t \qquad \begin{matrix}\text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0\end{matrix} \qquad t = 1,2,\ldots,n \tag{3.2.4}$$

where ε_t is $N(0,\sigma^2)$ white noise. This model is quite general, in that, the form of the seasonal function as well as the seasonal parameters can be different in each regime. Generalisations to three or more regimes with different seasonal fluctuations in each regime are obvious.

The relevance of this model lies in the empirical observation that for economic time series, the business cycle fluctuation does not appear to be independent of seasonal fluctuation (see Franses 1996). It is however unclear whether the parameter simply changes its value in different regimes, or the form of seasonal fluctuation function changes across the business cycle; say from stochastic to a more regular movement.

The results from the set of simulations based on the regime dependent seasonal SETAR(2,1,1) model are given in Table 3.2.3.1 (see Table Appendix). These assess whether original or seasonal adjusted data, simulated from the relevant model, is more appropriate to modelling with TSMARS. The length and number of simulated series generated is identical to that adopted for the seasonal SETAR(2,1,1) model in subsection 3.2.2. The parameters once again are $\mu = 0$, $\sigma = ¼$ with the remaining parameters based on the chosen model type.

Among the three model types adopted, Model 3 is simplest in that the seasonality is regular in both regimes. This accounts for the fact that it gives implied parameter estimates that are best. In contrast

Models 1 and 2 incorporate stochastic seasonality that changes across the regimes. This results in TSMARS producing poorer implied parameter estimates. The Seasonal Adjustment Method has performed poorest overall. More parameter estimates tend to be biased using this method. However, its Residual RMSE values are good. Therefore, unlike the Model 2 case in the previous subsection, neither method has affected the nonlinearity of the underlying SETAR(2,1,1) series. Once again, these observations support the view that the Direct Method is a slightly better modelling approach than prior seasonal adjustment.

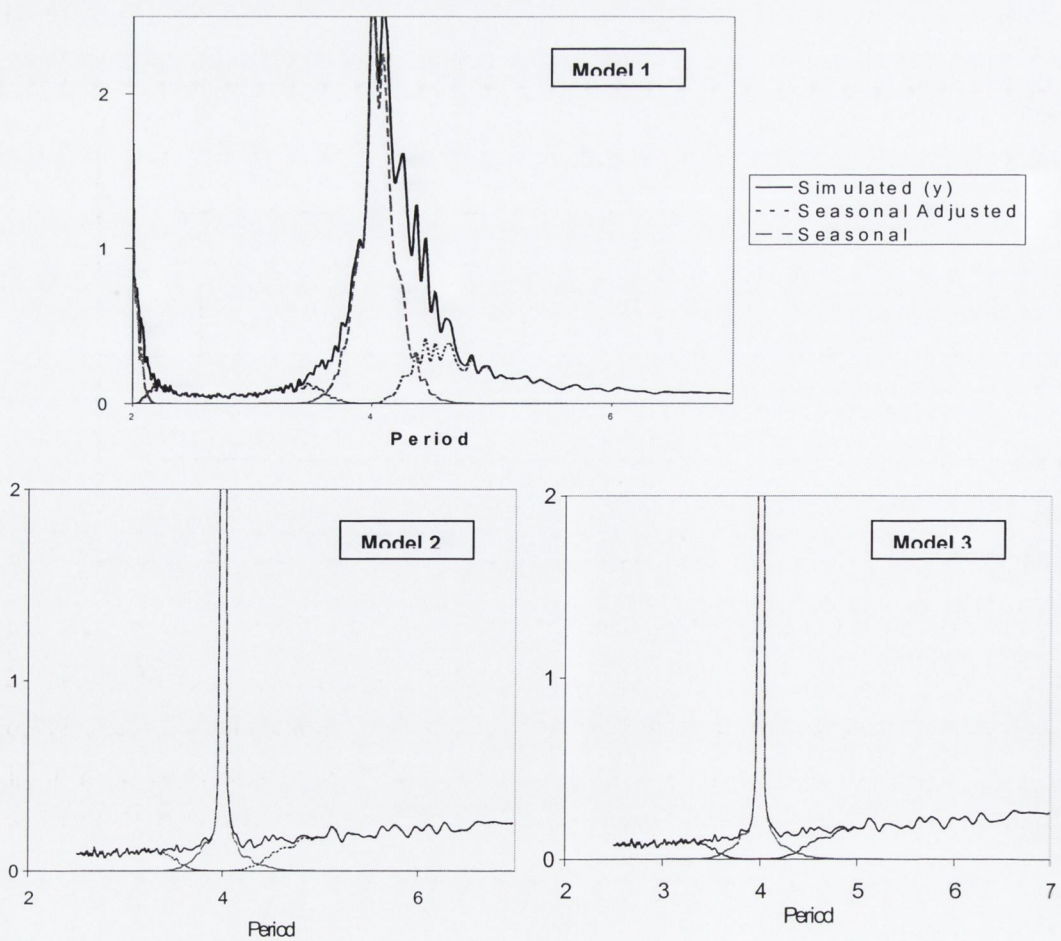
3.2.4 Orthogonality and nonlinearity of seasonal and non-seasonal time series

The validity of the simulation studies conducted in this section relies on two assumptions:

- The linear seasonal adjustment operator S_{28} will produce seasonal and seasonal adjusted series that are orthogonal. That is, they will each possess cycles whose periods are independent.
- The application of a linear seasonal adjustment operator, such as S_{28}, to a linear series will result in a seasonal adjusted series that is also linear, while its application to a nonlinear series will result in a seasonal adjusted series that is nonlinear.

The Gain function (see Ladiray & Quenneville 2001) of X11 and by implication S_{28} is sufficient to justify the claim that seasonal and seasonal adjusted series will be orthogonal when the input series is linear. A linear moving will produce a linear seasonal adjusted series but its autocorrelation pattern will, as mentioned earlier, be affected (see Hylleberg 1992). Therefore simulation results for the AR(1) models are well-founded since these two assumptions are not violated.

Figure 3.2.4.1: Mean Spectra of Seasonal SETAR(2,1,1) model ($\rho_2 = 0.75$, see Table 3.2.2.1)



Simulations results based on the nonlinear SETAR(2,1,1) models may however be open to question. Here, there is no guarantee that the seasonal and nonseasonal series will be orthogonal or that the seasonal adjusted series will retain threshold nonlinearity. Seasonal adjustment has been observed to interfere with the nonlinear characteristics of a time series. For example, Ghysels & Perron (1993) show that seasonal adjustment can smooth away level shifts while de Bruin (2002) has demonstrated that the smoothing constant in the SEASTAR model is deflated.

In order to check orthogonality a smoothed spectrum (using a Parzen window, see SAS/ETS User's Guide 1999) is computed for each of the 100 simulated series (with ρ_2 = 0.75) used to generate the results in Table 3.2.2.1 (see Table Appendix). The mean of these 100 spectra is computed across the frequency/period range and the resulting Mean Spectrum for each model type is displayed in Figure 3.2.4.1. The Mean Spectrum is computed for the original series, the seasonal adjusted series and the seasonal series.

It is clear from the spectral plots in Figure 3.2.4.1 that on average seasonal adjustment has separated quarterly (I.e. period = 4) cycles from longer cycles associated with the regular SETAR(2,1,1) part of the of the model. Since all plots show that the Seasonal Adjusted series has no power at Period = 4, while the Seasonal series has 100% of the power, these two series are orthogonal. That is, both of these series carry cycles that are largely independent of one another.

To see whether the seasonal adjustment operator S_{28} affects the nonlinear nature of the simulated series Tsay's F-test (Tsay 1989) is applied at lag 1. The test checks for the existence of one or more thresholds against a linear alternative (see Appendix, Statistical Test for further details). The test is applied to both the simulated series y_t and the corresponding seasonal adjusted series x_t. The number of series that remain nonlinear after seasonal adjustment is given in Table 3.2.4.1 (see Table Appendix). A small section of this table is reproduced here for discussion purposes, labelled Table 3.2.4.1(a).

Table 3.2.4.1(a): Nonlinearity of Seasonal Adjusted SETAR(2,1,1) Model with Seasonality Results

Original Series (y_t)		Seasonal Adjusted Series (x_t)					
		1%			5%		
		Linear	Non-linear	Total	Linear	Non-linear	Total
Seasonal Parameter (ρ_2)		Model 1					
0.25	Linear	44	7	51	21	6	27
	Nonlinear	14	35	49	21	52	73

To interpret the results, consider the top right-hand corner of Table 3.2.4.1(a) where ρ_2 = 0.25 and Tsay's F-test is applied at the 5% level. In this case threshold nonlinearity was detected in 73 of the 100 simulated series. On seasonal adjustment 52 of these remained nonlinear while 21 were found to be linear. At first sight, this would indicate that seasonal adjustment has changed the characteristics of 21 series. This however is not the case, since the test only has about 75% power. So, approximately 55 of

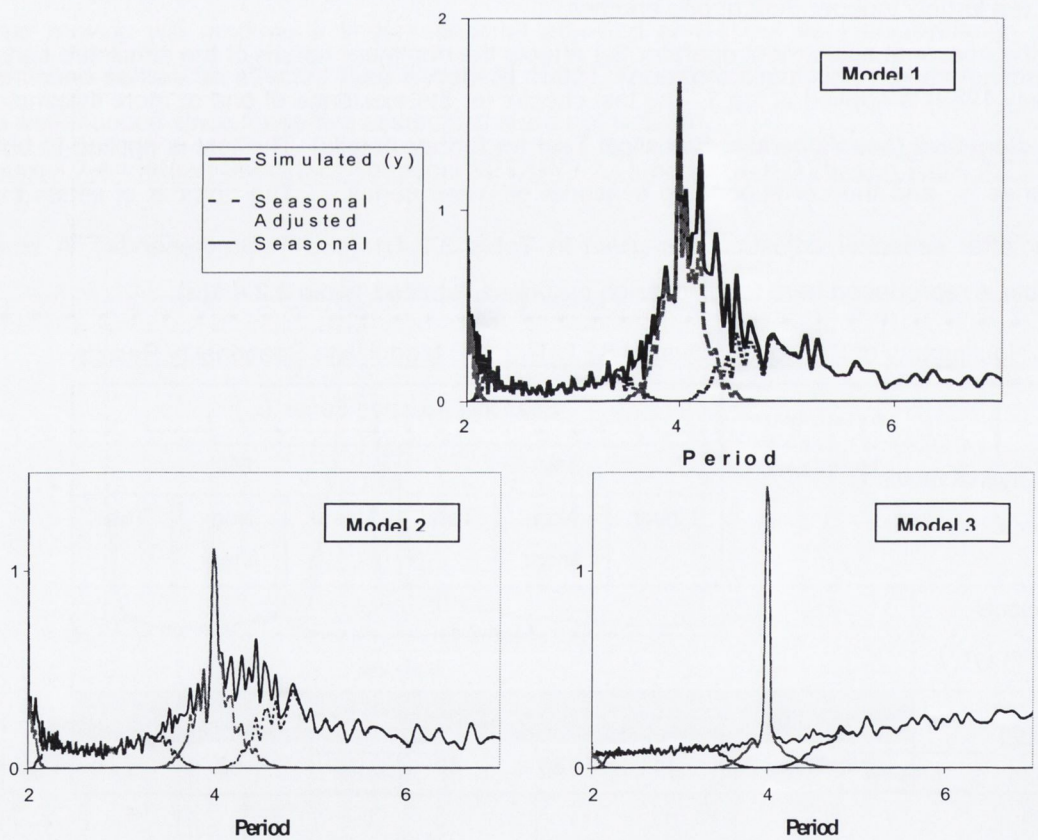
the 73 seasonally adjusted series will appear nonlinear after application of Tsay's test. Thus, 52 as against an expected 55, seasonally adjusted series were found nonlinear. Therefore, seasonal adjustment did not affect the nonlinear characteristics (i.e. did not remove the threshold) of the seasonal adjusted series.

Looking at the rest of the results for Model 1, those at the 5% level show that fewer nonlinear models are detected as ρ_2 is increased. This is due to the seasonal element of the series dominating the regular SETAR(2,1,1) part. This causes the test to pick up the linear seasonal element in preference to the nonlinear element. Other than this, there appears to be no discernible evidence that the nonlinear characteristics have been altered by seasonal adjustment.

The results for Models 2 and 3 at the 5% level are similar to those for Model 1. Once again about 70-75% of series tested show evidence of a threshold. Of these roughly 2/3$^{rd's}$ of the seasonal adjusted series retain the threshold. Based on the power of the test, the conclusion for these models is similar to that for Model 1.

The results at the 1% level for all three models are similar to those obtained at the 5% level except that more simulated series test as linear. This clearly is to be expected.

Figure 3.2.4.2: Mean Spectra of Regime Dependent Seasonal SETAR(2,1,1) model (ρ_{22} = 0.25)



Turning to the regime dependent seasonal SETAR(2,1,1) model, the mean spectral density where ρ_{22} = 0.25 is displayed in Figure 3.2.4.2. Model 3 spectra show that the seasonal and seasonal adjusted series are independent.

In contrast Models 1 and 2 possess stochastic AR seasonality. This causes the spectrum to fatten out around the seasonal period (i.e. Period = 4). It also produces a widening of the spectral peak in the simulated seasonal series. This however, is of relatively short duration. Outside this narrow band about the seasonal period there is very good separation of the seasonal and seasonal adjusted series. This indicates that both of these sets of simulated series are independent.

In Table 3.2.4.2 (see Table Appendix) the results of applying Tsay's F-test to the regime dependent seasonal SETAR(2,1,1) model are displayed. Once again, a small relevant section is displayed for discussion purposes, labelled Table 3.2.4.2(a). Model 1, with the test applied at the 5% level gives just 39 simulated series that appear to show a threshold while 61 appear to be linear. Of these 39, almost all remain nonlinear after seasonal adjustment. However, for the 61 that appear linear roughly a half or more become nonlinear after seasonal adjustment. Thus either the test is unreliable for this model or the characteristics of the simulated series are altered by seasonal adjustment.

However, the test appears reliable since the results at the 1% level show no inconsistency. That is, more series appear linear and fewer nonlinear, as expected. Therefore we infer that the nonlinear characteristics of the simulated series have been altered by seasonal adjustment.

Table 3.2.4.2: Nonlinearity Test Results of Regime Dependent Seasonal Adjusted SETAR(2,1,1)

Original Series (y_t)		Seasonal Adjusted Series (x_t)					
		1%			5%		
		Linear	Non-linear	Total	Linear	Non-linear	Total
Seasonal Parameter (ρ_{22})		Model 1					
0.25	Linear	41	37	78	20	41	61
	Nonlinear	2	20	22	3	36	39

The results for Model 2 show a similar pattern to those observed for Model 1. This is to be expected since both models possess a stochastic seasonal fluctuation. Once again, it must be inferred that seasonal adjustment altered the characteristics of the simulated series.

The results for Model 3 lie somewhere between those observed for the seasonal SETAR(2,1,1) model (3.2.3) and those already observed for the regime dependent model. That is, 2/3rd's or more simulated series show evidence of a threshold, as in the seasonal SETAR(2,1,1) case. But almost none of these appear linear after seasonal adjustment, as in the regime dependent Models 1 and 2 above. Moreover, of those that initially appear linear when tested, half or more appear nonlinear after seasonal adjustment; with the results at the 1% level being consistent with those observed at the 5% level. Thus in Model 3 seasonal adjustment has also altered the characteristics of the simulated series.

The results of this subsection show that seasonal adjustment, of both the seasonal SETAR(2,1,1) model and the regime dependent seasonal SETAR(2,1,1) model, result in seasonal and regular cycles whose periods are largely independent. Threshold nonlinearity testing of the seasonal SETAR(2,1,1) model

showed there is no evidence that the nonlinear characteristics of the simulated series were altered. In contrast, when the seasonality was regime dependent, there was evidence to support the claim that seasonal adjustment alters the nonlinear characteristics of the simulated series. Where the seasonality was stochastic the evidence supporting the conclusion was strongest, but even where the seasonality was deterministic there was also sufficient evidence to support the conclusion.

3.2.5 Concluding remarks

This section has contrasted the effect of directly modelling simulated time series with TSMARS against the alternative method based on prior seasonal adjustment. Using this approach implied parameter estimates were obtained and the two methods compared.

The simulations based on the seasonal AR(1) model showed that there was little difference between using the Direct Method and Seasonal Adjustment Method. This is to be expected since the data are linear. The results for the seasonal SETAR(2,1,1) model showed the Direct Method to be slightly more reliable. Simulation results from the regime dependent seasonal SETAR(2,1,1) model showed that the Seasonal Adjustment Method performed poorly. The Direct Method once again did best based on implied parameter estimates.

The results from Tsay's F-test (see Table 3.2.4.2) showed that seasonal adjustment affected the nonlinear characteristics of the regime dependent seasonal SETAR model. This conclusions agrees with de Bruin (2002) and Ghysels & Perron (1993), who also observed seasonal adjustment affected nonlinear characteristics of times series.

Taking all these observations and conclusions together TSMARS modelling using the Direct Method is preferred. However, the Seasonal Adjustment Method also gives reliable estimates when the seasonality is not regime dependent.

3.3 Conclusions

The purpose of this chapter was to check TSMARS in a variety of situations that occur frequently in time series. Specifically, attention was focussed on data transformations and seasonal adjustment. Both of these topics being particularly relevant to the empirical time series modelling that will be conducted in later chapters. The conclusion drawn from the simulation studies will therefore provide a useful guide in this forthcoming modelling exercise.

Simulation studies were conducted and showed that differencing did not influence the accuracy of TSMARS estimates but it can improve their precision. Therefore, to ensure precision, differencing is recommended. It was also found that using a Box-Cox transformation is futile. The log transformation is useful when modelling growth data to stabilise the variance.

In relation to seasonal adjustment versus direct modelling with TSMARS, this section showed that the latter approach is slightly better. However, for the regime dependent seasonal SETAR(2,1,1) model, the Seasonal Adjustment Method performed poorly. The impact of seasonal adjustment on the nonlinear characteristics of a time series was also examined. Only in the case of the regime dependent seasonal SETAR model was any effect observed.

The conclusions of this chapter are therefore clear. When TSMARS is applied, integrated data should be differenced and modelling the data without prior seasonal adjustment is preferred. This will ensure that estimates are the best possible and that nonlinear effects are not inadvertently filtered out.

Table Appendix

Table 3.1.1.1: AR Model Simulation Results

λ		$\Delta = 0$		$\Delta = 1$	
		Borderline Stationary			
		$\rho_1 = 1$	$\rho_2 = -0.5$	$\rho_1 = 1$	$\rho_2 = -0.5$
0	Mean	0.992	-0.408	1.005	-0.501
	Std. Dev.	0.127	0.113	0.090	0.096
	Valid Models	100	-	100	-
0.5	Mean	1.095	-0.291	1.008	-0.503
	Std. Dev.	0.160	0.080	0.090	0.097
	Valid Models	75	-	100	-
1	Mean	1.014	-0.515	1.009	-0.504
	Std. Dev.	0.085	0.083	0.090	0.097
	Valid Models	93	-	100	-
λ		Integrated			
		$\rho_1 = 1.5$	$\rho_2 = -0.5$	$\rho_1 = 1.5$	$\rho_2 = -0.5$
0	Mean	1.395	-0.408	1.475	-0.492
	Std. Dev.	0.375	0.148	0.08	0.086
	Valid Models	94	-	67	-
0.5	Mean	1.475	-0.510	1.490	-0.527
	Std. Dev.	0.170	0.094	0.060	0.056
	Valid Models	99	-	100	-
1	Mean	1.489	-0.514	1.486	-0.522
	Std. Dev.	0.084	0.081	0.058	0.055
	Valid Models	100	-	99	-

Table 3.1.1.2: AR Exponential Data Model Simulation Results

λ		$\Delta = 0$		$\Delta = 1$	
		Borderline Stationary			
		$\rho_1 = 1$	$\rho_2 = -0.5$	$\rho_1 = 1$	$\rho_2 = -0.5$
0	Mean	0.995	-0.480	1.009	-0.504
	Std. Dev.	0.138	0.184	0.090	0.097
	Valid Models	100	-	100	-
0.5	Mean	0.920	-0.423	0.524	-0.240
	Std. Dev.	0.224	0.148	0.283	0.133
	Valid Models	100	-	100	-
1	Mean	0.837	-0.250	0.541	-0.272
	Std. Dev.	0.549	0.664	0.337	0.235
	Valid Models	100	-	100	-
λ		Integrated			
		$\rho_1 = 1.5$	$\rho_2 = -0.5$	$\rho_1 = 1.5$	$\rho_2 = -0.5$
0	Mean	1.210	-0.326	1.486	-0.522
	Std. Dev.	0.393	0.205	0.058	0.055
	Valid Models	93	-	99	-
0.5	Mean	1.313	-0.363	1.023	-0.110
	Std. Dev.	0.242	0.200	0.279	0.255
	Valid Models	99	-	70	-
1	Mean	1.337	-0.375	0.791	-0.220
	Std. Dev.	0.205	0.201	0.548	0.361
	Valid Models	100	-	58	-

Table 3.1.2.1: SETAR Model Simulation Results

Borderline Stationary							
λ		Δ = 0			Δ = 1		
		$\rho_{11}=\rho_{21}=1$	$\rho_{12}=-0.7$	$\rho_{22}=-0.3$	$\rho_{11}=\rho_{21}=1$	$\rho_{12}=-0.7$	$\rho_{22}=-0.3$
0	Mean	0.973	-0.664	-0.375	1.000	-0.669	-0.285
	S.E	0.105	0.187	0.200	0.037	0.060	0.059
	Valid Models	69	-	-	100	-	-
0.5	Mean	1.078	-0.403	-0.393	1.000	-0.685	-0.293
	S.E	0.092	0.086	0.074	0.037	0.065	0.058
	Valid Models	63	-	-	100	-	-
1	Mean	0.985	-0.723	-0.434	1.001	-0.698	-0.300
	S.E	0.106	0.189	0.221	0.037	0.067	0.056
	Valid Models	56	-	-	100	-	-

Integrated								
λ		Δ = 0				Δ = 1			
		$\rho_{12}=1.7$	$\rho_{12}=-0.7$	$\rho_{12}=1.3$	$\rho_{22}=-0.3$	$\rho_{12}=1.7$	$\rho_{12}=-0.7$	$\rho_{12}=1.3$	$\rho_{22}=-0.3$
0	Mean	1.694	-0.694	1.312	-0.313	1.601	-0.601	1.355	-0.355
	S.E	0.067	0.067	0.116	0.116	0.058	0.058	0.124	0.124
	Valid Models	N/A	-	-	-	72	-	-	-
0.5	Mean	1.694	-0.694	1.313	-0.313	1.606	-0.607	1.364	-0.364
	S.E	0.067	0.067	0.116	0.116	0.056	0.056	0.135	0.135
	Valid Models	N/A	-	-	-	58	-	-	-
1	Mean	1.694	-0.694	1.313	-0.313	1.610	-0.610	1.356	-0.355
	S.E	0.067	0.067	0.116	0.116	0.055	0.055	0.129	0.129
	Valid Models	N/A	-	-	-	57	-	-	-

N/A = Not Applicable: In this case valid models are not distinguished as the raw data, without differencing applied.

Table 3.1.2.2: Exponential SETAR Data Model Simulation Results

Borderline Stationary							
λ		Δ = 0			Δ = 1		
		$\rho_{11}=\rho_{21}=1$	$\rho_{12}=-0.7$	$\rho_{22}=-0.3$	$\rho_{11}=\rho_{21}=1$	$\rho_{12}=-0.7$	$\rho_{22}=-0.3$
0	Mean	1.007	-0.731	-0.402	0.964	-0.819	-0.299
	S.E	0.120	0.219	0.189	0.044	0.114	0.051
	Valid Models	63	-	-	100	-	-
0.5	Mean	1.000	-0.731	-0.401	0.963	-0.854	-0.304
	S.E	0.132	0.299	0.193	0.044	0.127	0.052
	Valid Models	55	-	-	100	-	-
1	Mean	0.985	-0.822	-0.382	0.963	-0.903	-0.309
	S.E	0.111	0.369	0.168	0.044	0.137	0.051
	Valid Models	34	-	-	100	-	-

Integrated								
λ		Δ = 0				Δ = 1			
		$\rho_{12}=1.7$	$\rho_{12}=-0.7$	$\rho_{12}=1.3$	$\rho_{22}=-0.3$	$\rho_{12}=1.7$	$\rho_{12}=-0.7$	$\rho_{12}=1.3$	$\rho_{22}=-0.3$
0	Mean	1.694	-0.694	1.313	-0.313	1.610	-0.610	1.356	-0.356
	S.E	0.067	0.067	0.116	0.116	0.055	0.055	0.130	0.130
	Valid Models	N/A	-	-	-	57	-	-	-
0.5	Mean	1.626	-0.626	1.096	-0.096	1.651	-0.651	1.155	-0.155
	S.E	0.205	0.205	0.348	0.348	0.192	0.192	0.253	0.253
	Valid Models	N/A	-	-	-	26	-	-	-
1	Mean	1.783	-0.705	1.160	-0.157	1.893	-0.978	1.175	-0.127
	S.E	0.371	0.533	0.681	0.577	0.981	1.295	0.930	0.607
	Valid Models	N/A	-	-	-	19	-	-	-

Table 3.2.1.1: Implied Parameter Estimates for AR(1) Model with Seasonality

Method	Model 1: $y_t = \rho_1 y_{t-1} + \rho_2 y_{t-s} + \varepsilon_t$				
	Mean Residual RMSE	$\rho_1 = 0.15$		$\rho_2 = 0.25$	
		Mean	S.E	Mean	S.E
Direct	0.98	0.05	0.155	0.30	0.103
Seasonal Adjustment	0.78	0.02	0.095	0.64	0.117
				$\rho_2 = 0.5$	
Direct	0.98	0.07	0.112	0.51	0.082
Seasonal Adjustment	0.78	0.04	0.092	0.74	0.076
				$\rho_2 = 0.75$	
Direct	0.98	0.08	0.067	0.71	0.069
Seasonal Adjustment	0.78	0.06	0.063	0.85	0.064
	Model 2: $y_t = \rho_1 y_{t-1} + \rho_2 \begin{cases} 1 & t \bmod(s) = 2 \\ 0 & \text{otherwise} \end{cases} + \varepsilon_t$				
	Mean Residual RMSE	$\rho_1 = 0.4$		$\rho_2 = 0.25$	
		Mean	S.E	Mean	S.E
Direct	0.96	0.32	0.140	0.42	0.140
Seasonal Adjustment	0.81	0.33	0.154	0.43	0.093
				$\rho_2 = 0.5$	
Direct	0.97	0.38	0.119	0.55	0.114
Seasonal Adjustment	0.83	0.37	0.118	0.54	0.088
				$\rho_2 = 0.75$	
Direct	0.98	0.40	0.103	0.76	0.132
Seasonal Adjustment	0.84	0.39	0.106	0.75	0.118
	Model 3: $y_t = \rho_1 y_{t-1} + \rho_2 \ \text{Sin}(2\pi t / s) + \varepsilon_t$				
	Mean Residual RMSE	$\rho_1 = 0.20$		$\rho_2 = 0.25$	
		Mean	S.E	Mean	S.E
Direct	0.95	0.13	0.141	0.50	0.087
Seasonal Adjustment	0.81	0.07	0.188	0.47	0.090
				$\rho_2 = 0.5$	
Direct	0.98	0.17	0.138	0.59	0.115
Seasonal Adjustment	0.83	0.13	0.139	0.57	0.112
				$\rho_2 = 0.75$	
Direct	0.98	0.17	0.121	0.76	0.154
Seasonal Adjustment	0.82	0.16	0.127	0.75	0.153

Table 3.2.2.1: SETAR(2,1,1) Model with Seasonality Simulation Results

Method	Model 1: $y_t = \begin{Bmatrix} \rho_{11} y_{t-1} \\ \rho_{12} y_{t-1} \end{Bmatrix} + \rho_2 y_{t-s} + \varepsilon_t \quad \begin{pmatrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{pmatrix}$						
	Mean Residual RMSE	$\rho_{11} = 0.2$		$\rho_{12} = 0.1$		$\rho_2 = 0.25$	
		Mean	S.E	Mean	S.E	Mean	S.E
Direct	0.25	0.18	0.088	0.09	0.102	0.26	0.046
Seasonal Adjustment	0.23	0.03	0.076	-0.01	0.089	0.68	0.067
						$\rho_2 = 0.5$	
Direct	0.25	0.19	0.077	0.08	0.100	0.50	0.037
Seasonal Adjustment	0.24	0.04	0.066	0.0	0.073	0.77	0.039
						$\rho_2 = 0.75$	
Direct	0.25	0.19	0.045	0.08	0.083	0.74	0.032
Seasonal Adjustment	0.28	0.09	0.042	0.02	0.065	0.84	0.022
	Model 2: $y_t = \begin{Bmatrix} \rho_{11} y_{t-1} \\ \rho_{12} y_{t-1} \end{Bmatrix} + \rho_2 \begin{cases} 1 & \text{tmod}(s) = 2 \\ 0 & \text{otherwise} \end{cases} + \varepsilon_t \quad \begin{pmatrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{pmatrix}$						
	Mean Residual RMSE	$\rho_{11} = 0.7$		$\rho_{12} = 0.3$		$\rho_2 = 0.25$	
		Mean	S.E	Mean	S.E	Mean	S.E
Direct	0.25	0.62	0.124	0.36	0.065	0.25	0.015
Seasonal Adjustment	0.29	0.57	0.156	0.49	0.076	0.23	0.014
						$\rho_2 = 0.5$	
Direct	0.25	0.57	0.151	0.35	0.037	0.50	0.015
Seasonal Adjustment	0.43	0.55	0.200	0.37	0.041	0.47	0.018
						$\rho_2 = 0.75$	
Direct	0.25	0.59	0.157	0.33	0.021	0.75	0.014
Seasonal Adjustment	0.61	0.51	0.304	0.33	0.027	0.73	0.018
	Model 3: $y_t = \begin{Bmatrix} \rho_{11} y_{t-1} \\ \rho_{12} y_{t-1} \end{Bmatrix} + \rho_2 \text{Sin}(2\pi t / s) + \varepsilon_t \quad \begin{pmatrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{pmatrix}$						
	Mean Residual RMSE	$\rho_{11} = 0.7$		$\rho_{12} = 0.3$		$\rho_2 = 0.25$	
		Mean	S.E	Mean	S.E	Mean	S.E
Direct	0.25	0.63	0.061	0.36	0.136	0.25	0.015
Seasonal Adjustment	0.22	0.59	0.066	0.40	0.091	0.24	0.016
						$\rho_2 = 0.5$	
Direct	0.25	0.64	0.079	0.29	0.220	0.50	0.020
Seasonal Adjustment	0.22	0.60	0.049	0.42	0.066	0.49	0.016
						$\rho_2 = 0.75$	
Direct	0.25	0.63	0.113	0.37	0.250	0.74	0.021
Seasonal Adjustment	0.23	0.62	0.038	0.42	0.059	0.75	0.016

Table 3.2.3.1: SETAR(2,1,1) Model with Regime Dependent Seasonality Simulation Results

Method	Model 1: $y_t = \begin{cases} \rho_{11} y_{t-1} + \rho_{12} y_{t-s} \\ \rho_{21} y_{t-1} + \rho_{22} y_{t-s} \end{cases} + \varepsilon_t \quad \begin{pmatrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{pmatrix}$								
	Mean Residual RMSE	$\rho_{11} = 0.2$		$\rho_{21} = 0.1$		$\rho_{12} = 0.75$		$\rho_{22} = 0.25$	
		Mean	S.E	Mean	S.E	Mean	S.E	Mean	S.E
Direct	0.26	0.18	0.089	0.10	0.055	0.58	0.058	0.55	0.054
Seasonal Adjustment	0.21	0.18	0.098	0.03	0.119	0.58	0.077	0.35	0.066
								$\rho_{22} = 0.5$	
Direct	0.25	0.17	0.062	0.13	0.051	0.67	0.041	0.64	0.044
Seasonal Adjustment	0.20	0.19	0.064	0.08	0.102	0.66	0.054	0.51	0.054

Method	Model 2: $y_t = \rho_{11} y_{t-1} + \rho_{21} y_{t-1} + \rho_{22} \begin{cases} 1 & \text{tmod(s)} = 2 \\ 0 & \text{otherwise} \end{cases} \rho_{12} y_{t-s} + \varepsilon_t \quad \begin{pmatrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{pmatrix}$								
	Mean Residual RMSE	$\rho_{11} = 0.2$		$\rho_{21} = 0.1$		$\rho_{12} = 0.75$		$\rho_{22} = 0.25$	
		Mean	S.E	Mean	S.E	Mean	S.E	Mean	S.E
Direct	0.28	0.21	0.154	0.12	0.121	0.49	0.079	0.22	0.039
Seasonal Adjustment	0.24	0.08	0.211	0.05	0.070	0.53	0.076	0.22	0.020
								$\rho_{22} = 0.5$	
Direct	0.29	0.14	0.200	0.04	0.049	0.58	0.054	0.45	0.026
Seasonal Adjustment	0.26	0.02	0.200	0.07	0.041	0.64	0.038	0.43	0.022

Method	Model 3: $y_t = \rho_{11} y_{t-1} + \rho_{21} y_{t-1} + \rho_{22} \begin{cases} 1 & \text{tmod(s)} = 2 \\ 0 & \text{otherwise} \end{cases} \rho_{12} \left(\text{Sin}\left(2\pi t / s\right)\right) + \varepsilon_t \quad \begin{pmatrix} \text{if } y_{t-1} \geq 0 \\ \text{if } y_{t-1} < 0 \end{pmatrix}$								
	Mean Residual RMSE	$\rho_{11} = 0.7$		$\rho_{21} = 0.3$		$\rho_{12} = 0.75$		$\rho_{22} = 0.25$	
		Mean	S.E	Mean	S.E	Mean	S.E	Mean	S.E
Direct	0.26	0.67	0.079	0.30	0.073	0.71	0.057	0.23	0.042
Seasonal Adjustment	0.29	0.84	0.072	0.03	0.050	0.49	0.040	0.24	0.014
								$\rho_{22} = 0.5$	
Direct	0.30	0.74	0.073	0.22	0.050	0.70	0.075	0.45	0.019
Seasonal Adjustment	0.32	0.76	0.073	0.12	0.043	0.50	0.066	0.43	0.021

Table 3.2.4.1: Nonlinearity of Seasonal Adjusted SETAR(2,1,1) Model with Seasonality Results

Original Series (y_t)		Seasonal Adjusted Series (x_t)					
		1%			5%		
		Linear	Non-linear	Total	Linear	Non-linear	Total
Seasonal Parameter (ρ_2)	Model 1						
0.25	Linear	44	7	51	21	6	27
	Nonlinear	14	35	49	21	52	73
0.5	Linear	51	7	58	28	6	34
	Nonlinear	12	30	42	21	45	66
0.75	Linear	66	14	80	48	21	69
	Nonlinear	9	11	20	9	22	31
Seasonal Parameter (ρ_2)	Model 2						
0.25	Linear	47	0	47	21	3	24
	Nonlinear	13	40	53	21	55	76
0.5	Linear	40	3	43	18	2	20
	Nonlinear	15	42	57	19	61	80
0.75	Linear	44	7	51	19	6	25
	Nonlinear	35	14	49	22	53	75
Seasonal Parameter (ρ_2)	Model 3						
0.25	Linear	55	5	60	31	1	32
	Nonlinear	11	29	40	16	52	68
0.5	Linear	46	7	53	26	3	29
	Nonlinear	14	33	47	19	52	71
0.75	Linear	49	7	56	29	6	35
	Nonlinear	15	29	44	16	45	65

Table 3.2.4.2: Nonlinearity Test Results of Regime Dependent Seasonal Adjusted SETAR(2,1,1)

Original Series (y_t)		Seasonal Adjusted Series (x_t)					
		1%			5%		
		Linear	Non-linear	Total	Linear	Non-linear	Total
Seasonal Parameter (ρ_{22})	Model 1						
0.25	Linear	41	37	78	20	41	61
	Nonlinear	2	20	22	3	36	39
0.5	Linear	47	23	70	34	27	61
	Nonlinear	3	27	30	1	38	39
Seasonal Parameter (ρ_{22})	Model 2						
0.25	Linear	42	27	69	27	31	58
	Nonlinear	0	31	31	0	42	42
0.5	Linear	28	23	51	22	25	47
	Nonlinear	1	48	49	0	52	52
Seasonal Parameter (ρ_{22})	Model 3						
0.25	Linear	37	29	66	14	24	38
	Nonlinear	3	31	34	1	61	62
0.5	Linear	23	13	36	10	10	20
	Nonlinear	8	56	64	0	80	80

4 Modelling Empirical Economic Time Series with TSMARS

4.1 Introduction

Of particular interest in this thesis are empirical economic time series in official statistics published by the CSO. These series can display seasonal variation, independent effects (see below) and outliers. In addition some of the series may be nonlinear. In this chapter a number of empirical time series are modelled with TSMARS to search for evidence to support this claim. Furthermore, the size of nonlinear component is quantified based on the ANOVA decomposition of the TSMARS approximation.

To gauge nonlinearity a 'test bed' of 20 monthly economic flow, stock and index series is taken from CSO sources. After linear modelling with SARIMA+, statistical tests (see Appendix) on residuals from each series indicated the presence of nonlinear effects. These detailed test results are given in a Table Appendix at the end of this chapter (see Table 4.6.1.1). More importantly, they also provide a standard against which TSMARS modelling of these series can be judged. A description of SARIMA+ is given in an Appendix to this thesis.

Growth or seasonal effects in an empirical series can swamp nonlinear features. Therefore, the way these effects are modelled can confound evidence of nonlinearity. In section 4.2 four different modelling variations of TSMARS are set out; two of these have already appeared in the literature while the other two are novel. The difference between each variation is the mechanism for handling data transformations and seasonality; these will be described in the next section. The key purpose of these four variations is to check, whether evidence of nonlinearity and its level depend on the modelling approach. In addition, all four variations cater for independent effects (e.g. trading day). However, in contrast to SARIMA+, no attempt is made to deal with outliers in any TSMARS variation; this topic is taken up in Chapter 6.

In section 4.3 an individual scrutiny of the TSMARS approximation for two of the series is also conducted. The two series are the Imports of Power Machinery and the Number of Males on the Live Register in Nenagh adjusted for seasonal effects; recall the logged 1^{st} differences of the Power Machinery series was plotted in Fig 1.1.1. For both series, the four TSMARS approximations are examined. Evidence for a nonlinear component is evaluated through frame plots, the ANOVA decomposition and plots of the distribution of residuals. The purpose of this section is to show how outputs from TSMARS can be used to provide a qualitative assessment of the model fit to the data.

In section 4.4, the quality of the TSMARS approximation to each test-bed series is evaluated based on a battery of statistical tests (see Appendix, Relevant Statistical Tests). For each of the four model variations these tests indicate the presence of nonlinear effects in the regular and seasonal parts of the series. The level of nonlinearity is quantified directly from the ANOVA decomposition of the TSMARS approximation. Tables showing detailed modelling results for all 20 test-bed series, using all four TSMARS modelling variations and SARIMA+ are displayed in the Table Appendix at the end of the chapter. This systematic modelling of empirical economic time series and the evaluation of the resulting test statistics to check for nonlinearity is novel.

4.2 Economic Data Modelling Methods

There are several ways in which TSMARS can be applied to empirical economic data. This study confines itself to modelling univariate time series that may be non-stationary. In this thesis, changes in the level of an empirical series are handled by differencing while changes in variance are treated with a log transform (see Appendix, SAS/ETS User's Guide 1993). In addition, independent predictors are also incorporated to account for fixed effects such as trading day factors.

This section describes the modelling methods used to produce a TSMARS approximation. First, methods to incorporate fixed effects are briefly outlined. This is followed by a description of four different modelling variations. Each of these is designed to capture seasonality in an alternative way. The resulting set of four TSMARS approximations will be contrasted with those obtained by linear modelling (i.e. with SARIMA+) in later sections.

In the TSMARS literature, two of the four modelling variations adopted have already appeared. The first uses independent periodic predictors in TSMARS, giving the so-called Periodic Predictor SETAR(ASTAR) model of Lewis & Ray (1997, 2002). The variation, referred to simply as TSMARS in subsection 4.4.2 below.

The other variation that has already appeared uses periodic autoregressive predictors. Using these in TSMARS gives the Periodic SETAR(ASTAR) model of Lewis & Ray (2002). This modelling variation described in subsection 4.4.4 below and is called PTSMARS.

The remaining two variations of TSMARS described in this section are novel. One uses prior seasonal adjustment and is called SATSMARS. The other variation models a parsimonious set of appropriately transformed data. This variation is termed STSMARS.

4.2.1 Independent Predictor Effects

Independent predictors are incorporated in TSMARS models in this and subsequent chapters to account for to account for fixed effects. These fixed effects are length of month (MD), trading week length (TD) and Easter effects (the so-called trading effects).

The length of month effect is computed by subtracting 30.4375 (i.e. the average number of days in any month) from the number of days in that month (i.e. MD = month length – 30.4375). The trading week length effect is TD = number of work week days – 5 X number of weekend days/2 (see Pena et. al. 2000). The Easter effect is computed according to the *Corrected Immediate Impact* rule given in Ladiray & Quenneville (2001). These trading factors are entered as co-variates and allowed to fully interact with the selected set of lagged time series predictors.

4.2.2 TSMARS Model

In this case, the raw response values y_t are modelled in TSMARS. Thus, for example, no transformations are made to render the series stationary.

A set of s+1 monthly (s=12) lagged predictors $y_{t-1},\ldots,y_{t-(s+1)}$, a deterministic seasonal predictor $p_t = Sin(2\pi\, i/s)$, ($i=1\ldots s$) and a set of s categorical predictors (each having a 1 in month i and denoted by k_i) are computed and input to the TSMARS program. In this variation seasonality is identified either autoregressively through lagged values, or as a fixed effect through k_i, or as a periodic effect through p_t. The maximum interaction degree set to 3 and basis function threshold = 2 X 10^{-8}. After the first call, further calls are made to TSMARS with weights to account for heteroskedasticity among the residuals giving the basic TSMARS approximation

$$\hat{y}_t = f\left(y_{t-1},\ldots y_{t-(s+1)}, k_1,\ldots,k_s, p_t, \text{MD}_t, \text{TD}_t, \text{Easter}_t\right) \qquad (4.2.1)$$

where $f(\bullet)$ denotes the TSMARS model.

4.2.3 SATSMARS: Seasonal Adjusted TSMARS Model

Here, the time series is first seasonally adjusted with *Proc X11* (SAS/ETS) giving $x_t = S(y_t)$, with S denoting the monthly X11 seasonal adjustment operator. Then, as above, a set of lagged predictors $x_{t-1},\ldots,x_{t-(s+1)}$, categorical predictors and a deterministic seasonal predictor are computed. These are input to TSMARS with maximum interaction degree set to 3 and basis function threshold = 2 X 10^{-8}. No further calls are made to TSMARS. The seasonal factors are then reapplied giving the approximation

$$\hat{y}_t = sf_t * f\left(x_{t-1},\ldots x_{t-(s+1)}, k_1,\ldots,k_s, p_t, \text{MD}_t, \text{TD}_t, \text{Easter}_t\right)/100 \qquad (4.2.2)$$

where sf_t denotes the seasonal factor. In this variation seasonality is removed prior to modelling. Categorical predictors and the deterministic seasonal predictor are included to account for the possibility of residual seasonality left after seasonal adjustment; however, the results show that this in fact does not occur. Also, only the multiplicative model is used and so no results are available when the data possess negative values

4.2.4 STSMARS: Seasonal TSMARS

In this variation the time series is checked and adjusted, as appropriate, for a log, a constant and the set of difference transformations giving the transformed series denoted by

$$z_t = (1-B)^d\left(1-B^s\right)^D\left\{\log(y_t + c)\right\}$$

where B denotes the backward difference operator $(B\,y_t = y_{t-1})$, d (=0, 1, 2) denotes the regular difference operator, D (=0,1) denotes the seasonal difference operator, c is a constant adjustment. The parsimonious set of lagged predictors $z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}$ are computed and input into the TSMARS program, along with appropriately differenced trading effects' variables. The maximum interaction degree is set to 3 and basis function threshold = 2 X 10^{-8} with subsequent weighted calls to the TSMARS program to handle heteroscedasticity. On completion the sequence of transformations are applied in reverse giving the approximation based on the general form

$$\hat{y}_t = \exp\left[(1-B)^{-d} (1-B^s)^{-D} \left\{ f\left(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}, z_{\text{t,MD}}, z_{\text{t,TD}}, z_{\text{t,EASTER}}\right) \right\} \right] - c \qquad (4.2.3)$$

where $z_{t,MD}$ etc. denotes the appropriately chosen differenced value of MD etc. Note, that no independent categorical predictors or deterministic seasonal predictor are included in the model. Therefore, seasonality is only catered for autoregressively in this modelling variation. This variation is also closest in flavour to linear modelling methodology, in that, a parsimonious set of lagged predictors, along with differencing and a log transformation are used as identify the model.

4.2.5 PTSMARS: Periodic TSMARS

In this variation the time series is checked and adjusted, as appropriate, for log and the periodic (lags = 1 and s) differences. Appropriate transformations are applied giving the transformed series denoted by z_t. This series is transposed into tabular years (rows) X months (columns) form. Let i and j denote the year and month respectively and let the transposed value of z_t by $M_{i,j}$. Then each month, January for example, is estimated separately with response data $M_{i,1}$ and predictors $M_{i,2},\ldots,M_{i,12}$ giving the Periodic TAR (PTAR) model for each individual month. Trading effects for the relevant month (e.g. January in this case), are also added to the predictor set. In the length of month case MD is denoted by $M_{i,1/MD}$. One call is made to the TSMARS program for each month, with the maximum interaction degree is set to 1 and basis function threshold = 2 X 10^{-8}. On completion the sequence of transformations are applied in reverse giving the general form of approximation

$$\hat{y}_t = \exp\left[(1-B)^{-d} (1-B^s)^{-D} \begin{Bmatrix} f_1(M_2,\ldots,M_{12},M_{1/\text{MD}},\text{M}_{1/\text{TD}},M_{1/\text{EASTER}}) \\ \vdots \\ f_{12}(M_1,\ldots,M_{11},M_{12/\text{MD}},\text{M}_{12/\text{TD}},M_{12/\text{EASTER}}) \end{Bmatrix} \right] \qquad (4.2.4)$$

where M_1 denotes the vector of time series values in January and the seasonal lags are based on the seasonal integration test (see Appendix, Relevant Statistical Tests). Note, no independent categorical predictors or deterministic seasonal predictor are included in the model. The seasonality in this case is catered for autoregressively and separately in each period.

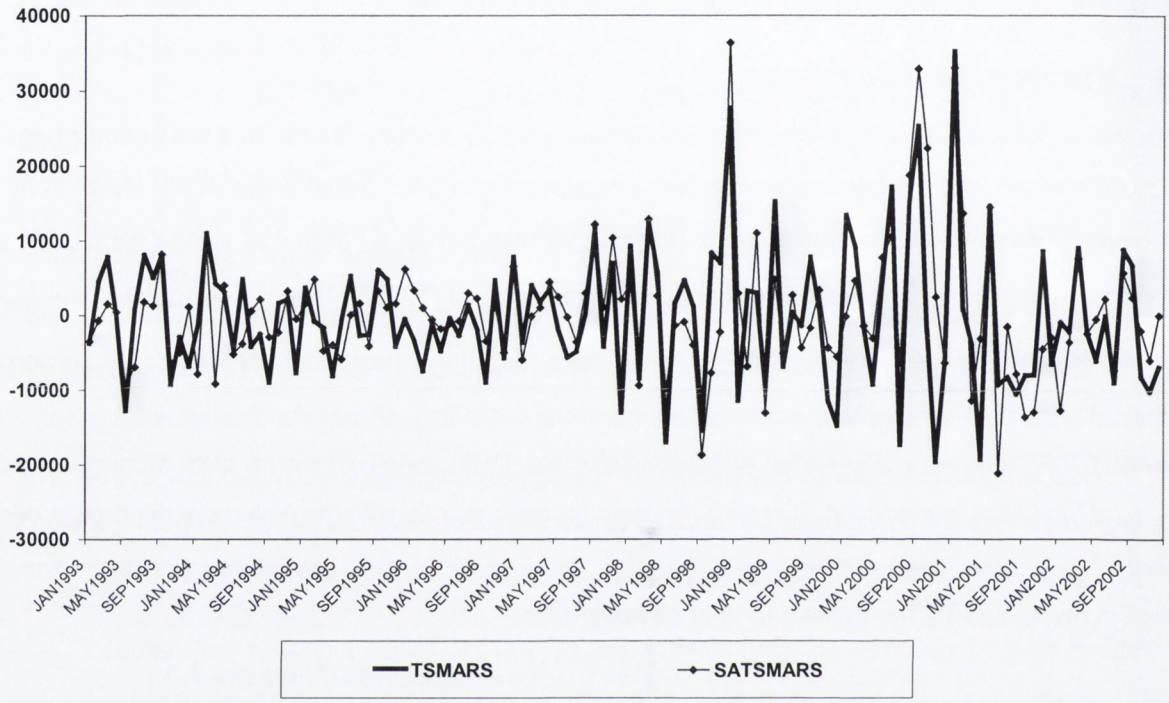
4.3 Individual Scrutiny of Two TSMARS Economic Time Series Approximations

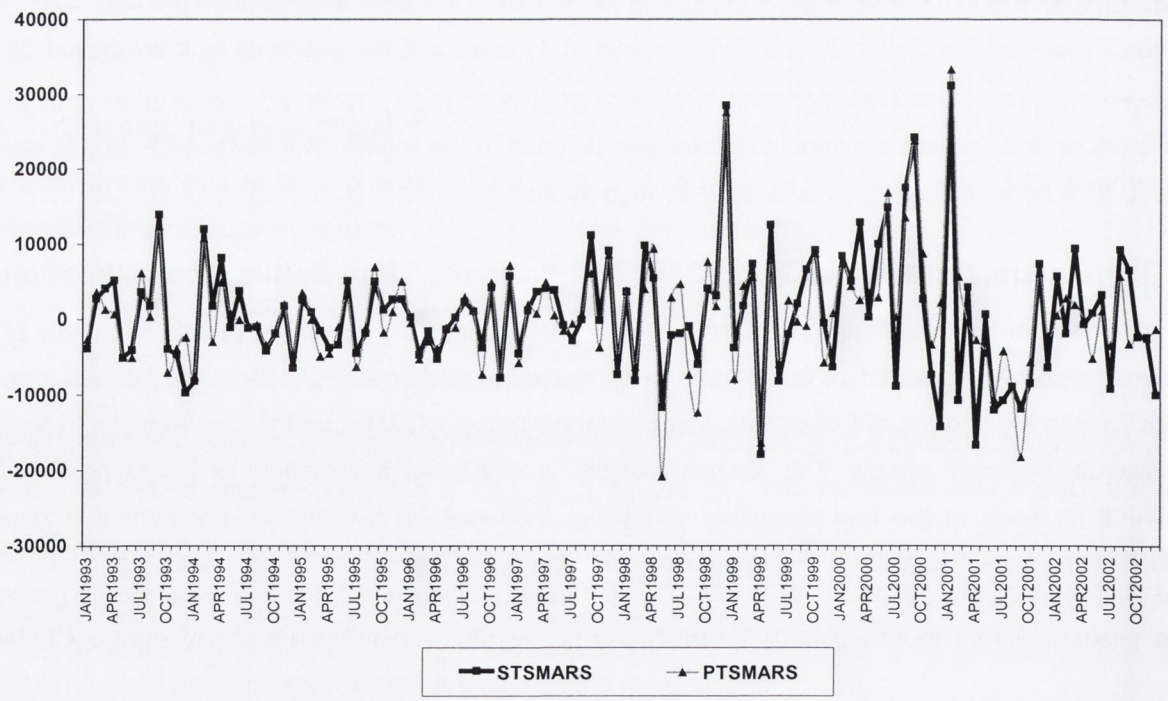
In this section the results obtained from two of the test time series are discussed in detail. Analysis of the other series, based on statistical test results, is conducted in section 4.4. As mentioned above, the two series chosen are Imports of Power Machinery and the Number of Males on the Live Register in Nenagh adjusted for seasonal effects. This section provides a qualitative assessment of the fit produced by TSMARS for each of the four modelling variations. Evidence for a nonlinear component is gauged through frame plots, the ANOVA decomposition and plots of the distribution of residuals.

4.3.1 Imports of Power Machinery

The results obtained from modelling this time series are displayed in Figure 4.3.1.1. Two separate residual plots covering the years 1993 to 2002 are provided for clarity. The first plot shows the both the TSMARS and SATSMARS (i.e. TSMARS applied to the seasonal adjusted series) residuals while the second shows the STSMARS and PTSMARS residuals.

Figure 4.3.1.1: Residual Plots of Imports of Power Machinery €000





The plots show that TSMARS and SATSMARS (i.e. TSMARS on seasonal adjusted series) provide very similar approximations. TSMARS tends to give a smooth approximation that tracks the local mean value of the series quite well. Both model variations however fail to pick up extra peaks in the years 2000 – 2002 as the residuals in this period are large. The plots for STSMARS and PTSMARS are also quite good. STSMARS modelled the differenced series. This plot is not as smooth but most of this local variation is accounted for by the adding back the differences. This would also account for STSMARS tending to follow the 2000 – 2002 peaks in the original series a little better than other methods.

The time series models produced by each approximation (reproduced from Table 4.6.1.1 with z_t denoting the 1st differenced series) respectively are

$$\hat{y}_t = 21{,}119 + 0.32y_{t-1} + 0.25y_{t-3} + 0.38y_{t-4} \qquad \text{(TSMARS)}$$

$$\hat{y}_t = 35{,}056 - 0.41y_{t-4} + 0.59(y_{t-1} - 52{,}932)_- \qquad \text{(SATSMARS)}$$

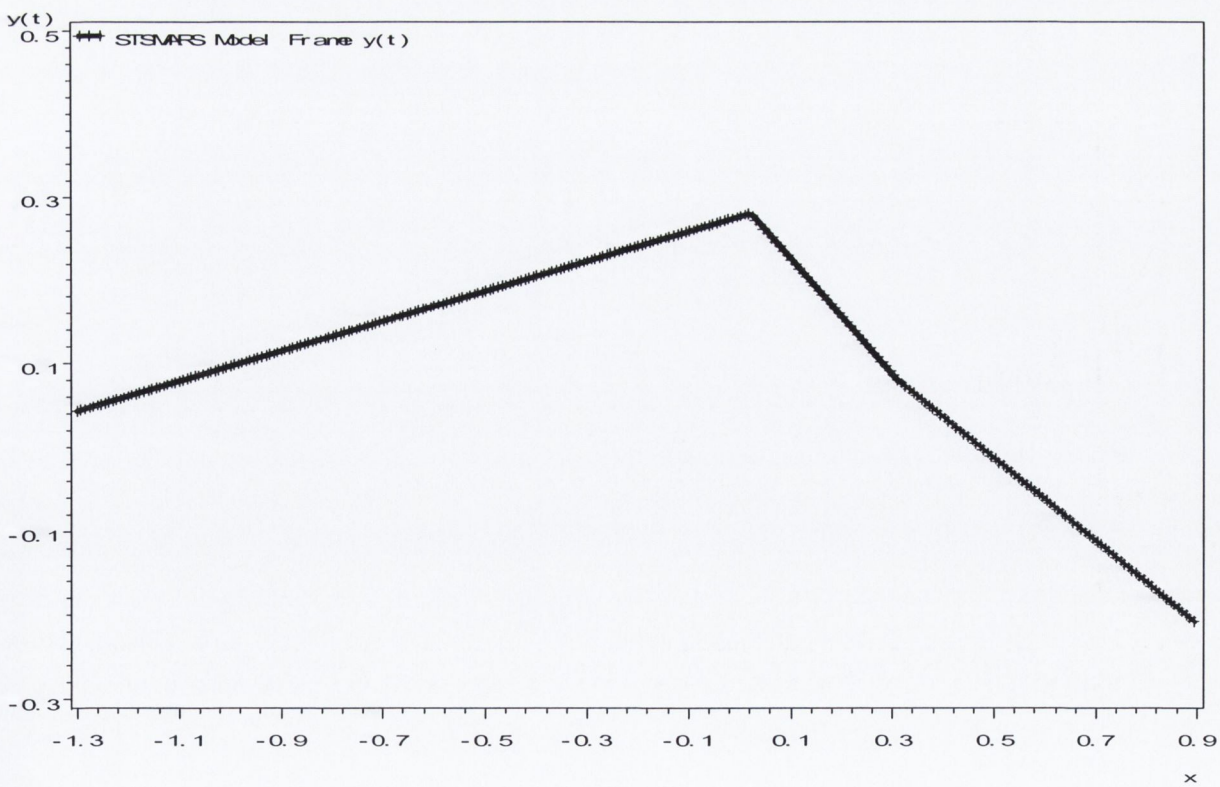
$$\hat{z}_t = 0.28 + 0.18z_{t-2} - 0.87(z_{t-1} - 1.28)_+ + 0.2(z_{t-12} - 1.05)_+ \qquad \text{(STSMARS)}$$

$$\hat{y}_t = \text{Mixed} \qquad \text{(PTSMARS)}$$

The term 'Mixed' is used to for PTSMARS where the 12 individual monthly models are different; most often being an AR model in one month and a SETAR model in another month.

The TSMARS approximation found is a nonseasonal linear model. This model is almost a simple moving average with coefficients of 1/3; to a close approximation. The SATSMARS is a nonseasonal SETAR model. In contrast the STSMARS model is a regime dependent seasonal SETAR model in the 1st differences. The frame for this model is plotted in Figure 4.3.1.2.

Figure 4.3.1.2: Imports of Power Machinery STSMARS Frame



It is clear from Figure 4.3.1.2 that the frame is not linear and moreover, it cannot be well approximated by a simple linear function. A quadratic function could be used indicating that higher order lags may help in modelling the data – this is the case with the linear TSMARS approximation.

However, the STSMARS frame in Figure 4.3.1.2 can be approximated quite well by the simple piecewise linear function

$$y = \begin{matrix} 0.1x & if\ x \le 0 \\ -0.3x & if\ x > 0 \end{matrix} \qquad (4.3.1)$$

This function in fact is the frame associated with the SETAR (2,1,1) model given in Chapter 1, equation (1.1.1). Recall, this SETAR model was suggested to explain the tendency of the differenced Imports of Power Machinery series, to stay negative after a negative value or to become negative after a positive value. Thus, STSMARS has found this asymmetric structure in the data and explained it with a regime dependent seasonal model.

However, the actual amount on nonlinearity is small, at only 2% according to the ANOVA decomposition given in Table 4.6.1.2. An extract from this table is given below for TSMARS and STSMARS approximations. Note: STSMARS figures are computed on the differenced series and the large residual in this case, is the residual from modelling this differenced series. In contrast, the TSMARS residual is computed from the raw series.

Extract of Table 4.6.1.2:

ANOVA Decomposition of Approximations to Imports of Power Machinery Series

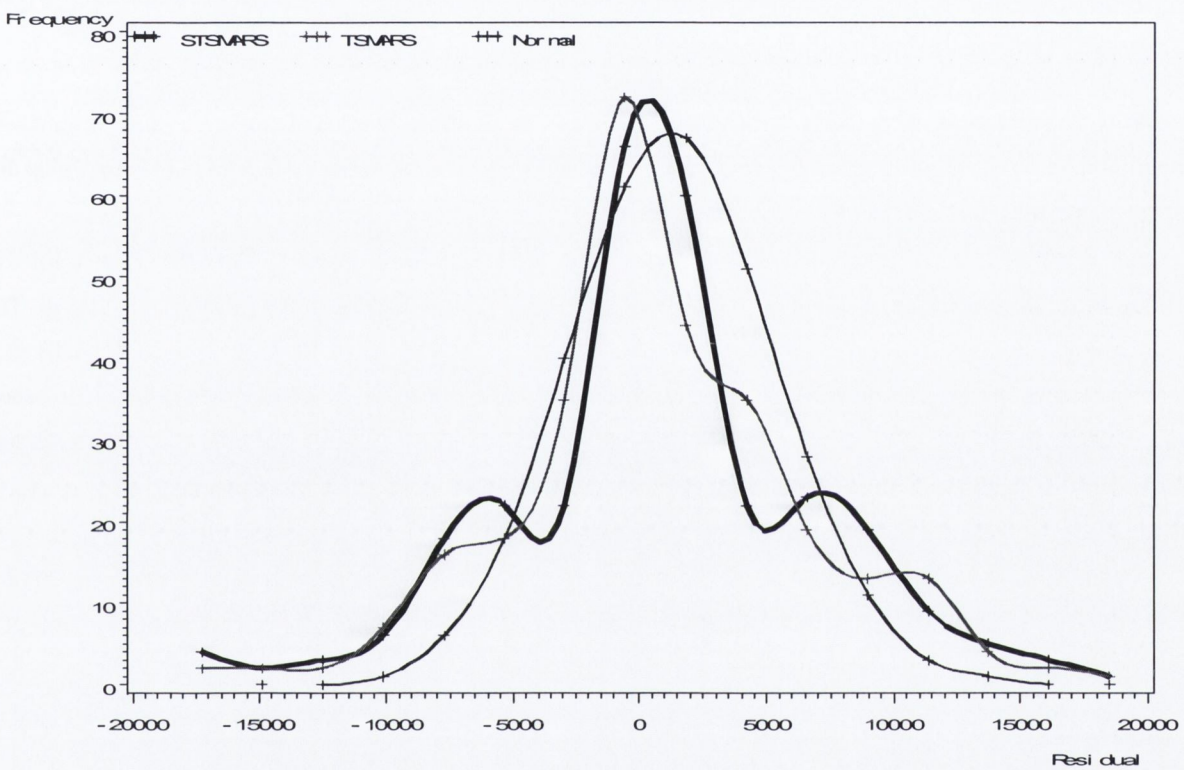
Method	Function Type	Function	% Variance
TSMARS	Mean	21,119	43
	Linear	$0.32y_{t-1} + 0.25y_{t-3} + 0.38y_{t-4}$	52
	Nonlinear	None	0
	Residual	-	5
STSMARS	Mean	0.28	15
	Linear	$0.18z_{t-2}$	13
	Nonlinear	$-0.87(z_{t-1} - 1.28)_+ + 0.2(z_{t-12} - 1.05)_+$	2
	Residual	-	70

Looking at the combined variance of the mean and linear functions, this is split in roughly in the same proportion over the mean function and linear function for both methods (i.e. 43/52 and 15/13). respectively. This indicates that the variance in the higher lags in the TSMARS approximation may be modelling curvature in the data. The STSMARS approximation shows that any curvature is likely to be small as only 2% of the variance is attributable to the nonlinear component. Thus, there is little difference between the TSMARS and STSMARS approximations. This conclusion is reinforced on taking 1st difference of the TSMARS approximation giving a model of the form (note, this function includes z_{t-2} which the usual 1st difference would not)

$$\hat{z}_t = -0.68z_{t-1} - 0.68z_{t-2} - [0.43z_{t-3} + 0.05z_{t-4} + O(z_{t-5})] \qquad (\Delta\text{TSMARS})$$

This is a function that approximates the negative slope in z_{t-1} on the right half of frame in Figure 4.3.1.2. The distribution of residuals from both of the TSMARS and STSMARS models fit are plotted in Figure 4.3.1.4. These show a high peak at the centre. The normal curve is also plotted and demonstrates that both TSMARS and STSMARS residual distributions have heavy tails. There is also evidence of two small peaks at about ±7,000. Both of these observations show that TSMARS did not manage to completely identify the underlying signal. In fact the heavy tails observed in the residual could be due outliers or to moving average components. This possible presence of moving average components is suggested by the fact that the SARIMA+ returned a 1st differenced first order moving average model for these data.

Figure 4.3.1.3: Imports of Power Machinery Residual distributions



4.3.2 Live Register Males Nenagh (seasonal adjusted)

The results obtained from modelling this time series are displayed in Figure 4.3.2.1. In this case only the STSMARS plot is given as all the other methods produced an approximation of the form

$$\hat{y}_t = 314 + 0.99 y_{t-1} \qquad \text{(TSMARS)}$$

Clearly, this approximation shows that the series should have been differenced before modelling. However, STSMARS is designed to handle data that should be differenced. In this case, 1st differences were applied giving the seasonal model approximation

$$\hat{z}_t = -0.03 + 0.17 z_{t-12} - 0.26(z_{t-3} + 0.009)_+ + 0.2(z_{t-13} + 0.02)_+ \qquad \text{(STSMARS)}$$

The plot in Figure 4.3.2.1 shows that the observed data are quite smooth and the STSMARS approximation is very close. The STSMARS model for the differenced series z_t has a threshold term at lag 3 and a linear and a threshold term at the seasonal lags. The extract below from ANOVA

decomposition (c.f. Table 4.6.1.2) shows that only 1% of the variance appears to be attributable to the threshold terms. The mean and linear components dominate the approximation.

Extract of Table 4.6.1.2:

ANOVA Decomposition of STSMARS Approximation to Live Register Series

Method	Function Type	Function	% Variance
STSMARS	Mean	-0.29	6
	Linear	$0.17z_{t-12}$	5
	Nonlinear	$-0.26(z_{t-3}-0.009)_+ + 0.2(z_{t-13}-0.02)_+$	1
	Residual	-	88

The frame for this model is plotted in Figure 4.3.2.2. It is clear from the plot that the STSMARS approximation is virtually linear over the range of the data. A small kink is visible and this explains the occurrence of the threshold terms.

Finally, in Figure 4.3.2.3 the residuals are plotted for both TSMARS and STSMARS. The TSMARS residuals are included to show the amount of frequency that is accounted for by differencing alone. The TSMARS plot is slightly skewed but otherwise approximates the Normal quite well. The STSMARS fit is very good except for the slight peak at residual values around 60. Also, there is no evidence for heavy tails in these distributions. Based on this analysis these data linear. Prior seasonal adjustment though has not removed the seasonal linear component that is discernible in the 1st differenced series.

Figure 4.3.2.1: Live Register No. Males Nenagh (sesaonal adjusted)

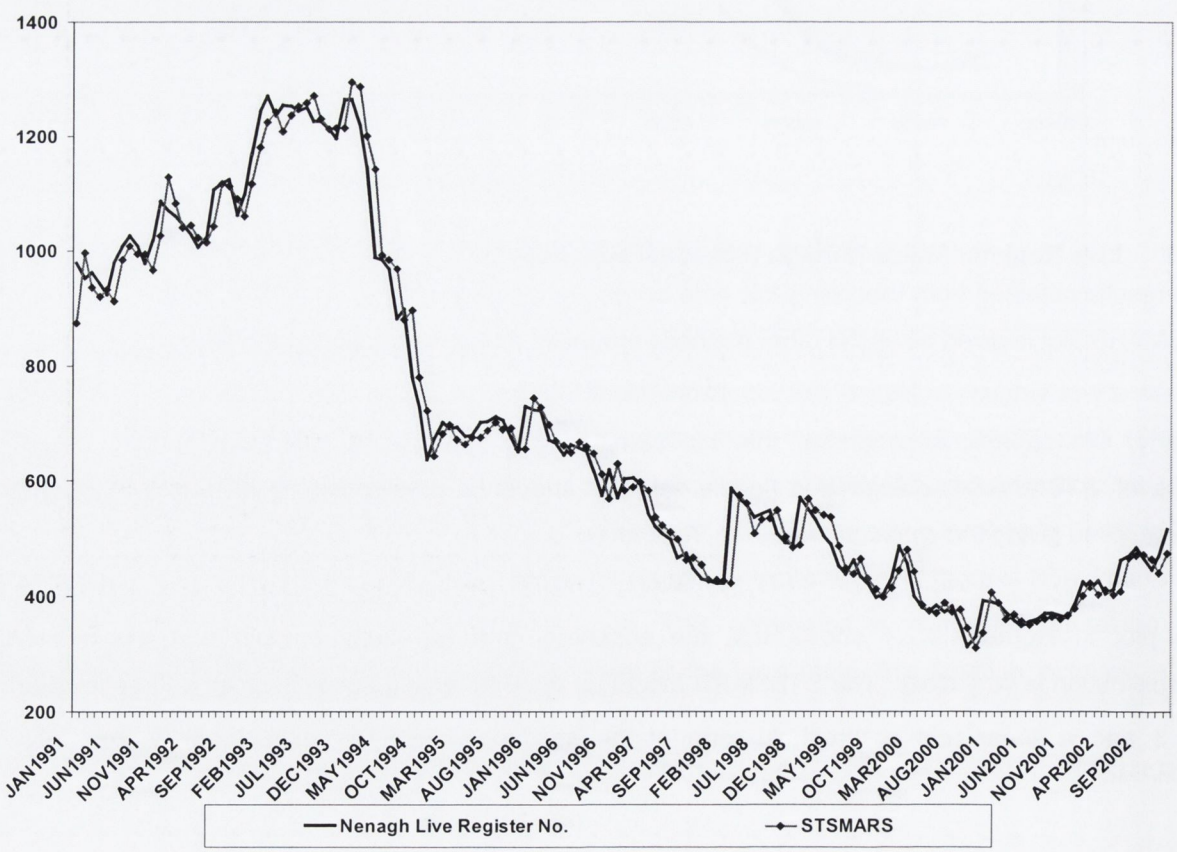


Figure 4.3.2.2: Live Register No. Males STSMARS Frame

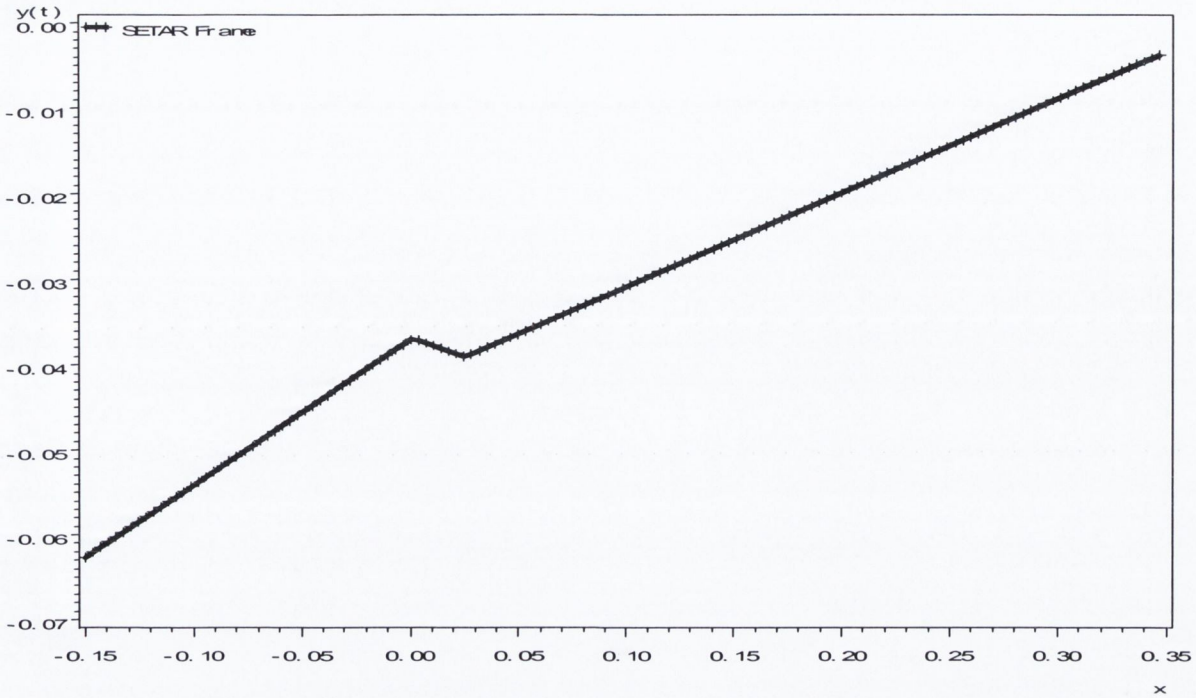
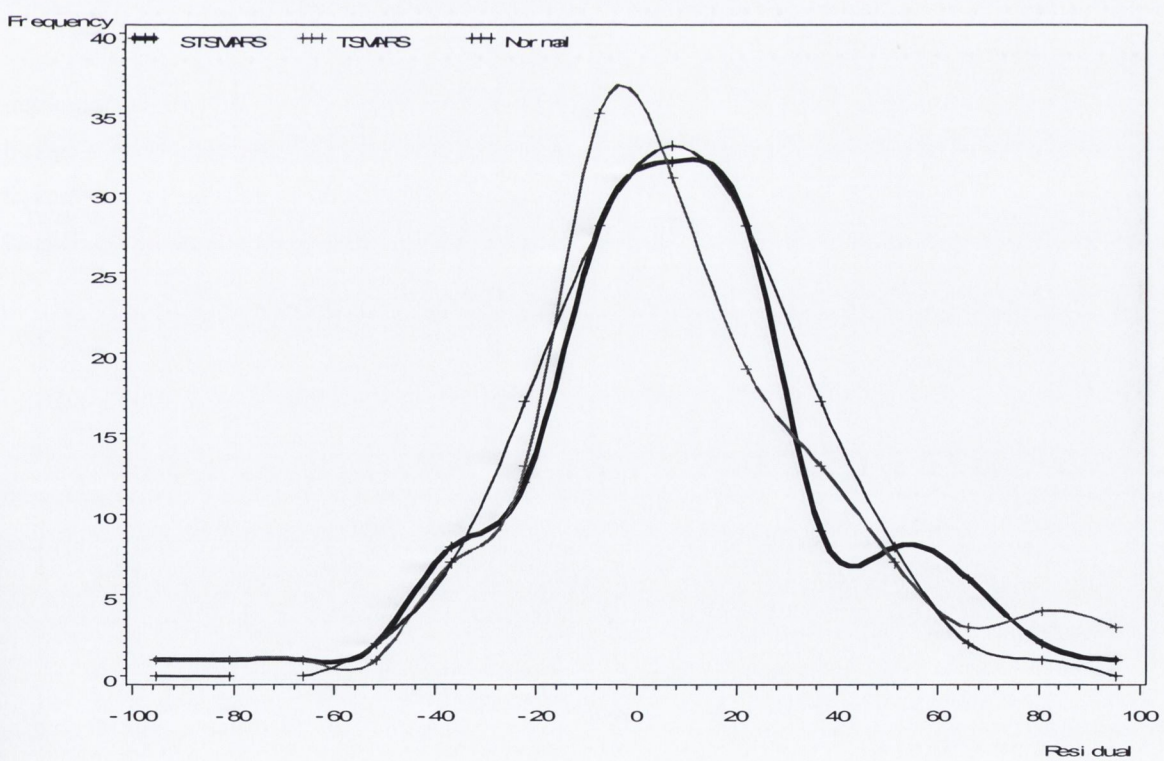


Figure 4.3.2.3: Live Register No. Males Residual distributions



4.3.3 Closing Remarks

The section contrasted four different variations of TSMARS approximations on two chosen series and gave qualitative assessment of the fit of each variation.

The analysis of the Imports of Power Machinery concluded that there was evidence for nonlinearity in the 1st differenced series of STSMARS. The size of this component was however quite small, at 2% according to the ANOVA decomposition. This however was sufficient to give a better residual fit than the linear approximation of TSMARS (i.e. without differencing). The residual distributions of both methods though did show evidence of heavy tails. This would indicate that there is potentially another nonlinear component, such as a departure from normality that STSMARS is unable to find; see, for example subsection 2.5.6, where TSMARS gave poor estimates where the data depated from normality.

The Live Register No. in Nenagh adjusted for seasonal effects was also analysed. Three of the four modelling variations produced a integrated order 1 approximation. STSMARS though produced a seasonal threshold model. On further analysis this model was ruled out, as the threshold effects only marginally perturbed the data away from linearity. In this case, it was concluded that this is a 1st order differenced series, with a small seasonal linear component driven by Gaussian error. This linear seasonal component however has not been removed by prior seasonal adjustment.

This qualitative assessment therefore shows that TSMARS produces accurate approximations to these data. However, even with sophisticated modelling variations, a departure from normal errors, such as heavy tails, can affect the TSMARS approximation.

4.4 Economic Data Modelling Results

In this section the modelling results are summarised and discussed. The summary has two components, a frequency table of significant test results and a frequency table of model types. The discussion is based on and an analysis of the MAPE statistics. The purpose of this examination is to assess the overall presence of nonlinear components, as well as their influence on the set of 20 'test bed' time series. Based on this, a verdict is made on the nature, type and influence of nonlinear effects in CSO series.

4.4.1 Explanation of Results Table

The results obtained from the time series modelling study of the test bed are displayed in Table 4.6.1.1. For each series in the table, the number of observation (N), the series code and its title are given. Then, for each TSMARS modelling variation an indicator is given to signify whether a log transformation or constant was applied. The TSMARS model approximating the data is also given, followed by statistics computed on the residual. These statistics are specified in the Appendix - Relevant Statistical Tests.
The model given in the SARIMA+ case has the following definition:

$$(p,d,q)(P,D,Q)_s\{MD, TD, EASTER\}[AO, LS, TS\ IO]$$

where p, d, q and P, D, Q specify the regular and seasonal AR order, level of differencing and MA order respectively. In *curly braces* the trading effects predictors are specified while in *square braces* the number of additive outliers, level shifts, transitional shifts and innovational outliers is given. Note that only significant trading effects predictors are incorporated in the final estimated SARIMA+ model.

In relation to the periodic PTSMARS method the model for each month is given when it is simple. However a number of the models found were complex with up to 12 different monthly additive models combined. When this occurred the model is classified as *Mixed* with independent predictors (when found) placed alongside in *curly braces*.

The column Cycles/Notes in Table 4.6.1.1 identifies whether a cycle (i.e. significant spike, see Brockwell & Davies 1991) was evident in the residual spectrum. Where a regular cycle is found and Tsay's F-test is also significant at a corresponding lag (or integer multiple thereof) then Tsay's F-test classifies it as a nonlinear effect. Evidence of a threshold can therefore only be accepted when there is no evidence of a cycle.

4.4.2 Discussion of Results

Table 4.6.1.1 provides a detailed set of results from the modelling exercise with statistics computed on the residual $y_t - \hat{y}_t$. To glean useful additional information from the statistical test results, Table 4.4.2.1 summarises the number of times each test produced a significant value at the 1% level (except for $t_{\mu,\sigma}$ at 5%). In the table the rank of each method is also given (best = 1) in braces and a final column shows the sum of the ranks. The purpose of Table 4.4.2.1 is to show any differences between the modelling variations in terms of results of test statistics. In this context a modelling variation will be judged better if it has the lower 'Sum of Ranks' value.

Table 4.4.2.1: Frequency of Significant Test Results

Method	Statistics									Sum of Ranks
	χ_1^2	χ_2^2	$t_{\mu,\sigma}$	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	
SARIMA+	10 (1)	18 (5)	1 (2)	5 (3)	18 (5)	1 (1)	0 (1)	0 (1)	1 (2)	(21)
TSMARS	17 (5)	11 (2)	4 (5)	7 (4)	14 (2)	14 (5)	13 (2)	13 (3)	4 (5)	(33)
SATSMARS	13 (3)	13 (3)	3 (4)	3 (1)	16 (3)	1 (1)	15 (4)	14 (4)	0 (1)	(24)
STSMARS	11 (2)	10 (1)	0 (1)	4 (2)	12 (1)	7 (4)	13 (2)	12 (2)	1 (2)	(17)
PTSMARS	13 (3)	15 (4)	1 (2)	9 (5)	16 (3)	4 (3)	17 (5)	15 (5)	3 (4)	(34)

Table 4.4.2.1 can be divided into two parts, on the left are the model adequacy tests; then χ^2, $t_{\mu,\sigma}$, Tsay F-test and BDS test. Where these are significant it indicates there is nonlinearity in the residuals. On the right are the seasonality tests.

In terms of model adequacy, the figures show that STSMARS is judged best with consistently low ranks. Among the others there is little difference. Therefore STSMARS has found more nonlinearity than any other modelling variation. STSMARS performs second best on seasonality test though the number of significant test statistics is only marginally lower than the other modelling variations.

With regard to seasonality, SARIMA+ out performs other methods. In fact, this method left almost no seasonality behind in the residuals. Moreover, the figures suggest that the seasonal component in these test-bed series is linear and does not interact with the regular component.

A particularly striking feature in Table 4.4.2.1 is that the $t_{\mu, \sigma}$ statistic and Periodic Variation F-test were significant in a small number of cases. This indicates that neither regular or seasonal heteroscedasticity is evident in the residual.

The MAPE statistic was analysed using the Kruskal-Wallis One-way test of ranks procedure treating each method as a separate group for the one-way analysis. The H-value that resulted was 0.82; this is not significant at the $\chi^2_{4,0.95}$ = 11.1, so there is no evidence of a difference in the MAPE values across methods. H-values were also computed pair wise. These ranged from 0.0 to 0.7 and once again none of these values are significant at the 5% level and so there does not appear to be any difference between the methods in terms of their MAPE.

Table 4.4.2.2 summarises the models observed in Table 4.6.1.1 according to model type. The frequency distribution shows that both TSMARS and SATSMARS returned an integrated I(1) model in 12 of the 20 cases. This is a reflection of the fact that the signal variance is dominated by the 'growth' component in many economic time series. In contrast roughly half of the models found by the STSMARS method displayed some nonlinear effect after the growth component was removed. This suggests that to identify nonlinear components, integrated effects should be removed prior to modelling with TSMARS. In relation to PTSMARS the models are simply classified as linear or nonlinear. Here again, the method seasonally differenced the data prior to modelling and the results show that over half of the models possessed some nonlinearity.

Table 4.4.2.2: Frequency of Different Model Types Observed in the Test Results

Method	Model Type							
	Mean only	Independent Predictors	Linear	Integrated I(1)	SETAR	Seasonal SETAR	Regime Dependent SETAR	Nonlinear
TSMARS	1	1	2	12	1	2	0	1
SATSMARS	0	0	1	12	4	1	1	0
STSMARS	0	2	5	2	2	4	5	0
PTSMARS	-	-	8	-	-	-	-	12

The table above indicates which methods found some nonlinearity. However, of greater interest is the number of test statistics that were not significant where a nonlinear model was found. Table 4.4.2.3 gives this number for the (five) model adequacy test. The presence of a '–' in the cell indicates that the model did not possess any nonlinearity. Looking at the figures it is clear that the test statistics from STSMARS result in fewer significant tests of nonlinearity. For test-bed series numbered 6 through 10, almost all nonlinear test were not significant. This verifies the claim that CSO series possess nonlinear features, since these models are nonlinear and these residual tests show no evidence of nonlinearity.

Table 4.4.2.3: Frequency of (five) Test Statistics that are not significant for Nonlinear Models only.

Method	Test Series Number									
	1	2	3	4	5	6	7	8	9	10
TSMARS	-	-	1	-	-	-	-	-	-	-
SATSMARS	-	-	1	-	-	-	-	-	-	-
STSMARS	-	1	-	1	-	5	5	5	4	5
Method	Test Series Number									
	11	12	13	14	15	16	17	18	19	20
TSMARS	-	-	3	-	-	-	0	1	-	-
SATSMARS	-	-	-	2	-	-	2	1	2	-
STSMARS	4	-	1	-	-	-	1	1	3	1

4.4.3 ANOVA Decomposition Results

The TSMARS program provides an ANOVA type analysis that breaks down the contribution of each basis function in the approximation to the overall variance. In Table 4.6.1.2 this breakdown is given grouped by basis function type. The breakdown is computed somewhat differently for each method in that the figures given, are those that arise solely from the call to the TSMARS program. This means that growth effects are excluded from the figures given for STSMARS and average figures are given for the 12 monthly approximations obtained in PTSMARS.

Table 4.4.3.1: Extract of Table 4.6.1.2 ANOVA Analysis

No	N	Series Code	Title	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
			Average	TSMARS	48	42	4	1	5
				SATSMARS	49	44	3	0	4
				STSMARS	25	19	6	0	50
				PTSMARS	67	18	8	0	7

The overall average figures given in Table 4.6.1.2 are extracted and displayed in Table 4.4.3.1. These show the mean and linear basis functions account for roughly 90% of the explained variance for three of the methods. These sources of variation largely account for the growth components of the time series that mask other characteristics such as nonlinearity. This masking is further emphasised by the fact that the residual variance is also relatively small for these three methods. The breakdown of the variance for STSMARS does appear different to that of the others in that the residual variance accounts for 50% of the overall variance. This however is an anomaly, since in this case the variance breakdown is given purely for the TSMARS program call. The approximation therefore does not always include the growth component. In general the nonlinear component appears to account for about 5% of the overall variance.

Only test-bed series numbered 3, 13, 14, 17, 18 and 19 show evidence of nonlinear signature in the residual using more than 1 method.

4.5 Conclusions

In this chapter a 'test bed' of 20 monthly economic flow, stock and index series were modelled with four TSMARS modelling variations. The purpose of this was to assess whether there is evidence for nonlinearity in each series. An individual scrutiny of the TSMARS approximation for two of the series was also conducted to glean further evidence in support of the nature of the nonlinearity.

The individual scrutiny showed that TSMARS modelled the data accurately. There was also evidence from frame and residual analysis to support the conclusion in favour of some nonlinearity. In particular, the 1st differenced Imports of Power Machinery series showed asymmetric structure. There was also evidence for some departure from normal errors.

The detailed results of Table 4.6.1.1 and the associated ANOVA breakdown in Table 4.6.1.2, showed that many of the TSMARS approximations were simply integrated order 1 growth models. This growth effect tended to mask other characteristics. The conclusions of Chapter 3 regarding the importance of differencing prior to TSMARS modelling therefore hold up. Where nonlinearity was found, invariably the STSMARS variation discovered it. Moreover, this was confirmed by the fact the test statistics for nonlinearity in the residual tended to be accepted. Analysis of the associated ANOVA decomposition showed that only about 5% of the variance was attributable to nonlinear effects. This however varied from 0% in most cases, to a maximum of 47% in one case.

The conclusions obtained from this modelling exercise show that nonlinearity was only present to a small degree. This conclusion is affirmed by the fact that linear SARIMA+ modelling gave results that were also good. Moreover, SARIMA+ incorporates both MA components and outliers. These may be responsible for heavy tails observed in the residual of the Imports of Power Machinery series. This in turn may be affecting the test statistics. If these effects are present then TSMARS will produce estimates of insufficient quality. The reasons for this are that TSMARS is not robust and does take account of dependent errors in it model estimates. These issues will be taken up in later chapters.

Table Appendix

Table 4.6.1.1: Time Series Test-bed Results

No	N	Series Code	Title	Method	Transformations		Model
					Log	Constant	
1	286	ASAM003	Cows Milk Protein Content (%)	SARIMA+	Y	Y	$(3, 1, 1)(1, 0, 0)_s$
				TSMARS	-	-	$y_t = 210{,}770 + 1.01\, y_{t-1}$
				SATSMARS	-	-	$y_t = 30{,}867 + 0.97\, y_{t-1}$
				STSMARS	Y	Y	$\Delta\left[y_t = -0.01 - 0.44\, y_{t-1} + 0.48\, y_{t-12}\right]$
				PTSMARS	Y	-	$M_1 = 10.3 + 1.04\, M_2$, $M_2 = 10.3 + 0.94\, M_1$, $M_3 = 10.4 + 0.92\, M_1$, $M_4 = 10.3 + 1.04\, M_1$, $M_5 = 10.3 + 0.98\, M_1$, $M_6 = 10.3 + 0.98\, M_3$, $M_7 = 10.3 + 0.96\, M_1$, $M_8 = 10.3 + 1.03\, M_2$, $M_9 = 10.3 + 1.00\, M_4$, $M_{10} = 10.4 + 1.02\, M_4$, $M_{11} = 10.4 + 1.01\, M_4$, $M_{12} = 10.3 + 0.97\, M_4$
2	286	ASAM206	Calves Slaughtering 000 Heads	SARIMA+	Y	N	$(1, 0, 1)(1, 0, 1)_s$ [0, 4, 3, 2]
				TSMARS	-	-	$y_t = 0.13 + 0.71\, y_{t-1}$
				SATSMARS	-	-	Series has zero values – inappropriate for multiplicative adjustment
				STSMARS	Y	N	$y_t = 1.08 + 0.18\, y_{t-1}(y_{t-2} - 1.2)_- - 1.2\, y_{t-1}(y_{t-2} - 1.2)_-(y_{t-12} - 2.1)_- + 0.5\, y_{t-1}(y_{t-2} - 1.2)_-(y_{t-12} - 0.8)_-$
				PTSMARS	Y	-	y_t = Mixed {TD}
3	286	ASAM305	Heifers Slaughtering 000 Tons	SARIMA+	Y	Y	$(0, 1, 1)(0, 1, 1)_s$ {TD}
				TSMARS	-	-	$y_t = 8.3 + 0.22 y_{t-2} + 0.42 y_{t-12} - 0.37 y_{t-13} - 0.56(y_{t-1} - 10.2)_- + 0.42(y_t - 10.2)_+$
				SATSMARS	-	-	$y_t = 8.48 - 0.78(y_{t-1} - 4.6)_- + 0.57(y_{t-3} - 4.6)_+ + 0.64 y_{t-12} - 0.4 y_{t-13} + 0.16(TD + 5)_+$
				STSMARS	Y	Y	$\Delta\left[y_t = 0.06 - 0.01(TD - 7)_- - 0.02(MD - 3)_+\right]$
				PTSMARS	Y	-	y_t = Mixed {TD}
4	250	FIAM023	Irish Currency in Circulation (€)	SARIMA+	Y	Y	$(1, 0, 1)(1, 0, 1)_s$ [0, 2, 2, 5]
				TSMARS	-	-	$y_t = 1{,}068.8 + 1.01\, y_{t-1}$
				SATSMARS	-	-	$y_t = 1{,}087 + 1.02\, y_{t-1}$
				STSMARS	Y	Y	$\Delta_2\left[\begin{array}{l} y_t = 0.11 + 0.2\, y_{t-1} + 0.57 y_{t-12} - 1.18(y_{t-2} - 0.11)_+ + \\ \quad 1.17(y_{t-3} - 0.12)_+ + \\ \quad 2.85\, y_{t-1}(y_{t-2} + 0.14)_+ - 2.63 y_{t-12}(y_{t-13} + 0.13)_+ \end{array}\right]$
				PTSMARS	Y	-	$M_1 = 6.9 + 0.99\, M_2$, $M_2 = 6.9 + M_1$, $M_3 = 6.9 + 0.98\, M_1$, $M_4 = 6.9 + 1.02\, M_1$, $M_5 = 6.9 + 1.04\, M_1$, $M_6 = 6.9 + 1.02 M_1$, $M_7 = 6.9 + 1.04\, M_1$, $M_8 = 6.9 + 1.01\, M_1$, $M_9 = 7.0 + 1.01\, M_1$, $M_{10} = 6.9 + 1.02\, M_1$, $M_{11} = 6.9 + 1.05\, M_1$, $M_{12} = 7.0 + 1.01\, M_1$
5	324	FIAM102	Exchange Rate $ £STR	SARIMA+	N	Y	$(1, 0, 1)(1, 0, 1)_s$ {EASTER}[0, 4, 1, 0]
				TSMARS	-	-	$y_t = 0.75 + 0.98\, y_{t-1}$
				SATSMARS	-	-	$y_t = 0.74 + 0.98\, y_{t-1}$
				STSMARS	N	Y	$\Delta\left[y_t = -0.02 + 0.37 y_{t-1}\right]$
				PTSMARS	N	-	$M_1 = 0.8 + M_2$, $M_2 = 0.8 + 0.97 M_1$, $M_3 = 0.8 + 0.93\, M_1$, $M_4 = 0.8 + 0.96\, M_1$, $M_5 = 0.8 + M_1$, $M_6 = 0.8 + 0.95 M_1$, $M_7 = 0.8 + 0.93\, M_1$, $M_8 = 0.8 + 0.92\, M_1$, $M_9 = 0.8 + 0.97\, M_1$, $M_{10} = 0.8 + 1.03\, M_1$, $M_{11} = 0.8 + 1.01 M_1$, $M_{12} = 0.8 + 0.96\, M_1$

No	Method	Statistics											
		χ_1^2	χ_2^2	$t_{\mu,\sigma}$	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
1	SARIMA+	0.01	0.01	-0.48	2, 3, 6	0.01	0.01	1.0	1.0	1.0	8.8	0.0	3
	TSMARS	0.01	0.01	1.63	1 – 3, 4 - 8	0.01	0.01	0.01	0.01	0.30	10.9	0.0	3
	SATSMARS	0.01	0.01	2.78	2 - 13	0.01	0.95	0.01	0.01	0.93	8.5	0.0	3
	STSMARS	0.01	0.01	-0.12	2, 4, 6 - 10	0.01	0.05	0.01	0.01	1.0	8.7	0.0	3
	PTSMARS	0.01	0.01	-0.99	2, 3, 5	0.01	1.0	0.01	0.01	0.22	11.5	0.0	3
2	SARIMA+	0.75	0.01	0.36	None	0.01	0.38	0.87	0.80	0.49	42.6	60.4	
	TSMARS	0.01	0.01	2.06	1, 2, 5, 6	0.01	0.01	0.01	0.01	0.13	51.4	58.6	
	SATSMARS	-	-	-	-	-	-	-	-	-	-	-	-
	STSMARS	0.01	0.01	-0.31	2, 6, 7	0.01	0.62	0.01	0.01	0.15	69.0	58.6	Outliers
	PTSMARS	0.62	0.99	-0.90	4	0.93	1.0	0.01	0.03	0.01	39.1	61.0	
3	SARIMA+	0.04	0.01	2.0	2	0.01	0.90	0.78	0.94	1.0	7.5	10.2	
	TSMARS	0.01	0.01	2.06	2	0.01	0.01	0.13	0.08	0.58	7.6	9.8	3
	SATSMARS	0.01	0.01	1.28	3	0.01	1.0	0.01	0.01	0.58	6.4	9.8	3
	STSMARS	0.01	0.01	1.13	2, 5	0.01	0.01	0.01	0.01	0.55	10.7	9.8	3
	PTSMARS	0.02	0.01	-0.08	None	0.01	1.0	0.06	0.15	0.22	5.5	10.3	3
4	SARIMA+	0.01	0.01	1.63	1, 2, 3, 7	0.01	0.72	1.0	1.0	0.24	1.6	0.0	3, 7
	TSMARS	0.01	0.12	2.32	1, 3, 10 - 12	0.01	0.01	0.01	0.01	0.01	3.0	0.0	3
	SATSMARS	0.01	0.74	1.34	1, 6, 10, 11	0.01	0.02	0.01	0.01	0.91	1.6	0.0	
	STSMARS	0.01	0.01	-0.94	2, 4, 6 - 10	0.01	0.05	0.01	0.01	1.0	2.6	0.0	3
	PTSMARS	0.01	0.01	1.73	All	0.01	1.0	0.01	0.01	0.02	2.8	0.0	
5	SARIMA+	0.21	0.01	-0.43	10	0.01	0.47	0.76	0.85	0.05	1.0	1.2	
	TSMARS	0.01	0.06	0.62	None	0.01	0.06	0.01	0.01	0.42	1.1	1.2	
	SATSMARS	0.01	0.16	0.68	None	0.01	0.93	0.01	0.01	0.46	1.1	1.1	
	STSMARS	0.40	0.45	0.75	None	0.01	0.08	0.15	0.15	0.70	1.1	1.1	Outliers
	PTSMARS	0.01	0.01	-0.19	None	0.01	1.0	0.01	0.01	0.03	1.6	1.1	

No	N	Series Code	Title	Method	Transformations		Model
					Log	Constant	
6	179	LRGM001	Live Register Total (No)	SARIMA+	Y	N	$(0, 1, 1)(0, 1, 1)_s$
				TSMARS	-	-	$y_t = 133{,}416 + y_{t-1}$
				SATSMARS	-	-	$y_t = 136{,}771 + y_{t-1}$
				STSMARS	Y	N	$\Delta\Delta_{12}\left[y_t = -4442 + 0.39y_{t-1} + 1.25(y_{t-12} - 6772)_-\right]$
				PTSMARS	Y	-	$M_1 = 141{,}806 + M_2$, $M_2 = 140{,}060 + M_1$, $M_3 = 135{,}591 + M_1$, $M_4 = 132{,}953 + M_1$, $M_5 = 129{,}394 + 0.98M_1$, $M_6 = 136{,}382 + 0.97M_1$, $M_7 = 141{,}736 + 0.97M_1$, $M_8 = 142{,}943 + 0.96M_1$, $M_9 = 129{,}204 + 1.01M_2$ $M_{10} = 126{,}680 + 0.99M_3$, $M_{11} = 130{,}268 + 0.98M_4$, $M_{12} = 135{,}869 + M_4$
7	179	LRGM111	Live Register/ Tara St. Total (No)	SARIMA+	Y	N	$(0, 1, 1)(0, 1, 1)_s[0, 2, 0, 0]$
				TSMARS	-	-	$y_t = 947 + 0.99y_{t-1}$
				SATSMARS	-	-	$y_t = 989 + y_{t-1}$
				STSMARS	Y	N	$\Delta\left[y_t = -0.02 - 0.48(y_{t-1} - 0.39)_+\right]$
				PTSMARS	Y	-	$M_1 = 6.9 + 0.97M_2$, $M_2 = 6.9 + 1.03M_1$, $M_3 = 6.9 + 1.01M_1$, $M_4 = 6.9 + 1.01M_1$, $M_5 = 6.9 + 1.04M_1$, $M_6 = 6.9 + M_1$, $M_7 = 6.9 + M_1$, $M_8 = 6.9 + 0.99M_1$, $M_9 = 6.9 + 0.98M_2$, $M_{10} = 6.9 + 1.01M_4$, $M_{11} = 8.2 - 1.01M_4$, $M_{12} = 7.0 + 0.96M_8$
8	179	LRGM438	Live Register/ Thomasown Males (No)	SARIMA+	Y	Y	$(0, 1, 1)(0, 1, 1)_s$
				TSMARS	-	-	$y_t = 181 + y_{t-1}$
				SATSMARS	-	-	$y_t = 186 + y_{t-1}$
				STSMARS	Y	Y	$\Delta\left[y_t = -0.05 + 0.31y_{t-12} - 0.5(y_{t-1} - 0.18)_+\right]$
				PTSMARS	Y	-	$M_1 = 5.3 + 0.97M_2$, $M_2 = 5.2 + 1.02M_1$, $M_3 = 5.2 + 1.05M_1$, $M_4 = 5.2 + 1.11M_2$, $M_5 = 5.2 + 0.98M_2$, $M_6 = 5.2 + 0.96M_2$, $M_7 = 5.3 + 0.80M_4$, $M_8 = 5.3 + 0.82M_4$, $M_9 = 5.3 + 0.86M_4$, $M_{10} = 5.3 + 0.83M_4$, $M_{11} = 5.2 + 1.02M_8$, $M_{12} = 5.3 + 1.05M_8$
9	179	LRGM515	Live Register/ Nenagh Males (No)	SARIMA+	Y	N	$(0, 1, 1)(0, 1, 1)_s$
				TSMARS	-	-	$y_t = 314 + 0.99y_{t-1}$
				SATSMARS	-	-	$y_t = 377 + 0.99y_{t-1}$
				STSMARS	Y	N	$\Delta\begin{bmatrix} y_t = -0.29 + 0.17y_{t-12} - 0.26(y_{t-3} + 0.009)_+ + \\ 0.2(y_{t-13} = 0.02)_+ \end{bmatrix}$
				PTSMARS	Y	-	$M_1 = 6.0 + 0.98M_2$, $M_2 = 5.9 + 1.02M_1$, $M_3 = 6.0 + 1.02M_2$, $M_4 = 5.9 + M_3$, $M_5 = 5.9 + 1.01M_3$, $M_6 = 5.9 + 0.98M_3$, $M_7 = 5.9 + 0.97M_3$, $M_8 = 5.9 + 0.96M_4$, $M_9 = 5.9 + 0.97M_5$, $M_{10} = 5.8 + 0.99M_5$, $M_{11} = 5.8 + 0.99M_7$, $M_{12} = 5.9 + 0.97M_8$
10	179	LRGM800	Live Register/ Newcastle West females (No)	SARIMA+	Y	N	$(0, 1, 1)(0, 1, 1)_s$
				TSMARS	-	-	$y_t = 286 + 0.98y_{t-1}$
				SATSMARS	-	-	$y_t = 301 + 0.99y_{t-1}$
				STSMARS	Y	N	$\Delta\left[y_t = -12.3 + 0.27y_{t-3} + 0.4y_{t-12} - 0.01y_{t-3}(y_{t-2} - 123)_+\right]$
				PTSMARS	Y	-	$M_1 = 335 + 1.08M_2$, $M_2 = 345 + 0.89M_1$, $M_3 = 336 + 0.96M_2$, $M_4 = 320 + 1.03M_2$, $M_5 = 306 + 1.03M_2$, $M_6 = 331 + 1.03M_3$, $M_7 = 358 + M_3$, $M_8 = 359 + 0.99M_6$, $M_9 = 285 + 1.05M_5$, $M_{10} = 295 + 1.04M_6$, $M_{11} = 290 + 1.03M_6$, $M_{12} = 315 + 1.13M_{11}$

No	Method	Statistics											
		χ_1^2	χ_2^2	$t_{\mu,\sigma}$	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
6	SARIMA+	0.01	0.01	0.55	2, 4	0.01	0.79	1.0	1.0	0.93	1.2	0	2
	TSMARS	0.01	0.01	-0.62	1 - 6	0.01	0.01	0.01	0.01	0.01	2.2	0	
	SATSMARS	0.01	0.01	0.18	2, 4, 5	0.01	1.0	0.01	0.01	0.20	1.0	0	2
	STSMARS	0.09	0.24	1.57	None	0.74	1.0	0.01	0.01	0.73	1.1	0	
	PTSMARS	0.01	0.01	-0.68	1, 12	0.01	1.0	0.01	0.01	0.01	3.6	0	
7	SARIMA+	0.01	0.01	1.76	4, 6	0.01	0.33	1.0	1.0	1.0	2.5	0	
	TSMARS	0.47	0.96	-0.02	8 - 12	0.01	0.01	0.35	0.44	0.55	3.2	0	
	SATSMARS	0.14	0.99	0.05	3	0.01	0.61	0.03	0.01	0.52	2.6	0	3
	STSMARS	0.65	0.86	-0.06	None	0.11	0.80	0.80	0.07	0.55	3.0	0	
	PTSMARS	0.01	0.01	-2.51	1	0.01	1.0	0.01	0.01	0.18	6.7	0	
8	SARIMA+	0.09	0.02	1.13	None	0.50	0.46	0.93	0.89	0.03	3.3	0.3	3
	TSMARS	0.01	0.01	0.50	1	0.59	0.01	0.22	0.34	0.04	3.2	0.3	
	SATSMARS	0.04	0.01	1.26	1, 4, 6	0.57	0.71	0.01	0.03	1.0	3.5	0.3	3
	STSMARS	0.55	0.06	-1.50	None	0.52	0.01	0.48	0.46	0.96	2.8	0.3	3
	PTSMARS	0.01	0.01	0.58	None	0.01	1.0	0.01	0.01	0.04	4.8	0.3	
9	SARIMA+	0.01	0.01	-0.89	2, 3	0.01	0.19	0.09	0.92	0.01	3.4	0.1	
	TSMARS	0.01	0.97	-0.75	5, 6	0.63	0.01	0.69	0.02	1.0	4.0	0.1	6
	SATSMARS	0.01	0.99	-1.77	None	0.01	0.95	0.92	0.05	0.78	3.1	0.1	2
	STSMARS	0.01	0.84	-1.14	None	0.53	0.01	0.93	0.04	0.92	3.8	0.1	
	PTSMARS	0.01	0.01	-0.96	2	0.01	1.0	0.01	0.01	0.23	5.6	0.1	
10	SARIMA+	0.05	0.01	-0.39	None	0.01	0.34	0.72	0.84	1.0	3.7	0.2	
	TSMARS	0.01	0.14	-1.50	2, 3	0.66	0.01	0.01	0.01	1.0	4.4	0.2	
	SATSMARS	0.01	0.01	0.13	7	0.69	1.0	0.01	0.02	0.98	2.7	0.2	7
	STSMARS	0.80	0.13	1.06	None	0.52	0.04	0.02	0.09	1.0	3.4	0.2	
	PTSMARS	0.01	0.01	-0.11	None	0.01	1.0	0.01	0.01	0.67	3.9	0.2	

No	N	Series Code	Title	Method	Transformations		Model
					Log	Constant	
11	288	MIAM014	Volume Index NACE 37 (Base 1985= 100)	SARIMA+	Y	N	$(0, 1, 1)(0, 1, 1)_s [3, 0, 1, 0]$
				TSMARS	-	-	$y_t = 35.7 + 0.99 y_{t-1}$
				SATSMARS	-	-	$y_t = 46.3 + y_{t-1}$
				STSMARS	Y	N	$\Delta_{12}\left[\begin{array}{l} y_t = -0.15 + 0.25 y_{t-1} + 0.12 y_{t-12} + 0.95 y_{t-1}(y_{t-12} - 0.52)_+ \\ \quad + 2.05(y_{t-3} - 0.28)_+ (y_{t-13} - 0.3)_+ - 0.77 y_{t-12}(y_{t-1} - 0.24)_+ \end{array}\right]$
				PTSMARS	Y	-	$\Delta_{12}[y_t = \text{Mixed } \{\text{EASTER}\}]$
12	288	MIAM051	Volume Index Manufacturing Industries (Base 1985 =100)	SARIMA+	Y	N	$(0, 1, 1)(0, 1, 1)_s \{TD\}$
				TSMARS	-	-	$y_t = 56.0 + y_{t-1}$
				SATSMARS	-	-	$y_t = 62.2 + y_{t-1}$
				STSMARS	Y	N	$y_t = 4.05 + 0.98 y_{t-1}$
				PTSMARS	Y	-	$\Delta[y_t = \text{Mixed } \{\text{TD, EASTER}\}]$
13	288	MIAM524	Volume Index NACE 429 Adjusted (Base 1985= 100)	SARIMA+	Y	Y	$(1, 1, 2)(0, 0, 0)_s$
				TSMARS	-	-	$y_t = 85.3 + 0.29 y_{t-8} + 0.67(y_{t-1} - 27.7)_+$
				SATSMARS	-	-	As above (seasonally adjusted series)
				STSMARS	Y	Y	$\Delta\left[\begin{array}{l} y_t = 0.15 - 0.3 y_{t-2} - 0.14 y_{t-3} + 4.85(y_{t-1} - 0.06)_+ - \\ \quad 0.18(y_{t-1} - 0.06)_- - 0.05 y_{t-12}(y_{t-1} - 0.06)_+ \end{array}\right]$
				PTSMARS	Y	-	$y_t = \text{Mixed}$
14	288	MTAM351	Dublin Airport Rainfall (mm)	SARIMA+	Y	Y	$(0, 1, 1)(0, 1, 1)_s [1, 0, 0, 0]$
				TSMARS	-	-	$y_t = 69.5 - 7.16 \text{Sin(t)}$
				SATSMARS	-	-	$y_t = 70.5 - 0.14 y_{t-12} - 0.005 (y_{t-5} - 144)_+ y_{t-12}$
				STSMARS	Y	Y	$y_t = 3.8 + 0.7(MD - 2.4)_+$
				PTSMARS	Y	-	$y_t = \text{Mixed } \{TD\}$
15	288	MTAM553	Mullingar Rainy Days (No.)	SARIMA+	N	Y	$(0, 1, 1)(0, 1, 1)_s$
				TSMARS	-	-	$y_t = 18.0$ - mean only fitted
				SATSMARS	-	-	$y_t = 19.5 - 0.1 y_{t-12}$
				STSMARS	N	Y	$y_t = 17.1 + 0.13 y_{t-12}$
				PTSMARS	N	-	$y_t = \text{Mixed } \{\text{TD, EASTER}\}$

No	Method	Statistics											
		χ_1^2	χ_2^2	$t_{\mu,\sigma}$	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
11	SARIMA+	0.02	0.01	-0.26	1, 4, 10	0.01	0.25	1.0	1.0	0.4	6.1	0.5	10
	TSMARS	0.01	0.01	-1.03	1, 2, 3, 12	0.01	0.94	0.03	0.01	0.81	12.6	0.5	3
	SATSMARS	0.02	0.01	-1.47	7	0.01	0.98	0.01	0.01	0.08	5.2	0.5	7
	STSMARS	0.80	0.13	1.06	1, 2	0.52	0.04	0.01	0.01	0.71	6.5	0.5	
	PTSMARS	0.04	0.01	-0.42	1, 2, 4	0.01	0.98	0.01	0.01	0.13	3.9	0.6	4
12	SARIMA+	0.01	0.01	0.07	6	0.01	0.65	0.92	0.97	0.16	3.0	0.5	6
	TSMARS	0.01	0.01	-0.77	2, 3	0.01	0.01	0.01	0.01	0.01	7.3	0.5	
	SATSMARS	0.01	0.01	0.84	6	0.01	0.98	0.01	0.01	0.74	2.6	0.5	7
	STSMARS	0.01	0.01	0.56	2, 3, 5, 9	0.01	0.01	0.01	0.01	0.01	7.3	0.5	2
	PTSMARS	0.01	0.01	-0.15	1, 4, 5	0.01	1.0	0.01	0.01	0.72	2.3	0.6	5
13	SARIMA+	0.05	0.01	1.17	1	0.01	0.92	0.72	0.90	0.99	5.5	1.0	
	TSMARS	0.38	0.08	-1.69	1	0.80	0.97	0.01	0.05	0.06	5.3	1.0	
	SATSMARS	"	"	"	"	"	"	"	"	"	"	"	
	STSMARS	0.01	0.01	1.17	6	0.01	0.67	0.03	0.04	0.99	5.4	1.0	10
	PTSMARS	0.02	0.04	-1.39	8	0.53	1.0	0.01	0.01	0.06	4.3	1.0	
14	SARIMA+	0.20	0.01	-0.85	1	0.01	0.21	0.58	1.0	0.16	91.7	1.3	10
	TSMARS	0.88	0.66	-0.08	3	0.51	0.01	0.47	0.27	0.16	76.9	1.2	3
	SATSMARS	0.77	0.90	2.0	2, 3	0.01	0.88	0.36	0.38	0.4	64.1	1.2	4
	STSMARS	0.75	0.58	-0.13	3	0.66	0.01	0.01	0.02	0.99	64.7	1.2	
	PTSMARS	0.89	0.98	0.27	None	0.51	0.99	0.01	0.01	0.72	51.4	1.2	8
15	SARIMA+	0.19	0.16	0.44	None	0.76	084	0.56	0.58	1.0	30.6	5.3	
	TSMARS	0.01	0.79	-0.74	None	0.62	0.01	0.37	0.60	0.01	29.1	5.1	
	SATSMARS	0.21	0.54	-0.75	None	0.69	0.94	0.43	0.30	1.0	23.6	5.1	
	STSMARS	0.33	0.80	0.21	None	0.55	0.01	0.54	0.72	1.0	28.7	5.1	12
	PTSMARS	0.50	0.39	-0.10	None	0.62	1.0	0.69	0.70	0.01	21.4	5.2	

No	N	Series Code	Title	Method	Transformations		Model
					Log	Constant	
16	380	RSAM501	Retail Sale Index: All Business Value Adjusted Base 1990 = 100	SARIMA+	N	Y	$(0, 0, 2)(0, 0, 1)_s[0, 4, 2, 0]$
				TSMARS	-	-	$y_t = 9.7 + y_{t-1}$
				SATSMARS	-	-	$y_t = 9.7 + y_{t-1}$
				STSMARS	N	Y	$y_t = 10.7 + y_{t-1}$
				PTSMARS	N	-	$\Delta\Delta_{12} y_t$ = Mixed
17	466	TRAM009	Exempt Vehicles New (No.)	SARIMA+	Y	Y	$(0, 1, 1)(0, 1, 1)_s$
				TSMARS	-	-	$y_t = 33.2 + 0.39y_{t-1} + 0.5y_{t-10} + 0.24y_{t-11} - 0.001y_{t-6}y_{t-10}$
				SATSMARS	-	-	$y_t = 287.7 + 0.82(y_{t-4} - 304)_-$
				STSMARS	Y	Y	$y_t = 2.6 + 0.4y_{t-1} + 0.221y_{t-3} + 0.24y_{t-12}$
				PTSMARS	Y	-	y_t = Mixed {TD, EASTER}
18	380	TSAM043	Imports SITC 59 Other Chemicals €000	SARIMA+	Y	Y	$(0, 1, 1)(0, 1, 1)_s$ {TD}
				TSMARS	-	-	$y_t = 36848 + 0.19y_{t-13} - 0.6(y_{t-1} - 41443)_-$
				SATSMARS	-	-	$y_t = 23{,}216 + 0.34y_{t-1} - 0.71(y_{t-3} - 31{,}470)_- + 0.41(y_{t-3} - 31{,}470)_+$
				STSMARS	Y	Y	$\Delta\left[\begin{array}{l} y_t = 0.12 - 0.44y_{t-1} - 0.29y_{t-3} + 0.19y_{t-12} \\ \quad + 0.23y_{t-13} + 0.24(MD-1)_+ - 0.23(MD-2)_+ \end{array}\right]$
				PTSMARS	Y	-	$\Delta\left[y_t = \text{Mixed}\{\text{TD, EASTER}\}\right]$
19	380	TSAM055	Imports SITC 71 Power Machinery €000	SARIMA+	Y	N	$(0, 1, 1)(0, 0, 0)_s$
				TSMARS	-	-	$y_t = 21{,}119 + 0.32y_{t-1} + 0.25y_{t-3} + 0.38y_{t-4}$
				SATSMARS	-	-	$y_t = 35{,}056 - 0.41y_{t-4} + 0.59(y_{t-1} - 52{,}932)_-$
				STSMARS	Y	N	$\Delta y_t = 0.28 + 0.18y_{t-2} - 0.87(y_{t-1} - 1.28)_- + 0.2(y_{t-12} - 1.05)_+$
				PTSMARS	Y	-	y_t = Mixed
20	380	TSAM601	Exports Adjusted €000	SARIMA+	Y	N	$(3, 2, 1)(0, 0, 1)_s[5, 1, 3, 0]$
				TSMARS	-	-	$y_t = 67{,}900 + y_{t-1}$
				SATSMARS	-	-	$y_t = 67{,}900 + y_{t-1}$
				STSMARS	Y	N	$\Delta_2\left[\begin{array}{l} y_t = -64.070 - 0.28y_{t-2} + 0.36y_{t-3} - 0.0001(y_{t-1} - \\ \quad 904{,}100)_+(y_{t-13} - 1{,}012{,}000)_- + \\ \quad 0.0001(y_{t-1} - 904{,}100)_+(y_{t-2} - 204)_-(y_{t-13} - 1{,}012{,}000)_- \end{array}\right]$
				PTSMARS	Y	-	$\Delta\Delta_{12}\left[y_t = \text{Mixed}\right]$

No	Method	Statistics											
		χ_1^2	χ_2^2	$t_{\mu,\sigma}$	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
16	SARIMA+	0.02	0.01	0.27	1, 3, 4	0.01	0.96	1.0	1.0	1.0	13.3	1.0	3
	TSMARS	0.01	0.01	-1.15	1, 2	0.01	0.98	0.01	0.01	0.38	1.8	1.0	2
	SATSMARS	"	"	"	"	"	"	"	"	"	"	"	"
	STSMARS	0.01	0.01	-0.87	1, 2	0.01	0.98	0.01	0.01	0.40	1.1	1.0	2
	PTSMARS	0.01	0.01	-1.38	2	0.01	1.0	0.01	0.01	0.74	1.1	1.1	
17	SARIMA+	0.01	0.01	-0.30	2, 4	0.01	1.0	1.0	1.0	0.03	51.6	0.5	4
	TSMARS	0.01	0.01	2.14	None	0.01	0.01	0.01	0.01	1.0	49.7	0.5	
	SATSMARS	0.01	0.01	0.25	None	0.01	0.91	0.01	0.01	0.74	45.8	0.5	
	STSMARS	0.01	0.01	-0.80	1, 4	0.01	0.01	0.01	0.01	0.55	41.8	0.5	
	PTSMARS	0.02	0.01	-0.53	1	0.01	1.0	0.63	0.10	0.27	29.5	0.5	
18	SARIMA+	0.01	0.01	-10.5	1, 4	0.01	0.64	1.0	1.0	0.14	16.9	0	4
	TSMARS	0.01	0.01	-0.74	None	0.01	0.01	0.01	0.01	0.38	22.7	0	4
	SATSMARS	0.01	0.01	-0.75	3	0.01	0.92	0.04	0.01	0.77	16.2	0	4
	STSMARS	0.01	0.01	0.40	2, 3, 4	0.01	0.78	0.01	0.01	0.89	19.6	0	3
	PTSMARS	0.01	0.01	-0.93	2, 6	0.01	1.0	0.01	0.01	0.71	11.9	0	
19	SARIMA+	0.01	0.01	0.90	1, 4	0.01	0.18	1.0	1.0	1.0	19.2	0	4
	TSMARS	0.01	0.08	0.88	None	0.01	0.25	0.01	0.01	0.49	21.0	0	
	SATSMARS	0.03	0.01	-1.0	3	0.01	1.0	0.01	0.01	0.99	17.8	0	3
	STSMARS	0.02	0.03	-0.01	2	0.01	0.4	0.01	0.01	1.0	18.6	0	2
	PTSMARS	0.01	0.04	-0.72	None	0.01	1.0	0.01	0.01	0.52	16.0	0	
20	SARIMA+	0.01	0.01	-1.37	2	0.01	0.32	1.0	1.0	0.25	7.1	0	2
	TSMARS	0.01	0.01	-0.34	2	0.01	0.92	0.01	0.01	0.42	7.2	0	2
	SATSMARS	"	"	"	2	"	"	"	"	"	"	"	2
	STSMARS	0.01	0.01	-0.65	1, 12	0.01	0.99	0.01	0.01	1.0	10.8	0	
	PTSMARS	0.01	0.01	0.22	None	0.01	1.0	0.01	0.01	0.07	7.9	0	

Table 4.6.1.2: ANOVA Analysis

No	N	Series Code	Title	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
1	286	ASAM003	Cows Milk Protein Content (%)	TSMARS	67	31	0	0	2
				SATSMARS	27	72	0	0	1
				STSMARS	24	30	0	0	46
				PTSMARS	91	9	0	0	0
2	286	ASAM206	Calves Slaughtering 000 Heads	TSMARS	18	46	0	0	36
				SATSMARS	-	-	-	-	-
				STSMARS	40	0	47	0	13
				PTSMARS[1]	35	39	22	0	4
3	286	ASAM305	Heifers Slaughtering 000 Tons	TSMARS	78	16	5	0	1
				SATSMARS[1]	66	33	0	0	1
				STSMARS[1]	4	4	1	0	91
				PTSMARS[1]	85	12	4	0	0
4	250	FIAM023	Irish Currency in Circulation (€)	TSMARS	45	55	0	0	0
				SATSMARS	44	56	0	0	0
				STSMARS	25	35	1	0	39
				PTSMARS	91	9	0	0	0
5	324	FIAM102[2]	Exchange Rate $ £STR	TSMARS	83	17	0	0	0
				SATSMARS	82	18	0	0	0
				STSMARS	7	7	0	0	86
				PTSMARS	85	15	0	0	0
6	179	LRGM001	Live Register Total (No)	TSMARS	58	42	0	0	0
				SATSMARS	60	40	0	0	0
				STSMARS	13	12	1	0	74
				PTSMARS	58	42	0	0	0
7	179	LRGM111	Live Register/ Tara St. Total (No)	TSMARS	35	65	0	0	0
				SATSMARS	35	65	0	0	0
				STSMARS	6	0	5	0	89
				PTSMARS	85	11	4	0	0
8	179	LRGM438	Live Register/ Thomasown Males (No)	TSMARS	52	48	0	0	0
				SATSMARS	53	47	0	0	0
				STSMARS	8	8	2	0	82
				PTSMARS	89	11	0	0	0

No	N	Series Code	Title	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
9	179	LRGM515	Live Register/ Nenagh Males (No)	TSMARS	45	55	0	0	0
				SATSMARS	49	51	0	0	0
				STSMARS	6	5	1	0	88
				PTSMARS[1]	90	10	0	0	0
10	179	LRGM800	Live Register/ Newcastle West females (No)	TSMARS	51	49	0	0	0
				SATSMARS[1]	54	46	0	0	0
				STSMARS[1]	17	18	2	0	63
				PTSMARS[1]	56	44	0	0	0
11	288	MIAM014	Volume Index NACE 37 (Base 1985= 100)	TSMARS	24	74	0	0	2
				SATSMARS	32	69	0	0	0
				STSMARS	26	13	29	0	32
				PTSMARS	38	0	0	0	64
12	288	MIAM051	Volume Index Manufacturing Industries (Base 1985 =100)	TSMARS	37	62	0	0	0
				SATSMARS	41	59	0	0	0
				STSMARS	83	17	0	0	0
				PTSMARS	32	27	19	0	22
13	288	MIAM524	Volume Index NACE 429 Adjusted (Base 1985= 100)	TSMARS	85	7	8	0	8
				SATSMARS	"	"	"	"	"
				STSMARS	11	8	4	0	77
				PTSMARS	96	3	1	0	0
14	288	MTAM351	Dublin Airport Rainfall (mm)	TSMARS	71	0	0	7	22
				SATSMARS	72	0	8	0	20
				STSMARS	93	0	5	0	2
				PTSMARS	83	4	11	0	2
15	288	MTAM553	Mullingar Rainy Days (No.)	TSMARS	93	0	0	0	7
				SATSMARS	88	6	0	0	6
				STSMARS	84	9	0	0	7
				PTSMARS	77	5	13	0	5
16	380	RSAM501	Retail Sale Index: All Business Value Adjusted Base 1990 = 100	TSMARS	13	87	0	0	0
				SATSMARS	"	"	"	"	"
				STSMARS	15	85	0	0	0
				PTSMARS[1]	35	40	13	0	12

No	N	Series Code	Title	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
17	466	TRAM009	Exempt Vehicles New (No.)	TSMARS	13	57	15	0	15
				SATSMARS[1]	56	0	29	0	15
				STSMARS[1]	55	44	0	0	1
				PTSMARS[1]	68	21	10	0	1
18	380	TSAM043	Imports SITC 59 Other Chemicals €000	TSMARS	57	5	31	0	7
				SATSMARS	83	0	12	0	5
				STSMARS	9	15	7	0	70
				PTSMARS	32	23	17	0	28
19	380	TSAM055	Imports SITC 71 Power Machinery €000	TSMARS	43	52	0	0	5
				SATSMARS	78	13	5	0	4
				STSMARS	15	13	2	0	70
				PTSMARS	82	14	4	0	0
20	380	TSAM601	Exports Adjusted €000	TSMARS	3	97	0	0	0
				SATSMARS	"	"	"	"	"
				STSMARS	10	16	3	0	71
				PTSMARS	33	23	34	0	10
			Average	TSMARS	48	42	4	1	5
				SATSMARS	49	44	3	0	4
				STSMARS	25	19	6	0	50
				PTSMARS	67	18	8	0	7

5 Forecasting Time Series with TSMARS

5.1 Introduction

Recall from Chapter 1 that a primary focus of this research is short term forecasting of a univariate time series using TSMARS. The literature on TSMARS shows that it has been used only on rare occasions to forecast future values of a time series (see for example, de Goojer et. al. 1998 and Lai & Wong 2001). In this chapter this gap is addressed. This research is novel in that it is the first time an in-depth study of this nature has been conducted for TSMARS.

The methodology adopted in this chapter is based on cross validation. Specifically, a time series is split into an estimation subset and a cross validation subset. The TSMARS model is computed from the estimation set. The model is then used to generate out-of-sample forecasts. These forecast values are compared with their true cross validation counterparts and residuals computed.

In this chapter, forecast errors for the simple time series models considered in Chapter 2 are computed. The purpose of this is to see if TSMARS, produces 'good' forecasts across a variety of time series models. Two quantities are examined, the Mean Forecast Error and more importantly on Root Mean Square Error (RMS) of the forecast. The percentage of forecast errors that lie outside the ±2 RMS prediction interval is computed to statistically test their accuracy.

The RMS is also compared with the true standard deviation; that is, the standard deviation of the predictive distribution. If, as the forecast horizon increases, the RMS value remains within 20% of the true value, then the predictive interval of TSMARS forecasts will be judged consistent. This will be referred to here as the 20% consistency rule. This figure is derived from the ratio of the RMS obtained in S simulation runs to the true standard deviation – that is $F^{100}_{\infty,0.05} = 1.2$, and so 20% reflects the difference due to chance in 100 simulations. It is important to emphasis that this is comparison of two predictive intervals and consistency refers to these two intervals being consistent over the forecast horizon. This comparison is novel for TSMARS forecasts.

In section 6.3 forecasting is conducted for the test-bed of 20 empirical series. Absolute annual residual forecast errors are reported for a five-year-ahead horizon. These values are standardised and compared across the four modelling variations from the previous chapter. The forecast errors are also compared to the MAPE statistics obtained in Chapter 4. The function of this comparison is to see if any modelling variation produces a better forecast than the other variations. Moreover, by comparison with SARIMA+ forecast errors, this study will show whether nonlinear effects contributed to better forecasts.

5.2 Out-of-sample Forecasting using TSMARS

Time series modelling techniques, whether parametric or nonparametric, all require a set of quantities (e.g. statistically estimated parameters or arbitrarily chosen fixed constants, such as, coefficient in an simple moving average) to be specified in order to provide out-of-sample forecasts. In TSMARS these parameters are the constants β_M of the regression splines, the set of variables $v(k,m)$, the associated signs $s(k,m)$ and knots $\xi(k,m)$. With these specified at time T, TSMARS can be used to give a one-step-ahead autoregressive forecast according to equation (2.3.1), namely

$$\hat{y}(T,1) = E(y_{T+1}|I_T) = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m} \left[s_{k,m} \left(y_{v(k,m)} - t^*_{\xi(k,m)} \right) \right]_+ \quad (5.2.1)$$

where $E(y_{T+1}|I_T)$ is the expectation of y_{T+1} conditional on the information set I_T, comprising the lagged values $y_T, y_{T-1}, y_{T-2}, \ldots, y_{T-p}$ and the specified TSMARS parameters.

To obtain forecasts for y_{T+h} $(h \geq 2)$ there are a several possibilities:

- First, a straightforward one-step-ahead forecast $\hat{y}_O(T,h)$ of length h (with O denoting the fact that the forecast is one-step) can be obtained using TSMARS with response y_T and predictors $y_{T-h}, y_{T-h-1}, y_{T-h-2}, \ldots, y_{T-h-p}$. That is, the forecast is based on the information set lagged h steps and so

  $$\hat{y}_O(T,h) = E(y_{T+h}|I_T)$$

  These forecasts can be inefficient since the most recently generated values are not used to obtain the most up-to-date TSMARS model.
- Second, the most recently generated forecast value $\hat{y}(T,h-1), (h \geq 2)$ is used in place of y_{T+h-1} and a sequence of h one-step-ahead forecasts generated. The remainder of the information set I_T is fixed at time T. The key point is that the TSMARS model is evaluated once at time T. It is then re-applied at each forecast step with the most recently generated forecast value plugged-in in place of the actual (unknown) value. The forecast obtained in this way is generally known as the *plug-in* forecast and is given by the nested sequence

  $$\hat{y}_{Pl}(T,h) = E(y_{T+h}|\hat{y}_{Pl}(T,h-1), \ldots, \hat{y}_{Pl}(T,2), E(y_{T+1} \mid I_T))$$

  where the subscript Pl indicates plug-in. The approach is called the *plug-in* principle and when the data are linear it gives consistent forecasts. The plug-in principle is the most common forecasting principle adopted in practice. This is also the principle that is used to generate forecasts in this thesis.
- Third, as above, a sequence of h one-step-ahead forecasts is generated with the information set appropriately updated at each step. This gives the following nested sequence of expectations

  $$\hat{y}_I(T,h) = E(y_{T+h}|E(\tilde{y}(t,h-1)|\ldots E(y_{T+1} \mid I_T)))$$

  where the forecasts $\tilde{y}(T,h-r) = \hat{y}(T,h-r) + \hat{e}_r$ $(1 \leq r \leq h-1)$ are computed by adding the 'appropriate' residual $\hat{e}_r$ to the model forecast value at time point r. This sequence in effect computes the 'law of iterated expectations' forecast denoted by subscript I (see Harvey 1993). It is consistent in that as $h \to \infty$ it gives the unconditional expectation (Harvey 1993).

Once forecast values are available from TSMARS, a key question is to measure their uncertainty or precision. To address this question, the simulation studies for nonlinear models conducted in Chapter 2 are re-worked to give out-of-sample forecasts. So, once again S = 50 or 100 (as appropriate) data sets of $n+h$ sample values are simulated from each nonlinear time series model. The TSMARS model is obtained based on the first n values. This model is then used to generate the actual forecast for the first step ($h=1$) and plug-in forecasts for the remainder of the forecast horizon ($h \geq 2$). Each forecasted value

is then subtracted from the corresponding actual simulated value taken from the cross validation set, giving the cross validation forecast error

$$e(n,k) = y_{n+k} - \hat{y}(n,k)$$

at each time point k, $(n+1 \le k \le h)$. The mean forecast error at each point k on the forecast horizon is then computed as

$$\text{Mean Forecast Error}\,(n,k) = \frac{1}{S}\sum_{i=1}^{S} e_i(n,k)$$

$$\text{Mean Forecast Error}\,(n,k) = \frac{1}{S}\sum_{i=1}^{S} e_i(n,k) \qquad (5.2.2)$$

The root mean square (RMS) forecast error at each point k on the forecast horizon is also computed as

$$\text{RMS}(n,k) = \sqrt{\frac{1}{S}\sum_{i=1}^{S} e_i^2(n,k)} \qquad (5.2.3)$$

This is the precision of the TSMARS cross-validation forecast error.

Using the computed RMS value a prediction interval of ±2 RMS about the mean forecast error is constructed. The percentage of cross validation forecast errors at each step-ahead that fall outside this interval is computed. Based on this the hypothesis

$$H_0: \text{Percentage of prediction errors} > 5\% \qquad (5.2.4)$$

is checked against the alternative. If the test is rejected then the forecast from TSMARS will be judged to be statistically accurate.

The RMS is also compared against the actual forecast precision derived either; (a) algebraically, or (b) by computing the standard deviation of the predictive distribution. The predictive distribution is obtained by numerical evaluation of Chapman-Kolmogorov integral equation using Gaussian quadrature (see Tong 1990 subsection 4.2.4.3) with conditional starting value of zero. The ratio of the RMS value to the true standard deviation will be assessed against the 20% consistency rule.

5.2.1 Forecasts for the SETAR(2,1,1) model

The SETAR(2,1,1) model (2.5.2) is re-examined to assess its forecasting performance. In this case the number of lagged predictors and the maximum interaction degree are restricted to one. Forecasting precision is only reported for the correctly identified SETAR(2,1,1) models (out of 100). The forecasting horizon is set to $h = 6$ steps ahead, as from that point on the standard deviation of the predictive distribution (i.e. true) standard deviation remains constant. The results are shown in Table 5.2.1.1, where the mean and RMS forecast error is given for different sample sizes. In the extreme right, the percentage of prediction errors lying outside the ±2 RMS bands is also given.

The mean forecast error for all samples sizes and at all steps ahead is good, being close to 0 in all cases; in fact, the absolute value in all cases is less than 0.112, which is roughly 1.9% of the data range. Furthermore, the TSMARS plug-in forecasts are statistically accurate as the percentage of errors that lie outside the ±2 RMS interval varies from two to seven. The hypothesis H_0 is therefore rejected. Therefore the TSMARS plug-in values are accurate – though this is borderline according to H_0.

Of greater interest is the RMS forecast error. Applying the 20% consistency rule, the RMS values for the sample size of 500 show initial deviation but settle after three steps. This suggests that the difference between the RMS and true standard deviation is due to different starting values. For the other two sample

sizes, all RMS values are within 20% of their true predictive distribution standard deviation values. Thus the predictive interval of the plug-in forecasts from TSMARS for this model are consistent.

Table 5.2.1.1: SETAR(2,1,1) Model Simulation Forecasting Results

Steps Ahead	Mean Forecast Error			RMS Forecast Error			Predictive Distribution Standard Deviation	Percentage of Errors outside ± 2 RMS Band		
	Sample size			Sample size				Sample size		
	250	500	750	250	500	750	-	250	500	750
1	0.031	-0.011	-0.004	0.244	0.249	0.244	0.264	6	4	7
2	0.032	0.078	-0.006	0.322	0.427	0.289	0.286	4	1	5
3	-0.021	0.006	-0.009	0.295	0.374	0.334	0.294	6	7	5
4	-0.062	-0.069	-0.041	0.309	0.310	0.322	0.298	6	6	6
5	-0.080	-0.072	-0.057	0.336	0.365	0.317	0.300	6	6	5
6	-0.112	-0.094	-0.065	0.341	0.327	0.326	0.301	2	4	5
Correct Models	49	71	86							

5.2.2 Forecasts for the EXPAR(1) and Additive Sine models

Forecasting simulations for these two nonlinear models are taken together. The forecast horizon is once again six steps-ahead. The results are shown in Table 5.2.2.1. Once again, the mean forecast error and RMS precision obtained by TSMARS are displayed. Also reported is the standard deviation of the predictive distribution as well as the percentage of errors outside the ±2 RMS band.

Looking at the results for the EXPAR(1) model (2.5.3) the mean forecast error is close to zero for all steps ahead. This is accurate, as the data range is from about –2.5 to 2.5. In this case the accuracy of the forecast is in line with expectations since the frame (Figure 2.5.3.1) is almost linear. For all steps other than one and three, the percentage of errors is less than five. Therefore the hypothesis H_0 is borderline and so, there is a small degree of doubt over TSMARS plug-in values in this case.

The RMS of the forecast error appears somewhat erratic. However, applying the 20% consistency rule shows, that only the third step RMS value is questionable when compared to the true value. Thus, except for this single instance, the predictive interval of plug-in forecasts from TSMARS is consistent for this model.

The mean forecast errors from Additive Sine model show some variability when compared to zero. The largest is –0.41 and the absolute value of this figure is about 7% of the data range [-3, 3]. The smallest is 0.023 and is about 0.4% of the data range. This variability in mean forecast errors reflects the cyclical nature of the underlying function. Indeed, for all steps the percentage of errors outside the ±2 RMS band is less than five. Therefore the hypothesis H_0 is rejected and TSMARS plug-in values are judged accurate.

The RMS value for the 1st step is accurate. This of course is the innovation error. From step two forward no RMS value is within 20% of the true standard deviation which settles down to 1.238. TSMARS

therefore has not produced a consistent predictive interval in this case. In particular, the poor RMS values are due to the point on the cycle that the forecast is made.

Table 5.2.2.1: EXPAR(1) and Additive Sine Model Simulation Forecasting Results

Steps Ahead	EXPAR(1)				Additive Sine Model			
	Mean Forecast Error	RMS Forecast Error	Predictive Distribution Standard Deviation	Percentage of Errors outside ± 2 RMS Band	Mean Forecast Error	RMS Forecast Error	Predictive Distribution Standard Deviation	Percentage of Errors outside ± 2 RMS Band
1	0.003	0.998	1.005	8	0.073	1.036	1.005	4
2	0.015	1.110	1.032	4	0.167	1.851	1.223	4
3	-0.036	1.549	1.032	6	0.135	1.990	1.235	2
4	0.010	0.802	1.032	4	-0.410	1.910	1.238	4
5	-0.087	1.159	1.032	4	0.023	2.170	1.238	4
6	0.016	1.218	1.032	4	-0.135	2.100	1.238	4

5.2.3 Forecasts for the ARCH(1) model

The conditional expectation of any future value for the ARCH(1) model (2.5.4)

$$y_t = \sigma_t \varepsilon_t \quad \text{with} \quad \sigma_t^2 = \alpha + \beta y_{t-1}^2.$$

is zero. The forecast Mean Square Error (MSE) can be found by applying the law of iterated expectations (Harvey 1993) and is

$$\text{MSE}(\hat{y}_{T+h} | y_T) = \alpha (1 + \beta + \beta^2 + \ldots + \beta^{h-1}) + \beta^h y_T^2$$

In this simulation study the parameters adopted are $\alpha = 0.7, \beta = 0.3$ and $\sigma^2 = 1$, with n = 300. Plug-in forecasts results obtained by applying TSMARS to the data simulated from this ARCH(1) model are given in Table 5.2.3.1. The table gives the mean forecast error, the RMS forecast error, the actual standard deviation obtained taking the square root of the MSE above and the percentage of errors outside the ±2 RMS band.

The mean forecast errors from ARCH(1) model are accurate when compared to zero. The largest is 0.14 and this figure is about 1% of the data range [-9, 7]. The percentage of errors outside the ±2 RMS band is more than five at steps one and four, though the actual values are only 6% and 8% respectively. Therefore the hypothesis H_0 is borderline and so, there is a small degree of doubt over TSMARS plug-in values in this case.

Applying the 20% consistency rule, the RMS of the forecast error is also accurate at all steps other than the sixth. However, repeating the simulations with other data simulated from the model did not show evidence of deviation at the sixth step. So, once again, the predictive interval of plug-in forecasts from TSMARS for this model is consistent.

Table 5.2.3.1: ARCH(1) Model Simulation Forecasting Results

Steps Ahead	ARCH(1)			
	Mean Forecast Error	RMS Forecast Error	Actual Standard Deviation	Percentage of Errors outside ± 2 RMS Band
1	0.016	0.802	0.837	6
2	0.066	0.886	1.000	2
3	0.023	0.838	1.044	4
4	0.065	0.944	1.057	8
5	0.145	1.286	1.061	2
6	0.004	0.504	1.062	4

5.2.4 Forecasts for the Markov Chain model

The forecasting performance of the Markov chain model of Lai & Wong (2001) given in subsection 2.5.6 is examined, once again using the six step-ahead forecast horizon. In this instance true forecast errors are not computed. Only the one-step-ahead error is computed from the variance function

$$V(y_t|y_{t-1}) = 1 + \exp(1 - y_{t-1}) \times (1/3 - y_{t-1}) - \exp(2(1 - y_{t-1}))/4 \qquad (5.2.5)$$

The results are shown in Table 5.2.4.1, where the mean forecast error and RMS precision obtained by TSMARS are displayed. Also given is percentage of cross validation forecast errors lying outside the ±2 RMS band, as well as the one-step forecast standard deviation computed from (5.2.5). Clearly, this forecast standard deviation is constant at all steps ahead.

Table 5.2.4.1: Markov Chain Model Simulation Forecasting Results

Steps Ahead	Markov Chain Model			
	Mean Forecast Error	RMS Forecast Error	One-step Forecast Standard Deviation	Percentage of Errors outside ± 2 RMS Band
1	0.010	0.245	0.243	8
2	-0.117	0.174	0.243	2
3	0.024	0.239	0.243	2
4	-0.009	0.218	0.243	6
5	-0.020	0.233	0.243	2
6	0.028	0.248	0.243	6

The mean forecast errors from this model are accurate for all steps except for step two. The data range is [0,1] so this step has an error of about 12%. The percentage of errors outside the ±2 RMS band is more

than five at steps one, four and six. However the actual percentages are only 8%, 6% and 6% respectively. Therefore the hypothesis H_0 is borderline and so, there is a doubt over TSMARS plug-in values in this case. This is in line with expectations as the TSMARS fit of this model displayed in Figure 2.5.6.1 is poor.

Applying the 20% consistency rule the RMS of the forecast error is accurate at all steps other than step two. However, repeating the simulations with other data simulated from the model did not show evidence of deviation at the sixth step. Thus the predictive interval of plug-in forecasts from TSMARS for this model are consitsent when compared to one-step ahead values.

5.2.5 Concluding Remarks

In this section, TSMARS forecast errors for the simple time series models considered in Chapter 2 were computed. The forecast errors were obtained by cross validation and the Mean Forecast Error and Root Mean Square forecast error were computed over a six-step horizon. Using the RMS, the percentage of errors lying outside the ±2 RMS band was computed at each step. Also computed was the standard deviation of the predictive distribution. The RMS value was compared to this using the 20% consistency rule.

For the time series models considered, Mean Forecast Errors were found to be close to zero. Moreover, the percentage of errors lying outside the ±2 RMS band was also generally less than five. This shows that TSMARS gives statistically accurate forecasts of future values.

The RMS values were also shown to be accurate according to the 20% consistency rule. This was true for all simulations except those conducted for the Additive Sine Model. Short-term forecasts from TSMARS are therefore judged to be both accurate and precise, with the proviso that care should be exercised where the underlying function possesses a turning point. Typically, when data have been modelled with TSMARS, a turning point can be identified by visual inspection of a plot of the frame over the data range (see, for example, the frame associated with the Additive Sine model displayed in Figure 2.5.4.1)

5.3 Forecasting Seasonal Economic Data with TSMARS

5.3.1 Introduction

In this section cross validation based forecasts errors for a five year time period are obtained for the 20 test-bed empirical time series modelled in Chapter 4. In particular, forecast errors from estimated TSMARS models are compared across the four modelling variations. The purpose of this is to see if a particular modelling variation gives superior forecasts. Based on the fact that there was no difference in MAPE statistics in Chapter 4, the initial expectation is that no modelling variation will prove superior.

Linear forecasts errors from the estimated SARIMA+ model are also computed using cross validation. These provide a standard to assess the value of TSMARS based forecasts. The focus of this comparison is to see whether nonlinear models contribute to improving the accuracy of forecasts over the five-year period studied. Here, once again based on results from Chapter 4, the initial expectation is that nonlinear TSMARS and in particular the STSMARS models, do not improve forecast accuracy over linear models.

5.3.2 Forecasting Methodology

As stated in the previous section, all time series methods use a set of quantities (e.g. statistically estimated parameters or arbitrarily specified fixed constants) to provide out-of-sample forecasts. In the

In TSMARS the one-step-ahead autoregressive plug-in forecasts are computed using equation (5.2.1). To obtain forecasts for y_{n+h} $(h \geq 2)$, the plug-in principle is adopted giving the nested forecast value

$$\hat{y}(n,h)_{Pl} = E\left(y(n,h)\middle|\hat{y}(n,h-1)_{Pl},\ldots,\hat{y}(n,2)_{Pl},k_{1,n+h},\ldots,k_{s,n+h},p_{n+h},\mathrm{MD}_{n+h},\mathrm{TD}_{n+h},\mathrm{Easter}_{n+h},E\left(y(n,1)\mid I_n\right)\right) \tag{5.3.1}$$

Here Pl denotes the fact that this is a plug-in forecast value. When h is small and the data are nonlinear then forecasts can be expected to be reasonably close to their true counterparts. Also note, in this forecast, it is assumed that the true values of the independent predictors are available at each point of the forecast horizon.

For the test bed of 20 series the plug-in forecast values are obtained for SARIMA+ and each of the four TSMARS modelling variations of Chapter 4. These monthly forecast values are generated for each year across a five-year time span. That is, for each of the last five years, parameters are estimated based on data up to the end of the previous year. Thus, if 1999 is taken as the last year, the parameters are estimated from data up to and including December 1998. The forecast horizon h is then taken as twelve steps (i.e. 1 year) ahead. An actual forecast value for January is generated from the model. Subsequent forecasts for February etc. are then generated using one-step ahead plug-in forecast values. Appropriately lagged plug-in values are added to account for differencing. Log and/or constant adjustments are applied to give the final forecast value $\hat{y}(n,h)$. This procedure gives twelve future values and the Mean Absolute Percent Forecast Error (MAPFE) over the twelve step horizon is computed in year i $(i = 1\ldots5)$ as:

$$MAPFE(i)\frac{\sum_{j=1}^{h}\left|y_{n+j,i}-\hat{y}(n,j)_i\right|}{\sum_{j=1}^{h}\left|y_{n+j,i}\right|}\times100\% \tag{5.3.2}$$

This repeated for each of the final 5 years 1999 to 2003 inclusive.

5.3.3 Summary of Forecasting Results

The forecasting results for the 20 test bed series are displayed in the Table 5.5.1.1 (see Table Appendix). The figures reported show the MAPFE for each of the 5 years using SARIMA+ and each TSMARS based modelling variation. The overall average of the MAPFE values – that is, the simple average of the MAPFE values for the five years considered, is also given in the extreme right column.

An extract from Table 5.5.1.1 for the Imports of Power Machinery series is given below for examination. The analysis of this series using STSMARS in Chapter 4 showed that a small amount of nonlinearity was present and this accounted for two percent of the overall variance. The constant and linear components accounted for most of the explained variance. This, predominantly linear nature of the series, is borne out by the accuracy of the SARIMA+ method forecasts. In contrast the TSMARS forecasts are all poorer. In particular STSMARS gives poorest forecasts in 2001, 2002 and 2003. In fact the MAPFE obtained by

STSMARS in 2001 is 81.6% while the SARIMA+ value is only 3.5%. Clearly, this difference is too large to be solely attributable to a two-percent nonlinear component. Inspection of Figure 1.1.1 in Chaper 1 shows that an outlier is present in 2001. This effect adversely influences the forecasts of all TSMARS methods, as this large value is propagated by the autoregressive nature of TSMARS. This does not occur with the SARIMA+ method, which returned a 1st differernced order one moving average model. This accounts for the better forecasts in this case.

Extract from Table 5.5.1.1: Forecast Error Analysis

No	N	Title	Method	% Mean Absolute Percent Forecast Error					
				1999	2000	2001	2002	2003	Mean
19	380	Imports SITC 71 Power Machinery €000	SARIMA+	15.1	22.8	3.5	11.6	12.2	13
			TSMARS	19.1	35.0	31.1	12.8	11.9	22
			SATSMARS	21.1	34.8	20.1	20.1	18.6	22.9
			STSMARS	21.7	25.2	81.6	39.4	21.4	37.9
			PTSMARS	23.8	24.6	27.2	8.2	17.6	20.3

As a first step to understanding the results given for the 20 empirical time series in Table 5.5.1.1, the standardised range of the mean column is computed. Thus, for the Mean figures in the extract table above, the standardised range is 1.1. This is computed as the range (i.e. 37.9 - 13) divided by the average of the five values in the Mean column (i.e. 23.2). The standardied ranges for all 20 test series are given in the following table:

Series	1	2	3	4	5	6	7	8	9	10
Range	0.8	1.2	0.3	2.3	0.8	0.6	0.9	0.5	0.4	0.2
Series	11	12	13	14	15	16	17	18	19	20
Range	0.3	0.7	1.4	0.1	0.3	1.9	0.6	2.1	1.1	0.4

These standardised ranges suggest there is a lot of variability between forecasting methods across these series. Only five series have a range of 0.3 (i.e. 30%) or less. This shows the modelling variations adopted are insufficient on their own to capture aspects of the signal that are important in forecasting these data.

A stated purpose given at the beginning of this section is to see whether any modelling variation gives better forecasts than any other variation. This is check by applying the Kruskal-Wallis ANOVA rank test to the mean errors. In this case we get a χ^2 test value of 0.64 >> 0.05. So, there is no evidence that any variation gives significantly smaller forecast errors than all others. Moreover, the test shows that no TSMARS based modelling variation gives smaller mean forecast errors the SARIMA+ method. This test is also applied pair wise to the mean errors obtained by each modelling variation across each year. The results are given in Table 5.3.3.1 and show that no pair-wise comparisons are significant. Thus, no method is better than any other method considered pair wise, nor is there any significant difference between the SARIMA+ approach and the four TSMARS modelling variations. Therefore the initial

hypothesis of no expected difference between TSMARS model variations and between these and linear models, based on the modelling results in Chapter 4 stands.

Table 5.3.3.1: Overall and Pair Wise χ^2 Rank Test Results

Year	Overall	SARIMA+				TSMARS		
		TSMARS	SATSMARS	STSMARS	PTSMARS	SATSMARS	STSMARS	PTSMARS
1999	0.64	0.34	0.76	0.58	0.37	0.18	0.61	0.97
2000	0.67	0.42	0.67	0.86	0.27	0.51	0.32	0.79
2001	0.57	0.61	0.95	0.82	0.10	0.73	0.75	0.34
2002	0.57	0.78	0.27	0.84	0.60	0.57	0.58	0.38
2003	0.37	0.46	0.89	0.42	0.22	0.37	0.90	0.08
Mean	0.80	0.60	0.87	0.50	0.56	0.39	0.82	0.89

Year	SATSMARS		STSMARS
	STSMARS	PTSMARS	PTSMARS
1999	0.37	0.27	0.68
2000	0.71	0.36	0.23
2001	0.81	0.17	0.25
2002	0.21	0.09	0.86
2003	0.42	0.35	0.12
Mean	0.25	0.35	0.97

A second key question posed at the start of this section is, whether forecasts for series that showed nonlinearity are better than their SARIMA+ counterparts. In Chapter 4 only the STSMARS method returned a sufficient number of nonlinear models to make an informed judgement on this question. Recall from Table 4.4.2.3 these were series numbered 2, 4, 6, 7, 8, 9, 10, 11, 13, 17, 18, 19 and 20. The difference between the Overall Mean MAPFE values for SARIMA+ and STSMARS is given in the Table 5.3.3.2.

In Table 5.3.3.2, where a value is positive it indicates STSMARS is better. A positive value occurs for six of the thirteen nonlinear models. Therefore, modelling nonlinearity has not improved the accuracy of the forecast over linear modelling. This conclusion is in line with expectations based on MAPE statistics given in Table 4.6.1.1. The difference in the overall mean MAPE values for corresponding series is also given in Table 5.3.3.2. In this case, there are seven positive values. Thus, the initial hypothesis that nonlinear STSMARS models do not improve forecast accuracy over linear models cannot be rejected.

Table 5.3.3.2: Differences between Overall Mean values for nonlinear models only

MAPFE value Table 5.5.1.1										
Series	1	2	3	4	5	6	7	8	9	10
Difference	-	19.5	-	-3.4	-	-0.4	2.7	-2.5	-3.4	1.9
Series	11	12	13	14	15	16	17	18	19	20
Difference	0.4	-	-2.9	-	-	-	-21.2	59.4	-24.9	2.8
Mean MAPE values Chapter 4 (Table 4.6.1.1)										
Series	1	2	3	4	5	6	7	8	9	10
Difference	-	-26.4	-	-1.0	-	0.1	0.5	0.5	-0.4	0.3
Series	11	12	13	14	15	16	17	18	19	20
Difference	-0.4	-	0.1	-	-	-	9.8	-2.7	0.6	-3.7

5.3.4 Concluding Remarks

In this section, cross validation plug-in forecasts were computed for the 20 empirical time series using each of the four TSMARS modelling variations and SARIMA+. Twelve one-step plug-in forecasts were generated from each model for each of the five years 1998 to 2003. Using the twelve forecast values the cross validation based MAPFE for each year was computed. The overall average of the five years was also computed.

The first question addressed was whether better forecast could be generated by any one five alternatives; namely, the four TSMARS variations and SARIMA+. The overall average MAPFE values were compared across these alternatives. No appreciable difference in MAPFE values was observed in terms of the range of the errors, or in terms of rank based tests. The hypothesis of no expected difference in forecast accuracy between TSMARS model variations and linear models was accepted.

The second question posed was whether forecasts for series that showed evidence of nonlinearity, are better with TSMARS than with their SARIMA+ counterparts. Overall average MAPFE statistics for SARIMA+ and STSMARS were compared. This showed that a nonlinear model did not generate more accurate forecast than an SARIMA+ model. This conclusion agrees with the conclusion of Chapter 4 that these empirical time series do not possess substantial nonlinearity.

Thus, the conclusion from this section for short term forecasting is clear. Sound linear modelling is sufficient to accurately forecast CSO series. Moreover, nonlinear models do not improve the quality of one-year-ahead forecasts.

5.4 Conclusions

In this chapter, cross validation based one-step-ahead plug-in forecast errors were computed. Initially data simulated from a nonlinear time series model was estimated using TSMARS. The estimated TSMARS model was used to generate one-step-ahead plug-in forecasts for a six-step-ahead horizon. Using these forecast values, cross validation forecast errors were computed at each step ahead. This procedure was repeated to create a set of (50 or 100) six-step-ahead forecast errors. The mean of this

set was computed giving the mean forecast error at each step ahead. These mean forecast errors showed that TSMARS generated accurate one-step-ahead plug-in forecast values.

The RMS of the forecast errors was also computed at each step ahead. This was compared to the standard deviation of the predictive distribution, using the novel 20% consistency rule derived from the 5% tails of the F-distribution. In most cases RMS values were consistent with the standard deviation of the predictive distribution. As a consequence, TSMARS was judged to give accurate forecasts with consistent forecast standard errors. It was also observed, that care should be exercised where the underlying function may possess a turning point.

Cross validation based forecasting, was also conducted for the 20 empirical series using the four different TSMARS modelling variations and SARIMA+. Comparisons showed there was no difference in forecasting performance between these alternative methods. Moreover, linear modelling produced forecasts equally as good even when a time series had shown evidence of being nonlinear in Chapter 4.

The implications of the research conducted in this chapter are clear. First, TSMARS will give consistent forecasts with correct forecast standard errors for nonlinear models. Second, forecasting seasonal economic data supports the belief that the data are mainly linear. This suggests that a robust linear method should be initially preferred for short term forecasting empirical economic time series. Third, inefficient forecasts from nonlinear models for empirical data may arise, because of outliers or dependent errors. Recall the Imports of Power Machinery series had a possible shock type outlier near the end of the data. TSMARS models, like all autoregressive models, will propagate this shock forward into the forecasts. TSMARS forecasts are therefore inefficient in the presence of an outlier. In contrast, moving average models are unaffected by shock type outliers. This may well account for the relatively good forecasts obtained by SARIMA+ on this series. In this case, a first order moving average model was fit to the data. Extending TSMARS to deal with shock outliers or dependent errors is explored in the next two chapters.

Table Appendix

Table 5.5.1.1: Forecast Error Analysis

No	N	Series Code	Title	Method	% Mean Absolute Percent Forecast Error					
					1999	2000	2001	2002	2003	Mean
1	286	ASAM003	Cows Milk Protein Content (%)	SARIMA+	7.3	12.1	12.0	11.9	8.4	10.3
				TSMARS	14.2	20.5	33.2	23.7	15.7	21.5
				SATSMARS	7.3	13.4	19.5	8.4	10.3	11.8
				STSMARS	7.3	15.5	13.6	21.3	8.4	13.2
				PTSMARS	12.7	12.9	15.6	12.5	13.9	13.5
2	286	ASAM206	Calves Slaughtering 000 Heads	SARIMA+	82.4	153.5	152.7	75.6	46.4	102.1
				TSMARS	90.7	86.5	64.0	51.2	68.8	72.2
				SATSMARS	-	-	-	-	-	
				STSMARS	82.9	175.1	53.0	49.6	51.8	82.5
				PTSMARS[1]	53.4	40.2	47.9	33.0	48.5	44.6
3	286	ASAM305	Heifers Slaughtering 000 Tons	SARIMA+[1]	16.1	17.6	15.2	10.5	16.1	13.6
				TSMARS	12.4	12.8	15.6	12.8	7.0	12.1
				SATSMARS[1]	11.9	13.9	8.4	12.3	10.0	11.3
				STSMARS[1]	13.8	27.0	13.3	11.4	9.0	14.9
				PTSMARS[1]	7.5	16.4	9.3	13.2	9.1	11.1
4	250	FIAM023	Irish Currency in Circulation (€)	SARIMA+	15.9	13.7	10.2	12.9	6.0	11.7
				TSMARS	10.9	22.0	13.5	13.6	9.8	14
				SATSMARS	15.3	12.9	7.4	16.2	4.3	11.2
				STSMARS	12.3	18.4	9.5	12.0	23.3	15.1
				PTSMARS	62.3	64.6	60.8	60.2	67.6	63.1
5	324	FIAM102[2]	Exchange Rate $ £STR	SARIMA+	6.0	2.7	1.8	0.2	0.2	2.2
				TSMARS	6.7	2.9	6.6	3.9	1.3	4.3
				SATSMARS	6.6	2.7	5.6	2.8	1.7	3.9
				STSMARS	7.1	2.9	5.8	2.5	2.2	4.1
				PTSMARS	1.7	1.6	1.9	2.3	1.0	1.7
6	179	LRGM001	Live Register Total (No)	SARIMA+	2.4	1.2	17.2	13.5	3.5	7.6
				TSMARS	13.4	10.4	5.1	10.0	9.9	9.8
				SATSMARS	10.0	12.9	3.4	12.7	6.8	9.2
				STSMARS	4.2	8.6	14.7	10.0	2.5	8
				PTSMARS	15.4	26.9	12.4	6.5	4.7	13.2

No	N	Series Code	Title	Method	% Mean Absolute Percent Forecast Error					
					1999	2000	2001	2002	2003	Mean
7	179	LRGM111	Live Register/ Tara St. Total (No)	SARIMA+	18.5	14.0	20.1	11.4	5.3	13.9
				TSMARS	27.8	23.4	10.3	10.2	10.0	16.3
				SATSMARS	10.5	26.5	8.9	11.4	4.7	12.4
				STSMARS	8.7	4.5	20.6	13.6	8.6	11.2
				PTSMARS	41.9	53.0	20.3	5.3	3.7	24.8
8	179	LRGM438	Live Register/ Thomasown Males (No)	SARIMA+	6.1	7.5	14.0	14.7	4.5	9.4
				TSMARS	19.1	20.1	8.1	9.1	4.6	12.2
				SATSMARS	13.7	13.3	10.1	12.3	5.1	10.9
				STSMARS	6.4	14.6	10.0	12.9	15.7	11.9
				PTSMARS	20.9	31.8	13.0	5.7	6.1	15.5
9	179	LRGM515	Live Register/ Nenagh Males (No)	SARIMA+	20.9	4.7	8.4	20.9	2.0	11.4
				TSMARS	22.8	12.5	3.5	16.2	10.1	13
				SATSMARS	26.7	8.6	5.6	18.7	2.9	12.5
				STSMARS	17.4	7.3	17.6	18.7	13.1	14.8
				PTSMARS	11.4	18.2	7.9	9.9	2.9	10.1
10	179	LRGM800	Live Register/ Newcastle West females (No)	SARIMA+	2.9	16.9	21.7	27.4	13.6	16.5
				TSMARS	12.7	35.4	6.8	14.2	10.2	15.9
				SATSMARS	8.8	32.8	3.5	16.9	7.3	13.9
				STSMARS	2.9	39.1	12.0	13.7	5.2	14.6
				PTSMARS	5.1	32.3	24.8	5.3	5.2	14.5
11	288	MIAM014[3]	Volume Index NACE 37 (Base 1985= 100)	SARIMA+	17.3	22.6	10.9	6.8	9.6	13.4
				TSMARS	11.2	12.8	15.1	19.5	12.8	14.3
				SATSMARS	16.2	14.3	10.0	7.3	8.5	11.3
				STSMARS	14.5	8.3	15.6	15.7	10.8	13
				PTSMARS[1]	10.1	20.2	23.6	8.3	14.3	15.3
12	288	MIAM051[3]	Volume Index Manufacturing Industries (Base 1985 =100)	SARIMA+	6.1	8.1	5.7	3.9	3.9	5.5
				TSMARS	15.3	6.7	14.7	15.2	4.6	11.3
				SATSMARS	5.4	5.9	7.5	6.2	4.7	5.9
				STSMARS	9.1	7.9	14.8	7.0	10.8	9.9
				PTSMARS[1]	7.6	10.7	9.2	3.4	7.6	7.7
13	288	MIAM524	Volume Index NACE 429 Adjusted (Base 1985= 100)	SARIMA+	5.5	6.8	5.0	8.7	10.5	7.3
				TSMARS	3.9	6.7	7.1	92	8.5	23.6
				SATSMARS	4.7	6.8	7.1	9.6	11.4	7.9
				STSMARS	9.5	7.6	5.5	19.5	9.0	10.2
				PTSMARS	5.6	7.2	6.9	8.8	12.2	8.1

No	N	Series Code	Title	Method	% Mean Absolute Percent Forecast Error					
					1999	2000	2001	2002	2003	Mean
14	288	MTAM351	Dublin Airport Rainfall (mm)	SARIMA+	40.9	36.9	47.4	46.7	75.5	49.5
				TSMARS	40.1	41.3	43.3	50.4	62.2	47.5
				SATSMARS	48.6	31.6	41.8	52.0	54.9	45.8
				STSMARS[1]	40.6	43.8	38.8	54.3	60.7	47.6
				PTSMARS[1]	53.1	41.4	44.9	52.1	58.2	49.9
15	288	MTAM553	Mullingar Rainy Days (No.)	SARIMA+	23.6	15.7	29.7	24.7	44.3	27.6
				TSMARS	17.3	20.7	24.4	22.7	33.0	23.6
				SATSMARS	19.3	14.6	25.0	27.6	38.6	25
				STSMARS	13.9	20.2	23.8	23.6	32.8	22.9
				PTSMARS[1]	37.0	21.7	19.8	32.0	39.9	30.1
16	380	RSAM501[2]	Retail Sale Index: All Business Value Adjusted Base 1990 = 100	SARIMA+	2.9	*	*	*	1.3	2.1
				TSMARS	1.5	2.7	2.9	1.1	3.4	2.3
				SATSMARS	1.5	2.6	3.0	1.2	2.9	2.2
				STSMARS	1.3	2.6	2.9	1.2	3.5	2.3
				PTSMARS	10.2	12.3	11.5	3.3	7.7	9
17	466	TRAM009	Exempt Vehicles New (No.)	SARIMA+	34.7	10.9	27.2	13.6	29.3	23.1
				TSMARS	44.1	34.4	53.7	55.8	25.8	42.8
				SATSMARS	19.4	35.2	56.3	19.5	17.6	29.6
				STSMARS	40.3	49.3	46.3	47.9	37.9	44.3
				PTSMARS[1]	21.8	23.5	34.1	20.4	42.5	28.5
18	380	TSAM043	Imports SITC 59 Other Chemicals €000	SARIMA+[1]	12.9	188.0	29.1	14.9	157.9	80.6
				TSMARS	10.8	15.6	20.1	20.5	15.3	16.5
				SATSMARS	13.0	15.8	22.7	15.4	12.3	15.8
				STSMARS[1]	21.7	19.6	36.6	14.4	13.5	21.2
				PTSMARS[1]	17.5	18.4	24.4	28.7	23.6	22.5
19	380	TSAM055	Imports SITC 71 Power Machinery €000	SARIMA+	15.1	22.8	3.5	11.6	12.2	13
				TSMARS	19.1	35.0	31.1	12.8	11.9	22
				SATSMARS	21.1	34.8	20.1	20.1	18.6	22.9
				STSMARS	21.7	25.2	81.6	39.4	21.4	37.9
				PTSMARS	23.8	24.6	27.2	8.2	17.6	20.3
20	380	TSAM601	Exports Adjusted €000	SARIMA+	3.3	9.3	40.5	7.7	10.1	14.2
				TSMARS	7.2	5.3	16.2	9.7	13.8	10.4
				SATSMARS	7.5	6.6	16.6	10.5	8.7	10
				STSMARS	7.2	8.2	16.2	14.1	11.4	11.4
				PTSMARS	13.6	7.5	20.2	7.2	10.9	11.9

1. Trading effects predictors incorporated in model.
2. Forecasting years cover the 5 year period 1997 – 2001.
3. Forecasting years cover the 5 year period 1995 - 1999

6 TSMARS Outlier Handling

6.1 Introduction

Time series data, like all statistical data, are often subject to outliers or discordant observations. In their presence, Friedman (1991a) remarks that MARS (and therefore TSMARS) is not robust. This is because the GCV model selection criterion is based on the squared error loss function. Therefore, the influence of an outlier can cause TSMARS to accept an incorrect model, say a nonlinear model, when linear model is appropriate.

In this chapter, the initial aim is to ensure consistent model selection; that is, to ensure that outliers do not cause TSMARS to choose an incorrect model, when it would otherwise have chosen a correct model. In this case, a correct model is the one MARS would have chosen had no outlier been present in the data. Consistent model selection is efficiently accomplished in MARS, by incorporating the outlier treatment mechanism directly into the GCV based basis function selection scheme. This is novel and is referred to below as the Current Model Outlier Treatment (CMOT) approach. In addition, three types of outlier adjustment methods are proposed; the first is suitable for independent data, the second is a robust method designed for independent hetroscedastic data, while the third is a novel extension to SETAR models of an existing method suitable for time dependent linear data. In addition, the implementation of each of these methods has novel aspects that ensure MARS retains it computational efficiency.

This chapter comprises three main sections. Section 6.2 is methodological. It sets out the CMOT approach as well as the three adjustment methods. In section 6.3 simulation studies are conducted that compare the adjustment methods. Specifically, interest is focussed on the number of correct models identified by TSMARS. The true number of correct models is computed based on a set of 'original' simulated time series. This is compared to the number obtained by TSMARS, 'with' and 'without' outlier adjustment, based on the same series contaminated with outliers. If, the 'with adjustment' number is closer to the true value, then the CMOT approach is deemed to be model selection consistent. If, in addition, parameter estimates are within two standard errors of their 'original' values, the approach will be defined as statistically acceptable.

In section 6.4 the emphasis returns to empirical series. In particular, for some of the empirical time series studied in Chapters 4, the model selection procedure may have been affected by outliers. This means that TSMARS cannot be relied upon to decide correctly whether a model is linear or nonlinear. This uncertainty casts doubt on the validity and extent of nonlinearity observed for the empirical series. However, with consistent outlier adjustment in place, unbiased conclusions, at least with respect to outliers, can be made regarding the existence and extent of nonlinearity for the set of 'test bed' series. This is the second purpose of this Chapter.

6.2 TSMARS Estimation Methods in the Presence of Outliers

In general two distinct approaches are used to handle outliers. The first is the diagnostic approach where outliers are identified from the residuals of a model. These are incorporated into the model and parameters re-estimated. The outlier treatment procedure implemented in SARIMA+ (see Appendix) is an example of the diagnostic approach.

The second is the robust approach. Here the estimation method is modified so that it is less affected or even unaffected by outliers. This allows robust parameter estimates to be computed and statistical tests based on the resulting robust model to be safely undertaken.

In this section we study the methodology to identify and efficiently adjust for outliers affecting TSMARS. Three adjustment methods are explored; these are a simple least squares based diagnostic approach, a robust procedure based on bounded influence and a time series based diagnostic approach. The performance of all three methods will be reviewed in later sections

6.2.1 Outlier Treatment

In this chapter we examine only the impact of additive outliers (AOs) - TSMARS is adaptive and so level and transitional shift effects will, in general, be directly modelled. The innovation effect is not considered because it propagates in the residuals; these are assumed Normal in standard TSMARS.

In the presence of an AO two difficulties can arise for TSMARS. The first occurs when a knot is located at the AO and TSMARS 'steps-into' it during its stepwise search. This causes TSMARS to locally over-fit as the basis functions adapt to the locally large derivative.

The second difficulty comes about when the knot step interval is greater than 1. Recall from Section 2.1 that every third order statistic is chosen as a potential knot point in the predictor space. This gives a default step interval of 3. In this case, TSMARS can 'step-over' the AO. The AO then affects the resulting estimates through its leverage and influence.

The 'greedy' forward stepwise search procedure in TSMARS, computes the decrease in the RSS at each step; that is, combination of parent basis function, available variable and available knot. In the presence of outliers one obvious possibility is to incorporate a diagnostic and treatment procedure at every step. This is likely to be expensive in terms of computational effort.

A second possibility is to use the approach of de Gooijer, Ray & Krager (1998). They applied Tsay's (1988) outlier detection approach to isolate potential additive outliers in the final TSMARS model. Each of these was incorporated as an independent categorical predictor and the TSMARS model re-estimated. de Gooijer et. al. (1998) concluded that the resulting estimates were much improved.

The outlier adjustment procedure of de Gooijer et. al. (1998) is reasonable. However the diagnostic procedure is based on the final model only. Outliers therefore are only outlying *w.r.t* that model. This means that the stepwise selection procedure used to choose basis functions is still affected by outliers. The TSMARS model may therefore be incorrect.

These two approaches are at opposite ends in terms of computational effort and their impact on the quality of the estimates. The first approach involves considerable work in terms of additional regressions at every step. If the diagnostic procedure adopted is appropriate this approach can be expected to yield consistent estimates. On the other hand, the second approach requires only one (or a small number of) additional regression(s) at each time point t (t = 1..n). In this case, even when the diagnostic procedure is appropriate, the TSMARS estimates can, as noted in de Gooijer et. al. (1998), fall short of optimal.

As a consequence of these drawbacks, neither of the two approaches described is adopted. A third approach that lies in between is chosen. It is called the 'Current Model Outlier Treatment' (CMOT) approach.

For TSMARS this is:

Forward stepwise search

Based on the current M-basis function TSMARS model $\hat{f}_t(M,\bullet)$ (equation (2.3.1)).

- **Apply** a diagnostic procedure to identify outliers for this model.
- **Re-estimate** the current model taking into account the effect of the outliers.
- **With** this augmented model **cycle** through the TSMARS forward stepwise search of each combination of parent basis function, available variable and available knot.
- **Enter** the minimum GCV combination is basis function into the basis function set and an updated TSMARS model $\hat{f}_t(M+1,\bullet)$ estimated. This model is based only on $\hat{f}_t(M,\bullet)$, that is without the outlier effect.

This cycle repeats itself until there is no further improvement in the GCV or the maximum number of basis functions is reached (i.e. as in standard TSMARS).

Backward Deletion.

Based on the current M-basis function TSMARS model $\hat{f}_t(M,\bullet)$.

- **Delete** the ith basis function giving the reduced model $\hat{f}_t(M,\bullet|J \text{ deleted})$ as in standard TSMARS.
- **Apply** a diagnostic procedure to identify outliers for this model.
- **Re-estimate** the reduced model $\hat{f}_t(M,\bullet|J \text{ deleted})$ taking into account the effect of the outliers and compute the resulting GCV.
- **Permanently delete** the minimum GCV basis function (denoted by J^*) from the basis function set.
- **Estimate** the reduced TSMARS model $\hat{f}_t(M,\bullet|J^* \text{ deleted})$ without the outlier effect and the GCV computed.

This cycle repeats itself until there is no further improvement in the GCV or no more basis functions remain to be deleted (i.e. the number of basis functions M = 1).

There are a number of advantages to adopting this approach. First, the effect of the outlier is estimated based on the current M-basis function TSMARS model. The resulting augmented model is therefore conditional on both the predictors and identified outliers. The forward and backward searches are based on this model. As each new basis function is temporarily added or deleted, the RSS and GCV are automatically corrected because the outlier has been incorporated. The associated basis function will therefore be more appropriate, since it gave rise to the minimum GCV. For independent cases we re-state this observation in the following theorem.

Proposition: For an additive outlier of size ω occurring at $x = x_j$ of the form

$$z_i = \begin{cases} y_i & i \neq j \\ y_i + \omega & i = j \end{cases}$$

the RSS computed via Gram-Schmidt (GS) of the agumented MARS model $z_i = f_i(M+1,\bullet)+\omega I(x_j)$ is estimated correctly by the CMOT procedure.

Proof: The proof is stepwise (i.e. by induction). First consider the simplest case where the MARS model being estimated is

$$z_i = \beta_0 + \beta_1 f_i(1,\bullet)+\omega I(x_j) \tag{6.2.0A}$$

that is, $M = 0$. In the CMOT procedure this model is estimated in sequence using GS as

$$z_i = \beta_0 + \omega I(x_j)$$

and then with $r_i = \hat{z}_i - \left(\hat{\beta}_0 + \hat{\omega} I(x_j)\right)$ the next model estimated is

$$r_i = \beta_1 f_i(1,\bullet)$$

The RSS from this sequence is identical to that obtained from OLS estimation of (6.2.0A) via the normal equations, since the sequence of predictors in GS procedure is orthogonal (see Hastie, Tibshirani & Friedman 2001). The RSS is therefore correctly estimated for the effect of the outlier.

Now consider the general situation where there are M basis functions. Incorporating the $M+1^{st}$ basis function we estimate the MARS model

$$z_i = \beta_0 + \sum_{k=1}^{M} \beta_k f_{k,i}(k,\bullet)+\beta_{M+1} f_{M+1,i}(M+1,\bullet)+\omega I(x_j) \tag{6.2.0B}$$

In the CMOT procedure this model is estimated via Gram-Schmidt in sequence as

$$z_i = \beta_0 + \sum_{k=1}^{M} \beta_k f_{k,i}(k,\bullet)+\omega I(x_j)$$

and then with $r_i = \hat{z}_i - \left(\hat{\beta}_0 + \sum_{k=1}^{M} \hat{\beta}_k f_{k,i}(k,\bullet)+\hat{\omega} I(x_j)\right)$ the next model estimated is

$$r_i = \beta_{M+1} f_{M+1,i}(M+1,\bullet)$$

Once again, since in this case the sequence of predictors is orthogonal, the RSS is identical to that obtained from OLS estimation of (6.2.0B) via normal equations. The RSS obtained by adding the $M+1^{st}$ basis function is therefore estimated correctly in the presence of the outlier.

Thus, since the approach is true for the $M = 0$ and also true for the M^{th} basis function model, it is true in general.

This completes the proof.

The CMOT is also computationally efficient, since the approach implements the diagnostic procedure once for each current M-basis function model. Therefore the computing time is only be increased at most by a factor MMAX (the maximum number of basis functions) times n^2. This is not a significant overhead when the overall computing time for the forward stepwise search is $O(n^4)$ (see Friedman 1991a)

The principles of the CMOT approach above do not specify the diagnostic procedure and treatment method to be adopted. Three methods are now described. Each is incorporated as an alternative method into the CMOT approach to outlier handling in TSMARS.

6.2.2 A Least Squares based Outlier Treatment Method

If we assume dependence in the data to be of short memory, then we can appeal to the 'whitening by windowing principle' (see Hart 1996). In essence this says that ordinary least squares is applicable when dependence in a time series is of short duration (i.e. strongly α-mixing, see Appendix). This suggests that it might prove worthwhile to use an outlier treatment method based on standard least squares.

The standard least squares methodology is straightforward (see Meyers 1989):

- **From the model** residuals $e_t = y_t - \hat{y}_t$, compute the standardised prediction residuals

$$r_t = \frac{e_t}{\sigma\sqrt{1-h_{tt}}}$$

  where $\sigma = 1.5\, med|e_t|$ is a robust estimate of the standard deviation (see Meyers 1989) and h_{tt} is the HAT matrix diagonal entry.

- **At each time point** t^* $(t = 1 \ldots n)$ where $r_i > c$ a constant, outliers are identified.

- **For each identified point** t^* an indicator function $I = \begin{cases} 1 \; if \quad t = t^* \\ 0 \; otherwise \end{cases}$ is added as a regressor and the model re-estimated.

Computing the HAT diagonals directly from the HAT matrix (i.e. $H = B\left(B^T B\right)^{-1} B^T$) is expensive as the matrix inversion is $O(n^3)$. However, in TSMARS the corresponding orthogonal matrix of basis function predictors $B^{\perp}$ is already available. Therefore each h_{tt} is simply the norm $(b_t^{\perp T} \bullet b_t^{\perp})$, of the t^{th} row of the $B^{\perp}$ (whose columns $\mathbf{b}_t$ are scaled by their norm). This follows directly from the QR decomposition of an arbitrary predictor matrix X

$$H = X\left(X^T X\right)^{-1} X^T = QR\left((QR)^T QR\right)^{-1} (QR)^T = QQ^T$$

since $Q^T Q = I$. Thus, each diagonal of the HAT matrix is simply $h_{ii} = \mathbf{q}_i \mathbf{q}_i^T$, where i labels the i^{th} row of Q.

With the HAT diagonals cheaply computed the Least Squares Outlier Treatment method is:

- **Compute** the standardised prediction residuals $r_t = \frac{e_t}{\sigma\sqrt{1-h_{tt}}}$.

- **Identify** outliers at each point t^* $(t = 1 \ldots n)$ where $r_t > c\,(= 4)$ a constant.

- **A sporadic indicator function** is constructed as follows:

  - **At each identified point** t^* the indicator function $I = \begin{cases} 1 \; if \quad t = t^* \\ 0 \; otherwise \end{cases}$ is successively built up by introducing a 1 at time point t^*.

- **This sporadic indicator function** is temporarily added as a regressor and the model re-estimated.
- **If this regressor** is linearly independent then the 1 introduced at this time point t^* is retained in the sporadic indicator function.

- **The sporadic indicator function** is added to the basis function matrix when no more outliers are identified.

The key point to note about this method is that, it is assumed all identified outliers arise from the same population – that is, outliers are hypothesised to have a common mean shift (see Meyers Chapter 5 1989). This explains why a single sporadic indicator function, having a 1 at each identified outlier location is added to the basis function matrix.

6.2.3 A Bounded Influence based Outlier Treatment Method

An alternative method of outlier treatment is to use robust regression. This technique is resistant to data values that exert a strong influence on results. The method adopted is based on the Huber M-estimator (see Meyers 1989).

Huber's influence function for the scaled residual $e_t^* = \frac{e_t}{\sigma}$ is given by

$$\psi\left(e_t^*\right) = \begin{cases} r & e_t^* > r \\ e_t^* & \left|e_t^*\right| \le r \\ -r & e_t^* < -r \end{cases}$$

This function is used to 'down weight' scaled residuals greater than a constant r in the basis function regression equation $y_t = \hat{f}_t\left(M, \bullet\right) = \mathbf{b}_t\,\beta$, where $\mathbf{b}_t$ is the t^{th} row of the (ordinary) basis function matrix B. This gives the weighted least squares equation

$$\sum_{t=1}^{n} w_t\, e_t^*\, \mathbf{b}_t = \mathbf{0} \qquad (6.2.1)$$

where the weighted residuals $w_t = \frac{\psi\left(e_t^*\right)}{e_t^*}$ are orthogonal to the regressors. Robust parameter estimates are then computed by iteratively re-weighted least squares (IRWLS), with the weights updated from the residuals obtained at each iteration (see Meyers 1989).

M-estimation, as described here, weights according to the size of the scaled residuals without regard to their leverage. In particular, when h_{tt} is close to unity (i.e. there is an outlier in the x-space) and the outlier is consistent with the fit to the data, then the influence diagnostic

$$DFFITS_t = \frac{\hat{y}_t - \hat{y}_{-t}}{\sigma_{-t}\sqrt{h_{tt}}}$$

where the subscript $-t$ denotes the fact that observation t is left out, will not be unduly large. *Bounded influence regression* results from replacing the scaled residuals in the weighted regression (6.2.1) with a measure that better uncovers outlying data. In this case DFFITS 'bounded influence' residuals (see Krasker & Welsch 1982) are adopted.

The DFFITS statistic is straightforward to compute based on standard formulae (see Meyers 1989). However, in the bounded influence case standard formulae are no longer available. This arises because h_{tt} and the leave-one-out estimate of the standard deviation of the residuals σ_{-t}, are weighted quantities that arise from the IRWLS procedure. To cater for this, the leave one out residuals from each IRWLS step are computed from scratch according to

$$e_{i,-t} = e_i + \frac{h_{i,t}}{1-h_{tt}} e_t \qquad (t,i = 1 \ldots n)$$

The standard deviation of these residuals is then computed using the robust estimate $\sigma_{i,-t} = 1.5\, med\left|e_{i,-t}\right|$.

The leave one out HAT matrix elements are computed efficiently, by noting that the upper triangular Cholesky factor R of the weighted covariance matrix (i.e. $B^T W B$), is available from each IRWLS step. Thus, the resulting bounded influence residuals can be computed using linear algebra (forward followed by back substitution steps respectively). For each bounded influence residual labelled by t^* that exceeds $u_{t^*} = c \times \sqrt{\frac{p}{n}}$ (see Belsley, Kuh & Welsch 1980) for some constant c = 4 and p predictors (i.e. columns of B), down-weighting factors are computed as:

$$w_t = \begin{cases} \dfrac{u_{t^*}}{DFFITS_{t^*}} & \text{if } u_{t^*} > 1 \\ 1 & \text{otherwise} \end{cases}$$

Having obtained the (down) weighting factors the weighted regression (6.2.1) is computed as follows:

- **Remove** each row $\mathbf{b}_{t^*}$ of the basis function matrix B and 'downdate' the associated Cholesky factor R .
- **Weight**, that is multiply each $\mathbf{b}_{t^*}$ by its corresponding weight w_{t^*}, replace this row in B and use this weighted row to 'update' the associated Cholesky factor R .
- **Use** linear algebra, forward/backward substitution, to obtain the weighted LS solution.

This procedure and the computation of the weighting factors is repeated until there is no further improvement in $\sum\left|e_i\right|$ the sum of the absolute residuals.

Note: 'downdate' and 'update' are standard techniques for efficiently obtaining a Cholesky factor R, when a row is removed or added respectively from a regression (see Golub & Van Loan 1996).

The weights that result from this bounded influence IRWLS are then used in the CMOT procedure to find the next minimum GCV basis function to be added/deleted respectively to the TSMARS model.

6.2.4 A Time Series based Outlier Treatment Method

The third method proposed for outlier treatment removes the independence restriction of the two previous methods. This method takes Tsay's regression based methodology (1988) for an ARMA model and applies it to the standard M-basis function TSMARS model $\hat{y}_t = \hat{f}_t(M, \bullet)$.

Consider the i.i.d. ε_t driven stationary AR(1) model

$$y_t = \beta\, y_{t-1} + \varepsilon_t$$

and the corresponding model with residual e_t and additive outlier assumed to exist at time point T, given by

$$z_t = \beta\, z_{t-1} + e_t = \left\{ \begin{matrix} y_t & t \neq T \\ y_t + \omega & t = T \end{matrix} \right\} + \varepsilon_t = y_t + \omega I_t(T) + \varepsilon_t$$

Note, in practice, e_t is estimated from an AR(1) model initially fit to the to the series z_t.

At time point $T+1$ therefore

$$y_{T+1} = z_{T+1} = \beta\, z_T + e_{T+1} = \beta\, y_T + \omega\, \beta + e_{T+1}$$

Using the lag operator B (i.e. $y_{t-1} = B\, y_t$) we get

$$e_t = \omega(1 - \beta B) I_t(T) + \varepsilon_t$$

This is a simple regression equation through the origin for an additive outlier of size ω at time T.

This procedure can also be applied to the SETAR(2,1,1) model (2.5.2)

$$y_t = \begin{matrix} \rho_1 y_{t-1} + \varepsilon_t & if\ y_{t-1} \leq 0 \\ \rho_2 y_{t-1} + \varepsilon_t & if\ y_{t-1} > 0 \end{matrix}$$

With left and right regime indicator functions

$$I_L = \begin{cases} 1 & if\ y_{t-1} \leq 0 \\ 0 & otherwise \end{cases} \text{ and } I_R = \begin{cases} 1 & otherwise \\ 0 & if\ y_{t-1} > 0 \end{cases}$$

this model can be written as

$$y_t = \rho_1 I_L y_{t-1} + \rho_2 I_R y_{t-1} + \varepsilon_t$$

If, as in the AR(1) case, an additive outlier occurs at time point T then

$$\begin{aligned} y_{T+1} = z_{T+1} &= \rho_1 I_L z_T + \rho_2 I_R z_T + e_{T+1} \\ &= \rho_1 I_L (y_T + \omega) + \rho_2 I_R (y_T + \omega) + e_{T+1} \\ &= \rho_1 I_L y_T + \rho_2 I_R y_T + \omega(\rho_1 I_L + \rho_2 I_R) + e_{T+1} \end{aligned}$$

giving

$$e_{T+1} = -\omega(\rho_1 I_L + \rho_2 I_R) + \varepsilon_{T+1}$$

In the SETAR(2,1,1) model it follows in general that

$$e_t = -\omega\left((1 - \rho_1 B) I_L + (1 - \rho_2 B) I_R\right) I_t(T) + \varepsilon_t \qquad (6.2.2)$$

This, once again, is a simple regression equation through the origin for the additive outlier of size ω occurring at time T in the SETAR(2,1,1) model. It leads to an obvious least squares estimator of ω when

T is known, e_t is available from the SETAR(2,1,1) model initially fitted to the series z_t and ε_t is assumed i.i.d.

This result can be straightforwardly extended to the case where an additive outlier occurs in a model where the AR order in each regime is 2 or higher. In this instance the lag polynomials can replaced by their general counterparts

$$\pi_L(B) = \left(1 - \rho_{L,1}B - \rho_{L,2}B^2 - \ldots - \rho_{L,p_L}B^{p_L}\right)$$
$$\pi_R(B) = \left(1 - \rho_{R,1}B - \rho_{R,2}B^2 - \ldots - \rho_{R,p_R}B^{p_R}\right)$$

giving the regression

$$e_t = -\omega\left(\pi_L(B)I_L + \pi_R(B)I_R\right)I_t(T) + \varepsilon_t \tag{6.2.3}$$

Equally, and without loss of generality, equation (6.2.2) can be extended to model the effect of an outlier on 3 or more regimes. In the case of 3 regimes, with M denoting the middle regime, the regression equation that follows is

$$e_t = -\omega\left(\left(1-\rho_1 B\right)I_L + \left(1-\rho_2 B\right)I_M + \left(1-\rho_3 B\right)I_R\right)I_t(T) + \varepsilon_t$$

leading in general l – regime case to the regression

$$e_t = -\omega\left(\left(1-\rho_1 B\right)I_1 + \left(1-\rho_2 B\right)I_2 + \ldots + \left(1-\rho_l B\right)I_l\right)I_t(T) + \varepsilon_t \tag{6.2.4}$$

Combining equations (6.2.3) and (6.2.4), we get the following result in general which is given as a theorem.

Theorem: For a single additive outlier occurring at time T of the form

$$z_t = \begin{cases} y_t & t \neq T \\ y_t + \omega & t = T \end{cases}$$

in the general stationary l -regime SETAR$(l, p_1, p_2, \ldots p_l)$ model, the regression equation for the additive outlier parameter of size ω is

$$e_t = -\omega\left(\pi_1(B)I_1 + \pi_2(B)I_2 + \ldots + \pi_l I_l(B)\right)I_t(T) + \varepsilon_t$$

where $\pi_K(B)$ is the K^{th} order lag polynomial occurring in a regime.

Proof: Combining equations (6.2.3) and (6.2.4) the result follows directly.

Note the stationary restriction on the time series forces all parameters to be less than 1 in absolute value. In the TSMARS context this means that the parameters must be tested before the method can be applied. In practice this makes the method sub-optimal in a nonparametric setting like TSMARS.

Based on this result the following time series based outlier method is proposed:

Given an l -regime SETAR$(l, p_1, p_2, \ldots p_l)$ model

- **At each time point** $t\,(=1,\ldots,n)$ an indicator function appropriate to the model is introduced at T as

$$I_t^*(T) = -\omega\left(\pi_1(B)I_1 + \pi_2(B)I_2 + \ldots + \pi_l I_l(B)\right)I_t(T)$$

- **Regress** this function on the model residuals e_t and the t-value t_T computed.
- **Define** $t_{T\max} = \max_t |t_T|$ where T denotes the time when the maximum occurs. If $t_{T\max} > c \quad (c = 4)$ then there is an additive outlier at time T with its effect estimated by $\hat{\omega}_{T\max}$.
- The outlier is incorporated into the model using Gram-Schmidt orthogonalisation of the indicator function and residuals re-computed.

The above steps are repeated until no further outliers are found.

Note, if k outliers are identified, there is no need to regress these together to eliminate masking. The reason for this is because each associated indicator function is already orthogonalised. This is equivalent multiple regression (see Hastie, Tibshirani & Friedman 2001).

6.2.5 Methodological Remarks

The key principle underlying the CMOT outlier treatment approach described is that the RSS (and therefore GCV) is corrected for the outlier effect before the search for a new basis function begins. Each basis function, is in fact a weak classifier that locally discriminates response values on the basis of a threshold (i.e. change point) in a predictor (see Hastie, Tibshirani & Friedman 2001, Chapter 9). With outliers present the selection of classification boundaries will be distorted. This distortion is corrected using the CMOT approach. This is not alone true for MARS but is also true for its relatives such as recursive partitioning, TURBO (see Friedman 1991) and GAMs (Hastie & Tibshirani 1990). In fact, even tree based weak classifiers such as CART (Brieman, Friedman, Olshen & Stone 1984) could be adapted to include the CMOT approach. This should improve selection of classification boundaries in the presence of outliers.

Another important feature of the CMOT approach is that no diagnostic procedure is specified to identify outliers. This makes the approach quite general, in that a diagnostic method appropriate to the data can be selected. Here three different methods have been proposed. The first of these is the straightforward and efficient LS method using only one sporadic indicator function for all outliers; the second is sophisticated but relatively inefficient Bounded Influence method; the third is appropriate for AR and SETAR time series. In this case, the key restriction is that parameter estimates remain in the stationary region. The usefulness and accuracy of these methods, within the CMOT approach to outlier treatment, is investigated in the next section using simulated time series.

6.2.6 Predictor Space Adjustments

The outlier adjustment procedures above are also augmented with adjustment in the predictor space. This takes two forms:

- Elimination of certain discordant knots from the knot space prior to the forward stepwise search.
- Elimination of certain discordant values in the orthogonal spline predictor space $B^{\perp}$

Elimination of discordant knots is accomplished by computing the (Euclidean/Mahalanobis) distance of each knot value from its predictor centre and excluding knot values outside 3 standard deviations. The alternative of computing the robust distances using Minimum Volume Ellipsoid (MVE) algorithm (see Rousseeuw 1985) available in SAS was tested. This method was slow and in contrast to the findings of

and Rousseeuw and van Zomeren (1990), did not give a discordant set (of knots) significantly different from that based on the (Euclidean/Mahalanobis) distance.

Elimination of the discordant values in the orthogonal predictor space is accomplished by computing the HAT diagonals of $B^{\perp}$ at the start of the outer loop of the forward stepwise search – note, these can be computed very cheaply as outlined for the LSAO method. The resulting locations identified are then set to zero for each new basis function that enters during forward stepwise search. Thus, that point on that basis function does not contribute to the regression fit and the computation of the lack-of-fit criterion.

6.3 Outlying Observation based Simulation Studies

TSMARS selects autoregressive linear and/or nonlinear structure by detecting the presence or otherwise of a threshold (i.e. knot) in the set of lagged predictors. When a series is contaminated TSMARS estimates may not be optimal. To see whether the CMOT approach corrects for this a set of simulations on univariate time series models is conducted. The CMOT procedure will be deemed to be model selection consistent, if the number of correct models obtained in the presence of outliers is in line with the number obtained on same data without outliers. If, in addition, parameter estimates are within two standard deviations of their true values, then the procedure will be considered to be statistically acceptable.

Each simulation initially involves generating a time series based on the model, with the first 100 generated sample values discarded to allow for "burn in". TSMARS estimates labelled 'Original' are then obtained from this simulated series. To the simulated series a single additive outlier is generated and introduced into the series. So called 'Contaminated' TSMARS estimates are then computed for this series. On this contaminated series TSMARS estimates are also computed using each of the three outlier treatment methods denoted by LS (least squares), BI (bounded influence) and TS (time series) respectively. This procedure is repeated 100 times and the average value of the parameter estimates, their standard errors and the number of correct models obtained are reported. The full set of simulations is repeated with 3 and 5 additive outliers respectively, placed at equally spaced time points. The detailed results are given in a Table Appendix at the end of this chapter.

6.3.1 Simulation of a linear AR(1) model

This simulation study examines the ability of TSMARS to identify a simple linear AR(1) model with known coefficients with data contaminated by outliers. The data are generated from a stationary AR(1) model (2.5.1) with autoregressive parameter ρ, $(|\rho|<1)$, driven by normally distributed noise $\varepsilon_t = N(0,\sigma^2)$

$$y_t = \mu + \rho\, y_{t-1} + \varepsilon_t \qquad t = 1,2,\ldots,n$$

Simulations were performed for $\rho = 0.5$ and 0.8 with $\mu = 0$ and $\sigma^2 = 1$. Outliers at time point T for this series are generated from the model by adding 3 standard errors to the series value at time T according to

$$y_{T,\text{Contaminated}} = y_T + 3 \times \sqrt{\frac{\sigma^2}{1-\rho^2}}$$

Table 6.3.1.1 shows the simulation results for $\rho = 0.5$. TSMARS was called with response y_t, one lagged predictor y_{t-1}, maximum interaction degree set to 1 and a sample size of n = 100. This experiment was repeated allowing 3 lagged predictors $y_{t-1}, y_{t-2}, y_{t-3}$ and maximum interaction degree of 3. All simulations were conducted with the smoothing parameter set to 3.

Displayed in Table 6.3.1.1 are the number of times an AR(1) model was correctly identified from the 100 simulation data sets. Also given is the average value of the estimated parameter $\hat{\rho}$ and its "Actual" standard error computed from the correctly identified models. The tolerance on the basis functions is $5.0X10^{-2}$.

With the 'Original' data, the results show the number of models correctly identified by TSMARS is 99 and 95, using 1 and 3 lagged predictors respectively. When contaminated with outliers the number of correctly identified models is affected. However, with adjustment, the number of correctly identified model is brought into line with the 'Original' number. In all but one set of simulations (of the 9 reported), this correction in the number of correct model is observed. The CMOT procedure is therefore model selection consistent in this case.

In general, all three adjustment methods (LSAO, BIF and TSAO) give very similar parameter estimates. These show only a slight improvement over the contaminated values. The outlier adjusted parameter estimates are well within two standard errors of the true values. Therefore the methodology is also statistically acceptable.

6.3.2 Simulation of a SETAR(2,1,1) model

The second simulation study is based on the SETAR(2,1,1) model (2.5.2)

$$y_t = \begin{matrix} \rho_1 y_{t-1} + \varepsilon_t & if\ y_{t-1} \le 0 \\ \rho_2 y_{t-1} + \varepsilon_t & if\ y_{t-1} > 0 \end{matrix}$$

driven by normally distributed noise $\varepsilon_t = N(0, 1/4)$. Outliers at time point T for this series are generated from the model by adding 3 standard errors to the series value at time T according to the regime

$$y_{T,\text{Contaminated}} = y_T + 3 \times \begin{cases} \sqrt{\sigma^2/(1-\rho_1^2)} & if\ y_{T-1} \le 0 \\ \sqrt{\sigma^2/(1-\rho_2^2)} & if\ y_{T-1} > 0 \end{cases}$$

TSMARS is applied to each simulated time series with the basis function tolerance set at $2X10^{-3}$ and smoothing parameter set at 3. A number of simulation studies were run on different series lengths and parameter values. The results for n = 500, with parameters values ρ_1 = 0.75 and ρ_2 = 0.25 are shown in Table 6.3.2.1 (see Table Appendix). In this table, the number of correctly identified SETAR models is given, along with the average values of the parameters ρ_1 and ρ_2, and standard errors labelled Std. Err. for the correct models. For the SETAR model the threshold/knot must also be estimated. Its average value and standard error are also given for the correctly identified models.

A scan of Table 6.3.2.1 gives the number of correct models for the 'Original' data at 82. When contaminated the number of correct models identified increases to nearly 100 in all cases. The stepwise selection mechanism in TSMARS is being affected by outliers, as there are about 18 false positive models. In contrast, the number of correct models obtained by each outlier adjustment method is much closer to the true number in every case. Moreover, there is little to choose between the methods, though

the BIF method does show slightly better performance overall. The methodology is therefore consistent for the model studied.

The parameters ρ_1 and ρ_2 and their standard errors have also been affected by outlier contamination. The adjusted values are not significantly different from their contaminated counterparts. They are however within two standard errors of their true values and so the methodology is statistically acceptable. Of particular interest are the knot estimates. In nearly every case these are more accurate than their contaminated counterparts. This is a significant improvement. It shows that the bias introduced by the 18 false positive models is removed by the CMOT procedure.

6.3.3 Simulation of the nonlinear additive sine model

This simulation revisits the nonlinear periodic time series equation (2.5.4)

$$y_t = 1.5\,\mathrm{Sin}\left(\frac{\pi\, y_{t-2}}{2}\right) - 1.0\,\mathrm{Sin}\left(\frac{\pi\, y_{t-3}}{2}\right) + \varepsilon_t$$

initially studied in subsection 2.5.4 and driven by Standard Normal noise ε_t. In this study, 100 time series of length n = 120 are generated and TSMARS is called twice with 3 predictors $y_{t-1}, y_{t-2}, y_{t-3}$ and maximum interaction 1. On the second call each outlier is replaced by its predicted value from the first call. The TSMARS model obtained from the second call, is used to generate the frame over a set of predictor (i.e. x-values) that correspond roughly to the range of values of y_t. The average frame response value denoted as "Original" along with the underlying nonlinear sine function $y(x)$ are displayed in Figure 6.3.3.1.

Also plotted in Figure 6.3.3.1 are the average outlier "Contaminated" as well as the treated LSAO, BIF and TSAO frame functions for the 100 simulated series. These average frame estimates have been obtained by introducing 5 randomly placed additive outliers into the data. This roughly corresponds to 1 shock occurring every second year if these were monthly data. Given past data $Y = (y_{t-1}, y_{t-2} \cdots)$ and standard deviation function $\sigma(y_t|Y)$ estimated via simulation, each additive outlier is computed at the randomly chosen time point T and added to the series value at that point according to

$$y_{T,\,\text{Contaminated}} = y_T \begin{cases} -3 \times \sigma(y_T|Y) & \text{if } y_T < 0 \\ +3 \times \sigma(y_T|Y) & \text{if } y_T \geq 0 \end{cases}$$

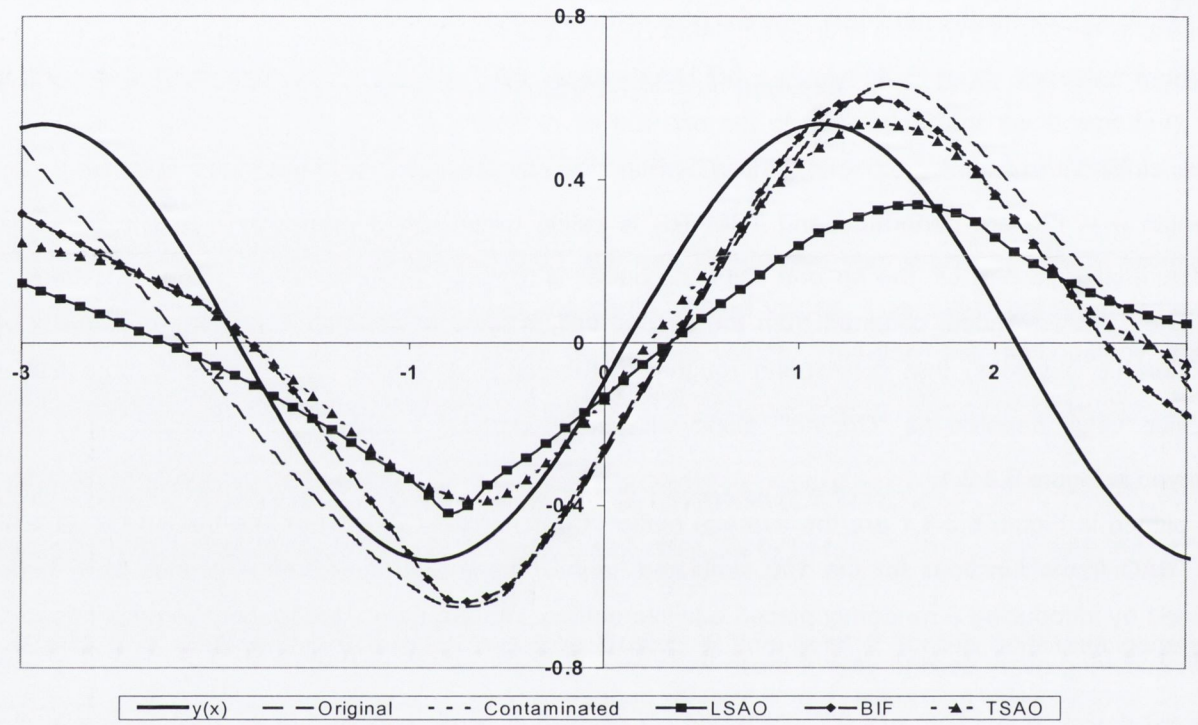
Looking at Figure 6.3.3.1, the frame obtained from modelling the "Original" data with TSMARS tracks the underlying additive sine function quite well. It will be noticed that this is quite an improvement over the frame obtained in Chapter 2, where the frame overshot the underlying function considerably. This improvement is due to the introduction of an additional parameter. This controls the ratio of the GCV at the start of the forward search, to the GCV obtained as each new basis is added. Using a value of 0.99 for this ratio gave the original frame of Chapter 2, while a value of 1.15 gives the frame in Figure 6.3.3.1.

When the original time series is contaminated, the resulting frame still displays periodicity and tracks the 'Original' series closely (compare with Figure 2.5.4.1 where standard errors are shown). The 'Contaminated results are computed from 100 acceptable simulation models. In contrast, the number of correct models obtained with 'Original' data is 82. Here an incorrect model is defined as one that results in

a single (i.e. the mean value) basis function. On this basis the 'Contaminated' data produces 18 false positive models.

For the adjustment methods, the LSAO method gave 96 correct models, the BIF method produced 100, while the TSAO method resulted in 88 correct models. The BIF method therefore is not consistent. The LSAO shows a small degree of consistency, while the TSAO method brings the number of correctly identified model more in line with the 'Original' number. It is therefore judged consistent within the CMOT procedure.

Figure 6.3.3.1: Frames of Nonlinear Additive Sine Model with Additive Outliers



Examining the frames, the LSAO (least squares) fit is poorest, reproducing a cycle whose period is acceptable but the resulting frame lacks symmetry. The TSAO (time series) method gives much better results but tends to undershoot the Sine function. The BIF (bounded influence) method gives very good estimates; these are in fact the same as the 'Contaminated' data. However, as the method is not consistent, the fit is not statistically acceptable. Therefore, based on the frame plot, only the TSAO methodology can be judged as statistically acceptable in this case

6.3.4 Simulation of a Markov model

In this last simulation study, the Markov model of Lai & Wong (2001) studied in subsection 2.4.6 is revisited. Recall the Markov chain $\{y_t\}$ has state space [0,1] and transition p.d.f.

$$p(y|x) = \begin{cases} e^{1-x} & \text{if } 0 \le y \le x \\ e^{1-x} - e^{y-x} & \text{if } y \le x \le 1 \end{cases}$$

This chain has a nonlinear regression function

$$E(y_t|y_{t-1}) = y_{t-1} - 1 + \exp(1 - y_{t-1})/2$$

variance function

$$V(y_t|y_{t-1}) = 1 + \exp(1 - y_{t-1}) \times (1/3 - y_{t-1}) - \exp(2(1 - y_{t-1}))/4$$

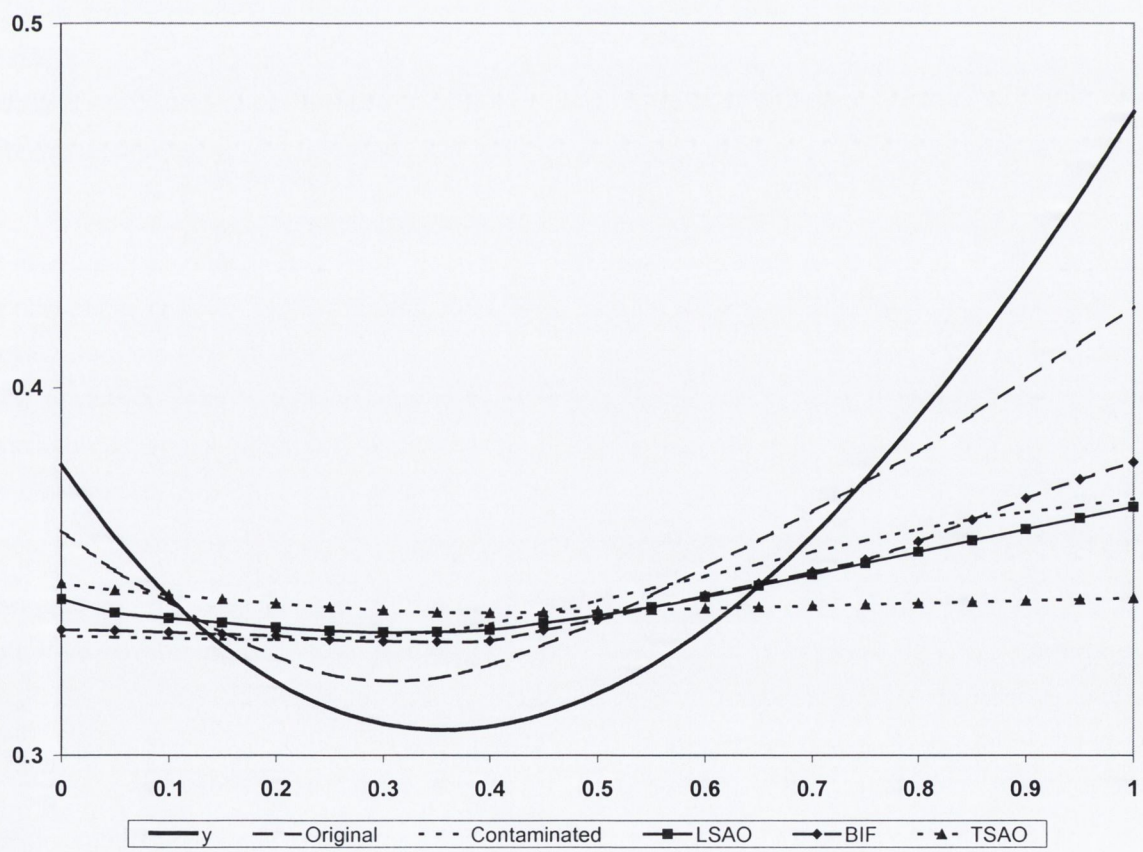
and the residuals are not i.i.d. and do not have normal marginal distributions.

The simulation study involves generating 100 time series data sets, each containing 300 sample values according to this Markov chain model. TSMARS is called for each data set with one lagged predictor y_{t-1} and the estimated model is used to compute the frame (as in the previous subsection). The average frame response value for uncontaminated data, denoted as "Original", along with the underlying nonlinear regression function $y(x)$ are displayed in Figure 6.3.4.1. Also plotted is the "Contaminated" and treated LSAO, BIF and TSAO frame functions.

The average frame estimates are obtained by introducing 5 randomly placed additive outliers into the data. Each additive outlier is computed at a randomly chosen time point T and added to the series value at that point according to

$$y_{T,\text{Contaminated}} = y_T \begin{cases} -3 \times \sqrt{V(y_T|y_{T-1})} & \text{if } y_T < E(y_T|y_{T-1}) \\ +3 \times \sqrt{V(y_T|y_{T-1})} & \text{if } y_T \geq E(y_T|y_{T-1}) \end{cases}$$

Figure 6.3.4.1: Frames for the Markov Chain Model Simulation with Additive Outliers



Looking at the frames in Figure 6.3.4.1 the original estimates track the curvature of the nonlinear regression function $y(x)$ reasonably well up to x = 0.8. After that TSMARS estimates tend to be too flat. The "Contaminated " data frame is badly effected by the five outliers added to the data. The fit obtained

from the outlier treatment methods does not improve on this. Therefore, the methodology does not appear to be robust when the outliers arise in data that are not Normal. Moreover, the number of correct models found was either 24 or 25 for both the contaminated data and all outlier treatment methods. The 'Original' data (see Chapter 2) gave 50 correct models. The CMOT methodology in this case has not improved consistency and has not reproduced statistically acceptable estimates.

6.3.5 Concluding Remarks

The simulation studies conducted have ranged across four different models that cover a broad spectrum of nonlinearity. The original uncontaminated estimates obtained by TSMARS were used as a basis for comparison. One or more additive outliers were introduced into the data and the impact on the estimates measured. Each simulation was then repeated with the three available methods, namely LSAO (least squares), BIF (bounded influence) and TSAO (time series) implemented within the CMOT procedure. If the methodology brought the number of correct models into line with 'Original' value, it was judged model selection consistent. If, in addition, the estimates were not affected (in terms of standard errors) the methodology was judged statistically acceptable.

The results of the simulation studies showed that when outliers were introduced the so-called "Contaminated" TSMARS estimates fell well short of optimal. This agrees with the observations of de Gooijer et. al. (1998). In general, the three treatment procedures improved the quality of the estimates. The simplest method based on least squares gave poorest improvement. Both the BIF and TSAO methods gave good results when residuals were normally distributed. Parameter estimates tended to be biased. However, this problem is easily corrected by using the estimated values to replace the outliers in a second call to TSMARS. If this is done the parameters will be almost identical to their original uncontaminated values.

The CMOT procedure is designed to ensure that model selection in TSMARS is consistent estimates in the presence of outliers. In particular, the objective is to ensure the number of correct models is in line with the number obtained on the same data without outliers. For the linear and threshold models the methodology was shown to be model selection consistent and to give statistically acceptable estimates. However, when the data was not suited to threshold modelling, results were more mixed. The methodology did ameliorate model selection consistency for the Sine model. However, for Markov model, the methodology did not show any degree of improvement in consistency.

Therefore the aim of correcting stepwise knot selection in the presence of additive outliers is achieved. This is true for both for data suited to TSMARS modelling and also true for curved data driven by normal disturbances.

6.4 Modelling Seasonal Economic Data with Outlier Adjusted TSMARS

As noted in Chapter 4, modelling real time series data is a necessary contrast to simulation studies based on an assumed model. A 'test bed' of 20 monthly economic flow, stock and index series was introduced that possess seasonal, independent effects as well as potential outliers and possibly nonlinearity. These

series were each studied using four alternative modelling model types of TSMARS. The key findings of the modelling study indicated:

- It was important to difference growth data prior to TSMARS modelling.
- Incorporating seasonal effect in the model was preferred to prior seasonal adjustment.
- There was evidence of nonlinear behaviour but the share (as a portion of the overall variance) was small.
- Outliers may have confounded statistical tests.

The last of these findings suggests that unbiased conclusions could not be deduced regarding the existence or extent of nonlinearity until TSMARS was enhanced. In simulation studies, the extensions to TSMARS outlined earlier in this chapter, made it robust against outliers when disturbances are Normally distributed. In particular, the simulations showed that model selection inconsistency brought about by outliers was removed. Therefore, TSMARS could be relied upon to decide correctly whether a model is linear or nonlinear. This means that the veracity or otherwise of the findings of Chapter 4, can now be re-examined with greater certainty using outlier handling in TSMARS.

6.4.1 TSMARS Data Modelling with Outlier Adjustment

As in Chapter 4, the test bed of 20 economic time series are each modelled as univariate data with independent predictors to account for length of month (MD), trading week length (TD) and Easter. Only two of the four alternative model types of Chapter 3 are studied; these are TSMARS and STSMARS. These two are chosen because they represent two extremes. TSMARS represents naive modelling in that it involves no data transformation. On the other hand, STSMARS represents intelligent modelling in that the data are appropriately transformed prior to modelling. Recall the descriptions are:

- **TSMARS**: Where no transformations are made to the time series values y_t. A set of s+1 monthly (s=12) lagged predictors $y_{t-1},\ldots,y_{t-(s+1)}$, as well as a deterministic seasonal predictor $p_t = Sin(2\pi\, i/s)$ (i=1...s) and a set of s categorical predictors (each having a 1 in month i and denoted by k_i) is computed and input to the TSMARS program. The maximum interaction degree set to 3 and basis function threshold = 2 X 10^{-8}. This gives the enhanced TSMARS approximation equation (3.4.1)

$$\hat{y}_t = f\left(x_{yt-1},\ldots y_{t-(s+1)}, k_1,\ldots,k_s, p_t, \mathrm{MD_t}, \mathrm{TD_t}, \mathrm{Easter_t}\right)$$

  where $f(\bullet)$ denotes the TSMARS model.

- **STSMARS**: In this model type the time series is checked and adjusted, as appropriate, for a log, a constant and the set of difference transformations (as set out in Chapter 1) giving the transformed series denoted by $z_t = (1-B)^d\left(1-B^s\right)^D\{\log(y_t + c)\}$

  where B denotes the backward difference operator $(B\,y_t = y_{t-1})$, d (=0, 1, 2) denotes the regular difference operator, D (=0,1) denotes the seasonal difference operator, c is a constant adjustment. The lagged predictors $z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}$ are then input into the TSMARS program along with

appropriately differenced trading effects predictors. The maximum interaction degree is set to 3 and basis function threshold = 2 X 10^{-8} with subsequent weighted calls to the TSMARS program to handle heteroscedasticity. On completion the sequence of transformations are applied in reverse giving the approximation based on the general form equation (3.4.3)

$$\hat{y}_t = \exp\left[(1-B)^{-d}(1-B^s)^{-D}\left\{f\left(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}, z_{t,MD}, z_{t,TD}, z_{t,EASTER}\right)\right\}\right] - c$$

where $z_{t,MD}$ etc. denotes the appropriately chosen differenced value of MD etc. Note that no independent categorical predictors or deterministic seasonal predictor are included in the model.

Note, the model definitions given above do not change when outlier adjustment is used. This is because the adjustment is incorporated into the TSMARS stepwise selection procedure.

6.4.2 Data Modelling Results

For both of the selected model types, TSMARS estimates are obtained using each of the three outlier adjustment methods; Least Squares Additive Outlier (LSAO), Bounded Influence (BIF) and Time Series Additive Outlier (TSAO). Thus, for each time series, a set of 6 models are computed with detailed results given in Table 6.4.2.1 (see Table Appendix).

For each series, as outlined in subsection 4.4.1, the number of observations (N), the series code and its title is given. Also, for each method, an indicator is given to signify whether a log or constant transformation is applied. The resulting model and associated statistics (specified in the Appendix) computed from the residual $y_t - \hat{y}_t$ are also displayed.

Table 6.4.2.2: Frequency of Significant Test Results

Model type	Method	Statistics								Sum of Ranks
		χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	
TSMARS										
	LSAO	16 (4)	11 (1)	6 (5)	15 (2)	12 (5)	9 (2)	8 (1)	3 (6)	(30)
	BIF	19 (5)	12 (3)	5 (4)	15 (2)	11 (4)	9 (2)	11 (4)	3 (6)	(32)
	TSAO	19 (5)	12 (3)	3 (2)	15 (2)	13 (6)	11 (5)	11 (4)	2 (4)	(27)
STSMARS										
	LSAO	13 (3)	12 (3)	3 (2)	16 (5)	19(3)	8 (1)	11 (4)	2 (4)	(25)
	BIF	12 (2)	13 (6)	5 (4)	15 (5)	7 (2)	11 (5)	12 (6)	0 (1)	(31)
	TSAO	10 (1)	12 (3)	4 (3)	16 (5)	6 (1)	9 (2)	10 (2)	1 (2)	(19)

Once again, the column called Cycles/Notes identifies whether a cycle (i.e. significant spike, see Brockwell & Davies 1991) was evident in the residual spectrum. Where a regular cycle is found and Tsay's F-test is also significant at a corresponding lag (or integer multiple thereof), then evidence for a

threshold is rejected. Experience in this research has shown the power of Tsay's F-test is small where there is evidence of a cycle.

Useful information is gleaned by summarising the statistical test results. Table 6.4.2.2 gives a summary of the number of times each test produced a significant value at the 1% level. The rank of each method is also given (best = 1) in braces and a final column shows the sum of the ranks.

The results in Table 6.4.2.2 show that STSMARS is better. Comparing the adjustment methods it is clear that the SATSMARS-TSAO method appears to be considerably better than the other methods in terms of rank. However, the Kruskal-Wallis One-way test of ranks showed no evidence of a difference in the MAPE values across methods; H-value = 0.23 < $\chi^2_{5,0.95}$.. Also, there is no difference between the methods in terms of their MAPE, as H-values ranging from 0.0 to 0.2, computed pair wise are not significant at the 5% level.

Of particular interest is a comparison of Table 6.4.2.2 with Table 4.4.2.1 of Chapter 4. Recall, Table 4.4.2.1 is corresponding set of figures obtained without the outlier adjustment procedure. The ranks obtained for both TSMARS and STSMARS with outlier adjustment (Table 6.4.2.2), show no clear improvement over their without adjustment counterparts (Table 4.4.2.1). Moreover, the summary observations of the last paragraph are similar to those of Section 4.4. Therefore, on the basis of statistical tests, there is no evidence to indicate outliers are confounding the conclusion from Chapter 4.

Table 6.4.2.3 summarises the models according to model type. Comparing the methods, it is clear that TSMARS estimates are dominated by the growth component, as in Chapter 4. In contrast, STSMARS found that roughly 75% of models are nonlinear. This is higher than corresponding figure, of 55% (11 out of 20), obtained in Chapter 4. Therefore, in contrast to the conclusion of the last paragraph, outliers have had an affect on the conclusions of Chapter 4.

Table 6.4.2.3: Frequency of Different Model Types Observed in the Test Results

	Method	Model Type							
		Mean only	Independent Predictors	Linear	Integrated I(1)	SETAR	Seasonal SETAR	Regime Dependent SETAR	Nonlinear
TSMARS	LASO	0	0	1	12	4	0	0	3
	BIF	0	0	2	12	4	0	0	2
	TSAO	1	1	2	12	3	0	0	1
STSMARS	LSAO	1	0	1	2	7	5	2	2
	BIF	0	0	1	2	6	2	2	7
	TSAO	2	0	1	2	3	2	2	8

The table above indicates which methods found some nonlinearity. However, of greater interest is the number of test statistics that were not significant where a nonlinear model was found. Table 6.4.2.4 gives this number for the (four) model adequacy test for STSMARS only – STSMARS is the only method with a reasonable number of nonlinear models. The presence of a '–' in the cell indicates that the model did not possess any nonlinearity. Looking at the figures, it is clear that the test statistics do not vary greatly

across the adjustment methods. However, where nonlinearity is found the BIF method has the cleanest residuals – that is, it gives more series where the residual tests are not significant (e.g. a results of 3 or 4). The figures are also compared with the corresponding figures from Chapter 4 (Table 4.4.2.3). Looking across the figures it s clear that more tests are not significant when these series were modelled without outlier adjustment. This reinforces the conclusion above that there is no evidence to imply that outliers are affecting the conclusion from Chapter 4.

Table 6.4.2.4: Frequency of Test Statistics that are not significant for STSMARS Nonlinear Models only.

Method	Test Series Number									
	1	2	3	4	5	6	7	8	9	10
LSAO	0	2	0	1	3	2	3	3	2	4
BIF	0	0	0	0	3	3	3	4	3	3
TSAO	0	2	0	0	3	3	3	2	3	4
STSMARS	-	1	-	1	-	4	4	4	3	3
Method	Test Series Number									
	11	12	13	14	15	16	17	18	19	20
LSAO	-	-	2	-	2	-	0	0	1	0
BIF	-	-	0	4	4	-	0	0	0	0
TSAO	-	-	2	-	2	-	0	0	0	1
STSMARS	3	-	1	-	-	-	1	1	2	1

6.4.3 ANOVA discussion

The ANOVA analysis provided by TSMARS breaks down the contribution of each basis function in the approximation to the overall variance. In Table 6.4.3.1 this breakdown is given for each method grouped by basis function type. For each model type the breakdown is computed somewhat differently, as the figures given are those arising solely from the call to the TSMARS program. This means that growth effects are excluded from the figures given for the STSMARS but are included in those for TSMARS.

The overall average figures given Table 6.4.3.1a, show the mean and linear basis functions account for roughly 90% of the explained variance for TSMARS. This accounts for the growth components of the time series. This result is similar to Chapter 4. This large growth component masks other characteristics such as nonlinearity and considerably diminishes the size of residual variance as a proportion of the overall variance.

The residual variance for all STSMARS based methods, accounts for over 50% of the overall variance. In this case, the variance break down is given purely for the TSMARS program call and so it does not always include the growth component. That stated the nonlinear component accounts for over $1/10^{th}$ of the overall variance. In contrast, the results in Chapter 4 showed only about $1/20^{th}$ of the variance

appearing as nonlinear. Therefore, outliers have had an affect on the extent of nonlinearity observed for this test-bed of empirical time series.

Table 6.4.3.1a: Extract of Overal Average Variance Breakdown in Table 6.4.3.1 (Table Appendix)

No	Series Code	Title	Model type	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
		Average	TSMARS	LSAO	47	38	8	3	4
				BIF	52	37	6	1	4
				TSAO	48	40	7	1	4
			STSMARS	LSAO	26	4	14	2	54
				BIF	28	5	14	2	50
				TSAO	23	5	12	2	58

6.4.4 Concluding Remarks

TSMARS has been adapted to handle outliers and the program run on the 'test-bed' of 20 economic time series. From the results obtained the following can be concluded:

- The analysis of the test statistics given in Table 6.4.2.2 demonstrates that the outlier adjustment methods did not result in any significant improvement in model adequacy or seasonal effects.
- The MAPE values across the methods were not significantly different to the values observed in Chapter 3.
- Table 6.4.2.3 showed that up to 25% more nonlinear basis functions were uncovered by STSMARS with outlier adjustment.
- Comparison of results on nonlinear models supported the view that outliers do not alter the conclusions of Chapter 4.
- The ANOVA analysis in Table 6.4.3.1a, showed that roughly $1/10^{th}$ of the variance of all models found using the STSMARS method was explained by nonlinear basis functions.

These conclusions appear contradictory. However, it must be kept in mind that the outlier adjustment mechanisms are designed to ensure the basis function selection procedure is consistent. No attempt is made to reduce the final RSS using independent predictors. Thus test statistics may still lack power as spikes will be evident in the residual. In certain cases, this also means that MAPE values may not show any improvement.

The outlier adjustment mechanisms do appear to show evidence of more nonlinearity. However, close scrutiny of Table 6.4.2.1 (see Table Appendix) shows that independent effects, namely length of month, trading day and Easter effects are much more evident. In about half of the 20 series an independent predictor was included in the model, while only four models in Chapter 4 possessed an independent effect. This is a cause for concern. Moreover, this implies that the 10% of the explained variance that was found to be nonlinear with outlier adjustment employed requires further scrutiny. In fact section 6.3 showed that the adjustment procedure is consistent only for additive outliers and independent innovation

error models. Therefore, since the 10% estimate of the explained variance being nonlinear is in doubt, the conclusion of Chapter 4 stands – that is, around 5% of the explained variance is nonlinear.

6.5 Conclusions

This chapter has addressed outlier treatment in TSMARS. There were two aims; namely, ensuring model selection in TSMARS is consistent and then using this to ensure the conclusions of Chapter 4 are well founded.

The consistency issue was addressed using the CMOT procedure. This procedure is both novel and general. It does not specify a treatment method and can be applied to any stepwise procedure. It ensures that the search for the next basis function is conducted with the effects of outliers ameliorated.

Three different adjustment procedures were also set out; namely, Least Squares (LSAO), Bounded Influence (BIF) and Time Series (TSAO). The LSAO method is highly efficient with virtually no computing overhead. It is ideal when residuals are independent. The BIF method is complex and can require substantial computing. It is suitable when the residuals are independent but may deviate from normality. The TSAO method is specific for AR, TAR and additive model time series. A theorem was proved that showed method modelled the error process correctly in these cases. Simulation studies compared the performance of the methods in situations where one or more additive outliers were introduced into the data. The results of the simulation studies showed the treatment procedures made TSMARS more consistent; that is, TSMARS was more likely to choose a correct model type in the presence of an outlier.

The outlier adjustment procedures were then run on the 'test-bed' of 20 economic time series. With consistent model selection in place, no evidence was found of any significant improvement in accuracy, model adequacy or seasonal effects over Chapter 4 values. In contrast, the ANOVA analysis showed that roughly $1/10^{th}$ of the variance of all models found was nonlinear. However, many models had trading effects. This is incompatible with earlier results and indicated there was a problem with model selection. One implication of this is that dependent errors may be more appropriate, to explain the 'test-bed' of 20 economic time series. The conclusion of Chapter 4, that the original estimate of around 5% of the explained variance is nonlinear, therefore stands. Of course, there is the possibility that dependent errors models may be better able to explain the processes underlying these empirical series. This issue is explored in the next Chapter.

Table Appendix

Table 6.3.1.1: AR(1) Model Simulation Results for $\rho = 0.5$

		1 Lagged Predictor			3 Lagged Predictors		
No. of outliers	Method	AR(1) Models found	$\hat{\rho}$	Standard Error	AR(1) Models found	$\hat{\rho}$	Standard Error
-	Original	99	0.506	0.081	95	0.506	0.080
1	Contaminated	98	0.431	0.083	98	0.432	0.081
	LSAO	100	0.431	0.089	98	0.428	0.086
	BIF	100	0.457	0.090	95	0.440	0.078
	TSAO	99	0.430	0.083	97	0.429	0.081
3	Contaminated	99	0.396	0.086	99	0.397	0.087
	LSAO	98	0.398	0.089	95	0.400	0.087
	BIF	100	0.410	0.087	92	0.419	0.096
	TSAO	99	0.396	0.086	94	0.401	0.085
5	Contaminated	91	0.350	0.070	93	0.341	0.096
	LSAO	96	0.341	0.081	97	0.342	0.098
	BIF	98	0.361	0.094	88	0.362	0.088
	TSAO	95	0.343	0.078	97	0.346	0.098

Table 6.3.2.1: SETAR(2,1,1) Model Simulation Results for $\rho_1 = 0.75$, $\rho_2 = 0.25$ and threshold (knot) = 0.0

No of Outliers	Method	Correct SETAR Models found	$\hat{\rho}_1$	Std. Err.	$\hat{\rho}_2$	Std. Err.	knot	Std. Err.
							No of lagged predictors = 1	
							Maximum Interaction Degree = 1	
1	Original	82	0.811	0.057	0.184	0.156	-0.046	0.075
	Contaminated	99	0.795	0.088	0.091	0.103	0.019	0.104
	LSAO	86	0.838	0.085	0.100	0.132	-0.055	0.142
	BIF	83	0.847	0.083	0.093	0.123	-0.041	0.081
	TSAO	85	0.856	0.089	0.070	0.153	-0.037	0.082
3	Contaminated	100	0.779	0.094	0.056	0.069	0.053	0.101
	LSAO	91	0.829	0.150	0.064	0.130	-0.146	0.322
	BIF	76	0.875	0.091	0.039	0.068	-0.051	0.082
	TSAO	74	0.867	0.099	0.004	0.083	-0.036	0.085
5	Original	82	0.813	0.058	0.181	0.157	-0.048	0.079
	Contaminated	100	0.770	0.102	0.057	0.061	0.047	0.091
	LSAO	87	0.819	0.173	0.050	0.092	-0.120	0.302
	BIF	72	0.891	0.118	0.027	0.053	-0.054	0.100
	TSAO	77	0.859	0.147	0.020	0.234	-0.048	0.121
							No of lagged predictors = 3	
							Maximum Interaction Degree = 3	
1	Contaminated	99	0.794	0.088	0.093	0.104	0.019	0.104
	LSAO	87	0.836	0.085	0.099	0.130	-0.052	0.143
	BIF	85	0.845	0.084	0.010	0.127	-0.042	0.079
	TSAO	86	0.856	0.089	0.141	0.400	-0.037	0.083
3	Contaminated	100	0.778	0.095	0.057	0.071	0.055	0.100
	LSAO	90	0.826	0.095	0.066	0.130	-0.145	0.324
	BIF	76	0.857	0.092	0.065	0.136	-0.060	0.118
	TSAO	79	0.867	0.096	0.058	0.138	-0.100	0.262
5	Contaminated	100	0.769	0.102	0.058	0.061	0.049	0.091
	LSAO	81	0.848	0.125	0.053	0.093	-0.115	0.298
	BIF	81	0.762	0.097	0.050	0.056	-0.048	0.099
	TSAO	77	0.864	0.116	0.010	0.124	-0.097	0.266

Table 6.4.2.1: Time Series Test-bed Results with Outlier Adjustment

No	N	Series Code	Title	Model	Method	Model
1	286	ASAM003	Cows Milk Protein Content (%)	TSMARS	LSAO	$y_t = 210{,}770 + 1.01(y_{t-1} - 180{,}102)_-$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta\Delta_{12}\left[\begin{array}{l} y_t = 0.2 - 0.4(y_{t-1} - 0.08)_+ + 0.25(y_{t-3} - 0.08)_+ + 0.25(y_{t-3} - 0.4)_+ - \\ 1.1(y_{t-12} - 0.2)_- \end{array}\right]$
					BIF	AS ABOVE
					TSAO	AS ABOVE
2	286	ASAM206	Calves Slaughtering 000 Heads	TSMARS	LSAO	$y_t = 0.12 + 0.57\,y_{t-1} + 11.2y_{t-1}(\text{Sin}(t - \pi/3)_+ - 14.5y_{t-1}(\text{Sin}(t - \pi/3)_+(y_{t-10} - 0.7)_-$
					BIF	$y_t = 0.14 + 0.59\,y_{t-1} + 12.0y_{t-1}(\text{Sin}(t - \pi/3)_+ - 19.8y_{t-1}(\text{Sin}(t - \pi/3)_+(y_{t-10} - 0.9)_- - 0.01MD$
					TSAO	$y_t = 0.28 + 0.57\,y_{t-1} + 25.1y_{t-1}(\text{Sin}(t - \pi/3)_+ - 34.7y_{t-1}(\text{Sin}(t - \pi/3)_+(y_{t-10} - 0.9)_-$
				STSMARS	LSAO	$y_t = 1.1 + 0.41\,y_{t-1} + 2.7y_{t-1}(y_{t-12} - 0.6)_- + 0.56\,y_{t-1}(y_{t-12} - 0.6)_+ - 2..3\,y_{t-1}(y_{t-12} - 0.6)_+(TD - 2.5)_- - 1..1\,y_{t-1}(y_{t-12} - 0.6)_+(TD - 2.5)_+$
					BIF	$y_t = 1.13 + 2.96\,y_{t-1}(y_{t-2} - 1.2)_- + 11.2y_{t-1}(\text{Sin}(t - \pi/3)_+ - 14.5y_{t-1}(\text{Sin}(t - \pi/3)_+(y_{t-10} - 0.7)_-$
					TSAO	$y_t = 0.95 + 0.58\,y_{t-1} + 0.91y_{t-1}(y_{t-2} - 1.2)_-$
3	286	ASAM305	Heifers Slaughtering 000 Tons	TSMARS	LSAO	$y_t = 8.9 - 0.52(y_{t-1} - 4.6)_- + 0.47(y_{t-3} - 4.6)_+ - 0.19y_{t-2} + 0.48y_{t-12} - 0.34y_{t-13} + 0.03TD(y_{t-4} - 2)_+ + 0.55(\text{Sin}(t - \pi/3)_+$
					BIF	$y_t = 9.1 + 0.45y_{t-12} + 0.37y_{t-13} - 0.99(y_{t-1} - 4.6)_- 0.5(y_{t-1} - 4.6)_-(y_{t-1} - 9.1)_- + 0.2MD$
					TSAO	$y_t = 11.2 - 1.23(y_{t-1} - 4.6)_-$
				STSMARS	LSAO	$\Delta\left[y_t = 0.05 - 0.01TD - 0.22(y_{t-1} - 0.69)_+ + 0.61(y_{t-12} - 0.97)_+\right]$
					BIF	$\Delta\left[\begin{array}{l} y_t = -0.07 - 0.24(y_{t-1} - 0.69)_+ - 0.69(y_{t-12} - 0.97)_+ \\ + 0.01MD(TD - 13.5)_- - 0.01MD(TD - 12.5)_- \end{array}\right]$
					TSAO	AS LSAO
4	250	FIAM023	Irish Currency in Circulation (€)	TSMARS	LSAO	$y_t = 1069 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta\left[y_t = 0.02 - 0.2(y_{t-3} - 0.06)_+ - 0.55(y_{t-12} - 0.07)_+ + 0.03(Easter - 1)_+\right]$
					BIF	$\Delta\left[\begin{array}{l} y_t = -0.002 - 2.2(y_{t-13} - 0.07)_- - 0.42(y_{t-13} - 0.07)_+ + 0.03(Easter - 1) \\ + 47(y_{t-13} - 0.07)_-(y_{t-1} - 0.06)_+(y_{t-3} - 0.06)_+ \end{array}\right]$
					TSAO	$\Delta\left[\begin{array}{l} y_t = 0.03 - 0.43(y_{t-13} - 0.07)_+ + 4.37(y_{t-12} - 0.13)_+(y_{t-13} - 0.07)_- \\ + 0.03(Easter - 1) \end{array}\right]$

No	Model	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
1	TSMARS	LSAO	0.01	0.01	2, 3, 5	0.01	0.01	1.0	1.0	0.3	10.9	0.0	3
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.01	0.01	2, 4, 12	0.01	1.0	0.01	0.01	0.87	9.9	0.0	3
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
2	TSMARS	LSAO	0.20	0.90	5	0.01	0.19	0.53	0.54	0.78	46.7	58.6	-
		BIF	0.01	0.01	5	0.01	0.22	0.01	0.01	0.13	45.3	58.6	-
		TSAO	0.01	0.99	2,4, 5-10	0.01	0.16	0.55	0.36	0.24	52.1	58.6	-
	STS MARS	LSAO	0.06	0.90	12	0.01	0.16	0.01	0.02	0.34	63.4	58.6	-
		BIF	0.01	0.01	2, 6	0.01	0.60	0.01	0.01	0.13	45.3	58.6	-
		TSAO	0.61	0.99	2, 6	0.01	0.07	0.01	1.0	0.49	58.4	58.6	-
3	TSMARS	LSAO	0.01	0.01	2	0.01	0.01	0.40	0.68	0.98	7.2	9.8	-
		BIF	0.01	0.01	1-3, 12	0.01	0.01	0.15	0.01	0.15	8.4	9.8	12
		TSAO	0.01	0.01	2	0.01	0.01	0.18	0.01	0.98	7.7	9.8	2
	STS MARS	LSAO	0.01	0.01	2	0.01	0.01	0.11	0.01	0.97	7.7	9.8	-
		BIF	0.01	0.01	2	0.01	0.01	0.33	0.05	0.98	7.8	9.8	-
		TSAO	0.01	0.01	1-3, 7	0.01	0.01	0.11	0.01	1.0	11.4	9.8	12
4	TSMARS	LSAO	0.01	0.12	1, 12	0.01	0.01	0.01	0.01	0.01	3.0	0.0	2
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.12	0.01	1, 6, 12	0.01	1.0	0.01	0.01	0.97	2.0	0.0	
		BIF	0.01	0.01	6, 12	0.01	0.96	0.01	0.01	0.34	1.9	0.0	12
		TSAO	0.01	0.01	3, 11	0.01	0.94	0.01	0.01	0.45	2.0	0.0	3

No	N	Series Code	Title	Model	Method	Model
5	324	FIAM102	Exchange Rate $ £STR	TSMARS	LSAO	$y_t = 0.75 + 0.98\, y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta_2 \left[\begin{array}{l} y_t = 0.02 + 0.35\ (y_{t-1} - 0.03)_+ + 0.001MD - \\ \quad 0.002(TD - 3)_+ \end{array} \right]$
					BIF	$\Delta_2 \left[\begin{array}{l} y_t = -0.01 + 0.38\ (y_{t-1} - 0.03)_+ - \\ \quad 0.15(y_{t-1} - 0.03)_+ (TD - 2)_- + \\ \quad 11(y2 - 0.05)_- (y_{t-3} - 0.04)_+ .(TD - 2)_+ \end{array} \right]$
					TSAO	$\Delta_2 \left[\begin{array}{l} y_t = -0.01 + 0.42\ (y_{t-1} - 0.03)_+ - 0.01(TD - 2)_+ - \\ \quad 0.02\,(y_{t-1} - 0.03)_+ (MD - 5)_- \end{array} \right]$
6	179	LRGM001	Live Register Total (No)	TSMARS	TSAO	$y_t = 133{,}416 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta\Delta_{12} \left[\begin{array}{l} y_t = -5{,}663 - 2.62\,(y_{t-2} - 3.460)_- + \\ 0.56\,(y_{t-2} - 3.460)_- \end{array} \right]$
					BIF	$\Delta\Delta_{12} \left[\begin{array}{l} y_t = -3{,}011 + 0.11(y_{t-1} - 3.460)_+ \\ + 0.38(y_{t-2} - 3.460)_- + 0.66(y_{t-2} - 3.460)_- - \\ 0.15(y_{t-12} - 6.421)_- \end{array} \right]$
					TSAO	$\Delta\Delta_{12} \left[\begin{array}{l} y_t = -4{,}387 - 2.41(y_{t-2} - 3.460)_- + 0.53(y_{t-2} - 3.460)_- - \\ 1.3(y_{t-2} - 6.055)_- + 1.34(y_{t-12} - 6.772)_- \end{array} \right]$
7	179	LRGM111	Live Register/ Tara St. Total (No)	TSMARS	LSAO	$y_t = 974 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta \left[y_t = -0.05 - 0.11(y_{t-1} - 0.35)_+ + 0.33(y_{t-12} - 0.35)_+ + 0.21(y_{t-13} - 0.40)_+ \right]$
					BIF	$\Delta \left[\begin{array}{l} y_t = -0.05 - 0.18(y_{t-1} - 0.35)_+ + 0.33(y_{t-12} - 0.35)_+ + 0.19(y_{t-13} - 0.42)_+ + \\ -0.16(y_{t-2} - 0.35)_+ (y_{t-3} - 0.35)_+ \end{array} \right]$
					TSAO	$\Delta \left[y_t = -0.03 + 0.11(y_{t-12} - 0.40)_+ \right]$
8	179	LRGM438	Live Register/ Thomasown Males (No)	TSMARS	LSAO	$y_t = 181 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta \left[\begin{array}{l} y_t = -0.03 + 0.04(y_{t-1} - 0.17)_+ - 14.5(y_{t-12} - 0.06)_+ + \\ 12.7(y_{t-12} - 0.13)_+ - 0.51(y_{t-12} - 0.11)_- \end{array} \right]$
					BIF	$\Delta \left[\begin{array}{l} y_t = -0.03 - 16.6(y_{t-3} - 0.17)_+ + 0.31(y_{t-12} - 0.06)_+ + \\ 162\,(y_{t-3} - 0.06)_- (y_{t-13} - 0.06)_+ \end{array} \right]$
					TSAO	$\Delta \left[\begin{array}{l} y_t = -0.04 - 8.4\,(y_{t-12} - 0.06)_- + 0.41(y_{t-12} - 0.06)_+ + 6.4(y_{t-12} - 0.12)_- \\ - 0.05\,(y_{t-1} - 0.12)_- (y_{t-12} - 0.06)_+ \end{array} \right]$

No	Model	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
5	TSMARS	LSAO	0.01	0.06	5	0.01	0.06	0.01	0.01	0.42	1.1	1.2	12
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.26	0.40		0.01	0.16	0.07	0.12	0.73	1.0	1.1	
		BIF	0.21	0.06		0.01	0.10	0.11	0.09	0.69	1.1	1.1	
		TSAO	0.34	0.20		0.01	0.09	0.24	0.19	0.63	1.1	1.1	
6	TSMARS	LSAO	0.01	0.01	1-3, 6	0.01	0.01	1.0	1.0	0.11	2.2	0.0	6
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.01	0.69		0.01	0.99	0.01	0.01	0.95	1.0	0.0	4
		BIF	0.7	0.45		0.01	1.0	0.01	0.01	0.52	1.0	0.0	4
		TSAO	0.11	0.24		0.01	1.0	0.01	0.01	0.74	1.0	0.0	4
7	TSMARS	LSAO	0.47	0.96	8 - 12	0.01	0.01	0.35	0.44	0.55	3.2	0.0	
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.55	0.86		0.01	0.18	0.89	0.29	0.35	2.9	0.0	4
		BIF	0.54	0.91		0.01	0.19	0.92	0.31	0.54	2.9	0.0	4
		TSAO	0.65	0.86		0.01	0.11	0.08	0.07	0.55	3.0	0.0	4
8	TSMARS	LSAO	0.01	0.01	1	0.59	0.01	0.22	0.34	0.04	3.2	0.3	6
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.42	0.01		0.01	0.01	0.77	0.76	1.0	3.2	0.3	
		BIF	0.10	0.38		0.54	0.02	0.49	0.46	1.0	3.3	0.3	
		TSAO	0.24	0.01	1	0.73	0.02	0.17	0.51	1.0	3.3	0.3	2

No	N	Series Code	Title	Model	Method	Model
9	179	LRGM515	Live Register/ Nenagh Males (No)	TSMARS	LSAO	$y_t = 314 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta[y_t = -0.22\,y_{t-1} - 0.14y_{t-3} + 0.15\,y_{t-12} + 1.77y_{t-13}(y_{t-1} - 0.05)_+]$
					BIF	$\Delta[y_t = -0.04 + 0.12\,y_{t-1} - 0.11y_{t-3} + 0.24\,y_{t-12} - 0.36y_{t-1}(y_{t-2} - 0.07)_+]$
					TSAO	$\Delta\begin{bmatrix} y_t = -0.05 + 0.18\,y_{t-12} + 0.15\,y_{t-13} + 67y_{t-12}(y_{t-2} - 0.06)_-(y_{t-3} - 0.13)_+ \\ + 41y_{t-12}(y_{t-2} - 0.06)_-(y_{t-3} - 0.26)_- \end{bmatrix}$
10	179	LRGM800	Live Register/ Newcastle West females (No)	TSMARS	LSAO	$y_t = 286 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STSMARS	LSAO	$\Delta[y_t = -11.7 - 0.27\,y_{t-3} + 0.43\,y_{t-12} - 0.01y_{t-3}(y_{t-2} - 114)_+]$
					BIF	$\Delta\begin{bmatrix} y_t = -23.7 + 0.39\,y_{t-12} - 0.41(y_{t-3} - 75)_+ + 0.14y_{t-12}(y_{t-1} - 131)_+ - \\ 0.05(y_{t-3} - 84)_+(y_{t-2} - 132)_+ \end{bmatrix}$
					TSAO	$\Delta[y_t = -11.7 - 0.27\,y_{t-3} + 0.43\,y_{t-12} - 0.01y_{t-3}(y_{t-2} - 123)_+]$
11	288	MIAM014	Volume Index NACE 37 (Base 1985= 100)	TSMARS	LSAO	$y_t = 35.7 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STS MARS	LSAO	$\Delta_{12}[y_t = -0.11 + 0.5y_{t-1} - 0.25\,y_{t-2} - 0.39y_{t-12} + 0.24y_{t-13} + 0.01TD]$
					BIF	$\Delta_{12}[y_t = -0.10 + 0.53y_{t-1} - 0.23\,y_{t-2} - 0.37y_{t-12} + 0.24y_{t-13}]$
					TSAO	$\Delta_{12}[y_t = -0.11 + 0.48y_{t-1} + 0.26\,y_{t-2} - 0.31y_{t-12} + 0.24y_{t-13} + 0.01TD]$
12	288	MIAM051	Volume Index Manufacturing Industries (Base 1985 =100)	TSMARS	LSAO	$y_t = 56.0 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STS MARS	LSAO	$y_t = 4.05 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
13	288	MIAM524	Volume Index NACE 429 Adjusted (Base 1985= 100)	TSMARS	LSAO	$\Delta[y_t = 85.7 + 0.27\,y_{t-8} + 0.67(y_{t-1} - 36.5)_+]$
					BIF	$\Delta[y_t = 91.7 + 0.88(y_{t-1} - 36.5)_+ - 0.13(y_{t-8} - 10.5)_+]$
					TSAO	mean only fitted
				STS MARS	LSAO	$\Delta\begin{bmatrix} y_t = 0.21 - 0.69(y_{t-1} - 0.03)_+ - 0.27(y_{t-2} - 0.03)_+ - 0.02TD + \\ 0.11MD(y_{t-1} - 0.03)_+ \end{bmatrix}$
					BIF	$\Delta[y_t = 0.03 + 1.44(y_{t-1} - 0.06)_- - 0.21(y_{t-1} - 0.06)_+]$
					TSAO	$\Delta\begin{bmatrix} y_t = 0.01 - 1.09MD(y_{t-1} - 0.13)_- - 0.03MD(y_{t-1} - 0.13)_- \\ + 0.16MD(y_{t-1} - 0.03)_-(TD - 14)_- + 2.75MD(y_{t-1} - 0.03)_-(TD - 1)_- - \\ 4.0MD(y_{t-1} - 0.03)_-(TD - 2)_- \end{bmatrix}$

No	Model	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
9	TSMARS	LSAO	0.01	0.97	5, 6	0.63	0.01	0.69	0.02	1.0	4.0	0.1	6
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.01	0.87	None	0.01	0.01	0.81	0.15	0.89	3.8	0.1	3
		BIF	0.02	0.84	None	0.01	0.05	0.96	0.13	0.96	3.9	0.1	3
		TSAO	0.02	0.89	None	0.01	0.18	0.02	0.15	0.76	3.9	0.1	6
10	TSMARS	LSAO	0.01	0.14	2, 3	0.66	0.01	0.01	0.01	1.0	4.4	0.2	
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.80	0.13	None	0.55	0.04	0.02	0.09	0.76	3.4	0.2	
		BIF	0.82	0.01	None	0.62	0.03	0.04	0.01	0.87	3.5	0.2	
		TSAO	0.80	0.13	None	0.55	0.04	0.02	0.09	0.87	3.4	0.2	
11	TSMARS	LSAO	0.01	0.01	1 - 3	0.01	0.01	0.01	0.01	0.01	12.6	0.5	4
		BIF			None	AS	ABOVE						
		TSAO			None	AS	ABOVE						
	STS MARS	LSAO	0.01	0.01	1, 7	0.01	0.98	0.02	0.01	0.96	6.6	0.5	2
		BIF	0.02	0.01	1, 7	0.01	0.98	0.01	0.01	0.90	6.6	0.5	
		TSAO	0.01	0.01	1, 7	0.01	0.99	0.02	0.01	0.93	6.7	0.5	
12	TSMARS	LSAO	0.01	0.01	6	0.01	0.65	0.92	0.97	0.16	3.0	0.5	6
		BIF			None	AS	ABOVE						
		TSAO			None	AS	ABOVE						
	STS MARS	LSAO	0.01	0.01	2, 3, 5	0.01	0.01	0.01	0.01	1.0	7.3	0.5	3
		BIF			None	AS	ABOVE						
		TSAO			None	AS	ABOVE						
13	TSMARS	LSAO	0.42	0.04	1	0.01	0.83	0.01	0.06	0.89	5.4	1.0	
		BIF	0.01	0.19	5, 8	0.01	0.95	0.93	0.31	0.96	6.2	1.0	
		TSAO	0.01	0.01	2 - 18	0.01	1.0	0.01	0.01	0.93	10.7	1.0	10
	STS MARS	LSAO	0.01	0.01	None	0.76	0.96	0.16	0.02	0.89	5.7	1.0	
		BIF	0.01	0.01	2	0.01	0.80	0.01	0.01	0.96	5.8	1.0	
		TSAO	0.02	0.02	2	0.01	0.90	0.04	0.01	0.93	5.5	1.0	

No	N	Series Code	Title	Model	Method	Model
14	288	MTAM351	Dublin Airport Rainfall (mm)	TSMARS	LSAO	$y_t = 66.1 + 1.95MD - 0.04y_{t-1} - 6.25\text{Sin(t)}$
					BIF	$y_t = 65.4 + 2.44MD - 0.08y_{t-3} - 5.12\text{Sin(t)}$
					TSAO	$y_t = 53.2 + 3.7MD$
				STS MARS	LSAO	$y_t = 4.0$ mean only fitted
					BIF	$y_t = 4.1 - 0.07(y_{t-1} - 0.96)_+$
					TSAO	$y_t = 4.0$ mean only fitted
15	288	MTAM553	Mullingar Rainy Days (No.)	TSMARS	LSAO	$y_t = 18.0 - 1.03(y_{t-12} - 2)_-$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STS MARS	LSAO	$y_t = 19.0 - 1.03(y_{t-12} - 2)_-$
					BIF	$y_t = 16.1 + 0.13Y_{T-2} + 0.59MD + 0.07MDy_{t-12}y_{t13} - 0.06MD(y_{t-2} - 15)_+(y_{t-13} - 20)_+$
					TSAO	AS LASO
16	380	RSAM501	Retail Sale Index: All Business Value Adjusted Base 1990 = 100	TSMARS	LSAO	$y_t = 9.7 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STS MARS	LSAO	$y_t = 10.7 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
17	466	TRAM009	Exempt Vehicles New (No.)	TSMARS	LSAO	$y_t = 27.5 + 0.37y_{t-1} + 0.15y_{t-4} + 0.3y_{t-5} + 0.21y_{t-12} + 0.001(y_{t-7} - 301)_- - 0.002(y_{t-7} - 301)_+$
					BIF	$y_t = 25.3 + 0.5y_{t-1} + 14.6MD$
					TSAO	$y_t = 28.5 + 0.5y_{t-1} + 0.36y_{t-8}$
				STS MARS	LSAO	$y_t = 3.4 + 0.41(y_{t-1} - 0.94)_+ + 0.21(y_{t-3} - 0.94)_- + 0.24(y_{t-12} - 0.94)_+$
					BIF	$y_t = 3.3 + 0.5(y_{t-1} - 0.94)_+ + 0.11(y_{t-3} - 0.94)_- + 0.23(y_{t-12} - 0.94)_+$
					TSAO	$y_t = 4.7$ mean only fitted

No	Model	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
14	TSMARS	LSAO	0.95	0.64	None	0.88	0.01	0.48	0.28	0.16	75.9	1.2	
		BIF	0.93	0.56	None	0.59	0.02	0.60	0.38	0.16	77.1	1.2	
		TSAO	0.75	0.66	3	0.66	0.01	0.52	0.47	0.16	76.7	1.2	4
	STS MARS	LSAO	0.89	0.59	None	0.56	0.01	0.01	0.02-	0.99	63.8	1.2	4
		BIF	0.70	0.62	3	0.79	0.01	0.01	0.02-	0.99	66.2	1.2	4
		TSAO				AS	ABOVE						
15	TSMARS	LSAO	0.01	0.78	12	0.68	0.01	0.38	0.62	0.01	29.1	5.1	6
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.01	0.78	12	0.68	0.01	0.38	0.62	0.01	29.1	5.1	6
		BIF	0.10	0.71	None	0.72	0.01	0.39	0.47	1.0	28.0	5.1	6
		TSAO				AS	LSAO						
16	TSMARS	LSAO	0.01	0.01	1, 5, 11	0.01	0.86	0.01	0.01	0.3	2.5	1.0	4
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.01	0.01	1, 2, 9 -11	0.01	0.01	0.98	0.01	0.4	1.8	1.0	4
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
17	TSMARS	LSAO	0.01	0.01	1, 5, 6	0.01	0.01	0.01	0.01	0.93	46.0	0.5	
		BIF	0.01	0.01	1, 2 - 5	0.01	0.01	0.01	0.01	0.55	52.8	0.5	4
		TSAO	0.01	0.01		0.01	0.01	0.01	0.01	0.95	51.4	0.5	
	STS MARS	LSAO	0.01	0.01	3, 11	0.01	0.01	0.01	0.01	0.54	41.4	0.5	4
		BIF	0.01	0.01	4, 11	0.01	0.01	0.01	0.01	0.43	39.1	0.5	4
		TSAO	0.01	0.01	2 - 16	0.01	0.28	0.01	0.01	0.98	67.4	0.5	4

No	N	Series Code	Title	Model	Method	Model
18	380	TSAM043	Imports SITC 59 Other Chemicals €000	TSMARS	LSAO	$y_t = 26{,}445 + 0.34y_{t-1} - 0.82(y_{t-12} - 31{,}602)_- + 5.06(y_{t-13} - 39{,}651)_+$
					BIF	$y_t = 28{,}571 - 0.10(y_{t-1} - 27{,}095)_- + 0.35(y_{t-1} - 27{,}095)_+ + 0.81(y_{t-12} - 29{,}190)_-$
					TSAO	$y_t = 44{,}669 - 1.05(y_{t-1} - 43{,}443)_-$
				STS MARS	LSAO	$\Delta\left[y_t = 0.96 - 0.56(y_{t-1} - 0.09)_+ - 0.26(y_{t-2} - 0.09)_+ + 0.47(MD - 4)_= + 0.43(MD - 5)_+\right]$
					BIF	$\Delta\left[y_t = 0.52 - 0.52(y_{t-1} - 0.09)_+ - 0.22(y_{t-2} - 0.09)_+ + 0.18(MD - 4)_=\right]$
					TSAO	$\Delta\left[y_t = 0.55 - 0.52(y_{t-1} - 0.09)_+ + 0.3(y_{t-2} - 0.09)_+ + 0.13(y_{t-3} - 0.09)_+ + 0.2(MD - 4)_+\right]$
19	380	TSAM055	Imports SITC 71 Power Machinery €000	TSMARS	LSAO	$y_t = 3{,}367 + 0.27y_{t-2} + 0.26y_{t-3} + 0.17y_{t-5} + 0.32(y_{t-1} - 640)_+ + 0.0001y_{t-2}(y_{t-6} - 23{,}306)_-$
					BIF	$y_t = 41{,}169 - 0.1y_{t-8} - 0.83(y_{t-1} - 22{,}795)_- + 0.27(y_{t-1} - 22{,}795)_+ - 25.2TD + 76.6TD\ MD$
					TSAO	$y_t = 4{,}883 + 0.42y_{t-1} + 0.52y_{t-12}$
				STS MARS	LSAO	$\Delta\left[y_t = 0.36 - 0.73(y_{t-1} - 1.13)_+ - 0.27(y_{t-2} - 0.57)_+ + 0.12(y_{t-3} - 0.57)_+ + 0.11(y_{t-12} - 0.57)_+ + 0.02(Easter - 0.5)_+\right]$
					BIF	$\Delta\left[y_t = 0.13 - 0.61(y_{t-1} - 1.13)_+\right]$
					TSAO	$\Delta\left[y_t = 0.44 + 0.37(y_{t-1} - 1.26)_- - 1.22(y_{t-1} - 1.26)_+ - 0.35(y_{t-2} - 0.57)_+ + 0.28(y_{t-3} - 0.57)_+ + 0.14(y_{t-3} - 0.57)_+(y_{t-13} - 0.55)_+ + 0.67(y_{t-1} - 1.26)_+(Easter - 0.5)_+\right]$
20	380	TSAM601	Exports Adjusted €000	TSMARS	LSAO	$y_t = 67{,}900 + y_{t-1}$
					BIF	AS ABOVE
					TSAO	AS ABOVE
				STS MARS	LSAO	$\Delta_2\left[y_t = 238{,}107 - 0.44(y_{t-3} - 1{,}809{,}900)_- + 326{,}222(Easter - 0.5)_- + 0.42(y_{t-3} - 1{,}809{,}000)_-(Easter - 3)_-\right]$
					BIF	$\Delta_2\left[y_t = 185{,}191 - -1.89(y_{t-3} - 877{,}000)_- + 0.44(y_{t-3} - 877{,}000)_+ + 2.35(y_{t-3} - 877{,}000)_-(Easter - 0.5)_-\right]$
					TSAO	$\Delta_2\left[y_t = 185{,}191 - 0.7(y_{t-1} - 877{,}000)_+ - 1.89(y_{t-3} - 877{,}000)_- + 0.44(y_{t-3} - 877{,}000)_+ + 3.25(y_{t-3} - 877{,}000)_-(Easter - 0.5)_-\right]$

No	Model	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
18	TSMARS	LSAO	0.01	0.01	12	0.01	0.12	0.30	0.02	0.90	19.9	0	
		BIF	0.01	0.01	1, 11	0.01	0.21	0.11	0.01	0.66	27.2	0	
		TSAO	0.01	0.01	1	0.01	0.01	0.01	0.01	0.44	23.9	0	2
	STS MARS	LSAO	0.01	0.01	2, 5, 10	0.01	0.01	0.01	0.01	0.90	19.5	0	4
		BIF	0.01	0.01	2, 5, 10	0.01	0.01	0.01	0.01	0.85	20.0	0	4
		TSAO	0.01	0.01	2, 5, 10	0.01	0.01	0.01	0.01	0.81	19.9	0	4
19	TSMARS	LSAO	0.01	0.01	None	0.01	0.20	0.01	0.01	1.0	19.7	0	4
		BIF	0.01	0.01	2- 5, 7, 8	0.01	0.66	0.01	0.01	0.73	22.3	0	4
		TSAO	0.01	0.01	None	0.01	0.29	0.01	0.01	1.0	22.3	0	2
	STS MARS	LSAO	0.01	0.01	None	0.01	0.28	1.0	0.01	0.49	19.9	0	
		BIF	0.01	0.01	2, 4, 6, 12	0.01	0.08	0.30	0.02	0.90	19.9	0	
		TSAO	0.01	0.01	6	0.01	0.12	0.30	0.02	0.90	19.9	0	
20	TSMARS	LSAO	0.01	0.01	2 – 5, 9 – 14	0.01	0.92	0.01	0.01	0.42	7.2	0	2
		BIF				AS	ABOVE						
		TSAO				AS	ABOVE						
	STS MARS	LSAO	0.01	0.01	2, 6, 12	0.01	0.93	0.01	0.01	0.95	10.6	0	4
		BIF	0.01	0.01	1, 2, 6	0.01	0.99	0.01	0.01	0.94	10.9	0	2
		TSAO	0.01	0.01	1, 2, 6	0.01	0.97	0.01	0.01	0.97	8.4	0	3

Table 6.4.3.1: ANOVA Analysis

No	Series Code	Title	Model type	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
1	ASAM003	Cows Milk Protein Content (%)	TSMARS	LSAO	67	0	31	0	2
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	13	0	10	29	46
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
2	ASAM206	Calves Slaughtering 000 Heads	TSMARS	LSAO	21	0	23	39	17
				BIF	51	0	25	0	16
				TSAO	61	17	15	0	7
			STSMARS	LSAO	47	0	38	0	7
				BIF	40	0	47	0	13
				TSAO	44	0	40	0	16
3	ASAM305	Heifers Slaughtering 000 Tons	TSMARS	LSAO	76	13	9	0	1
				BIF	71	18	7	0	1
				TSAO	87	0	11	0	2
			STSMARS	LSAO	24	0	25	0	51
				BIF	18	0	31	0	51
				TSAO	24	0	25	0	51
4	FIAM023	Irish Currency in Circulation (€)	TSMARS	LSAO	45	55	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	7	16	10	0	67
				BIF	8	19	11	0	59
				TSAO	19	15	4	0	62
5	FIAM102[2]	Exchange Rate $ £STR	TSMARS	LSAO	83	17	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	7	1	6	0	86
				BIF	8	7	1	0	84
				TSAO	8	8	1	0	83

No	Series Code	Title	Model type	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
6	LRGM001	Live Register Total (No)	TSMARS	LSAO	58	42	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	18	0	19	0	63
				BIF	19	0	19	0	62
				TSAO	15	0	14	0	71
7	LRGM111	Live Register/ Tara St. Total (No)	TSMARS	LSAO	35	65	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	7	0	6	0	87
				BIF	8	0	7	0	87
				TSAO	6	1	5	0	89
8	LRGM438	Live Register/ Thomasown Males (No)	TSMARS	LSAO	52	48	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	9	0	10	0	81
				BIF	6	0	9	0	85
				TSAO	8	0	7	0	85
9	LRGM515	Live Register/ Nenagh Males (No)	TSMARS	LSAO	45	55	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	0	6	5	0	89
				BIF	4	4	1	0	91
				TSAO	7	6	1	0	86
10	LRGM800	Live Register/ Newcastle West females (No)	TSMARS	LSAO	51	49	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	17	18	2	0	63
				BIF	19	18	4	0	59
				TSAO	17	18	2	0	63

No	Series Code	Title	Model type	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
11	MIAM014	Volume Index NACE 37 (Base 1985= 100)	TSMARS	LSAO	24	74	0	0	2
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	23	26	17	0	34
				BIF	22	25	18	0	34
				TSAO	23	26	17	0	34
12	MIAM051	Volume Index Manufacturing Industries (Base 1985 =100)	TSMARS	LSAO	37	62	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	83	17	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
13	MIAM524	Volume Index NACE 429 Adjusted (Base 1985= 100)	TSMARS	LSAO	84	0	8	8	0
				BIF		AS	ABOVE		
				TSAO	-	-	-	-	-
			STSMARS	LSAO	7	0	7	3	80
				BIF	7	0	8	0	85
				TSAO	6	0	16	0	78
14	MTAM351	Dublin Airport Rainfall (mm)	TSMARS	LSAO	69	6	0	3	22
				BIF	70	6	1	1	22
				TSAO	67	0	0	11	22
			STSMARS	LSAO	-	-	-	-	-
				BIF	95	0	3	0	2
				TSAO	-	-	-	-	-
15	MTAM553	Mullingar Rainy Days (No.)	TSMARS	LSAO	93	0	0	0	7
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	93	0	0	0	7
				BIF	79	3	8	4	6
				TSAO	93	0	0	0	7

No	Series Code	Title	Model type	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
16	RSAM501	Retail Sale Index: All Business Value Adjusted Base 1990 = 100	TSMARS	LSAO	13	87	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	22	0	22	0	56
				BIF	28	0	30	0	42
				TSAO	25	0	25	0	50
17	TRAM009	Exempt Vehicles New (No.)	TSMARS	LSAO	15	60	11	0	15
				BIF	19	62	11	0	15
				TSAO	17	67	0	0	16
			STSMARS	LSAO	71	0	28	0	1
				BIF	71	0	28	0	1
				TSAO	-	-	-	-	-
18	TSAM043	Imports SITC 59 Other Chemicals €000	TSMARS	LSAO	53	13	27	0	6
				BIF	66	0	29	0	5
				TSAO	59	0	33	0	8
			STSMARS	LSAO	16	0	16	0	68
				BIF	15	0	14	0	71
				TSAO	15	0	14	0	71
19	TSAM055	Imports SITC 71 Power Machinery €000	TSMARS	LSAO	9	27	59	0	5
				BIF	84	4	7	0	5
				TSAO	17	43	35	0	5
			STSMARS	LSAO	18	0	18	0	64
				BIF	21	0	21	0	48
				TSAO	18	0	18	0	64
20	TSAM601	Exports Adjusted €000	TSMARS	LSAO	3	97	0	0	0
				BIF		AS	ABOVE		
				TSAO		AS	ABOVE		
			STSMARS	LSAO	13	0	18	1	68
				BIF	3	0	13	0	84
				TSAO	3	0	13	0	84

No	Series Code	Title	Model type	Method	% Variance				
					Mean	Linear	Nonlinear	Independent	Residual
		Average	TSMARS	LSAO	47	38	8	3	4
				BIF	52	37	6	1	4
				TSAO	48	40	7	1	4
			STSMARS	LSAO	26	4	14	2	54
				BIF	28	5	14	2	50
				TSAO	23	5	12	2	58

7 Threshold Moving Average Estimation with TSMARS

7.1 Introduction

In this thesis, studies of empirical economic time series are emphasised. As these series may possess dependent errors, modelling with moving average (MA) components is often desirable. However, TSMARS is based on autoregression splines and cannot distinguish an MA component. So, where a time series possesses an MA component, TSMARS will always identify an unsuitable model.

The main purpose of this chapter is to resolve this shortcoming of TSMARS. A novel extension is set out which builds on existing TSMARS methodology. It enables TSMARS adaptively to discover Self-exciting Threshold Moving Average (SETMA) models, Adaptive Spline Threshold Moving Average (ASTMA) and Adaptive Spline Autoregressive Threshold Moving Average (ASTARMA) models; both the ASTMA model and its generalisation, the ASTARMA model are new. In fact these two models are instances of model (1.2.1), the general nonlinear dependent error model with additive disturbances. This extension to TSMARS is therefore a significant development of the program. Moreover, it is the first instance where MA components are estimated within a nonparametric procedure. It is also the first procedure that estimates model (1.2.1) in a systematic manner.

Section 7.2 gives the theory and associated methodology that enable TSMARS to estimate MA time series. In this extension to TSMARS attention is focused on the residual from the current TSMARS model. Essentially, an innovative application of conditional least squares (CLS) is used to fit a parsimonious threshold model to this residual. In particular, at each each knot in the lag variable y_{t-v}, a parsimonious SETMA(2, v, v) model

$$r_t = \begin{matrix} \varepsilon_{L,t} + \theta_L \varepsilon_{L,t-v} & \text{if } y_{t-v} \le r \\ \varepsilon_{R,t} + \theta_R \varepsilon_{R,t-v} & \text{if } y_{t-v} > r \end{matrix} \qquad (7.1.1)$$

with $\varepsilon_{L,t}, \varepsilon_{R,t} \sim N(0, \sigma^2_{L/R})$, is fit to the residual at each knot; that is, a model for the residual r_t of the form $r_t = \varepsilon_t + \theta \varepsilon_{t-v}$, is fit in each regime. This procedure identifies a left/right lagged innovation function $\varepsilon_{L,t-v}$. The parent basis function is multiplied by this innovation function and the resulting new basis function is added to the TSMARS model by orthogonalisation. The GCV of this model is then used to decide if there is an improvement in the lack-of-fit. By proceeding stepwise through the set of lagged predictors, higher order parsimonious ASTMA and ASTARMA basis function models can be successively built up. This procedure is efficient, as the basis function selection is restricted to a sequence of parsimonious MA models. Note, using the residual to identify the basis function makes the procedure a mixed stepwise (for AR) and stagewise (for MA) component estimation method. By stagewise we mean a method that regresses predictors on the current model residual. In contrast, stepwise regresses the current model on the response.

In section 7.3 the MA estimation routine is tested on simulated data. The purpose is to 'prove' the routine; that is, ensure that it gives accurate, precise and consistent results. This is the second focus of this chapter. The routine will be judged correct if at least three-fifths of models are 'correct' and this fraction increases with the sample size. If, in addition, parameter estimates are within two standard errors of their true values, then the routine will be deemed statistically correct. The three-fifths figure is slightly lower

than the two-thirds figure adopted in Chapter 2, reflecting that mixed AR and MA models are harder to estimate than straightforward AR models.

With the MA estimation in place and proven, section 7.4 returns to study the empirical series. This extended version of TSMARS is applied, to see if MA component estimation provides greater insight into the nature and extent of nonlinearity of these series. This is the third focus of this chapter.

7.2 *Estimation of Moving Average Series with TSMARS*

7.2.1 Conditional Least Squares Estimation of the MA(1) model

A simple 1st order moving average model for a time series y_t having n observations, a single parameter θ and Normal error ε_t is

$$y_t = \theta\,\varepsilon_{t-1} + \varepsilon_t \tag{7.2.1}$$

This can be estimated by either conditional least squares (CLS) or using the Kalman filter (see Harvey 1993). CLS estimation is accomplished by assuming $\varepsilon_1 = \frac{\partial \varepsilon_1}{\partial \theta} = 0$ and minimising the residual sum of squares (RSS)

$$S(\theta) = \sum_{i=1}^{n} \varepsilon_t^2 = \sum_{i=1}^{n} (y_t - \theta\,\varepsilon_{t-1})^2 \tag{7.2.2}$$

with respect to θ.

When θ is known, observe that, since $\varepsilon_1 = 0$, the innovation can be computed by recursion as $\varepsilon_2 = y_2$ and $\varepsilon_t = y_t - \theta\,\varepsilon_{t-1}$. Thus the innovations ε_t and the parameter θ are dependent and so the quantity $S(\theta)$ is nonlinear. We seek an iterative minimisation of this nonlinear function based on an initial guess $\theta^{(0)}$.

Clearly the minimum of (7.2.2) occurs when

$$\frac{\partial S}{\partial \theta} = S'(\theta) = \sum_{i=1}^{n} \frac{\partial \varepsilon_t}{\partial \theta}\,\varepsilon_t = 0 \tag{7.2.3}$$

Iteratively solving this nonlinear equation gives an estimate $\hat{\theta}$ of θ that minimises (7.2.2) and provides via recursion the estimated innovations $\hat{\varepsilon}_t = \hat{\varepsilon}_t(\hat{\theta})$.

Specifically, $S'(\theta)$ is expanded in a Taylor approximation about θ_0 as

$$S'(\theta) \approx S'\left(\theta^{(0)}\right) + \Delta\theta\, S''\left(\theta^{(0)}\right) \tag{7.2.4}$$

where $\Delta\theta = \vartheta - \theta^{(0)}$ is the change in θ and $S''(\theta) = \frac{\partial^2 S\left(\theta^{(0)}\right)}{\partial^2 \theta^{(0)}}$. This suggest the following iterative scheme for $\hat{\theta}$

$$\Delta\left(\hat{\theta}^{(k+1)}\right) = \hat{\theta}^{(k+1)} - \hat{\theta}^{(k)} = S'\left(\hat{\theta}^{(k)}\right) \Big/ S''\left(\hat{\theta}^{(k)}\right) \tag{7.2.5}$$

whence, at convergence $\hat{\theta}^{(k)} = \hat{\theta}$. The 1st derivative is evaluated from (7.2.3) by substituting $\hat{\varepsilon}_t^{(k)}$ for ε_t while the 2nd derivative term can be simply computed as

$$S''(\hat{\theta}^{(k)}) = 2\sum_{i=1}^{n}\left[\left(\hat{\varepsilon}_t'^{(k)}\right)^2 + \hat{\varepsilon}_t^{(k)}\,\hat{\varepsilon}_t''^{(k)}\right] \approx 2\sum_{i=1}^{n}\left(\hat{\varepsilon}_t'^{(k)}\right)^2$$

on keeping second order terms only. Thus (7.2.5) becomes

$$\Delta\left(\hat{\theta}^{(k+1)}\right) = \sum_{i=1}^{n}\hat{\varepsilon}_t^{(k)}\,\hat{\varepsilon}_t'^{(k)} \Big/ \sum_{i=1}^{n}\left(\hat{\varepsilon}_t'^{(k)}\right)^2 \tag{7.2.6}$$

Observe that (7.2.6) can be thought of as arising from the OLS solution of the regression, through the origin, of $\hat{\varepsilon}_t\left(\hat{\theta}^{(k)}\right)$ on $\hat{\varepsilon}_t'\left(\hat{\theta}^{(k)}\right)$. So this minimisation can be efficiently computed by:

- Using the current estimate $\hat{\theta}^{(k)}$ to generate $\hat{\varepsilon}_t^{(k)}$ and in parallel the derivative $\hat{\varepsilon}_t'^{(k)} = -\hat{\theta}^{(k)}\,\hat{\varepsilon}_{t-1}'^{(k)} - \hat{\varepsilon}_{t-1}^{(k)}$.
- Then by regressing $\hat{\varepsilon}_t^{(k)}$ on $\hat{\varepsilon}_t'^{(k)}$ to get a new estimated of $\hat{\theta}^{(k+1)}$.

Iterating this yields the Gauss-Newton approximation to the CLS estimate of θ. Crucially, it also gives an estimate of the estimate of lagged innovation $\hat{\varepsilon}_{t-1}$.

7.2.2 Estimation of the SETMA(2,1,1) model

The CLS estimation methodology can be extended to estimate the 2-regime Self Exciting Threshold Moving Average SETMA(2,1,1) model, having threshold lag equal to 1 and constant threshold value r

$$y_t = \varepsilon_t + \begin{cases}\theta_{1,1} + \theta_{1,2}\,\varepsilon_{t-1} & if\ y_{t-1} \le r \\ \theta_{2,1} + \theta_{2,2}\,\varepsilon_{t-1} & if\ y_{t-1} > r\end{cases} \tag{7.2.7}$$

This model is a special case of the 2-regime SETMA(2,q,q) model of order q studied by de Gooijer (1998). Note, a slightly different and more complex threshold moving average model also exists where the threshold lag variable is ε_{t-1}, namely

$$y_t = \varepsilon_t + \begin{cases}\theta_{1,1} + \theta_{1,2}\,\varepsilon_{t-1} & if\ \varepsilon_{t-1} \le r \\ \theta_{2,1} + \theta_{2,2}\,\varepsilon_{t-1} & if\ \varepsilon_{t-1} > r\end{cases} \tag{7.2.8}$$

This model is new. It is referred to as the Innovation Excited Threshold Moving Average IETMA(2,1,1) model and will not be pursued further here; its generalisation to order q models with threshold lags greater than 1 are obvious.

Taking the SETMA(2,1,1) model (7.2.7), without loss of generality set the constants $\theta_{1,1}$ and $\theta_{2,1}$ = 0; for notational convenience let $\theta_{1,2} = \theta_L$ and $\theta_{2,2} = \theta_R$ denote left and right sided regime parameters respectively. In the left regime we can write $\varepsilon_t = \varepsilon_{t,L}$ while in the right $\varepsilon_t = \varepsilon_{t,R}$ and so using indicator functions (7.2.7) becomes

$$y_t = \left(\varepsilon_{t,L} + \theta_L\,\varepsilon_{t-1,L}\right)I\left(y_{t-1} - r\right)_- + \left(\varepsilon_{t,R} + \theta_R\,\varepsilon_{t-1,R}\right)I\left(y_{t-1} - r\right)_+ \tag{7.2.8}$$

In this model version y_t can be considered to be made up of two distinct left and right regime predictors. Moreover, the left predictor is non zero only when the right predictor is zero and vice versa. This suggests that the RSS in this case can be split across these regimes as follows:

$$\sum_{i=1}^{n}\varepsilon_t^2 \;=\; \sum_{i=1}^{n_L}\varepsilon_{t,L}^2 + \sum_{i=1}^{n_R}\varepsilon_{t,R}^2$$

with the left and right RSS components orthogonal. Therefore the minimisation of the RSS can be accomplished by separately minimising the RSS in the left and right regimes. That is, the CLS estimates are obtained from

$$\text{Min}\left(\sum_{i=1}^{n}\varepsilon_t^2\right) = \begin{cases} \text{Min}\sum_{i=1}^{n_L}\varepsilon_{t,L}^2 = \sum_{i=1}^{n_L}\left(y_t-\theta_L\,\varepsilon_{t-1,L}\right)^2 & \text{if } y_{t-1}\le 0 \\ \text{Min}\sum_{i=1}^{n_R}\varepsilon_{t,R}^2 = \sum_{i=1}^{n_R}\left(y_t-\theta_R\,\varepsilon_{t-1,R}\right)^2 & \text{if } y_{t-1}>0 \end{cases}$$

Thus, the MA(1) estimation procedure, can be applied to each regime separately; that is, regressing $\varepsilon_{t,L}$ on $\frac{\partial\varepsilon_{t,L}}{\partial\theta_L}$ to estimate θ_L and regressing $\varepsilon_{t,R}$ on $\frac{\partial\varepsilon_{t,R}}{\partial\theta_R}$ to estimate θ_R, .

In practice this 'Gauss-Jacobi' type procedure gives $\hat{\theta}_L$ and $\hat{\theta}_R$ that yield a 'satisfactory' estimate of minimum RSS when both the threshold lag and threshold value are specified. However, these parameter estimates are inefficient. Furthermore, the estimated innovation ε_t and more importantly its lagged value ε_{t-1} are also inefficient. The reason is that, at a regime change point (e.g. when the threshold lag value y_{t-1} changes sign) say moving from left to right, $\varepsilon_{t,R}$ is dependent on $\varepsilon_{t-1,R}$, which at the change point is in fact $\varepsilon_{t-1,L}$. Thus the left and right RSS components are dependent at the change point. So, to get accurate values of θ_L and θ_R the following 'Gauss-Seidel' type estimation procedure is necessary.

Gauss-Seidel Estimation:

1. **Generate** the vectors $\hat{\varepsilon}_t \equiv \hat{\varepsilon}_t(\hat{\theta})$ and $\hat{\varepsilon}'_t \equiv \hat{\varepsilon}'_t(\hat{\theta})$ with $\hat{\theta} = \hat{\theta}_L, \hat{\theta}_R$ depending on the regime for t = 1..n.
2. **Extract** from $\hat{\varepsilon}_t$ and $\hat{\varepsilon}'_t$ left column vectors $\hat{\varepsilon}_{t,L} = \hat{\varepsilon}_t$ and $\hat{\varepsilon}'_{t,L} = \hat{\varepsilon}'_t$ if $y_{t-1} \le r$ and 0 otherwise
3. **Regress** $\hat{\varepsilon}_{t,L}$ on $\hat{\varepsilon}'_{t,L}$ to get a new estimate $\hat{\theta}^*_L$
4. **Regenerate** the vectors $\hat{\varepsilon}_t$ and $\hat{\varepsilon}'_t$ with $\hat{\theta} = \hat{\theta}^*_L, \hat{\theta}_R$ depending on the regime for t = 1..n.
5. **Extract** from $\hat{\varepsilon}_t$ and $\hat{\varepsilon}'_t$ right column vectors $\hat{\varepsilon}_{t,R} = \hat{\varepsilon}_t$ and $\hat{\varepsilon}'_{t,R} = \hat{\varepsilon}'_t$ if $y_{t-1} > r$ and 0 otherwise
6. **Regress** $\hat{\varepsilon}_{t,R}$ on $\hat{\varepsilon}'_{t,R}$ to get a new estimate $\hat{\theta}^*_R$.

Repeat steps 1 – 6 until there is no further improvement in the parameter estimates.

Using these estimates of θ_L and θ_R the innovation ε_t and left $\varepsilon_{t-1,L}$ and right $\varepsilon_{t-1,R}$ lag innovation vectors can be accurately computed. To fit this into the MARS framework, the SETMA(2,1,1) parameters $\theta_{1,2}$ and $\theta_{2,2}$ are then collectively re-estimated in the standard OLS type regression model

$$y_t = \theta_{1,2}\,\varepsilon_{t-1,L} + \theta_{2,2}\,\varepsilon_{t-1,R} + \eta_t \qquad (7.2.9)$$

with error $\eta_t \sim N(0,\sigma^2)$. In this manner, left and right lagged innovations act like moving average spline basis function predictors for y_t. This is a similar to ordinary left and right splines based y_{t-1} in the SETAR(2,1,1) case.

So, depending on the purpose, there are two slightly different estimation procedures. First, if an approximation to the minimum of the RSS is required, then it is sufficient to treat each regime independently and use the Gauss-Jacobi iteration. Second, where it is necessary to estimate parameter values and accurate left and right lagged innovations, then the dependent approximation to the RSS based on Gauss-Seidel iteration is necessary. The combination of these two estimation methods forms the core procedure proposed here for identifying and modelling MA components within TSMARS.

7.2.3 Modifications to TSMARS to Identify Moving Average Elements

The forward stepwise search in TSMARS (Friedman 1991) computes the improvement in the lack-of-fit criterion at each step based on a parent basis function, lagged predictor variable and knot (i.e. variable values) combination.

To find MA components each step is replaced by:

- **Compute** the residual from the current TSMARS model $\hat{f}_t(M,\bullet)$ (i.e. equation (2.3.1)) based on M basis functions, that is $s_t = y_t - \hat{f}_t(M,\bullet)$
- **Apply** the Gauss-Jacobi procedure to the residual s_t giving the proposed new (left or right) basis function b_{M+1}
- **Take** this basis function temporarily taken the current set using Gram-Schmidt orthogonalisation (i.e. as if it were an ordinary AR type basis function) and compute the reduction in the RSS.
- **Compute** the GCV and compare it to the 'best' GCV found in the current round of the forward stepwise search (including the GCV based on AR-type basis functions).
- **Treat** the knot k, variable –v and current basis function M as candidates to be entered into the current TSMARS model if the GCV is lowest found so far,
  Note: –v is used to distinguish that an MA basis function is added to the basis function set (in the standard AR case the variable is denoted by +v).

When the Gauss-Jacobi procedure is adopted, the basis function is found by iteartion using the current residual s_t, the lag variable v and the current knot. However, for this v^{th} variable, the iteration scheme is used to fit a parsimonious (threshold) MA(v) model of the form

$$s_t = \varepsilon_t + \theta\,\varepsilon_{t-v} \qquad \text{if } y_{t-v} \le r, \text{ or } > r$$

to the residual. The parent basis function is then multiplied by resulting lagged threshold innovation $\hat{\varepsilon}_{t-v}$ giving the new basis function b_{M+1}. This is added to the current TSMARS model. Thus, an MA(q) model is built up in an additive manner in a similar fashion to the AR(p) model in standard TSMARS. Moreover, products of lagged variables can be combined both additively and multiplicitively with lagged innovations. These can also be combined multiplicitively among themselves to generate ASTMA and ASTARMA models.

When the current round of the forward search procedure is complete the candidate basis function must then be added to the current basis function set.

This is accomplished as follows:

- **Compute** the residual from the current TSMARS model $f_t(M,\bullet)$, that is $s_t = y_t - f_t(M,\bullet)$
- **Apply** the Gauss-Seidel procedure to the residual s_t, giving the proposed new (left or right) basis function b_{M+1}
- **Add** this basis function to the current set using Gram-Schmidt orthogonalisation (as in (7.2.9))

With the forward step complete, TSMARS applies a one-at-a-time basis function deletion strategy. This is unchanged where basis functions involve MA components. The resulting model generalises the ASTAR model of Lewis & Stevens (1991) to Adaptive Spline Threshold Moving Average (ASTMA)

$$\hat{y}_t = \hat{f}_t = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{l=1}^{L_m} I\left\{\left[s_{l,m}\left(y_{v(l,m)} - y^*_{\xi(l,m)}\right)\right]_+\right\}\varepsilon_{v(l,m)} \tag{7.2.10}$$

where the indicator function $I(\bullet)$ is used to select a 1 for each positive value of that spline function. We also have the mixed AR and MA form, namely the Adaptive Spline Threshold Autoregressive Moving Average (ASTARMA) model

$$\hat{y}_t = \hat{f}_t = \beta_0 + \sum_{m=1}^{M} \beta_m \prod_{k=1}^{K_m}\left[s_{k,m}\left(y_{v(k,m)} - y^*_{\xi(k,m)}\right)\right]_+ \prod_{l=1}^{L_m} I\left\{\left[s_{l,m}\left(y_{v(l,m)} - y^*_{\xi(l,m)}\right)\right]_+\right\}\varepsilon_{v(l,m)} \tag{7.2.11}$$

with basis function comprising K_v product AR splines and L_v product MA type spline functions.

7.2.4 Concluding Remarks

In this section the methodological elements of MA component estimation in TSMARS have been set out. The forward stepwise search is augmented with a computationally efficient CLS estimate using a Gauss-Newton procedure. Knot identification is accomplished with a Gauss-Jacobi routine while parameter estimation uses the more accurate Gauss-Seidel procedure, which is computationally twice as expensive. These procedures were incorporated into a novel extension of TSMARS based on parsimonious SETMA models. This extension enables TSMARS to estimate SETMA, ASTMA and ASTARMA models. Furthermore, both ASTMA and ASTARMA are novel model forms.

7.3 Simulation Studies based on Moving Average Models

This section is devoted to 'proving' the MA estimation routine. This will ensure that it gives accurate, precise and consistent results. Recall, the routine will be judged correct if at least three-fifths of models are 'correct' and this fraction increases with the sample size. If, in addition, parameter estimates are within two standard errors of their true values, then, as in Chapter 2, the routine will be deemed statistically correct. Three different types of model are simulated with varying parameter values. The detailed results of these simulations are reported in a Table Appendix at the end of this chapter.

7.3.1 Simulation of an MA(1) model

The first simulation study is based on the linear 1st order MA model driven by normally distributed noise $\varepsilon_t = N(0,\sigma^2)$

$$y_t = \theta\,\varepsilon_{t-1} + \varepsilon_t \qquad t = 1,2,\ldots,n \tag{7.3.1}$$

Simulations are conducted for a range of values of the parameter θ and $\sigma^2 = 1$. In each simulation 100 data sets are generated. TSMARS is called with response y_t, one lagged predictor y_{t-1} and maximum

interaction degree set to 1. Simulation experiments are conducted for two sample sizes of 100 and 200 respectively. These experiments are repeated allowing 3 lagged predictors $y_{t-1}, y_{t-2}, y_{t-3}$ and maximum interaction degree of 3. All simulations are conducted with the smoothing parameter set to 3. The basis function threshold parameter is set at $1.5\text{x}10^{-1}$ and convergence in the iterative estimation schemes is set at $1.0\text{x}10^{-2}$. The resulting estimates are displayed in the Table Appendix (Table 7.5.1.1); an extract from Table 7.5.1.1 showing the poorest simulation results is given here for reference.

Extract from Table 7.5.1.1: MA(1) Model Simulation Results

No of lagged predictors	Maximum Interaction Degree	Parameter θ	n	Parameter Estimate	$Std.Err.(\hat{\theta})$		Number of MA(1) Models found	Total Simulation Time (sec)
					Actual	True		
3	3	-0.8	100	-0.69	0.113	0.060	62	438
			200	-0.70	0.085	0.042	69	3,068

In Table 7.5.1.1 the number of times an MA(1) model was correctly identified, from the 100 simulation data sets is given. Also given is the average value of the estimated parameter $\hat{\rho}$ and its "Actual" standard error, computed from the correctly identified models. For comparison the true asymptotic standard error $\sqrt{(1-\theta^2)/n}$ for an MA(1) model with n observations is shown (see Harvey 1993 Chapter 3).

The number of correct models found is about 75 on average, though it is clear that this falls to minimum of 62 (see Extract table) in the worst case. The number of correct models found increases when the sample size is 200. The MA extension is therefore correct. Moreover, the parameter estimate is within two standard errors of the true value in all cases except when $\theta = \pm 0.8$ and $n = 200$; though even here it is on the two standard error boundary. Therefore this MA extension of TSMARS is statistically correct for this models.

7.3.2 Simulation of an ARMA(1,1) model

The second simulation study is based on the ARMA(1,1) model driven by normally distributed noise $\varepsilon_t = N(0, \sigma^2)$

$$y_t = \phi\, y_{t-1} + \theta\, \varepsilon_{t-1} + \varepsilon_t \qquad t = 1, 2, \ldots, n \tag{7.3.2}$$

Once again 100 simulations runs are conducted with a combination of AR and MA parameter values and $\sigma^2 = 0.5$. TSMARS is called with response y_t, one lagged predictor y_{t-1}, smoothing parameter set to 3 and maximum interaction degree set to 1. Simulation runs were repeated for two sample sizes of n = 100 and 200. The basis function threshold parameter is set at $1.5\text{x}10^{-2}$ and convergence in the iterative estimation schemes is set at $1.0\text{x}10^{-2}$. The resulting parameter estimates and their standard errors are displayed in Table 7.5.1.2 (see Table Appendix). An extract from Table 7.5.1.2 showing the poorest simulation results is given here for reference.

In Table 7.5.1.2 the number of times an ARMA(1,1) model was correctly identified, the average value of the estimated AR parameter ϕ and the MA parameter θ, as well as the standard errors is shown. The true 'Actual' asymptotic standard error is also given for both parameters (see Chapter 3, Harvey 1993).
The extract shows that only 35 correct models are identified at the smaller sample size. This is a combination of a sample size effect and a small MA parameter effect. In all other cases 57 or more correct models are found (see Table 7.5.1.2). So, when there is a reasonable sample size the extension performs correctly. Furthermore, the parameter estimates are all within two standard errors of the true values so the extension is also statistically correct for this model.

Extract from Table 7.5.1.2: ARMA(1,1) Model Simulation Results

					Standard Error			
No of lagged predictors	Maximum Interaction Degree	n	Parameter	Parameter Estimate	Actual Estimate	True	Number of ARMA(1,1) Models found	Total Simulation Time (sec)
1	1	100	ϕ = -0.5	-0.49	0.101	0.102	35	220
			θ = -0.25	-0.29	0.129	0.114	-	-
		200	ϕ =- 0.5	-0.60	0.079	0.072	76	1,315
			θ = -0.25	-0.13	0.163	0.081	-	

7.3.3 Simulation of an SETMA(2,1,1) model

The third and final simulation study is based on the SETMA(2,1,1) model

$$y_t = \varepsilon_t + \begin{cases} \theta_1 \varepsilon_{t-1} & if\ y_{t-1} \le r \\ \theta_2 \varepsilon_{t-1} & if\ y_{t-1} > r \end{cases} \qquad t = 1, 2, \ldots, n \tag{7.3.3}$$

driven by normally distributed noise $\varepsilon_t = N(0, \sigma^2)$

Simulations are conducted on 100 data sets generated from the SETMA(2,1,1) model with a combination parameter values θ_1 and θ_2. The threshold value (i.e. knot) r = 0 and $\sigma^2 = 0.25$. TSMARS is called with response y_t, one lagged predictor y_{t-1}, smoothing parameter set to 3 and maximum interaction degree set to 1. Simulation runs were repeated for two sample sizes of n = 300 and 500. The basis function threshold parameter and convergence in the iterative estimation schemes is set at 1.5×10^{-3}. The resulting parameter estimates and their standard errors are displayed in Table 7.5.1.3 (see Table AppendixIn Table 7.5.1.3 the number of times an SETMA(2,1,1) model was correctly identified, the average value of the estimated parameters, as well as the standard errors is shown.). An extract from Table 7.5.1.3 showing the poorest simulation results is given here for reference.
The extract showing the poorest results has fewer than 20 correct models identified. However, the parameters in each regime are equal in size but opposite in sign. In this situation, the regimes tend to cancel out and the process is hard to distinguish from WN. In all other cases reported in Table 7.5.1.3 the number of correct models is 46 or more and this figure increases with sample size. The extension is

therefore correct. In addition, parameter estimates are within two standard errors in every case, other than the poorest given in the extract. Looking at the threshold, it is also accurately estimated as it falls

Extract from Table 7.5.1.3: SETMA(2,1,1) Model Simulation Results

No of lagged predictors	Maximum Interaction Degree	n	Parameter	Parameter Estimate	Parameter Estimate Standard Error	Number of TMA(2,1,1) Models found	Total Simulation Time (secs)
1	1	300	$\theta_1 = 0.5$	0.41	0.139	18	937
			$\theta_1 = -0.5$	-0.29	0.065	-	-
			r = 0	-0.09	0.087		
		500	$\theta_1 = 0.5$	0.41	0.090	15	2,484
			$\theta_2 = -0.5$	-0.32	0.067	-	-
			r = 0	-0.11	0.100	-	-

well within its two standard error bound of zero. In particular, this shows that the Gauss-Jacobi (i.e. regime independent CLS estimation) is appropriate for knot identification in the forward stepwise search The extension can therefore be judged as statistically correct for this model. However, care should be taken to ensure that MA components do not interact to give WN like sequences; this topic is not explored futher in this thesis.

7.3.4 Concluding Remarks

In this section, the goal was to show that the MA extension to TSMARS returned correct models and gave accurate, precise and consistent parameter estimates based on data simulated from a model. The models used were; an MA(1) model, an ARMA(1,1) model and a SETMA(2,1,1) model. In each case the extension was 'proven' to provide an acceptable number of correct models with accurate parameter values. The routine is statistically correct and therefore it can be reliably used to identify nonlinear MA components in time series in general.

7.4 Modelling Seasonal Economic Data using Moving Average TSMARS

The study conducted in Chapter 6 found that outliers did not alter the degree of nonlinearity found in test-bed empirical time series. In this section these series are re-modelled using TSMARS with MA component estimation. The purpose of this is to see if further insight can be gained into the nature and extent of nonlinearity in these series.

7.4.1 Moving Average TSMARS Models

The test bed of 20 economic time series are each re-modelled as univariate data with independent predictors to account for length of month (MD), trading week length (TD) and Easter. In this study, only two of the four alternative model variations given in Chapter 4 are applied; namely TSMARS and STSMARS. Both of these are applied with MA estimation and in addition, results are obtained 'with' and 'without' outlier adjustment. Their breif descriptions are:

- **TSMARS MA**: Where no transformations are made to the time series values y_t the TSMARS with MA approximation is

$$\hat{y}_t = f\left(y_{t-1},\dots y_{t-(s+1)},\varepsilon_{t-1},\dots\varepsilon_{t-(s+1)},k_1,\dots,k_s,p_t,\text{MD}_t,\text{TD}_t,\text{Easter}_t\right) \tag{7.4.1}$$

where $f(\bullet)$ denotes the TSMARS-MA model. The maximum interaction degree is once again set to 3 and basis function threshold = 2 X 10^{-8}.

- **TSMARS AO-MA**: This is identical to the above except that Least Squares Additive Outlier (LSAO) adjustment is incorporated into the estimation.

- **STSMARS MA**: Here, as before the transformed series $z_t = (1-B)^d\left(1-B^s\right)^D\{\log(y_t+c)\}$ is modelled. The lagged predictors $z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}$ are input into the TSMARS program along with appropriately differenced trading effects predictors $z_{t,MD}$ etc.. Also included are the estimated innovations giving the approximation based on the general form

$$\hat{y}_t = \exp\left[\left(1-B\right)^{-d}\left(1-B^s\right)^{-D}\left\{f\left(z_{t-1},z_{t-2},z_{t-3},z_{t-s},z_{t-(s+1)},\varepsilon_{t-1},\dots\varepsilon_{t-(s+1)},z_{\text{t,MD}},z_{\text{t,TD}},z_{\text{t,EASTER}}\right)\right\}\right]-c \tag{7.4.2}$$

Note, no independent categorical predictors or deterministic seasonal predictor are included in the model. The maximum interaction degree is set to 3 and basis function threshold = 2 X 10^{-8}.

STSMARS AO-MA: Here, once again this is identical to the STSMARS MA approximation above except that Least Squares Additive Outlier (LASO) adjustment is incorporated into the estimation.

Models (7.4.1) and (7.4. 2) are based on autoregressive and/or innovation spline functions. Therefore, TSMARS with iterative MA estimation generates ASTMA and ASTARMA models. The striking feature of (7.4.1) and (7.4.2) is that both, with additive innovation ε_t, are instances of the general nonlinear dependent error model (1.2.1). Moreover, this is the first systematic procedure to approximate model (1.2.1). Indeed, it appears that this is the first instance of nonparametric procedure that estimates MA components.

7.4.2 Moving Average TSMARS Results

The results for both of the selected model types obtained using MA estimation, both with or without LSAO outlier adjustment, are displayed in the Table Appendix (Table 7.5.1.4). As before, for each series the number of observations (N), the series code and its title is given. The resulting model and associated statistics (specified in the Appendix) computed from the residual are also given. Note, as in Chapters 4 and 6, that evidence of a threshold is accepted only when there is no evidence of a cycle in the residual spectrum.

As in previous chapters, the test statistics obtained are summarised and displayed in Table 7.4.2.1. Based on the 'Sum of Ranks', it is clear that STSMARS-MA is the best. However, this is somewhat misleading as the actual counts in the body of the table, show there is little difference. Furthermore, there no evidence of a difference in the MAPE values across methods, as a Kruskal-Wallis One-way test of

ranks was not significant (H-value = 0.04 < $\chi^2_{5,0.95}$). Therefore, no statistical difference is apparent among the model types employed and as observed in Chapter 6, outliers do not appear important.

It is of greater interest to compare the results in Table 7.4.2.1. with those obtained in section 4.5. Most results are quite similar, except for Tsay's F-test for evidence of a lag threshold. In Table 7.4.2.1 significant results for this test are virtually eliminated. This occurs because the spectrum in virtually all cases shows a significant spike at the threshold lag. In this case evidence for a threshold is rejected while evidence for a cycle is favoured. This choice is reinforced by the fact that periodic autoregression is also evident; furthermore, these effects will appear as cycles in the residual spectrum. On this basis, it is inferred that the residual does not possess any threshold autoregression effects. A second inference is, model building should include periodic predictors in model types adopted.

Table 7.4.2.1: Frequency of Significant Test Results

Model type	Method	Statistics								Sum of Ranks
		χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	
TSMARS										
	MA	16 (3)	13 (3)	1(2)	15 (2)	13(3)	11 (2)	11 (1)	1(2)	(18)
	AO-MA	16 (3)	13 (3)	1(2)	15 (2)	13 (3)	11 (2)	11 (1)	1(2)	(18)
STSMARS										
	MA	13 (1)	12 (1)	0 (1)	14 (1)	7(1)	10 (1)	12 (2)	0(1)	(9)
	AO-MA	14 (2)	13 (3)	0 (1)	14 (1)	8 (2)	11(2)	15 (3)	0 (1)	(15)

Tables 7.4.2.2(a) and (b) summarise the models observed in Table 7.5.1.4 according to model type. In contrast to previous analysis, Tables 7.2.2.2(a) and (b) give a 2-way analysis. That is, the models obtained for each series in Chapter 4 (see Table 4.6.1.1) are cross-classified according to the model obtained for the same series given in Table 7.5.1.4. This 'mobility' analysis is conducted for the TSMARS vs. TSMARS-MA in Table 7.4.2.3(a) and STSMARS vs. STSMARS-MA in Table 7.4.2.3(b).

The figures in Table 7.4.2.2(a) are dominated by integrated of order 1 models. Of the 12 observed in Chapter 4, 11 are unchanged integrated order 1 models. The diagonal elements account for 14 of the 20 models and so there is little mobility. Computing the K-coefficient of agreement (see Bishop Fienberg & Holland 1975) gives a value of K = 0.5 which is not significant. The K statistics compares the actual agreement in a table with 'chance agreement' that occurs when row and column variables are orthogonal. There is some small movement to off diagonal elements. In particular, two MA type models were found that in Chapter 4, had been independent predictor models. There is though little movement from more complex to less complex models. This is deduced based on the fact that two models remain above the main diagonal while two remain below.

Table 7.4.2.2(a): Mobility analysis of TSMARS models found

	Model Type	TSMARS MA						
		Model Type						
		Mean + Independent Predictors	Linear (all types)	Integrated I(1)	SETAR (all types)	Regime Dependent SETAR	Nonlinear	MA (all types)
Original TSMARS (Ch 4)	Mean + Independent Predictors	0	0	0	0	0	0	2
	Linear (all types)	0	1	0	0	0	1	0
	Integrated I(1)	0	0	11	1	0	0	0
	SETAR (all types)	0	0	0	2	0	0	0
	Regime Dependent SETAR	0	0	0	1	0	0	0
	Nonlinear	0	1	0	0	0	0	0

The results for STSMARS are given in Table 7.4.2.2(b). Here 10 of the 20 models stay on the main diagonal indicating a good level of agreement; the K value = 0.4 bears this out. Mobility to off diagonal elements is once again of more interest. Specifically, four of the models found have been reclassified as MA type models. Looking more closely at Table 7.4.1.1, it is observed that all these MA models are of a

Table 7.4.2.2(b): Mobility analysis of Different Model Types Observed in the STSMARS Test Results

	Model Type	STSMARS MA						
		Model Type						
		Mean + Independent Predictors	Linear (all types)	Integrated I(1)	SETAR (all types)	Regime Dependent SETAR	Nonlinear	MA (all types)
Original STSMARS (Ch 4)	Mean + Independent Predictors	0	0	0	0	0	0	2
	Linear (all types)	0	3	1	0	0	0	0
	Integrated I(1)	0	0	2	0	0	0	1
	SETAR (all types)	0	1	0	2	0	1	0
	Regime Dependent SETAR	0	0	0	0	1	0	1
	Nonlinear	0	2	1	1	0	2	0

relatively simple form. The off diagonal elements of Table 7.4.2.2(b) also show there is a movement to simpler models, with four of the original six nonlinear models reclassified to a simpler type of model. Therefore, when MA component estimation is adopted, TSMARS tends to produce more simple and stable short-term economic time series models. This is in line with expectations as fluctuations in many economic series are often due to the propagation of errors.

It is important to stress that many of the models found did have an MA component in the forward stepwise TSMARS estimation. These were subsequently removed by the backward elimination procedure or because they contributed little to the overall variance. Therefore, the modelling of these components has noticeably affected the final TSMARS model in most cases leading to simpler fitting models.Once again it is also of interest to see the results of the test statistics on nonlinear models found. This analysis is given in Table 7.4.2.3. It shows the number (out of four) of not significant model adequacy tests taken from Table 7.5.1.4 as well as the corrsponding numbers for the STSMARS taken from Table 4.4.2.3.

Table 7.4.2.3: Frequency of Test Statistics that are not significant for STSMARS Nonlinear Models only.

Method	Test Series Number									
	1	2	3	4	5	6	7	8	9	10
STSMARS MA	-	1	-	-	-	4	4	-	4	4
STSMARS AO-MA	-	1	-	-	-	4	4	-	3	4
STSMARS	-	1	-	1	-	4	4	4	3	3
Method	Test Series Number									
	11	12	13	14	15	16	17	18	19	20
STSMARS MA	0	-	0	3	2	-	-	0	0	0
STSMARS AO-MA	-	-	0	3	2	-	-	0	1	0
STSMARS	3	-	1	-	-	-	1	1	2	1

Comparing the results in table 7.4.2.3 there is a good deal of agreement. This is to be expected as MA components give smoother, more parsimonous approximations to the data, but will not greatly affect the size of residuals in a growth series. However, there is a difference for series numbered 14 and 15; these are, Dublin Airport Rainfall (mm) and Mullingar Rainy Days (No.) respectively. This result is interesting because these are the only non-economic series included in the test-bed. Moreover, both of these series were modelled with SETMA(1,0,12) (i.e. right regime only parsimonous twelve month lag) models. This intriguingly suggests that naturally occurring weather processes depend on a one-year threshold and the longer-term trend lagged by one year. For example, the implication of this for rainfall at Dublin Airport is, if there was rainfall one year ago, then rainfall should be expected now that is equal to the average rainfall

plus a constant times the deviation about the long-term trend one-year ago. In addition, if there was no rainfall one year ago, then no rainfall should be expected now.

7.4.3 Concluding Remarks

TSMARS has been adapted to handle MA component estimation and the program run on the 'test-bed' of 20 empirical time series. The program was used both with and without outlier adjustment.

From the results obtained the following can be concluded:

- The analysis of the test statistics given in Table 7.4.2.1 demonstrates that the inclusion of MA components did not result in any dramatic improvement in most model adequacy or seasonal effects.
- The existence of threshold nonlinearity and hetroscedastic seasonality was eliminated.
- Cycles persist in the data and these appear to be attributable to periodic autoregressive effects.
- The MAPE values across the methods were not significantly different though some of these were improved over Chapter 4 values.
- The mobility analysis tables 7.4.2.2(a) and (b) showed that models were different when MA component estimation was included; this change was more evident for STSMARS.
- The MA type estimation tended to produce simpler models in line with the expectation that economic series tend to be dominated by innovation errors.

In contrast to the results of Chapter 6, where outlier adjustment was studied, the results obtained here are not contradictory. Simpler more stable models were found. Moreover, the MA type models ignored independent effects. This result is in line with Chapter 4 (and SARIMA+ modelling results). Therefore, the frequency of independent predictors found in Chapter 6 was due to MA components that upset the outlier adjustment procedure. Also, since the outlier adjusted STSMARS-MA estimates were not too different from their unadjusted counterparts, the conclusion of Chapter 6, that these data are not dominated by outliers stands. Therefore, the Chapter 4 conclusion that around 5% of the explained variance is nonlinear also stands. Moreover, without MA component estimation, models fit by TSMARS will be inappropriate and may give rise to misleading out of sample forecasts. This fact was particularly evident for the two weather series (numbers 14 and 15).

7.5 Conclusions

The purpose of this chapter was to enable TSMARS to model dependent error processes. A novel extension that builds on existing TSMARS methodology was set out. This is based on CLS estimation of a parsimonious MA model. This extension allows TSMARS to identify SETMA, ASTAR and ASTARMA models; both ASTMA and ASTARMA are novel model forms.

The MA extension to TSMARS was then shown to be statistically correct on data simulated from three different models. In each of these simulation studies, the extension provided an acceptable number of correct models with accurate parameter values. This conclusion being subject to the proviso, that the MA component is sufficiently large to ensure it can be distinguished.

The main drawback of this approach to identifying MA-type components is the amount of computing time involved. This means the method is best used where the number of observations is moderate and where standard TSMARS has already been used for preliminary analysis.

Finally, attention returned to TSMARS with MA component estimation of the 'test-bed' of 20 economic time series. The results showed that no significant improvement was observed in accuracy and in most model adequacy or seasonal tests. However, the existence of threshold nonlinearity and hetroscedastic seasonality was eliminated, as cycles were due to periodic effects. In general, simpler more stable models that ignored independent effects were found. It was inferred from this that MA components had upset the outlier adjustment procedure in Chapter 6. Moreover, since MA type models found both 'with' and 'without' outlier adjustment were similar, the outlier adjustment procedure can be judged sound. Therefore, the extent of nonlinearity was unaltered by outliers. The conclusion, that the original estimate of around 5% of the explained variance was nonlinear was therefore accepted. Accordingly, MA component estimation has provided greater insight into the nature and extent of nonlinearity of these series than is otherwise possible.

Table Appendix

Table 7.5.1.1: MA(1) Model Simulation Results

No of lagged predictors	Maximum Interaction Degree	Parameter θ	n	Parameter Estimate	$Std.Err.(\hat{\theta})$		Number of MA(1) Models found	Total Simulation Time
					Actual	True		
1	1	0.8	100	0.73	0.110	0.060	75	201
			200	0.74	0.089	0.042	87	772
		0.5	100	0.50	0.101	0.087	71	228
			200	0.51	0.075	0.061	78	1,139
		-0.5	100	-0.45	0.142	0.087	76	218
			200	-0.48	0.076	0.061	80	1,135
		-0.8	100	-0.70	0.142	0.060	70	210
			200	-0.71	0.084	0.042	79	1,712
3	3	0.8	100	0.73	0.106	0.060	67	418
			200	0.74	0.085	0.042	73	2,721
		0.5	100	0.51	0.099	0.087	66	452
			200	0.51	0.076	0.061	86	3,090
		-0.5	100	-0.46	0.088	0.087	65	718
			200	-0.48	0.077	0.061	84	3,687
		-0.8	100	-0.69	0.113	0.060	62	438
			200	-0.70	0.085	0.042	69	3,068

Table 7.5.1.2: ARMA(1,1) Model Simulation Results

					Standard Error			
No of lagged predictors	Maximum Interaction Degree	n	Parameter	Parameter Estimate	Actual Estimate	True	Number of ARMA(1,1) Models found	Total Simulation Time (sec)
1	1	100	ϕ = 0.75	0.73	0.086	0.054	73	256
			θ = 0.25	0.31	0.110	0.080	-	-
		200	ϕ = 0.75	0.78	0.067	0.039	96	1,153
			θ = 0.25	0.25	0.143	0.057	-	-
		100	ϕ = 0.5	0.57	0.078	0.067	80	306
			θ = 0.5	0.43	0.114	0.067	-	-
		200	ϕ = 0.5	0.56	0.072	0.047	92	1,114
			θ = 0.5	0.44	0.097	0.047	-	-
		100	ϕ = -0.75	-0.87	0.117	0.160	57	270
			θ = 0.25	0.42	0.183	0.234	-	-
		200	ϕ = -0.75	-0.73	0.172	0.113	92	1,133
			θ = 0.25	0.25	0.223	0.165	-	-
		100	ϕ = -0.5	-0.49	0.101	0.102	35	220
			θ = -0.25	-0.29	0.129	0.114	-	-
		200	ϕ = 0.5	-0.60	0.079	0.072	76	1,315
			θ = -0.25	-0.13	0.163	0.081	-	
		100	ϕ = -0.50	-0.57	0.084	0.067	70	287
			θ = -0.50	-0.40	0.110	0.067	-	-
		200	ϕ = -0.50	-0.57	0.076	0.047	90	1,333
			θ = -0.50	-0.40	0.092	0.047	-	-

Table 7.5.1.3: SETMA(2,1,1) Model Simulation Results

No of lagged predictors	Maximum Interaction Degree	n	Parameter	Parameter Estimate	Parameter Estimate Standard Error	Number of TMA(2,1,1) Models found	Total Simulation Time (secs)
1	1	300	$\theta_1 = 0.75$	0.75	0.113	95	1,158
			$\theta_2 = 0.25$	0.30	0.108	-	-
			r = 0	-0.08	0.161	-	-
		500	$\theta_1 = 0.75$	0.75	0.101	99	2,945
			$\theta_2 = 0.25$	0.30	0.093	-	-
			r = 0	-0.07	0.124	-	-
		300	$\theta_1 = 0.5$	0.46	0.123	69	995
			$\theta_2 = -0.25$	-0.16	0.110	-	-
			r = 0	-0.08	0.127	-	-
		500	$\theta_1 = 0.5$	0.50	0.089	73	2,567
			$\theta_2 = -0.25$	-0.18	0.061	-	-
			r = 0	-0.08	0.057	-	-
		300	$\theta_1 = 0.5$	0.41	0.139	18	937
			$\theta_1 = -0.5$	-0.29	0.065	-	-
			r = 0	-0.09	0.087		
		500	$\theta_1 = 0.5$	0.41	0.090	15	2,484
			$\theta_2 = -0.5$	-0.32	0.067	-	-
			r = 0	-0.11	0.100	-	-
		300	$\theta_1 = 0.75$	0.69	0.142	46	1,080
			$\theta_2 = -0.25$	-0.10	0.102	-	-
			r = 0	-0.10	0.114	-	-
		500	$\theta_1 = 0.75$	0.71	0.108	54	2,918
			$\theta_2 = -0.25$	-0.14	0.058	-	-
			r = 0	-0.09	0.090	-	-

Table 7.5.1.4: Time Series Test-bed Results with Moving Average Estimation

No	N	Series Code	Title	Model type	Method	MA evidence	Model
1	286	ASAM003	Cows Milk Protein Content (%)	TSMARS	MA	N	$y_t = 210{,}770 + 1.01(y_{t-1} - 180{,}102)_-$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta\Delta_{12}[y_t = -0.03 - 0.4\,y_{t-1} - 0.17\,y_{t-2} + 0.5\,y_{t-12} + 0.001TD]$
					AO-MA	Y	AS ABOVE
2	286	ASAM206	Calves Slaughtering 000 Heads	TSMARS	MA	Y	$y_t = 0.14 + 0.57\,y_{t-1} + 18.8y_{t-1}(\text{Sint} - \pi/3)_+ - 24.2y_{t-1}(\text{Sint} - \pi/3)_+(y_{t-10} - 0.9)_-$
					AO-MA	Y	AS ABOVE
				STSMARS	MA	N	$y_t = 0.95 + 0.58\,y_{t-1} + 0.91y_{t-1}(y_{t-2} - 1.2)_-$
					AO-MA	N	$y_t = 1.03 + 0.58\,y_{t-1} + 1.29y_{t-1}(y_{t-2} - 0.8)_-$
3	286	ASAM305	Heifers Slaughtering 000 Tons	TSMARS	MA	N	$y_t = 11.2 - 1.23(y_{t-1} - 4.6)_-$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta[y_t = -0.08 - 0.33\,y_{t-1} + 0.54y_{t-12} - 0.25\varepsilon_{t-1}]$
					AO-MA	Y	AS ABOVE
4	250	FIAM023	Irish Currency in Circulation (€)	TSMARS	MA	N	$y_t = 1069 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta_2[y_t = -0.08 + 0.61\,y_{t-12}]$
					AO-MA	N	$\Delta_2[y_t = -0.08 + 0.68\,y_{t-12}]$
5	324	FIAM102	Exchange Rate $ £STR	TSMARS	MA	N	$y_t = 0.75 + 0.98\,y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta[y_t = -0.02 + 0.37\ y_{t-1}]$
					AO-MA	Y	$\Delta[y_t = -0.01 + 0.34\ y_{t-1}]$
6	179	LRGM001	Live Register Total (No)	TSMARS	MA	N	$y_t = 133{,}416 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta\Delta_{12}[y_t = -4{,}442 + 0.39\ y_{t-2} + 1.25(y_{t-12} - 6{,}372)_-]$
					AO-MA	Y	AS ABOVE
7	179	LRGM111	Live Register/ Tara St. Total (No)	TSMARS	MA	N	$y_t = 974 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	N	$\Delta[y_t = -0.02 + 0.48(y_{t-12} - 0.4)_+]$
					AO-MA	Y	$\Delta[y_t = -0.06 + 0.2\ y_{t-1} + 0.42\ y_{t-12}]$
8	179	LRGM438	Live Register/ Thomastown Males (No)	TSMARS	MA	N	$y_t = 181 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta[y_t = -0.05 + 0.31y_{t-12}]$
					AO-MA	Y	AS ABOVE

No	Model Type	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
1	TSMARS	MA	0.01	0.01	2, 3, 5	0.01	0.01	1.0	1.0	0.3	10.9	0.0	3
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.01	0.01	2, 4	0.01	0.05	0.01	0.01	1.0	8.0	0.0	2
		AO-MA				AS	ABOVE						
2	TSMARS	MA	0.25	0.99	6 - 8	0.01	0.01	0.98	0.39	0.22	48.6	58.6	-
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.01	0.73	3, 6	0.01	0.11	0.01	0.01	0.53	61.3	57.4	12
		AO-MA				AS	ABOVE						
3	TSMARS	MA	0.01	0.01	1 - 3	0.01	0.01	0.01	0.01	1.0	11.4	9.8	3
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.01	0.01	2	0.01	0.01	0.29	0.01	1.0	8.0	9.8	3
		AO-MA				AS	ABOVE						
4	TSMARS	MA	0.01	0.12	1, 12	0.01	0.01	0.01	0.01	0.01	3.0	0.0	2
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.01	0.01	10, 11	0.01	0.90	0.01	0.01	1.0	9.8	0.0	2
		AO-MA	0.01	0.01	1 - 3	0.01	0.94	0.01	0.01	0.45	2.0	0.0	3
5	TSMARS	MA	0.01	0.06	5	0.01	0.06	0.01	0.01	0.42	1.1	1.2	12
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.40	0.45		0.01	0.08	0.15	0.15	0.70	1.1	1.1	
		AO-MA	0.01	0.01	2 - 12	0.01	1.0	0.01	0.01	0.70	5.6	1.1	10
6	TSMARS	MA	0.01	0.01	1-3, 6	0.01	0.01	1.0	1.0	0.11	2.2	0.0	6
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.09	0.24		0.74	1.0	0.01	0.01	0.73	1.1	0.0	4
		AO-MA				AS	ABOVE						
7	TSMARS	MA	0.47	0.96	8 - 12	0.01	0.01	0.35	0.44	0.55	3.2	0.0	
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.65	0.86		0.01	0.11	0.08	0.07	0.55	3.0	0.0	4
		AO-MA				AS	ABOVE						
8	TSMARS	MA	0.01	0.01	1	0.59	0.01	0.22	0.34	0.04	3.2	0.3	6
		AO-MA				AS	ABOVE						
	STSMARS	MA	0.11	0.01	1	0.50	0.-01	0.20	0.32	1.0	3.3	0.3	2
		AO-MA				AS	ABOVE						

No	N	Series Code	Title	Model type	Method	MA evidence	Model
9	179	LRGM515	Live Register/ Nenagh Males (No)	TSMARS	MA	N	$y_t = 314 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta[y_t = -0.01 - 0.27(y_{t-3} - 0.14)_+ + 0.16\varepsilon_{t-2}]$
					AO-MA	Y	$\Delta[y_t = -0.01 + 0.27\varepsilon_{t-3} + 26.5(y_{t-2} - 0.12)_+(\varepsilon_{t-13} - 0.12)_-]$
10	179	LRGM800	Live Register/ Newcastle West females (No)	TSMARS	MA	N	$y_t = 286 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$y_t = -12.3 - 0.27\,y_{t-3} + 0.43\,y_{t-12} - 0.01 y_{t-3}(y_{t-2} - 115)_+$
					AO-MA	Y	AS ABOVE
11	288	MIAM014	Volume Index NACE 37 (Base 1985= 100)	TSMARS	MA	N	$y_t = 35.7 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	N	$\Delta_{12}[y_t = -0.12 + 0.25\,y_{t-1} + 0.28\,y_{t-2} + 0.77\,y_{t-1}(y_{t-12} - 0.53)_-]$
					AO-MA	N	$\Delta_{12}[y_t = -0.07 + 0.47\,y_{t-1} + 0.29\,y_{t-2} - 0.19\,y_{t-12}]$
12	288	MIAM051	Volume Index Manufacturing Industries (Base 1985 =100)	TSMARS	MA	N	$y_t = 56.0 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	N	$y_t = 4.05 + y_{t-1}$
					AO-MA	N	AS ABOVE
13	288	MIAM524	Volume Index NACE 429 Adjusted (Base 1985= 100)	TSMARS	MA	N	$y_t = 102 - 1.4(y_{t-1} - 26.5)_-$
					AO-MA	N	AS ABOVE
				STSMARS	MA	N	$\Delta\begin{bmatrix} y_t = 0.16 - 0.3\,y_{t-2} - 0.15\,y_{t-3} + \\ 4.8(y_{t-1} - 0.62)_- - 0.3(y_{t-1} - 0.62)_- \end{bmatrix}$
					AO-MA	N	$\Delta[y_t = 0.05 - 0.29\,y_{t-2} + 0.77(y_{t-1} - 0.14)_-]$
14	288	MTAM351	Dublin Airport Rainfall (mm)	TSMARS	MA	N	$y_t = 62.2 + 0.1(\varepsilon_{t-12} - 64.6)_-$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$y_t = 3.99 + 0.07(\varepsilon_{t-12} - 2.55)_-$
					AO-MA	Y	$y_t = 4.0 + 0.07(\varepsilon_{t-12} - 2.03)_-$
15	288	MTAM553	Mullingar Rainy Days (No.)	TSMARS	MA	Y	$y_t = 18.0 + 0.22(\varepsilon_{t-12} - 14)_+$
					AO-MA	Y	AS ABOVE
				STSMARS	MA	Y	$y_t = 19.0 - 0.05(\varepsilon_{t-12} - 16)_+$
					AO-MA	Y	AS ABOVE
16	380	RSAM501	Retail Sale Index: All Business Value Adjusted Base 1990 = 100	TSMARS	MA	N	$y_t = 9.7 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	N	$y_t = 10.7 + y_{t-1}$
					AO-MA	N	AS ABOVE

No	Model type	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
9	TSMARS	LSAO	0.01	0.97	5, 6	0.63	0.01	0.69	0.02	1.0	4.0	0.1	6
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.09	0.90	None	0.74	0.01	0.91	0.02	0.90	3.9	0.1	6
		TSAO	0.04	0.98	13	0.81	0.01	0.69	0.01	0.29	3.7	0.1	6
10	TSMARS	LSAO	0.01	0.14	2, 3	0.66	0.01	0.01	0.01	1.0	4.4	0.2	
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.80	0.13	None	0.55	0.04	0.02	0.09	0.76	3.4	0.2	
		TSAO				AS	ABOVE						
11	TSMARS	LSAO	0.01	0.01	1 - 3	0.01	0.98	0.01	0.01	0.99	6.8	0.5	2
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	1	0.01	0.98	0.01	0.01	0.99	6.8	0.5	2
		TSAO	0.01	0.01	1 - 4	0.01	0.01	0.01	0.01	0.03	16.9	0.5	2
12	TSMARS	LSAO	0.01	0.01	6	0.01	0.65	0.92	0.97	0.16	3.0	0.5	6
		TSAO				AS	ABOVE						
	STSMARS	LSAO											
		TSAO				AS	ABOVE						
13	TSMARS	LSAO	0.01	0.01	All	0.01	0.98	0.01	0.01	1.0	10.2	1.0	10
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	14	0.01	0.79	0.05	0.04	0.99	5.6	1.0	10
		TSAO	0.01	0.01	3, 4	0.01	0.72	0.01	0.01	1.0	5.9	1.0	10
14	TSMARS	LSAO	0.81	0.74	3	0.89	0.01	0.51	0.28	0.16	75.9	1.2	12
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.70	0.57	3	0.69	0.01	0.01	0.01	1.0	65.3	1.2	12
		TSAO	0.82	0.62	3	0.61	0.01	0.01	0.01	1.0	66.3	1.2	12
15	TSMARS	LSAO	0.32	0.83	None	0.82	0.01	0.62	0.74	0.01	28.7	5.1	6
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.76	12	0.72	0.56	0.01	0.39	0.62	29.1	5.1	6
		TSAO				AS	ABOVE						
16	TSMARS	LSAO	0.01	0.01	1, 5, 11	0.01	0.86	0.01	0.01	0.3	2.5	1.0	4
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	1, 2, 9 -11	0.01	0.01	0.98	0.01	0.4	1.8	1.0	4
		TSAO				AS	ABOVE						

No	N	Series Code	Title	Model type	Method	MA evidence	Model
17	466	TRAM009	Exempt Vehicles New (No.)	TSMARS	MA	Y	$y_t = 28.5 + 0.5\,y_{t-1} + 0.36y_{t-10}$
					AO-MA	Y	AS ABOVE
				STSMARS	MA	Y	$y_t = 3.1 + 0.64y_{t-1}$
					AO-MA	Y	AS ABOVE
18	380	TSAM043	Imports SITC 59 Other Chemicals €000	TSMARS	MA	N	$y_t = 44{,}669 - 1.05\,(y_{t-1} - 43{,}443)_-$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta\left[y_t = 0.38 - 0.45\,y_{t-1} + 1.43(y_{t-2} - 0.45)_- - 0.38(y_{t-3} - 0.91)_+\right]$
					AO-MA	Y	$\Delta\left[\begin{matrix} y_t = 0.21 - 0.51\,y_{t-1} + 0.12y_{t-2} + 0.2y_{t-12} + 0.22y_{t-13} - \\ 0.03y_{t-2}\varepsilon_{t-13} \end{matrix}\right]$
19	380	TSAM055	Imports SITC 71 Power Machinery €000	TSMARS	MA	N	$y_t = 4{,}883 + 0.42\,y_{t-1} + 0.52\,y_{t-12}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta\left[y_t = 0.28 - 0.18\,y_{t-1} - 0.87(y_{t-1} - 1.39)_+ + 0.22(y_{t-13} - 1.04)_+\right]$
					AO-MA	Y	$\Delta\left[y_t = 0.3 - 0.22\,y_{t-2} - 0.68(y_{t-1} - 0.68)_+ + 0.23(y_{t-12} - 0.84)_+\right]$
20	380	TSAM601	Exports Adjusted €000	TSMARS	MA	N	$y_t = 67{,}900 + y_{t-1}$
					AO-MA	N	AS ABOVE
				STSMARS	MA	Y	$\Delta_2\left[y_t = -123{,}839 - 0.27\,y_{t-2} + 0.39y_{t-3}\right]$
					AO-MA	Y	$\Delta_2\left[y_t = -155{,}705 - 0.27\,y_{t-2} + 0.4y_{t-3}\right]$

No	Model type	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
17	TSMARS	LSAO	0.01	0.01	14	0.01	0.01	0.01	0.01	0.92	51.4	0.5	3
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	1, 5, 6	0.01	0.01	0.02	0.01	0.34	45.9	0.5	3
		TSAO				AS	ABOVE						
18	TSMARS	LSAO	0.01	0.01	1, 15	0.01	0.01	0.01	0.01	0.44	23.9	0	12
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	1 - 2, 5	0.01	0.01	0.01	0.01	0.46	20.8	0	12
		TSAO	0.01	0.01	1 - 5	0.01	0.01	0.01	0.01	0.59	20.4	0	4
19	TSMARS	LSAO	0.01	0.01	None	0.01	0.29	0.01	0.01	1.0	22.3	0	4
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	2	0.01	0.40	0.01	0.01	0.44	18.6	0	4
		TSAO	0.01	0.01	None	0.01	0.26	0.01	0.01	0.80	17.9	0	4

No	Model type	Method	Statistics										
			χ_1^2	χ_2^2	Tsay F-test lags	BDS test	Seasonality F-test	PAR(1) F-test	PAR(2) F-test	Periodic Variation F-test	MAPE	% Error	Cycles/ Notes
20	TSMARS	LSAO	0.01	0.01	2 – 5, 9 – 14	0.01	0.92	0.01	0.01	0.42	7.2	0	2
		TSAO				AS	ABOVE						
	STSMARS	LSAO	0.01	0.01	6, 12	0.01	0.94	0.01	0.01	1.0	9.6	0	8
		TSAO	0.01	0.01	2, 6	0.01	0.93	0.01	0.01	1.0	7.8	0	8

8 Bootstrapped TSMARS Forecast Errors

8.1 Introduction

In time series analysis, a forecast of an unknown future value is called a point prediction. Its distribution is known as the predictive distribution and this plays an important role when data are not normal. The statistical uncertainty of the point prediction is referred to as the future prediction interval, or simply the prediction interval. This interval will contain the future value with high probability, say 95% or 99%. In this chapter realistic prediction intervals for some TSMARS models are computed using novel adaptations of existing bootstrap methods.

The predictive interval for a stationary time series with normal errors is given as the mean plus and minus a multiple of the standard deviation of the predictive distribution. Predictive intervals constructed in this way have been used for some special nonlinear models including the SETAR model (see de Bruin 2001). However, for most nonlinear data, where the noise may be asymmetric and/or multimodal, this approach is unsuitable. In this chapter an alternative approach based on bootstrap re-sampling (see Efron & Tibshriani 1990) is adopted.

Bootstrap methods have been used with success for computing intervals for parameter estimates (see Buhlmann 2002) and autocorrelations (see Romano & Thombs, 1996). In this chapter our focus is on computing bootstrap future prediction intervals for TSMARS models.

Intervals for the AR(p) model driven by non-Normal disturbances are constructed in Thombs & Schucany (1990), Breidt et. al. (1995), Romano & Thombs (1996), Hansen (1999) and Kim (2002). The method set out in Thombs & Schucany (1990) is reviewed in section 8.3 and adapted to TSMARS models. The key element of this adaptation is that the back casting device of Thombs & Schucany (1990) is avoided. The method is called the TSMARS Sieve Bootstrap. Sieve type bootstraps are parametric as they fit a parametric model to the observations. Residuals from the model are re-sampled to generate bootstrap time series realisations. Based on these a predictive interval is computed.

Also implemented is the Vectorised Block Bootstrap of Buhlmann (2002). This method is nonparametric and was developed for computing intervals for parameter estimates. It is adapted in section 8.3 for computing prediction intervals. In general nonparametric methods re-sample blocks of the observations directly. These blocks are joined together to generate a bootstrap time series realisation.

In section 8.4 both the Sieve and Vectorised block bootstrap methods are contrasted in simulations studies. These studies show that the predictive intervals are accurate and consistent with increasing sample size for models driven by noise from normal, exponential and normal mixture models respectively. For data simulated from a linear model the results obtained are compared with those obtained elsewhere.

In section 8.5 the bootstrap forecast methods are applied to a subset of the empirical test-bed series. Independent effects in these data are incorporated directly into the models used. In the Sieve Bootstrap case, seasonality is also catered for in the model, while in the Block Bootstrap case, the block length is chosen so that it does not interfere with seasonality. These methods are applied to the test-bed series

and prediction intervals computed for a forecast horizon of 12 steps. One important aspect of this study is that the prediction intervals can be judged against the cross-validation estimates of Chapter 5.

8.2 Preliminaries

The overall objective is to obtain a prediction interval for a future value for a time series. In Chapter 5 forecast errors from a time series of length n were computed for $k = 5$ steps ahead with a TSMARS model computed on data up to $n-k$. The precision of the forecast was computed as the RMS forecast error over S simulated datasets. This procedure computed the predictive distribution

$$p(y_{n-k+i} \mid p(y_{n-k})) \tag{8.2.1}$$

at each time point $n-k+i\,(i=0\ldots k-1)$. This distribution is not the true conditional predictive distribution, since y_{n-k} is not fixed for every simulated dataset. However, formula (8.2.1) computes an approximation to the unconditional density as S $\rightarrow \infty$. Clearly, this predictive distribution is less relevant as k increases.

The (true) conditional predictive distribution at k steps ahead given an infinite realisation $\mathbf{Y}_n = y_n, \ldots$ is

$$p(y_{n+k} \mid \mathbf{Y}_n) = \int_{-\infty}^{\infty} p(y_{n+k} \mid y_{n+1})\, p(y_{n+1} \mid \mathbf{Y}_n)\, dy_{n+1} \quad k = 1\ldots \tag{8.2.2}$$

This is a recursive equation for the conditional predictive density. It is known as the Chapman-Kolmogorov (C-K) integral equation and can be evaluated using Gaussian quadrature (see Tong 1990 subsection 4.2.4.3). Recall this numerical approach was used to compute intervals in Chapter 5. Based on (8.2.2) the distribution the predictive interval, denoted by PI_{n+k}, that covers a future value y_{n+k} with probability $\beta = 100(1-\alpha)\%$ is

$$PI_{n+k}(L_n, U_n) = P(L_n < y_{n+k} \leq U_n \mid \mathbf{Y}_n) \tag{8.2.3}$$

where L_n and U_n are the upper and lower quantiles of the interval and $P(L_n < y_{n+k} \leq U_n \mid \mathbf{Y}_n)$ is the corresponding distribution function. If as S $\rightarrow \infty$ $E\{PI_{n+k}(L_n, U_n)\} \rightarrow 1-\alpha$ then $PI_{n+k}(L_n, U_n)$ is also a $\beta = 100(1-\alpha)\%$ unconditional interval for y_{n+k}. Predictive intervals constructed in this was are known as parametric intervals – the standard normal interval (e.g. Box-Jenkins 1976 approach for the linear model) is the most commonly used.

For autoregressions, the distribution of $y_{n+k} \mid \mathbf{Y}_p$ where $\mathbf{Y}_p = y_n, y_{n-1} \cdots y_{n-p+1}$ is the same as the distribution of $y_{n+k} \mid \mathbf{Y}_n$. Therefore it makes sense to write the conditional predictive interval for autoregressive processes as

$$PI_{p,n+k}(L_n, U_n) = P(L_n < y_{n+k} \leq U_n \mid \mathbf{Y}_p) \tag{8.2.4}$$

Once again, if as S $\rightarrow \infty$ $E\{PI_{p,n+k}(L_n, U_n)\} \rightarrow 1-\alpha$ then $PI_{p,n+k}(L_n, U_n)$ is also a $\beta = 100(1-\alpha)\%$ unconditional interval for y_{n+k}. A key purpose of this chapter is to construct intervals of the type (8.2.4) and to show by simulation that $E\{PI_{p,n+k}(L_n, U_n)\} \rightarrow 1-\alpha = \beta$. In simulation studies we take $\alpha = 0.1$ and

so we consider 90% intervals – simulation results for other values of β not reported here are similar to the $\beta = 90\%$ case.

In this chapter the effort is on generating bootstrap critical quantile values L_n and U_n. Bootstrapping is different to other re-sampling methods such as cross-validation in that the data are subtly randomised. The full sample of data is used and randomly reordered so as to preserve the underlying characteristics of the data generating process.

The first approach is known as the parametric bootstrap. It relies on finding L_n and U_n based on a parametric model for the data and then 'pivoting' on these critical values to isolate y_{n+k}. The method adopted here follows Thombs & Schucany (1990) in that future values are used to define the root quantity $R_n = y_{n+1}$ (see Breidt et. al 1995). Letting $\hat{f}(y_n)$ denote the nonlinear (e.g. TSMARS) model fit to the data, the root may be written as

$$R_n = \hat{f}(y_n) + \{y_{n+1} - \hat{f}(y_n)\} = \hat{f}(y_n) + \hat{\varepsilon}_t \tag{8.2.5}$$

where $\hat{\varepsilon}_t$ are the residuals from the model fit. This root quantity can then be used to generate replicate samples of the observed time series as

$$R_t^* = \hat{f}(y_t^*) + \hat{\varepsilon}_t^* \tag{8.2.6}$$

where y_t^* and $\hat{\varepsilon}_t^*$ are replicates of y_t and ε_t respectively. The predictive interval (8.2.4) is then computed from this replicate series. In practice a large number (say 100) bootstrap replicate series are used to estimate the predictive interval. The Thombs & Schucany (1990) method of constructing the predictive interval is described in detail in the next section. In general parametric bootstrap methods preserve the underlying characteristics of the data generating process by preserving correlation structure in the observations.

The second method is wholly nonparametric. Here the observations $y_1, y_2 \ldots y_n$ are randomly selecting in blocks and these blocks are then joined end to end to create a replicate series y_t^*. The predictive interval (8.2.4) is then computed from this replicate series. Once again a large number (say 100) bootstrap replicate series are used to estimate the predictive interval. Block bootstrap methods rely on a suitable value for block length to preserve the underlying characteristics of the data generating process.

8.3 Bootstrapping Methods for Time Series

This section is methodological in that two algorithms for bootstrapping time series future values are set out; namely, the Parametric Bootstrap Method and the Nonparametric Bootstrap Method. The section begins with a review of the Thombs & Schucany (1990) algorithm for computing predictive intervals of the linear AR(p) model. Next this algorithm is adapted to handle NLAR(p) problems – this method is called the TSMARS Sieve Bootstrap. Finally the Nonparametric Vectorised Block Bootstrap of Buhlmann (2002) is described and adapted for computing predictive intervals.

8.3.1 Linear AR(p) Parametric Bootstrapping

For linear time series data with finite 2nd order moments, the parametric bootstrap method has been used with success by among others (Thombs & Schucany 1990, Breidt et. al 1995, Romano & Thombs 1996, Hansen 1999 and Kim 2002). In general, the method has been used to generate prediction intervals based on quantiles of the predictive distribution where the disturbances are not Normal. The Thombs & Schucany (1990) method is set out here for the AR(1) model; generalisations to AR(p) are straightforward.

Consider the stationary AR(1) time series model defined by

$$y_t = \phi_f \, y_{t-1} + \varepsilon_t \tag{8.3.1}$$

where ϕ_f is an unknown constant, $\{\varepsilon_t\}$ is a sequence of zero mean independent errors with common distribution function F_ε having finite 2nd order moments and $t = 0,\pm1,\pm2,\ldots$. Model (8.3.1) is called the forward model and associated with it is the backward model where

$$y_t = \phi_b \, y_{t+1} + e_t \tag{8.3.2}$$

These two models have the same correlation structure (see Box & Jenkins 1976) endowing the time series with a useful time-reversible property. This property is particularly useful as it allows replicate series to be generated that have the same last value (or last p values as appropriate) and, in addition, have the same correlation structure. Clearly, from the definition of the predictive interval (8.2.4) replicate series having the same last value is a fundamental requirement.

The steps involved in generating a bootstrapped prediction interval are (see, Thombs & Schucany 1990):

1 Estimate the AR(1) forward model and associated centred residuals $\{\hat{\varepsilon}_t\}$, and let $\hat{F}_\varepsilon$ be their distribution function.

2 Estimate the AR(1) backward model and associated centred residuals $\{\hat{e}_t\}$ and let $\hat{F}_e$ be their distribution function.

3 Compute the bootstrap replicate y_n^* by setting $y_n^* = y_n$ (i.e. the last data value) and use the estimated backward model to compute

$$y_{t-j}^* = \phi_b \, y_{t-j+1}^* + \hat{e}_{t-j}^*$$

where $\hat{e}_{t-j}^*$ are random i.i.d. draws from $\hat{F}_e$.

4 Compute new forward model estimates $\hat{\phi}_f$ from the bootstrap replicate.

5 Compute future values y_{n+k}^* using the new forward model parameter estimates and forward residuals drawn i.i.d. from $\hat{F}_\varepsilon$. Note, these will be conditional on $y_n^* = y_n$.

6 Repeat steps 3 – 5 until B bootstrap future values y_{n+k}^* have been defined at each step ahead k.

7 Let G_B^* be the cdf of the future value y_{t+k}^*, then the endpoints of the prediction interval are given by the quantiles of G_B^*.

Note that a prediction interval constructed in this way is a percentile interval in the sense described in Hall (1992).

In general there are two concerns with the above method. The first lies in the fact that the forward F_ε and backward F_e residual distributions are not the same when the innovation is correlated. As a consequence the re-sampling scheme should only be used when the innovation sequence is i.i.d. (see Thombs & Schucany 1990). Fortunately for TSMARS models this is the case as the innovation is assumed i.i.d.
A second concern is that the method relies on the assumption of time reversibility. As noted in Chapter 1 this cannot be assumed for TSMARS models. Even the simple SETAR(2,1,1) model has a correlation structure that is regime dependent. Moreover, recall from Chapter 1 that y_t can be generated from y_{t-1} in the forward problem, but this is impossible in the backward problem, as the regime depends on y_{t-1} which is unknown. This asymmetry is fundamental and means that back casting cannot be used for TSMARS models.

8.3.2 Nonlinear AR(p) Parametric Bootstrapping

In this subsection a novel alternative to back casting is proposed that guarantees the last p values across every replicate series are the same. This involves removing the last p values from the observed time series and appending them to end of each replicate series (of length $n-p$). If this is done in a sensible manner, then the distribution of future values will be conditional on these last p values of the data. This adaptation is now implemented within the so-called sieve bootstrap algorithm (see Buhlmann 2002).

TSMARS Sieve Bootstrap

1. Given an observed time series y_t $(t=1,2,\ldots,n)$ estimate the TSMARS model $f_t(\bullet)$ and compute the estimated innovation errors $\hat{e}_t = y_t - f_t(\bullet)$. The prediction interval will be computed for this model. This model also fixes the lag order p. The last p values of the observed series are retained for appending to each replicate series.
2. Set m = 1,000.
3. Start with $y_{-m}^*, y_{-m+1}^* \cdots, y_{-m+p-1}^*$ as a randomly selected subseries from y_t and simulate y_t^* for $t=-m+p,\ldots,0,1,2,\ldots,l, l>n$ from $f_t(\bullet)$ as $y_t^* = f_t(\bullet)+\hat{e}_t$ where $\hat{e}_t$ is a random draw from $\hat{F}_e$ the distribution of the estimated innovation errors rescaled by $\sqrt{(n-p)/(n-2p)}$ (see Thombs & Schucany 1990) to compensate for deflation in the innovation variance due model fitting.
4. Select $r>n-p$ such that $\left|y_{n-p+1}-y_r^*\right|, \left|y_{n-p+2}-y_{r+1}^*\right|, \ldots \left|y_n - y_{r+p-1}^*\right|$ is a minimum.
5. Select the subseries of length $n-p$ from $\{y_t^*\}$ as $y_{r-n}^*, y_{r-n+1}^* \cdots, y_r^*$.
6. To this subseries join the last p values of the original series giving the sieve bootstrap replicate series $y_{s-n}^*, y_{s-n+1}^* \cdots, y_s^*, y_{n-p+1},\ldots, y_n$ of length p.

7. Estimate the TSMARS sieve bootstrap model $f_t^{SB}(\bullet)$ and compute the innovation errors $\hat{e}_t^* = y_t^* - f_t^{SB}(\bullet)$.

   Here the sieve bootstrap replicate series is used to estimate the TSMARS model $f_t^{SB}(\bullet)$.

8. TSMARS models that do not have the same form as the original model are rejected as invalid.

9. Future values are computed using the plug-in rule (see Chapter 5), the last p original values and a valid TSMARS sieve bootstrap model $f_t^{SB}(\bullet)$. The forecast is given by $y_{n+k}^* = f_{n+k}^{SB}(\bullet) + \hat{e}_{t+k}^*$ where $\hat{e}_{t+k}^*$ (k > 0) is a random draw from $\hat{F}_e$, the distribution of the estimated sieve bootstrap replicate innovation errors.

10. Repeat steps 3-10 until B bootstrap future values of each y_{t+k}^* are available.

11. Let G_B^* be the cdf of the future value y_{t+k}^* then the endpoints of the prediction interval are given by the quantiles of G_B^*.

A couple of points are worth noting about this algorithm. First, the predictive set of last p values from the original series is retained to generate the conditional predictive distribution. This set is used both in the bootstrap replicate series and to start off the (Markovian) forecast sequence in step 9. The resulting interval therefore approximates (8.2.4) and gives the unconditional interval as $B \to \infty$.

Second, two different sets of innovations are used. The original sequence $\hat{e}_t$ is used to generate the bootstrap replicate series while the bootstrapped sequence $\hat{e}_t^*$ is used in forecasting. This provides sufficient mixing for every bootstrap replicate series. Both of these points are designed to ensure that bootstrap replicate series and forecasts provide a good approximation to the underlying predictive stationary distribution. Moreover, it is reasonable to conjecture based on Theorem 3.1 of Thombs & Schucany (1990) that $y_{t+k}^* \to y_{t+k}$ in distribution provided the $f_t^{SB}(\bullet) \to f_t(\bullet)$ in probability. Simulation studies to be conducted in the next section indicate this conjecture to be true. It is worth mentioning that this procedure is general - that is, fitting the TSMARS model at steps 1 and 7 of the algorithm could in principle be replaced by any parametric modelling method. This option is not explored further in this thesis.

8.3.3 Nonparametric Bootstrapping

A standard nonparametric method for generating bootstrap replicates for time series data is the so-called Moving Block Bootstrap (or block bootstrap for short). This method makes no assumptions about the underlying distribution of data generating process. It is described in Efron & Tibshriani (1990) and also in Buhlmann (2002).

For an observed time series y_t $(t = 1,2,\ldots,n)$, the basic idea behind the block bootstrap (see Efron & Tibshriani 1990) is to build k overlapping blocks of consecutive vectors $(y_{1,\ldots,}y_l)$, $(y_{2,\ldots,}y_{l+1}),\ldots,(y_{n-l+1,\ldots,}y_n)$, each of length l; where for simplicity, it is assumed that $n = kl\,(k \in Z+)$. The block length l is a tuning parameter of the method. Then, randomly resample with replacement, k blocks

and join them end-to-end to generate a bootstrap replicate series of the original data. Using this replicate series, estimate the relevant time series model and associated statistics. Repeat this until sufficient bootstrap estimators are available to compute final estimates.

This method of time series bootstrapping has been called the naive block bootstrap by Buhlmann (2002). The key flaw in the approach is the method is not adapted to the problem. So, for example, computing the lag(1) autocorrelation coefficient using the naïve block bootstrap, ignores the fact that the autocorrelation coefficient is a functional of the p=2 dimensional distribution of (y_t, y_{t-1}). This, as demonstrated in Buhlmann (2002), produces clusters of 'bad' points in the scatter plot of y_t vs. y_{t-1}.

The alternative proposed in Buhlmann (2002) to overcome these difficulties is the Vectorised Block Bootstrap. This method is adapted here for estimating the prediction interval of an observed time series. In this adaptation TSMARS is initially called with the original series to provide an estimate of the lag order p. Therefore, this adaptation is semi-parametric. This call determines the length of the sequence of retained last values required for forecasting. It also provides an estimate of the dimension of the predictive distribution $(y_t, y_{t-1}, \ldots, y_{t-p})$. This is important because the future value and the predictive interval, are functionals of the underlying $p+1$ dimensional predictive distribution. By retaining the last p values and sampling vectors of dimension p the method set out in the following algorithm ensures that the underlying (nonlinear) autoregressive structure of the observations is retained in replicate series.

Time Series Vectorised Block Bootstrap

1. Given an observed time series y_t $(t = 1,2,\ldots,n)$ estimate the TSMARS model $f_t(\bullet)$ to fix model and lag order p.
2. Select and appropriate value for block length l.
3. Construct the $p+1$ dimensional lagged data vector $Y_t = (y_t, y_{t-1}, \ldots y_{t-p})$ for each $t = p+1, \ldots, n$.
4. Select the vector subseries $Y_p, Y_{p+1}, \ldots Y_{n-1}$ and Y_n and retain the last as the fixed prediction vector.
5. Build overlapping blocks of consecutive vectors $(Y_p, Y_{p+1}, \ldots Y_{p+l-1}), (Y_{p+1}, Y_{p+2}, \ldots Y_{p+l}), \ldots$ $(Y_{n-1-l+1}, \ldots Y_{n-1})$ of length l and assume for simplicity that $n - p = kl$ $(k \in Z+)$.
6. Then, resample k blocks independently with replacement and join end-to-end to form a series of length $n - p$.
7. To this series append the retained prediction vector giving the bootstrap replicate series Y_t^*.
8. Estimate the TSMARS vectorised block bootstrap model $f_t^{VBB}(\bullet)$ based on this series.
   Here the vectorised bootstrap replicate replaces the original response and p predictors in the re-estimated TSMARS model.
9. TSMARS models that do not have the same form as the original model are rejected as invalid.
10. Future values are computed using the last p and a valid TSMARS vectorised block bootstrap model $f_t^{VBB}(\bullet)$. The plug-in forecast is given by $y_{n+k}^* = f_{n+k}^{VBB}(\bullet) + \hat{e}_{t+k}^*$ where $\hat{e}_{t+k}^*$ (k > 0) is a random

draw from $\hat{F}_e$, the distribution of the estimated vectorised block bootstrap replicate innovation errors and $y_t^* \in Y_t^*$.

11. Repeat steps 6-10 until B bootstrap sets of future values are available.

12. Let G_B^* be the cdf of the future value y_{t+k}^* then the endpoints of the prediction interval are given by the quantiles of G_B^*.

In this algorithm we assume the block length tuning parameter is fixed at $l = \max\left(4, \sqrt[3]{n}\right)$ for simple time series models; this figure is suggest by Buhlmann (2002) based on simulation studies. However, in this thesis no studies have been undertaken to find the relationship between l and G_B^*.

8.4 Bootstrapped Predictive Intervals for Simulated Models

8.4.1 Models and Testing Procedure

To ascertain the quality of the bootstrap methods outlined in the previous section, simulations studies are conducted based on the linear AR(1) model and the SETAR(2,1,1) model. Both models are considered under three different i.i.d. noise distributions. These distributions are taken from Thombs & Schucany (1990) and are normal, exponential and a mixture of normals respectively. The AR(1) model is

$$y_t = -0.8\, y_{t-1} + \varepsilon_t \quad \text{where } \varepsilon_t \begin{cases} N(0,1) \\ Exp(1) \\ N(-1,1)\, U < 0.9;\ N(9,1)\, U \geq 0.1 \end{cases} \tag{8.4.1}$$

where $U \sim$ Uniform[0,1]. Simulations are conducted based on each of these three models with n = 25, 50 and 100 observations respectively.

The SETAR(2,1,1) model is

$$y_t = \begin{matrix} 0.7 y_{t-1} \\ 0.3 y_{t-1} \end{matrix} + \varepsilon_t \quad \begin{matrix} if\ y_{t-1} \leq 0 \\ if\ y_{t-1} > 0 \end{matrix} \tag{8.4.2}$$

with the same noise distributions as in (8.4.1) and n = 100, 250 and 500 respectively.

For each of these model combinations and sample sizes a time series realisation is generated. Forecast values are also generated for k steps ahead. To estimate the probability content $\beta = 100(1-\alpha)\%$ and average length of interval for these series we adopt the following test procedure.

1. Simulate a series of length n according to a specific model combination and generate R=100 future values y_{n+k}^R at each step ahead k.

2. Use the bootstrap procedure to obtain a 90% prediction interval G_n^* based on $\left(L_n^*, U_n^*\right)$.

3. Estimate the conditional coverage by $\left(\beta_k^* = \#\left\{L_n^* \leq y_{n+k}^R \leq U_n^*\right\}\right)$ and interval length $Len(k) = U_n^* - L_n^*$.

Steps 1 – 3 are repeated 100 times to get a collection of summary measures $\{\beta_k^*, Len(k)\}$ In the Table Appendix we report the average value of the conditional coverage and its standard error, and also the average interval length and it standard error for the models considered. We mention that each simulation run can take several hours as TSMARS is called 10,000 times. For this reason we have not reported the standard symmetric intervals of Thombs & Schucany (1990). However, simulation studies based on fitting an exact model for the simulated series show the methods outlined are consistent when the number of bootstrap replicate series is increased to 999.

8.4.2 Discussion of Results

For both the Parametric Sieve Bootstrap (SB) and the Vectorised Block Bootstrap (VBB) methods results for a 11-step ahead prediction horizon are displayed in Tables 8.4.2.1 and 8.4.2.2 (see Table Appendix) for the AR(1) and SETAR(2,1,1) models respectively.

An extract from Table 8.4.2.1 for up to 3 – steps ahead is displayed below for the AR(1) model driven by exponential noise. In this extract the prediction interval for the smallest sample size (n=25) appears a little disappointing as the conditional coverage falls short of 90% for both bootstrap methods. However, these figures compare well with the conditional coverage figures in Thombs & Schucany (1990) – these were lower than nominal but nevertheless within a few percentage points. The differences between their figures and those given here are explained by the fact that Thombs & Schucany (1990) only fit an AR(1) model. In contrast TSMARS models do not always match this correct form. However, as the sample size is increased the conditional coverage approaches the nominal 90% showing the methods are consistent. This is in keeping with expectations and moreover the accuracy of the intervals as n increases is equal to that obtained by Thombs & Schucany (1990).

Extract of Table 8.4.2.1: 90% Bootstrapped forward prediction estimates for the AR(1) Model

	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
n=25								
1	86.4	20	3.3	30	87.6	10	2.9	32
2	82.4	30	4.0	20	85.1	10	3.9	40
3	80.7	30	4.5	30	83.3	10	4.2	40
n=50								
1	90.4	10	2.9	30	89.0	9	3.0	25
2	86.4	10	3.9	30	86.3	8	4.0	23
3	85.7	10	4.2	30	85.6	8	4.2	22
n=100								
1	90.3	10	2.9	20	89.0	7	2.9	19
2	87.2	10	4.0	20	87.4	6	3.9	18
3	86.7	10	4.3	20	86.2	7	4.2	17

Contrasting the estimates from both methods it is clear that they give very similar coverage probabilities and interval lengths. This is a somewhat surprising as it was expected the model free Vectorised Block Bootstrap would give poorer performance than the Sieve Bootstrap.

The corresponding detailed table in the Appendix shows that similar results are obtained for this AR(1) model with normal and normal mixture noise models respectively. It can be observed interval lengths at each step for normal noise are slightly narrower than their asymptotic counterparts of 3.3, 4.2 and 4.7 respectively. Taking all of these observations together we conclude that both bootstrap methods give accurate and consistent predictive intervals for this model.

An extract from Table 8.4.2.2 for up to 3 – steps ahead is displayed below for the SETAR(2,1,1) model driven by exponential noise. In this extract the prediction intervals obtained for all sample sizes is good using the Sieve method. This is also true for the block method except that in this case the method fails smallest sample size (n=100). The improvement in accuracy of the interval as the sample size increases is also evidence for consistency. In the Table Appendix this pattern is also repeated for the normal errors case.

However, it will be seen that the methods behave differently when the error is bimodal. The Sieve method produces reasonable intervals, albeit slightly narrow at the smallest sample size. In contrast the block method fails. Here TSMARS is the problem as it 'over smooths' the smaller mode of the distribution particularly at the smaller sample sizes. This results in poor coverage probabilities and unsatisfactory interval lengths. We mention this problem does not occur when the exact threshold model is fit to the data – the block bootstrap method is therefore reliable.

Extract of Table 8.4.2.2: 90% Bootstrapped forward prediction estimates for the SETAR(2,1,1) Model

	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
n=100								
1	82.5	10	1.5	37	-	-	-	-
2	83.6	10	1.5	40	-	-	-	-
3	84.0	10	1.5	40	-	-	-	-
n=250								
1	90.1	10	1.4	20	90.0	6	1.5	16
2	88.2	10	1.5	20	89.4	6	1.6	18
3	88.6	10	1.6	20	88.6	6	1.6	19
n=500								
1	89.3	10	1.5	20	88.9	7	1.4	16
2	89.4	30	1.6	10	90.0	5	1.6	16
3	89.3	20	1.6	10	89.6	5	1.6	14

8.4.3 Concluding Remarks

In this section two well known bootstrapping methods were adapted for time series modelled with TSMARS. The key feature of both methods is that the last p values of the time series are retained to generate the conditional predictive interval. The particularly novel aspect of this alteration to the parametric bootstrap was that back-casting could be avoided.

Both the parametric and block bootstrap methods were then used to generate predictive intervals for data simulated from AR(1) and SETAR(2,1,1) models respectively. In nearly all instances the methods produced credible prediction intervals. Moreover, the coverage probabilities and interval lengths were shown to converge to nominal values as the sample size increased. A surprising result was that the block bootstrap performed as well as the parametric bootstrap for the models studied. A second appealing feature of the methods is that the nonlinearity in the SETAR model was captured resulting in excellent coverage probabilities and interval lengths. We can therefore conclude that these bootstrap methods can be expected to produce reliable and consistent intervals for both linear and SETAR data in general.

8.5 Bootstrap Intervals for Short-term Economic Time Series

In this section the two bootstrap methods are applied to some of the test bed series. Bootstrap methods are only applicable when data are stationary. Therefore attention is focussed on models arising from the Seasonal TSMARS method, STSMARS. The purpose of this section is to generate prediction intervals for those stationary STSMARS models found in Chapter 4. These are contrasted against their cross-validation counterparts obtained in Chapter 5. Any unexplained divergence between these will indicate that the bootstrap methodology is defective and render the prediction interval useless.

Recall STSMARS takes a time series y_t and after appropriate transformations gives the series z_t. The lagged predictors $z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}$ are then input into the TSMARS program along with appropriately differenced trading effects predictors. The maximum interaction degree is set to 3 and basis function threshold = 2 X 10^{-8}. Bootstrap future values are generated for the transformed series z_t. On completion of the TSMARS call, the sequence of transformations are applied in reverse, giving predictions for the model (see Chapter 4 for definition of terms)

$$\hat{y}_t = \exp\left[\left(1-B\right)^{-d}\left(1-B^s\right)^{-D}\left\{f\left(z_{t-1}, z_{t-2}, z_{t-3}, z_{t-s}, z_{t-(s+1)}, z_{\text{t,MD}}, z_{\text{t,TD}}, z_{\text{t,EASTER}}\right)\right\}\right] - c$$

Bootstrap prediction intervals are then computed from these predicted values.

There are two key differences between bootstrapping the simple models of the last section and STSMARS models. First, the independent trading day predictors have to be included. These are generated as fixed effects based on the date of the future values n+k (k=1,...12). They are reused in every bootstrap replicate to generate the set of future values. Specifically, they are included in the model $f_{n+k}^{SB}(\bullet)$ in step 9 of the Sieve Bootstrap and in $f_{n+k}^{VBB}(\bullet)$ in step 10 of the Vectorised Block Bootstrap.

A second consideration is seasonality. This is not a problem for the Sieve Bootstrap as replicate series are generated directly from the STSMARS model. However, in the Vectorised Block Bootstrap, an inappropriate choice of block length will induce seasonality that is out of phase compared to the orginal

series. To ensure this does not happen, the block length is set equal to the periodicity and the block starting point is chosen as the first day of each year. This constraint ensures the seasonal effects are not tampered with by the bootstrap method. Therefore, the mixing properties of the sample innovation distribution $\hat{F}_e$ will be unaffected.

Only eight of the 20 test bed series studied in Chapter 4 are used to generate predictive intervals. The remaining models all share similarities with these eight (for example, a number of models are linear). Therefore intervals computed from the remaining series will not provide any additional useful information.

The eight test bed series chosen are numbers 1, 2, 6, 9, 11, 13, 19 and 20 from Table 5.5.1.1. These are representative of a number different statistical areas. Predictive statistics, for up to 12 steps ahead, computed from the associated STSMARS model is given in Table 8.5.1.1 (see Table appendix at end of chapter). The statistics reported are the mean predictive value, its coefficient of variation and inter quartile range (divided by the median predicted value) denoted by Mean, CV. and SIRQ respectively. The mean value actually quoted is the percentage difference of the mean predicted value from the mean value of the original data. This indicates how far the predictive value is away from the centre of the data. A 'good' starting value, for forecasting purposes, being one that is close to the mean.

The results in Table 8.5.1.1 shows a good degree of consistency across all problems considered. In particular the predictive mean value is consistent. It does not tend to the mean value as the forecasting horizon increases, but remains stable reflecting the fact that the data are differenced. There is also a good degree of consistency between the CV value at each step ahead and the SIRQ value. The latter figure, as expected, being generally larger. This trend is not maintained in problem 2, where the SIRQ values grow rapidly for the VBB method. Outliers in these data close to the end of the series cause this problem. Overall, it is clear that both methods give very similar results.

The results in Table 8.5.1.1 are useful, but of limited value. Of greater interest is their comparison with cross-validation forecast errors given in Table 5.5.1.1. Recall the figures in Table 5.5.2.1 were obtained by retaining the final years data as a cross-validation set. Errors were computed based on the deviation of the forecast value from the true retained value. The % average residual error given in Table 5.5.1.1 is compared with the average CV obtained (over the 12 –steps ahead) for each problem in Table 8.5.1.1. These figures are displayed in Table 8.5.1.2.

The figures in Table 8.5.1.2 show the bootstrap methods tended to give % errors that are close to cross validation errors of Table 5.5.1.1. Differences do exist, in particular for problems 6 and 9. For both of these problems only a small number of correct models were used to generate bootstrap replicates. This accounts for the narrower interval. Generating more bootstrap replicates will ameliorate this flaw, as more correct models will be found.

For Problem 2, outliers near the end of the series induce nonstationarity. This adversely affects the size of predictive interval. This difficulty was evident in Table 5.5.1.1 where the maximum error was 175%.

Allowing for these difficulties, both methods have worked well and give reasonably good predictive distributions. This allied to the fact the methods worked well for the simple simulated models,

demonstrates that these bootstrapping predictive intervals can be relied on. It is also nice to see that the cross validation intervals are in line with the bootstrap figures.

Table 8.5.1.2: Comparison of Errors

	Problem Number							
	1	2	6	9	11	13	19	20
% Error in Table 5.5.1.1	13.2	82.5	8.0	11.9	13.0	10.2	37.9	11.4
Sieve Bootstrap CV	14.1	193.2	1.5	4.7	12.7	7.9	27.1	3.1
Vectorised Block CV	12.7	-518.8	1.8	4.1	9.9	8.9	29.1	11.3

8.6 Closing Remarks

Novel variations of two bootstrap methods, the parametric sieve type and vectorised block have been set out. Their merits were tested on data simulated from simple time series models driven by normal innovations and on a number of empirical time series.

Three modifications to existing bootstrap scheme were implemented. First, both methods endeavoured to recreate the true predictive distribution by retaining the last (or last p) values of the series in every bootstrap replicate sample. Second, the methods ensured sufficient mixing is retained in the bootstrapped replicate series. This was accomplished using two distinct sequences of innovations; namely those from the original model to generate bootstrap samples, and those from the bootstrap model for forecasting. Third, only correct models were used to build forecasted values. With these three modifications, the Sieve and Block methods of Buhlmann (2002) have been adapted to compute predictive intervals for TSMARS models. Tests on simple simulated models found the methods produced consistent and accurate predictive intervals.

The methods were also applied to a subset of the test-bed problems. Once again consistent predictive distributions were obtained. The predictive intervals were also compared to cross-validation based intervals. This comparison showed that the bootstrap intervals were similar. As a consequence the predictive intervals obtained are accurate and reliable estimates for these empirical series. Moreover, the consistency demonstrated indicates the methods should work for other time series modelling methods.

The main limitations of this approach to generating the predictive distribution are; first, when the retained last p-values of the series are added back to the bootstrap replicate a discontinuity is introduced. In the Sieve method, the effect of this discontinuity is minimised by finding a suitable point at which to join the retained last p-values to the bootstrap replicate (step 4 of the algorithm). However, even this is not done for the VBB method.

Table Appendix

Table 8.4.2.1: 90% Bootstrapped forward prediction estimates for the AR(1) Model

	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
Normal Noise								
n=25								
1	84.7	20	3.3	30	87.6	10	2.9	32
2	84.2	20	4.0	20	85.1	10	3.9	40
3	83.8	30	4.5	30	83.3	10	4.2	40
4	83.4	20	4.6	30	87.6	10	2.9	32
5	82.3	20	4.7	30	83.3	10	4.2	40
n=50								
1	88.1	10	3.3	10	87.5	7	3.2	20
2	87.4	10	4.1	10	87.5	7	3.8	20
3	86.6	10	4.6	20	86.0	8	4.1	20
4	85.9	10	4.8	20	84.8	9	4.4	20
5	86.6	10	5.0	20	85.7	9	4.5	20
n=100								
1	88.4	10	3.2	10	88.1	5	3.2	11
2	87.9	10	4.1	10	88.1	6	4.1	11
3	87.1	10	4.5	10	88.0	7	4.6	13
4	86.7	10	4.7	10	87.4	7	4.9	13
5	87.5	10	5.0	10	87.2	8	5.0	15
Exponential Noise								
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
n=25								
1	86.4	20	3.0	40	87.6	10	2.9	32
2	82.4	30	3.7	40	85.1	10	3.9	40
3	80.7	30	4.1	40	83.3	10	4.2	40
4	80.5	30	4.3	40	87.6	10	2.9	32
5	78.7	30	4.3	40	83.3	10	4.2	40
n=50								
1	90.4	10	2.9	30	89.0	9	3.0	25
2	86.4	10	3.9	20	86.3	8	4.0	23
3	85.7	10	4.2	20	85.6	8	4.2	22
4	85.3	10	4.5	20	85.2	9	4.5	25
5	85.6	10	4.6	20	85.9	8	4.8	24
n=100								
1	90.3	10	2.9	20	89.0	7	2.9	19
2	87.2	10	4.0	20	87.4	6	3.9	18
3	86.7	10	4.3	20	86.2	7	4.2	17
4	86.4	10	4.6	20	86.3	8	4.5	18
5	86.2	10	4.7	20	86.3	8	4.8	20

	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
Normal Mixture Noise								
n=25								
1	87.2	20	10.5	40	87.5	10	10.9	40
2	84.7	20	14.3	40	84.9	14	14.3	40
3	83.0	20	14.7	40	83.6	14	14.6	40
4	81.5	20	15.1	40	80.6	17	14.7	40
5	81.6	20	15.4	40	80.9	19	15.4	50
n=50								
1	89.1	10	10.8	30	89.2	9	10.8	25
2	86.4	10	14.6	30	87.4	8	15.0	31
3	86.4	10	15.4	30	86.2	11	15.4	28
4	84.6	20	15.6	30	85.4	13	16.0	27
5	85.6	20	16.8	30	85.9	12	17.0	29
n=100								
1	89.6	7	10.8	25	89.0	7	10.6	25
2	86.6	8	15.0	27	87.1	7	14.9	28
3	86.9	7	15.8	20	87.5	7	15.8	20
4	87.0	7	16.4	18	87.0	7	16.4	15
5	87.3	7	16.8	19	87.2	8	16.8	19

Table 8.4.2.2: 90% Bootstrapped forward prediction estimates for the SETAR(2,1,1) Model

	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
Normal Noise								
n=100								
1	89.1	16	1.7	10	-	-	-	-
2	88.9	15	1.9	10	-	-	-	-
3	89.6	15	1.9	10	-	-	-	-
4	88.5	16	1.9	10	-	-	-	-
5	89.0	16	2.0	20	-	-	-	-
n=250								
1	89.1	10	1.7	10	89.4	5	1.7	9
2	89.1	5	1.9	10	89.3	5	1.9	11
3	88.6	10	1.9	10	88.2	6	2.0	11
4	87.6	10	1.9	10	88.9	5	2.0	11
5	88.7	10	2.0	10	89.8	5	2.0	11
n=500								
1	89.0	5	1.7	9	89.2	6	1.7	8
2	89.1	5	1.9	8	89.3	6	1.9	10
3	89.4	5	1.9	10	89.2	6	2.0	12
4	88.6	5	2.0	10	88.8	6	2.0	11
5	89.3	5	2.0	9	89.6	5	2.0	10
Exponential Noise								
	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
n=100								
1	82.5	10	1.5	37	-	-	-	-
2	83.6	10	1.5	40	-	-	-	-
3	84.0	10	1.5	40	-	-	-	-
4	84.0	10	1.6	41	-	-	-	-
5	84.6	10	1.6	42	-	-	-	-
n=250								
1	90.1	10	1.4	20	90.0	6	1.5	16
2	88.2	10	1.5	20	89.4	6	1.6	18
3	88.6	10	1.6	20	88.6	6	1.6	19
4	88.5	10	1.6	20	88.9	6	1.6	22
5	88.7	10	1.6	20	89.4	6	1.6	20
n=500								
1	89.3	10	1.5	20	88.9	7	1.4	16
2	89.4	30	1.6	10	90.0	5	1.6	16
3	89.3	20	1.6	10	89.6	5	1.6	14
4	88.8	10	1.5	10	90.2	5	1.6	15
5	88.6	10	1.5	10	89.9	5	1.6	16

	Sieve Bootstrap				Vectorised Block Bootstrap			
Steps Ahead k	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$	Mean β_k^*	% $S.E.(\beta_k^*)$	Mean $Len(k)$	% $S.E.(Len(k))$
Normal Mixture Noise								
n=100								
1	73.4	49	4.8	54	-	-	-	-
2	73.3	48	5.1	52	-	-	-	-
3	72.8	48	5.2	53	-	-	-	-
4	72.7	47	5.3	52	-	-	-	-
5	72.3	47	5.4	52	-	-	-	-
n=250								
1	86.4	19	5.3	26	-	-	-	-
2	85.8	19	5.7	22	-	-	-	-
3	65.1	18	5.7	21	-	-	-	-
4	84.7	18	5.8	23	-	-	-	-
5	84.5	19	5.9	21	-	-	-	-
n=500								
1	89.0	7	5.6	17	69.2	44	5.7	14
2	86.8	9	5.7	14	73.3	8	6.5	54
3	87.5	7	6.0	12	63.4	24	13.6	406
4	86.8	8	5.9	15	61.9	10	111.1	758
5	88.1	6	6.1	14	58.5	17	1577.1	816

Table 8.5.1.1: Bootstrapped forward prediction estimates for the selected Test Bed Series

No	Method	Statistic	1	2	3	4	5	6	7	8	9	10	11	12
1	Code: ASAM003		$\Delta[y_t = -0.01 - 0.44\, y_{t-1} + 0.48\, y_{t-12}]$											
			Correct Models: Sieve = 51, Block = 39											
	Sieve	Mean	75.9	29.1	24.9	50.1	32.4	26.8	45.6	33.0	24.0	51.6	27.9	30.7
		CV	0.0	12.0	12.4	15.6	16.4	15.3	15.3	12.0	13.0	13.3	15.5	14.5
		SIRQ	0.0	13.1	17.5	15.3	20.1	23.2	19.8	15.8	15.6	18.5	19.5	21.5
	Vector	Mean	75.9	26.4	19.3	75.9	17.8	30.6	49.9	38.7	16.4	70.0	16.8	38.5
		CV	0.0	12.3	11.1	14.5	11.6	11.6	10.5	11.9	12.1	12.0	11.5	11.0
		SIRQ	0.0	18.4	14.7	21.8	15.2	15.5	17.2	14.1	14.9	17.6	16.7	14.4
2	Code: ASAM206		$y_t = 1.08 + 0.18\, y_{t-1}(y_{t-2} - 1.2)_- - 1.2\, y_{t-1}(y_{t-2} - 1.2)_-(y_{t-12} - 2.1)_- + 0.5\, y_{t-1}(y_{t-2} - 1.2)_-(y_{t-12} - 0.8)_-$											
			Correct Models: Sieve = 27, Block = 21											
	Sieve	Mean	-3.2	-1.9	-0.7	-1.5	-1.5	-0.6	-0.7	-1.4	-0.8	-1.3	-0.7	-0.9
		CV	0.0	100.0	129.4	153.7	129.1	136.3	153.4	180.3	255.5	319.5	261.1	307.0
		SIRQ	0.0	111.1	177.4	282.9	123.6	141.1	207.4	195.6	255.1	236.3	261.3	272.7
	Vector	Mean	-3.2	-1.8	-1.3	-0.5	0.0	-0.2	-1.3	-2.9	-4.5	-6.7	-7.8	-11.5
		CV	0	96	205	125	110	129	221	531	1949	-4318	-2955	-1745
		SIRQ	0	94	183	135	69	90	92	114	91	116	110	81
6	Code: LRGM001		$\Delta\Delta_{12}[y_t = -4442 + 0.39 y_{t-1} + 1.25(y_{t-12} - 6772)_-]$											
			Correct Models: Sieve = 19, Block = 37											
	Sieve	Mean	-40.5	-33.5	-34.0	-35.0	-40.3	-35.5	-32.8	-26.6	-25.4	-33.1	-36.9	-37.5
		CV	0.0	1.5	1.2	1.1	1.9	1.6	1.2	1.5	1.4	1.2	1.6	1.8
		SIRQ	0.0	1.9	1.6	1.1	3.1	2.1	1.6	1.2	1.5	1.9	2.9	2.5
	Vector	Mean	-40.5	-33.8	-35.2	-35.1	-40.9	-34.5	-32.6	-26.2	-25.8	-32.9	-36.7	-35.8
		CV	0.0	1.6	2.0	2.1	2.0	1.7	1.8	1.6	1.5	1.8	2.1	1.8
		SIRQ	0.0	1.7	2.3	2.5	2.1	2.2	2.6	1.7	1.1	2.6	2.4	2.3
9	Code: LRGM438		$\Delta[y_t = -0.04 + 0.28 y_{t-12} - 0.18(y_{t-1} - 0.18)_- - 0.49(y_{t-1} - 0.18)_+]$											
			Correct Models: Sieve = 15, Block = 7											
	Sieve	Mean	-26.7	-28.6	-27.2	-27.8	-29.2	-29.0	-26.8	-24.8	-28.0	-30.3	-33.0	-37.8
		CV	0.0	3.7	4.2	5.2	4.1	4.6	5.5	3.0	3.6	4.0	6.3	7.5
		SIRQ	0.0	6.5	6.3	5.9	5.7	8.7	5.6	5.5	7.2	3.5	7.0	11.0
	Vector	Mean	-26.7	-23.7	-24.1	-28.6	-28.6	-27.2	-28.6	-23.6	-26.4	-29.4	-33.9	-37.7
		CV	0.0	2.8	4.8	3.7	5.4	3.9	3.8	4.1	3.5	3.6	4.5	4.6
		SIRQ	0.0	4.2	6.6	6.4	9.0	7.7	4.2	6.4	7.2	5.8	5.8	8.6
11	Code: MIAM014		$\Delta_{12}[y_t = -0.16 + 0.24 y_{t-1} + 0.26 y_{t-2} + 0.2(y_{t-13} - 0.14)_+ + 1.04 y_{t-1}(y_{t-12} - 0.52)_+]$											
			Correct Models: Sieve = 35, Block = 29											
	Sieve	Mean	53.0	28.4	44.7	38.0	41.3	51.5	35.5	51.0	63.3	61.5	66.1	70.5
		CV	0.0	11.2	15.5	11.7	11.9	20.5	13.9	10.1	10.1	10.3	15.2	9.8
		SIRQ	0.0	16.0	17.3	15.6	12.8	22.0	23.8	9.2	11.8	10.4	6.7	9.6
	Vector	Mean	53.0	24.9	39.4	36.9	40.7	56.0	35.3	57.6	65.2	62.3	64.2	69.4
		CV	0.0	6.3	8.0	10.3	11.4	11.9	10.7	11.1	11.2	10.6	8.1	9.2
		SIRQ	0.0	8.9	7.2	6.1	15.2	16.1	14.0	9.3	18.4	16.5	7.0	15.9

No	Method	Statistic	1	2	3	4	5	6	7	8	9	10	11	12
13	Code: MIAM524		$\Delta\left[y_t = 0.15 - 0.3y_{t-2} - 0.14y_{t-3} + 4.85(y_{t-1} - 0.06)_+ - 0.18(y_{t-1} - 0.06)_-\right]$											
			Correct Models: Sieve = 41, Block = 18											
	Sieve	Mean	-2.9	-14.5	-31.2	-18.5	-10.3	-11.4	-31.5	-51.9	-19.1	6.8	-16.7	-28.7
		CV	0.0	7.6	7.3	7.8	7.8	7.9	9.7	6.1	7.1	8.3	8.8	8.5
		SIRQ	0.0	7.7	10.1	8.8	9.7	10.4	10.7	10.1	7.7	10.6	11.3	9.9
	Vector	Mean	-2.9	-13.7	-33.2	-7.1	-15.0	-16.4	-32.8	-51.2	-22.3	7.0	-22.8	-27.9
		CV	0.0	7.9	10.2	17.3	5.0	8.3	13.3	4.1	5.4	8.6	6.9	10.7
		SIRQ	0.0	4.0	15.3	14.8	5.5	13.1	14.5	3.7	7.0	5.7	11.6	12.2
19	Code: TSAM055		$\Delta\left[y_t = 0.28 + 0.18y_{t-2} - 0.87(y_{t-1} - 1.28)_- + 0.2(y_{t-12} - 1.05)_+\right]$											
			Correct Models: Sieve = 8, Block = 9											
	Sieve	Mean	17.9	50.7	17.3	16.1	12.0	32.8	5.8	20.9	46.3	26.0	22.4	29.0
		CV	0.0	20.9	24.2	29.2	20.7	31.4	21.6	18.0	36.5	49.5	28.3	18.3
		SIRQ	0.0	30.8	42.1	24.4	40.5	58.9	24.2	25.3	48.1	25.7	31.3	31.1
	Vector	Mean	17.9	44.1	27.5	13.3	16.5	22.2	10.5	16.1	35.3	22.6	38.4	27.5
		CV	0.0	29.8	55.1	17.8	36.4	24.8	17.8	29.6	29.9	18.3	18.7	41.4
		SIRQ	0.0	50.6	45.2	19.6	36.9	41.4	12.4	40.5	38.1	24.5	29.4	38.3
20	Code: TSAM601		$\Delta_2\left[\begin{array}{l} y_t = -64.070 - 0.28y_{t-2} + 0.36y_{t-3} - 0.0001(y_{t-1} - 904{,}100)_+(y_{t-13} - 1{,}012{,}000)_- + \\ \quad 0.0001(y_{t-1} - 904{,}100)_+(y_{t-2} - 204)_-(y_{t-13} - 1{,}012{,}000)_- \end{array}\right]$											
			Correct Models: Sieve = 79, Block = 22											
	Sieve	Mean	63.4	61.2	55.6	61.2	57.2	51.5	54.8	47.8	54.7	51.6	50.4	52.2
		CV	0.0	1.7	2.8	2.8	3.0	2.7	3.5	3.9	3.4	3.7	3.1	3.4
		SIRQ	0.0	1.3	1.6	1.8	1.7	2.3	2.8	2.1	2.4	3.5	2.5	2.2
	Vector	Mean	63.4	61.6	55.9	60.3	55.6	52.3	55.3	48.0	54.4	52.0	72.2	24.5
		CV	0.0	2.0	2.6	2.2	2.9	2.7	3.0	4.9	2.8	3.6	28.1	69.0
		SIRQ	0.0	0.8	1.0	2.8	3.5	2.2	2.7	1.4	2.6	2.6	48.3	94.6

9 Contributions and Conclusions

9.1 Background

This thesis studied nonlinearity in time series. Our focus was on estimation and short term forecasting using TSMARS, a time series extension of the Multivariate Adaptive Regression Splines (MARS) procedure of Friedman (1991a). MARS was chosen because it is model free and can discern nonlinearity in terms of asymmetry in the predictors. The method also gives a precise measure of the degree of nonlinearity. This facility is particularly valuable in measuring the degree of nonlinearity in empirical time series published by the Central Statistics Office (CSO).

Several aspects arise in the study of time series, such as seasonality, outliers, and dependent errors. Each of these require extensions that are novel to TSMARS. These extensions constitute an important contribution of this thesis. A new version of TSMARS has been implemented in SAS/IML that incorpoates these extensions. This platform makes TSMARS much more accessible to researchers in many fields.

9.2 Contributions

In Chapter 1 nonlinear models relevant in this thesis were reviewed. First the statistical and dynamic aspects of a number of nonlinear time series models were set out. It was shown that even simple linear non-Gaussian process give rise to asymmetry. The properties of some well-known simple nonlinear models were also examined. Threshold models and their generalisations were also set out. These models tend to be suitable in situations where there is asymmetry and are also particularly suited to modelling with TSMARS. The concept of a frame was introduced; this is a graphical tool that enables visualisation of nonlinear SETAR models. In this chapter seasonal models both linear and nonlinear were also set out.

In Chapter 2 the MARS algorithm was described in detail. Based on this description, the first contribution of this research was the development of a new version of the MARS program written in SAS/IML. This program was assessed against some of the results given for Friedman's original version (1991a). We showed that the SAS/IML version gave statistically equivalent result to Friedman's original. The program was adapted for time series and the resulting TSMARS program benchmarked against identical studies reported in the literature.

Having developed the TSMARS program and tested it the emphasis then switched to empirical analysis. An in-depth analysis was conducted of TSMARS under various settings. First, the question of whether data should be transformed prior to modelling was addressed. Using Box-Cox transformations and integrated models, simulations were conducted that showed transforming the data prior to modelling could improve the precision of the estimates.

An analysis of the impact of seasonal adjustment on TSMARS was also conducted Chapter 3 using the concept of 'implied parameters'. Three different model types incorporating fixed seasonality, stochastic seasonality and regime dependent seasonality respectively were used to conduct the analysis. These models and the studies conducted are novel. The studies demonstrated that modelling the seasonal effects directly as part of the TSMARS model proved better than the alternative of prior seasonal adjustment. Moreover, for the regime dependent seasonal model, which was not trivial to assess, the

results indicated that seasonal adjustment did affect the nonlinear characteristics of the simulated series. This is an important finding.

In Chapter 4 an extensive analysis was then conducted on twenty empirical CSO time series to see how much nonlinearity could be observed. Four different and sophisticated TSMARS modelling variations were developed for this study. Novel aspects of these variations included seasonal adjustment prior to TSMARS modelling and parsimonious modelling with variables lagged at 1, 2, 3, 12 and 13 past periods. First, we found that it was important to difference the data to remove growth effects before modelling. Second, directly incorporating seasonal effects was preferable to methods that involved prior seasonal adjustment. The results showed that growth effects tended to mask other characteristics and when removed, as for example in the STSMARS method, the test results tended to show evidence of nonlinear behaviour. Third, ANOVA analysis showed that about 5% of the overall variance was explained by nonlinear effects, though one series (ASAM206) had some 47% of its variance nonlinear. Moreover, this finding was in agreement with the results reported in Chapter 4 for sophisticated linear modelling using SARIMA+ (see Appendix).

In Chapter 5 TSMARS was used for out of sample forecasting. Cross validation based forecast errors were computed. The distrbution of errors for models discussed in Chapter 2 was computed. The associated predictive interval was found to be in line with theoretical values. Cross-validation based plug-in forecasting was also applied to the empirical series. In this case the errors were found to be in line the values of the MAPE statistic observed in Chapter 4.

Chapter 6 implemented a novel outlier treatment methodology in TSMARS. This is called the Conditional Model Outlier Treatment (CMOT) procedure. We proved that this approach ensures that the model selection mechanism in TSMARS is consistent in the presence of outliers.

Three different adjustment procedures are also set out; namely, Least Squares (LSAO), Bounded Influence (BIF) and Time Series (TSAO). The LASO method is ideal when residuals are independent. The BIF method is suitable when the residuals are independent but may deviate from normality. The TSAO method is specific for autoregressive, threshold and additive model time series. We proved this method modelled the error process correctly in these cases. Simulation studies show that these treatment procedures make TSMARS consistent; that is, TSMARS is more likely to choose a correct model type in the presence of an outlier. Both the CMOT procedure and these three outlier treatment mechanisms are important contributions to TSMARS and to the subject of robust methods generally.

The outlier adjustment procedures are also run on the 'test-bed' of twenty economic time series and we found no evidence to alter the conclusions of Chapter 4 - that is, nonlinearity is only present to a small degree. Moreover, we find that the nature of the models found suggest that dependent errors may be more appropriate to model the twenty empirical CSO series.

In Chapter 7 we extend TSMARS to incorporate moving average (MA) components. This is a particularly significant development of the program as many economic time series are better modelled with MA rather than AR terms. This extension allows TSMARS to identify SETMA, ASTAR and ASTARMA models; both ASTMA and ASTARMA are novel model forms.

To gain efficiency we implement parsimonious MA estimation using conditional least squares (CLS) based on a Gauss-Newton procedure. We used two variations of the methodology; namely, Jacobi and

Seidel iteration schemes. There are a number of important innovations. The Jacobi iteration de-couples regimes and estimates each separately. This Jacobi iteration is only used for finding the threshold as the residual sum of squares (RSS) in this step is not too sensitive to the method. Final estimation uses the regime dependent Seidel iteration to ensure accurate estimates. In simulation studies this new methodology is shown to be statistically sound.

Attention then returned to the study of empirical CSO series using TSMARS with MA estimation. The results showed that no improvement over earlier estimates could be discerned. However, simpler more stable models were found demonstrating that MA components better explain the nature and extent of nonlinearity of CSO series. This is a significant finding.

In Chapter 8 predictive intervals for TSMARS models are computed. We use two novel variations of existing parametric and nonparametric bootstrap methods. These ensure that the intervals account for explicit dependence of forecast on the last p values of a p^{th} order model of a time series. We conduct simulation studies that show the predictive intervals obtained for simple linear models are close to those obtained elsewhere. Moreover, we apply our methods to simple threshold models driven by three different forms of noise. Once again forecast intervals are shown to be accurate and consistent.

These bootstrapping methods were also used to compute predictive intervals for some of the empirical CSO Series. The results showed the intervals are generally small for these data. Moreover, these intervals were in close agreement with cross validation intervals obtained in Chapter 5 when large fluctuations do not occur near the end of a time series.

9.3 Conclusions

The contributions outlined above are an important step in understanding nonlinearity in seasonal time series. Substantial gaps in understanding how these series should be modelled with TSMARS have been addressed. However, the analysis for all that is incomplete.

One key finding of the thesis is the evidence of periodicity in the form of PAR effects in the residuals. In a sense a PAR model is a kind of vector model approach. This is the main reason why these models have not been considered in more detail here in this thesis. However, without incorporating these effects more fully into the models used it cannot be said that the nonlinearity identified is present with absolute certainty. Further work however remains to be done on this aspect.

The statistical tests used throughout represent a useful subset of tests for seasonal data that are available in the statistical literature. However, many more test are available for different types of nonlinearity (see Pena 2000). While some of these were examined the evidence showed that the power of these was poor. This is the main reason the set chosen was fixed upon. However, an important weakness of the testing was the tests did not assess level of nonlinearity in the data, only that the effect was evident.

In Chapter 3 one seasonal series (called Model 2) had a spike in every second season that acted like an additive outlier. This recurring spike effect caused problems. This however may be handled by calling TSMARS twice and using weights on the second call to adjust for any excessive variance arising in the seasonal adjusted series; this approach was however not examined. What is less obvious is that the

bounded influence outlier methodology is implicitly implementing this by down-weighting large residuals in the IRWLS estimation. This method and the TSAO method are expensive, as the computing time required roughly doubles. However, computing times associated with the estimates were not reported as they still remain reasonable (i.e. within minutes).

In Chapter 7 the approach to identifying MA-type components is also computing intensive. Once again experience shows it more than doubles depending on the number of iterations allowed in the Gauss-Newton procedure. This means the method is best used where the number of observations is moderate and where standard TSMARS has already been used for preliminary analysis. Clearly the method could be reviewed to attempt to improve some aspects of its efficiency.

In Chapter 8 the predictive distribution was examined for TSMARS models. This study was quite limited but nevertheless informative. A problem with the method is that when the retained last p-values of the series are added back to the bootstrap replicate a discontinuity is introduced. While this was addressed it remains an issue for further study. For the block bootstrap the dependence of interval length on the chosen block length also need to be investigated. A third limitation of the approach is that seasonality has to be handled for the predictive distributions. This was not such a problem for the sieve method but for the block bootstrap it is. We avoided the problem somewhat by limiting the block length to the period of the data. This meant that the seasonality was not disturbed but of course any underlying nonlinearity may have been.

Finally, it should be mentioned that the TSMARS modelling was univariate only. Versions of vector TSMARS are available (Hastie 1996) and other related methods such as PolyMARS have been used for vector time series modelling (see De Goojer & Ray 2002). The SAS/IML version of MARS used in this thesis also includes a reliable multivariate modelling option. Research on modelling and forecasting using VASTAR models or VASTARMA models that incorporate seasonality remains to be done.

10 Appendix

10.1 Linear Time Series Models and SARIMA+

In this thesis linear time series models are used to provide benchmark estimates against which other methodologies are judged. The key value of linear models is their simplicity. In addition they are a first order approximation to nonlinear models. For short term forecasting therefore, they tend to give excellent results that are comparable to nonlinear models.

A linear model for a given a time series variable y_t, observed at a time point $t = 1 \ldots n$, is assumed to be driven by an unobserved error process ε_t called the innovation. In addition, it is pressumed that the innovation is white noise (WN), that is

$$\begin{aligned} E(\varepsilon_t) &= 0 && \forall t \\ E(\varepsilon_t^2) &= \sigma_\varepsilon^2 && \forall t \\ E(\varepsilon_s \varepsilon_t) &= 0 && \forall\, s, t \text{ and } s \neq t \end{aligned}$$

where $E(\cdot)$ is the expectation function. This definition specifies that the errors are uncorrelated. If it is further assumed that the errors are independent and identically distributed (i.i.d.) the error process is distinguished as strict WN. Therefore, normally distributed errors are strict WN random variables.

The empirical data used in this thesis are mainly seasonal time series. Benchmark linear estimates of these series are obtained from a Multiplicative Seasonal ARIMA model labelled SARIMA+. The key elements of the SAS/IML program to automatically estimate this model are now outlined with further methodological details available in Pena et. al. (2000) and Wei (1990).

The Multiplicative Seasonal ARIMA model denoted by SARIMA(p,d,q)X(P,D,Q)$_s$ of Box & Jenkins (1976) is implemented. This model captures the within period, so called regular, relationships through a nonseasonal ARIMA(p,d,q) model and the between period seasonal relationships with the ARIMA(P,D,Q)$_s$ model. In back shift notation (i.e. with $y_{t-1} = By_t$) the full SARIMA model is given by

$$\Phi_P(B)(1-B)^D \phi_p(B)(1-B)^d y_t = \Theta_Q(B)\, a_t \qquad (10.1.1)$$

where a_t is residual strict WN. The p, P order regular and seasonal autoregressive polynomials are denoted by ϕ_p and Φ_P respectively. The corresponding degrees of differencing are d (usually 0, 1 or 2) and D (usually 0 or 1) respectively while the q, Q order regular and seasonal moving average polynomials are denoted by θ_q and Θ_Q respectively.

Algorithms that allow for automatic model selection and outlier treatment of a given time series modelled with (10.1.1) have also been implemented. Automatic model selection is done using the Model Identification Algorithm given in section 7.3 of Pena et. al. (2000). This algorithm picks the 'best' model using the smallest BIC (Schwarz's Bayesian Information Criterion).

Model outliers are identified based on standard intervention analysis for ARIMA models, see Wei, (1990) Chapter 9. This device handles additive outliers (i.e. shocks), level shifts, transitional shifts and innovation type outliers for the SARIMA model for y_t as well as masking via multiple regression. The methodology

that comprises automatic SARIMA model selection followed by outlier treatment is referred to here as SARIMA+.

10.2 Stationarity, Mixing and Invariants

The NLAR(p) model (1.3.8) of order p from Chapter 1, repeated here, has the general form

$$y_t = f(y_{t-1},\ldots,y_{t-p})+\varepsilon_t \tag{10.2.1}$$

and quite naturally its study often begins with an analysis of its stochastic stability properties; that is, whether the model is stationary. A second related area of study is the dependence structure of the model; for time series models this is generally known as mixing.

To establish whether a time series defined by a linear model is stationary is not easy even for simple nonlinear models. One method (see Tong 1990) is to apply stochastic Lyapounov equations to establish stationarity. This method borrows directly from the Lyapounov 'energy' equation approach of dynamical systems. Finding a suitable energy function is not trivial and this is one of the reasons why the approach has been largely superseded by the ergodic method in the statistical literature.

For (10.2.1) the ergodic method involves expressing the time series model as a Markov chain over $\Re^p$ and then establishing that the chain is ergodic. That is, for $\mathbf{y}_t = (y_t, y_{t-1},\ldots,y_{t-p+1})^T$ and $\boldsymbol{\varepsilon}_t = (\varepsilon_t,0,\ldots,0)^T$ in $\Re^p$ there exists a stationary distribution (for simplicity a density) $h(\mathbf{y}_t)$ and a given innovation transition probability density $g(\boldsymbol{\varepsilon}_t) = g(\mathbf{y}_t|\mathbf{y}_{t-1})$ such that

$$h(\mathbf{y}_t) = \int h(\mathbf{x})g(\mathbf{y}_t - f(\mathbf{x}))\,d\mathbf{x} \tag{10.2.2}$$

Starting off the chain with $h(\mathbf{x})=1$ the stationary distribution can be found by iteratively solving this homogeneous integral equation. This clearly is an almost impossible task but in general stationarity can be established if the total variation $\rho^{-n}\|h_n(\mathbf{y}_t|\mathbf{x})-h(\mathbf{y}_t)\|\to 0$ for any $\mathbf{x}$. In this case the sequence $\{\mathbf{y}_t\}$ is ergodic if $\rho=1$ and geometrically ergodic if $\rho<1$ (see Fan & Yao 2003).

Closely linked to the concept of ergodic sequences and stationarity for a time series model is the notion of *mixing*. As alluded to above mixing measures the degree of dependence between different parts of the time series. Specifically, a mixing time series has the property that the past and distant future are asymptotically independent. For mixing sequences both the law of large numbers (i.e. ergodic theorem) and a central limit theorem can be established (see Fan & Yao 2003).

For strictly stationary processes (i.e. those driven by i.e. innovations) the idea is to define mixing coefficients to measure the strength of dependence for two segments of a time series that are apart from each other in time. The most commonly used mixing coefficients are the so-called strong mixing or α-mixing sequence of coefficients. To define this sequence let $\{y_t, t=0,\pm1,\pm2,\ldots\}$ be a strictly stationary time series. For n = 1, 2, ... define the sequence

$$\alpha(n) = \sup|P(A)P(B)-P(AB)| \tag{10.2.3}$$

where A is the σ-algebra (i.e. set of events over the basic sample space) generated by the sequence $\{y_t, i \le t \le j\}$ and B is the is the σ-algebra generated by the sequence $\{y_t, n \le t \le \infty\}$. Intuitively, α(n) measures the frequency of values shared by the sequence in A that are also in B. A time series then is defined to be strongly mixing if the sequence α(n) tends to 0 with n going to infinity. Strong mixing is in fact the weakest form of mixing, essentially based on frequency counts. Other forms of mixing based on different norms can also be defined (see Fan & Yao 2003).

The notion of mixing sequences and dependence implies that some form of autocorrelation structure may persist in a time series; that is, there may be an invariant sequence (or more generally an invariant set) that is recurrent. In deterministic nonlinear systems the invariant set lies on a so-called attractor (e.g. limit cycle) that is often characterised by three (or variations thereof) interrelated invariants. These are Lyapounov exponents, correlation dimension and entropy. A key ingredient in the study of the general dynamical system

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) \tag{10.2.4}$$

is that the system is dissipative. This guarantees that the solutions remain bounded or equivalently that the trajectories in phase space converge to an 'attracting set' known as an attractor. When the attractor is not a stable equilibrium or saddle point, limit cycle or torus it is called as a 'strange attractor'; a term coined by Ruelle & Takens (1971). A well-known example of a simple dynamical system that exhibits a strange attractor is the Henon map (see Tong 1990)

$$\begin{aligned} x_{t+1} &= 1 - a\,x_t^2 + y_t \\ y_{t+1} &= b\,x_t \end{aligned} \quad \Leftrightarrow \quad x_{t+1} = 1 - a\,x_t^2 + b\,x_{t-1} \tag{10.2.5}$$

What is important about this system is that even though the solutions (i.e. realisations) are bounded, any two solutions that initially start off close soon diverge. That is, the solutions will end up on the attractor but will not be close as time evolves. As a consequence the study of nonlinear dynamical systems has not emphasised the properties of solutions but rather invariant measures that can be used to characterise the attractor. The most popular of these measures among time series analysts is the correlation dimension and correlation entropy as they can be estimated relatively easily from the correlation integral.

The correlation integral is usually estimated using the Grassberger-Procaccia Method (see Diks 1999 among other places). For time delay τ and lag vectors $\mathbf{y}_t^m = (y_t, y_{t-\tau}, \ldots y_{t-(m+1)\tau})$ of length m (known as m-histories or reconstruction vectors), the correlation integral is computed according to

$$C_m(\varepsilon) = \frac{2}{\binom{n-T+1}{2}} \sum_{k=1}^{n-1} \sum_{t=k+1}^{n} H\left(\varepsilon - \left\|\mathbf{y}_k^m - \mathbf{y}_t^m\right\|_\infty\right) \tag{10.2.6}$$

where H(•) is the Heaviside function and denominator incorporates the Theiler correction (see Diks 1999) for dependence among lagged vectors that are close in time. Roughly speaking, the correlation integral counts the frequency of repeated lagged patterns in the data as a function of a closeness parameter ε. For deterministic time series, the correlation integrals for small ε and large m behave according to the scaling relation

$$C_m(\varepsilon) \sim e^{-m\tau K_2} \varepsilon^{D_2} \tag{10.2.7}$$

where D_2 is the correlation dimension and K_2 is the correlation entropy per unit time (or correlation entropy for short). The correlation dimension is the basis of the so-called BDS nonlinearity test (see Brock et. al. 1996) given in next section of the Appendix. It can be interpreted as the dimension of the attractor and need not be an integer but is bounded by m. The correlation entropy is a measure of the rate of divergence of solution that initially start off close together. Both of these measures are independent of the initial conditions of the dynamic equation (10.2.4).

Invariant measures are powerful tools for understanding deterministic nonlinear systems and are used in the presence of stochastic noise. This approach to time series analysis is popular in engineering and adopted for statistics by, among others, Krantz & Schreiber (2004). However, in the presence of noise computing the invariants can be unreliable. The problem stems from the fact that even if the underlying nonlinear function is dissipative there is no reason why the same function (called a skeleton by Tong 1990) with even additive noise, as in (10.2.1) should be stationary. In fact whether or not the stochastic time series model remains bounded depends on the 'admissible size' of the noise and whether this is swamped by the geometry of the attractor (see Chen & Tong 1994). For example, they show that the EXPAR(1) model

$$y_t = \phi_1 y_{t-1} + \phi_2 e^{-y_{t-1}^2} y_{t-1} + \varepsilon_t \tag{10.2.8}$$

is ergodic if $|\phi_1| < 1$ and the half-width of the support of ε_t is larger than $\max\left\{\phi_2 e^{-y_{t-1}^2} y_{t-1}\right\}$ Thus distinguishing the 'size' of the noise from that of attractor is critical. It explains why the study of invariant sets tends to be secondary to understanding the stochastic evolution based on ergodic and mixing properties.

10.3 Relevant Statistical Tests

Testing of seasonal time series data can be broadly divided into tests for independent effects (e.g. trading day patterns) seasonal effects and other nonseasonal or regular tests of the residuals arising from a fitted model. The regular tests generally comprise tests of randomness and tests for specific characteristics such as some form on nonlinearity. A set of regular and seasonal statistical tests arising in the literature is now outlined. These tests will be used throughout this thesis to characterise the lack of fit of residuals from a fitted model at the 1% level.

There are some well-known omissions to the test list. These include the Bi-spectrum test of Shubbo-Rao (see Tong 1990) and as generalisations of Tsay's F-test described in Pena (2000). The reason for these omissions is that the power of the Bi-spectrum test in applications is poor (see Pena 2000 for example). The other tests are omitted because they specify a particular nonlinear model, such as the bilinear model. These tests are somewhat less relevant as evidence, for example, of a y_t^2 component will show up as a v-shaped threshold at the origin. Tsay's F-test, which has power, is therefore likely to pick up this effect and indicate nonlinearity.

10.3.1 Regular statistical test

Box-Ljung χ_1^2 test: This is a test of randomness for a seasonal time series model with period s and having n residuals. It is computed according to the following formula

$$\chi_1^2 = n(n+2)\sum_{k=1}^{2s}\frac{r_k^2}{n-k} \tag{10.3.1}$$

where r_k is the autocorrelation coefficient (*c.f.* equation (1.2.4)) at lag k. This test value is computed at lags from 2 through 24 and if statistically significant at 90% or more of the 23 lag values the hypothesis of randomness is rejected.

Box-Ljung χ_2^2 test: This is a test of randomness that is virtually identical to the χ_1^2 except that the squares of the residuals and used in formula (7.2.1). This test is known to be useful for indicating the presence of hetroscedasticity (i.e. possible ARCH effects) in the residual series (see Harvey 1993).

Smoothed coefficient of variation $t_{\mu,\sigma}$: Based on a s+1 point centred moving average the sliding t-value is computed to check if the mean of the residual changes with its variance. This may indicate a regime change in the innovation sequence.

Tsay's F-test (Tsay 1989): This is an ordered autoregression test for threshold nonlinearity. Generalisations of the test for other types of nonlinearity are given in Pena (2000). The original test is adopted as it has been found to be robust for the data arising in this thesis.

Basically, the test checks for the existence of threshold autoregression by examining the standardised predictive residuals that arise through recursive regression. These predictive residuals are ordered according to the values of the threshold variable. By doing so a threshold model is transformed into a linear model. If there is a regime shift at some point the standardised residuals from the ordered autoregression will deviate from normality. Plots of the standardised residuals against the series at specific lags may then show evidence of a threshold. The method can be implemented using a Kalman Filter but the approach adopted here follows Tsay's original recursive regression implementation. In this implementation of the test, the threshold lag order is tested from 1 through 24 and if significant for 21 of the 24 lags tested, then evidence for threshold lag autoregression is accepted. Note however, that experience using this test has demonstrated that it is particularly susceptible in the presence of dominant frequencies in the spectral density of the residual series. In practice that is, the presence of a cycle in the residual series is often misinterpreted as a threshold and so evidence of a threshold is only accepted when the spectrum is flat.

It is worth noting that this notion of ordering the residuals according to the threshold value is essentially the mechanism underlying TSMARS, the subject of much of the remainder of this thesis.

BDS test (Brock et. al 1996): This is the nonlinear analogue of the Box-Ljung χ_1^2 autocorrelation test on the residuals. It is based on the correlation integral (1.2.13) with the closeness parameter ε chosen at 1 standard deviation of the data, see Pena (2000)). The test statistic is based on the quantity $\sqrt{n}\left[C_{m,n}(\varepsilon) - C(\varepsilon)^m\right]$ where $\lim_{n\to\infty} C_{m,n}(\varepsilon) = C(\varepsilon)^m$; the statistic is asymptotically normal (see Brock et. al.

1996). In the implementation of this test in this thesis lag lengths of m = 1 to 10 are used; evidence of nonlinearity is accepted if the BDS statistic is significant for at least half of these m-histories.

10.3.2 Tests for seasonality

The seasonality tests described here rely on the seasonal dummies models outlined in section 1.4. Here once again the test is applied to a residual series where it is assumed that the alternative for the test is the mean.

Seasonality F-test: This test uses the seasonal dummies directly in regression model (1.4.2) giving the model

$$e_t = \delta_1 D_{1,t} + \delta_2 D_{2,t} + \delta_3 D_{3,t} + \delta_4 D_{4,t} + a_t \qquad (10.3.2)$$

(a_t is white noise). The associated F-statistic based on the corrected regression sum of squares is used to test whether all the seasonal means $\delta_i = 0$.

Periodic Autoregression of order 1 (PAR(1)) test (Franses 1996): This test is a direct application the PAR model equation (1.4.3) to a residual series using lag order 1 AR polynomials in each season. In this case the regression model is

$$e_t = \delta_1 B_s D_{1,t} + \delta_2 B_s D_{2,t} + \delta_3 B_s D_{3,t} + \delta_4 B_s D_{4,t} + a_t \qquad (10.3.3)$$

where B_s is the appropriate seasonal lag operator. Once again the F-test is used to test whether all the seasonal regression coefficient $\delta_i = 0$.

Periodic Autoregression of order 2 (PAR(2)) test: This test is identical to the above test except that order 2 AR polynomials are used in each season.

Periodic Variation (VPAR) test (Franses 1996): This test checks for periodic hetroscedasticity in the residuals. The test is identical to Seasonality F-Test based on equation (10.3.2) except that the residuals e_t are replaced by their squares giving the regression model

$$e_t^2 = \delta_1 D_{1,t} + \delta_2 D_{2,t} + \delta_3 D_{3,t} + \delta_4 D_{4,t} + a_t \qquad (10.3.4)$$

10.3.3 Independent predictor tests

Tests are carried out for three independent predictor effects, namely length of month (MD), trading day (TD) or more precisely trading week length, and Easter effects. The length of month effect is computed by subtracting 30.4375 (i.e. the average number of days in any month) from the number of days in that month while the trading week length effect = number of work week days – 5 X number of weekend days/2 (see Pena 2000). The Easter effect is computed according to the *Corrected Immediate Impact* rule given in Ladiray & Quenneville (2001). These 3 predictors are regressed against the residual and the likelihood-ratio statistics computed for each parameter in the multiple regression model. The level for these test is taken as 1%.

10.3.4 Error measures

Two error measures are adopted and these are used throughout the thesis.

Residual Sum of Squares (RSS): This, of course is the usual uncorrected formula based on the errors $e_t = y_t - \hat{y}_t$, where $\hat{y}_t$ is the estimated value;

$$\text{RSS} = \frac{1}{n}\sum_{t=1}^{n} e_t^2 \tag{10.3.5}$$

Mean Absolute Percent Error (MAPE): This measure is often more meaningful for assessing forecasting performance in that a portion of the end of the series is used to compute it. Specifically, the formula covering the last 3 years of data is

$$\text{MAPE} = \frac{100}{n}\sum_{t=0}^{3s-1} \frac{|e_{n-t}|}{|y_t|} \tag{10.3.6}$$

11 Bibliography

Bellman R. E. (1961), Adaptive Control Processes, Princeton Univ. Press.

Box G. E. P. and Jenkins G. M. (1976), Time Series Analysis forecastin and Control, 2nd ed., Holden-Day, San Franscisco.

Brieman L., Friedman J. H., Olshen R. and Stone C. J. (1984), Classification and Regression Trees (CART), Wadsworth, Belmont, CA.

Brock W. A., Dechert W. D., Scheinkman J. A. and LeBaron B. (1996), A Test for Independence Based on the correlation Dimension, Econometric Reviews, 15.

Brockwell P. J. and Davies A. D. (1991) Time Series: Theory and Methods 2nd Edition, Springer, NY.

Buhlmann P. (2002), Bootstraps for Time Series, Statistical Science, Vol 17, No 1, p52-72.

Buja A., Hastie T. and Tibshirani R. (1989), Linear smoothers and additive models (with discussion), Ann. Stat., 17, 453-555.

Chen R. and Tsay R. S. (1993a), Nonlinear Additive ARX Models, JASA, 88, 955-967.

Chen R. and Tsay R. S. (1993b), Functional-Coefficeint Autoregressive Models, JASA, 88, 298-308.

Cleveland W. S. (1979), Robust locally-weighted regression and smoothing scatterplots, JASA, 74, 829-836.

Cleveland R. B., Cleveland W. S., Mc Rae J. E. and Tarpening I. (1990), STL: A seasonal trend decomposition procedure based on Loess, JOS 3-73.

Craven P. and Wabha G. (1979), Smoothing noisy data with spline functions, Numer. Math, 31, 377-403.

de Boor C. (1978), A Practical Guide to Splines, Springer-Verlag.

de Bruin P., (2002) Essays on Modelling Nonlinear Time Series, Ph. D. Thesis, No 291 Tinbergen Institute Research Series, Amsterdam.

Dechert W. D. (1996), Testing time series for nonlinearities: The BDS approoach, in Nonlinear Dynamics and Economics, Proceedings of the 10th International Symposium in Economic theory and Econometrics, Barnett W. A. ed, Cambridge Univ Press.

de Goojer J. A., Ray B. K. and Krager H. (1998) Forecasting exchange rates using TSMARS, J. of Int. Money and Finance, 17, 513-534.

de Goojer J. A. and Ray B. K. (2002) Modelling Vector Nonlinear Time Series using PolyMARS, unpublished.

Diks C. (1999), Nonlinear Time Series Analysis, Methods and Applications (Nonlinear Time Series and Chaos Vol. 4 – ed H. Tong), World Scientific, London.

Efron B. and Tibshirani R. J. (1998), An Introduction to the Bootstrap, Chapman & Hall, London, UK.

Fan J. (2000), Prospects in Non-parametric Modelling, JASA, 95, 1296-1300.

Fan J. and Yao Q. (2003), Nonlinear Time Series Nonparametric and Parametric Methods, Springer, NY.

Findlay D. F., Monsell B. C., Bell W. R., Otto M. C. and Chen B. (1996), X12regARIMA Seasonal Adjustment Program, Census Bureau, Washington D.C..

Franses P. H. (1996), Periodicity and Stochastic Trends in Economic Time Series, Oxford University Press, New York.

Friedman J. H.(1991a), Multivariate Adaptive Regression Splines (with discussion), Ann. Stat. 19, 1-141.

Friedman J. H.(1991b), Estimating functions of mixed ordinal and categorical variables using adaptive splines, Dept. of Statistics, Stanford Univ, Tech. Report LCS 108.

Friedman J. H.(1991c), Adaptive spline networks. In advances in Neural Information Processing Systems, 3, Morgan Kaufmann, San Mateo, CA.

Friedman J. H.(1993), Fast MARS, Dept. of Statistics, Stanford Univ, Tech. Report LCS 110.

Friedman J. H., Hastie T. and Tibshirani R. (2000), Additive Logistic Regression: A statistical view of Boosting (special invited paper), Ann. Stat., 28, 337-407.

Friedman J. H. and Silverman B. W. (1989), Flexible parsimonious smoothing and additive modelling (with discussion), Technometrics, 31, 3-39.

Golub G. H. & Van Loan C. F., (1996), Matrix Computations 3rd edition, John Hopkins University Press, Baltimore, Maryland, USA.

Gomez V. and Maravall A. (1997), TRAMO SEATS Trend Estimation & Seasonal Program, Ministerio de Economia y Hacienda, Madrid, Eurostat Luxembourg.

GSS Methodology Series no. 2 (1996), Report of the Task Force on Seasonal Adjustment, London.

Hall P. (1992), The Bootstrap and Edgeworth Expansion, Springer-Verlag, London, UK.

Hansen B. E., (1999), The Grid bootstrap and the Autoregressive Model, The Review of Economics and Statistics, 81(4), 594-607.

Hardle W., Lutkepohl H and Chen R (1997), A review of non-parametric time series analysis, Int. Stat. Rev., 65, 49-72.

Hart J. D. (1996), Some automated methods of smoothing time dependent data, J. Nonparametric Statistics, 6, 115-142.

Harvey A. C. (1993), Time Series Models, MIT Press.

Hastie T., Tibshirani R. and Friedman J. H. (2001), The Elements of Statistical Learning, Springer.

Hastie T., and Tibshirani R. (1990), Generalised Additive Models, Chapman & Hall.

Hastie (1996), MARS program in the R language, http://cran.r-project.org/

Henderson R. (1916), Notes on Graduation by Adjusted Average, Transactions (Actuarial Society of America), 17.

Hyllberg S. (eds.), (1992), Modelling Seasonality, Oxford Univ. Press.

Ihaka R. and Gentleman R. (1996), R: a language for data analysis and graphics. J. Comp. and Graphical Stats., 5, 299-314.

Jay Breidt F., Davis R. A. and Dunsmuir W. T. M. (1995), Improved Bootstrap Prediction Intervals for Autoregressions, Journal of Time Series Analysis, Vol 16, No 2, 177-200.

Kim, J. H., (2002) Bootstrap Prediction Intervals for Autoregressive Models of Unknown or Infinite Lag Order, J Forecast, 21, 265-280.

Kim, C. & Nelson C. R. (1999) State-space Models with Markov Switching, MIT Press, USA.

Krantz H. and Schreiber T. (2004), Nonlinear Time Series Analysis, Cambridge Univ. Press, UK.

Ladiray D. and Quenneville B (2001), Seasonal Adjustment with the X-11 Method, Springer.

Lai T. L. and Wong S. P. (2001), Stochastic Neural Networks with applications to nonlinear times series, JASA, 96, 968-981.

Lewis P. A. W. and Stevens J. G. (1991), Nonlinear modelling of time series using Multivariate Adaptive Regression Splines (MARS), JASA, 86, 864-877.

Lewis P. A. W. & Ray B. K. (1997), Modelling Long-Range Dependence, Nonlinearity and Periodic Phenomena in Sea surface Temperatures using TSMARS, JASA Vol 92 No 439, p881-893.

Lewis P. A. W. & Ray B. K. (2002), Nonlinear Modelling of Periodic Threshold Autoregressions using TSMARS, J. of Time Series Analysis, Vol. 23, No 4.

Li W. K. and Lam K (1995) Modelling asymmetry in stock returns by a threshold ARCH model, The Statistician, 44, 333-341.

Melard G. and Roy R, (1988) Modeles de series chronologique avec seuils, Reveues de Statistiques Appliques, 36, 5 –24.

Morgan J. N. and Sonquist J. A. (1963), Problems in the analysis of survey data , and a proposal, JASA, 58, 415-434.

Meyers R. H. (1989), Classical and Modern Regression with Applications (2nd Edition), Duxbury Press.

Pearsons W. M. (1919), Indices of Business Conditions, Review of Economic Statistics, 1, 5-107.

Pena D. Tiao G. C. and Tsay R. S. (2000), A course in Time Series Analysis, Wiley.

Planas C. (1997), Applied Time Series Analysis, Modelling, forecasting, unobserved components analysis and the Wiener-Kolmogorov filter, Eurostat, Luxembourg.

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P. (2002), Numerical Recipes in C, 2nd edition, Cambridge Univ. Press, UK.

Rosenblatt M. (2000), Gaussian and Non-Gaussian Linear Time Series and Random Fields, Springer Series in Statistics, Springer-Verlag, New York, USA.

Ruelle D. and Takens F. (1971), On the Nature of Turbulence, Comm. Math. Phys., 20, 167-192.

Romano R. P. and Thombs L.A., (1996) Inference for Autocorrelations Under Weak Assumptions, JASA Vol 91 No 434, 590-600.

SAS Institute Inc., (1993), SAS/ETS User's Guide, Version 6, Cary, NC, USA

SAS Institute Inc., (1999), SAS/ETS User's Guide, Version 8, Cary, NC, USA

Shiskin J., Young A. H., and Musgrave J. C. (1967). The X-11 variant of the Census method II seasonal adjustment program, Technical Report 15, Bureau of the Census, U.S. Dept. of Commerce, Washing D.C., U.S.

Stenseth N.C. , Falck W., and Chan K.S. (1999) From ecological patterns to ecological processes: Phase and density dependencies in the Candaian Lynx cycle. Procedings of the National Acadamy of Sciences USA , 95, 15430-15435.

Stone C., Hansen M., Kooperberg C. and Truong Y. (1997), Polynomial splines and their tensor products (with discussion), Ann. Stat., 25, 1371-1470.

Thombs L.A. and Schucany W. R., (1990) Bootstrap Prediction Intervals for Autoresression, JASA Vol 85 No 410, 486-492.

Tiao G.C. and Tong H. (1994), Someadvances in nonlinear adaptive modelling in time series. J. of Forecasting, 13, 109 – 131.

Tong H. (1990), Non-linear Time Series – A Dynamical Systems Approach, Oxford Science.

Tsay R. S. (1988), Outliers, level shifts and variance changes in time series, J. of Forecasting, 7, 1-20.

Tsay R. S. (1989), Testing and Modeling Threshold Autoregressive Processes, JASA 84.
Tsay R. S. (2000), Time Series and Forecasting: Brief History and Future Research, JASA 95, 638-643.

Wei W. S. (1990), Time Series Analysis - Univariate and Multivariate Methods, Addison & Wesley, UK.