Pacific-Rim Real Estate Society Conference

University of Technology Sydney (UTS)

January 24 — 27, 2000

Modelling Private New Housing Starts In Australia

John Flaherty
RMIT University
School of Marketing (Property Group)
Email: john.flaherty@rmit.edu.au

Ric Lombardo
RMIT University
School of Marketing (Property Group)
Email: ric.lombardo@rmit.edu.au

Abstract

Kew words: Housing starts, completions, economic aggregates residential property cycles, causal relationships, time series models.

A number of causal and non - causal approaches to forecasting private new housing starts in Australia are considered in this paper. Simple time series models are developed in Excel and more complex ARIMA models are estimated using the popular econometric software SHAZAM. Combining models is also discussed.

Interrelationships between housing starts and the state of the economy is briefly examined.

Introduction

This paper attempts to model quarterly private new housing starts in Australia using AUSSTATS historical data[1] for a sample period that spans the quarters 1970(1) to 1998(2) inclusive - a total of 114 observations. An *out of sample* period that spans the 4 quarters :1998(3) to 1999(2) is also used to gauge the relative forecasting performance of various alternative models. Note that even though historical data on private new housing starts were available going back to 1955(3), it was decided to ignore this data on the grounds that the behaviour of this variable was quite different prior to 1970. See Chart 1 below for some visual confirmation of this assertion.

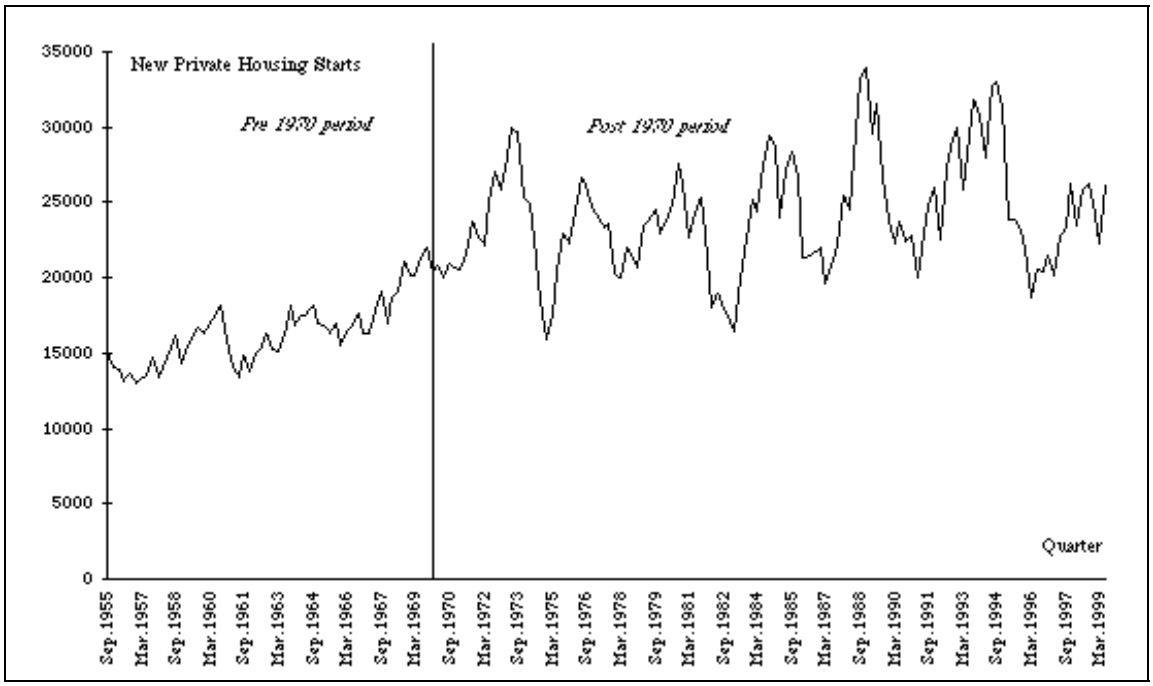

Chart 1 : The Behaviour of Private New Housing Starts - Pre and Post 1970

The models considered in this paper are primarily non-causal models. These non-causal forecasting models, range from the fairly naive to the more complex. A complete listing of the estimated models is as follows:

- The single exponential smoothing model (**sxsm**)
- Brown's double exponential smoothing model (**dxsm**)
- Holt's exponential smoothing model (**hxsm**)
- Winter's exponential smoothing model (**wxsm**)
- The classical decomposition of time series model (**cdtsm**)
- The linear multiple regression trend model with seasonal dummy variables (**sdtm**)
- The trigonometric seasonal forecasting model (**tsfm**)
- ARIMA

In addition a number of causal relationships are explored to investigate more general economic relationship

The remaining part of this paper is divided into 4 sections. Section 1 is devoted to a discussion of non-causal forecasting models of the non-ARIMA variety. Section 2 is concerned with the estimation of an appropriate ARIMA model. Combining models is considered next in Section 3 and section 4 discusses causal relationships. Finally, Section 5 provides a conclusion to the work undertaken in this paper and offers suggestions for future research in this area.

Section 1 Non Causal Models other than Arima

All the models considered in this section were estimated within the Excel spreadsheet environment. The various parameters of the **sxsm**, **dxsm**, **wxsm** and **hxsm** models were obtained with the use of the *Solver* tool available in Excel[2]. Use was also made of Excel's *Regression Analysis* tool to arrive at estimates of the cdtsm, sdtm and tsfm models[3]. Finally, the Shazam software program was used to test for the existence of autocorrelation in the residuals of all the models mentioned in this paragraph.

In the discussion of all models considered in this paper, the following notation will be employed.

Y_t denotes private new housing starts at quarter t

$\hat{Y}_t$ denotes predicted or forecasted private new housing starts at quarter t

Single Exponential Smoothing model (sxsm)

The single exponential smoothing model is normally used for situations where the data are not subject to a trend. The general form of the **sxsm** model is given by :

$$\hat{Y}_{t+1} = \alpha Y_t + (1 - \alpha)\hat{Y}_t \text{ or } \hat{Y}_{t+1} = \hat{Y}_t + \alpha(Y_t - \hat{Y}_t) \qquad (1.0)$$

where $0 \leq \alpha \leq 1$ is a smoothing constant.

Excel's Solver tool was used to find the α value that minimised mean square error (MSE) for expression (1.0) over the sample period. This value was found to be $\alpha = 1$, so that the eventual estimated model took the form :

$$\hat{Y}_{t+1} = Y_t \qquad (1.1)$$

Brown's Double Exponential Smoothing model (dxsm)

The double exponential smoothing model is used to forecast time series data that enjoy a linear trend. The dxsm forecasting model may be written as:

$$\hat{Y}_{t+\tau} = a_t + b_t\tau \qquad (2.0)$$

where the analyst wishes to make a forecast τ periods into the future based on the most current estimate of the intercept a_t and slope coefficient b_t at time t. Both a_t and b_t are functions of S_t - an exponentially smoothed value of Y_t at time t as well

as SS_t - an exponentially smoothed value of S_t at time t.[4] In estimating the above model, initial estimates of the intercept and slope - a_o and b_o - were arrived at by using a standard least squares procedure applied to the sample data set as recommended by Hanke and Reitsch (1998, p161). In turn, (a_o, b_o) are used to generate the initial values for S_t and SS_t at t = 0. Finally, the smoothing constant $0 \leq \alpha \leq 1$ that is employed for both the single and double exponentially smoothed Y-values is chosen in such a way as to minimise MSE over the sample period.

The optimal value of the smoothing constant turned out to be $\alpha = .54$. By the final observation of the sample period the dxsm forecasting model was given by :

$$\hat{Y}_{t+\tau} = 25561.03 + 695.70\tau \qquad t = 114 \tag{2.1}$$

Holt's Exponential Smoothing Model (hxsm)

Like the **dxsm** model (discussed above), the **hxsm** model is used to forecast time series data characterised by a linear trend. Whilst the eventual **hxsm** forecasting model has much the same outward appearance as the one set out in expression (2.0), the actual mechanism by which the (a_t b_t) pairs are adjusted at each time period is quite different from that of the **dxsm** model [5].

By the final observation of the sample period, the **hxsm** forecasting model was given by:

$$\hat{Y}_{t+\tau} = 25838.0 + 250.6\tau \qquad t = 114 \tag{3.0}$$

As with the **dxsm** model, the **hxsm** model is sensitive to the starting values given to the intercept and slope values. The approach taken here is one recommended by Hanke and Reitsch (1997, p.163 - 164). In particular, they suggest that a_1 be taken as the average of a few past observations and that b_1 may be estimated by using the slope of the trend equation obtained from past data [6]. For this study, the data for the 8 quarters preceding 1970(1) were used to estimate (a_1, b_1).

Winter's Exponential Smoothing Model (wxsm)

Winter's exponential smoothing model is specifically designed to forecast time series that are subject to a linear trend as well as multiplicative seasonal influences. The **wxsm** forecasting model may be written as :

$$\hat{Y}_{t+\tau} = \{a_t + b_t\tau\}S_{t-L+\tau} \tag{4.0}$$

where the parenthesised term in expression (4.0) has an analogous interpretation to the right hand side of expression (2.0); namely it represents a straight line trend forecast. The remaining term - $S_{t-L+\tau}$ - represents the latest available seasonal index number that is used to adjust the parenthesised forecast for seasonality. The only remaining term to explain is the "L" that appears in the subscript of the seasonal index variable; it represents the length of seasonality or 4 in the case of the present application which deals with quarterly data.

As with most smoothing models that are quite sensitive to initial settings, several different approaches are available for providing the commencing values required for the model estimation process[7]. Two approaches were considered; the first being a variant of one suggested by Flaherty et al (1999, pp. 454-455) and the other by Hanke and Reitsche (1997, pp. 166). As will be seen subsequently in this paper, the former (latter) generated superior results in the out of sample period (sample period).

By the final observation of the sample period, the first of these forecasting models (referred to as **wxsm1**)was given by:

$$\hat{Y}_{t+\tau} = (25551 + 43\tau)S_{t-L+\tau} \qquad t = 114 \tag{4.1}$$

and somewhat surprisingly, no exponential smoothing adjustment to the *initial* seasonal indices ensued throughout the sample period[8]. The unvarying seasonal indices were as follows :

Mar Qtr Index	0.9655
Jun Qtr Index	1.1070
Sep Qtr Index	1.0540
Dec Qtr. Index	0.9698

The second of these forecasting models (referred to as **wxsm2**) was given by :

$$\hat{Y}_{t+\tau} = (26280 + 64\tau)S_{t-L+\tau} \qquad t = 114 \tag{4.2}$$

Again, no exponential smoothing adjustment to the *initial* seasonal indices eventuated throughout the sample period[9]. The unvarying seasonal indices were as follows :

Mar Qtr Index	0.9621
Jun Qtr Index	0.9828
Sep Qtr Index	1.0725
Dec Qtr. Index	0.9826

The Classical Decomposition (cdtsm) and Seasonal Dummy (sdtm) models

Like the previous model, both the classical decomposition model (**cdtsm**) as well as the multivariate seasonal dummy and trend model (**sdtm**), attempt to forecast a time series that is subject to a linear trend and seasonal influences.

The **cdtsm** model was obtained using a very traditional approach illustrated by Levin (1987, Ch.14). The general form of the quarterly forecasting model is given by :

$$\hat{Y}_t = \{a + bt\}S_q \tag{5.0}$$

where it is assumed that the analyst wishes to obtain a seasonally adjusted trend forecast for t periods from a pre-determined base period (in the case of the present

application this is 1969(4); the time period immediately preceding that of the first observation of the sample period). The *a* and *b* appearing in expression (5.0) are the unvarying intercept and slope coefficient of a linear trend line and S_q is an appropriate multiplicative seasonal index for the quarter q that corresponds to time period t.

The actual **cdtsm** model [10] that was estimated over the sample period is given by :

$$\hat{Y}_t = \{22526.58 + 27.37t\}S_q \qquad (5.1)$$

where t = 0 at 1969(4) and the multiplicative seasonal indices $\{S_q\}$ were as follows:

Mar Qtr Index	0.9361
Jun Qtr Index	1.0128
Sep Qtr Index	1.0384
Dec Qtr. Index	1.0127

Estimation of the **sdtm** model proceeded along conventional lines illustrated by Flaherty et al. (1999, p.467-468). The general form of this model is given by:

$$\hat{Y}_t = \{a + bt\} + b_1S_1 + b_2S_2 + b_3S_3 \qquad (5.2)$$

where the parenthesised portion of expression (5.2) performs much the same role as that which appears in expression (5.0). Namely, it is designed to capture the long term secular growth of the series. On the other hand, S_q appearing in expression (5.2) denotes a zero-one *seasonal dummy variable* for the q^{th} quarter (q = 1 to 3). Moreover, its slope coefficient b_q is designed to capture the seasonal impact that this quarter has on quarterly housing starts.[11] Note that there is no dummy variable for the fourth quarter. The seasonal impact of this quarter - which acts as a benchmark quarter - is incorporated within the intercept term a.[12]

Reproduced below is the actual **sdtm** model that was estimated over the sample period through the application of multiple linear regression. The estimated model is accompanied by diagnostics, with bracketed t-values appearing beneath the estimated coefficients.

$$\hat{Y}_t = \underset{[24.91]}{22804.09} + \underset{[2.67]}{27.73t} - \underset{[-1.87]}{1811.98S_1} - \underset{[-.02]}{24.02S_2} + \underset{[.65]}{637.55S_3} \qquad (5.3)$$

$R^2 = .12 \quad \bar{R}^2 = .08 \qquad F = 3.60 \qquad DW = .31$

The above model yields a disappointingly low adjusted $\bar{R}^2$, and although the F statistic indicates that the overall relationship is significant at the 1% significance level, it appears that none of the seasonal dummy variables have any significant impact on housing starts at the 5% significance level. Finally, the DW statistic indicates the presence of significant first order serial correlation in the disturbance term.

Trigonometric Seasonal Forecasting model (tsfm)

The final model considered in this section of the paper is a trigonometric seasonal forecasting model whose estimation is illustrated by Flaherty et al. (1999, pp. 469-470). The general form of this forecasting model[13] - as applied to quarterly data - is reproduced below:

$$\hat{Y}_t = a + b_1 t + b_2 \text{Cos}\left(\frac{2\pi t}{4}\right) + b_3 \text{Sin}\left(\frac{2\pi t}{4}\right) + b_4 \left\{t\text{Cos}\left(\frac{2\pi t}{4}\right)\right\} + b_5 \left\{t\text{Sin}\left(\frac{2\pi t}{4}\right)\right\} + b_6 Y_{t-1} \quad (6.0)$$

Although the full model in expression (6.0) was estimated, a number of explanatory variables were found to be statistically insignificant and were subsequently dropped from a more compact model whose estimate is presented below :

$$\hat{Y}_t = \underset{[2.96]}{3536.21} \; \underset{[-3.95]}{-1044.23} \left\{\text{Cos}\left(\frac{2\pi t}{4}\right)\right\} \; \underset{[-5.52]}{-21.63} \left\{t\text{Sin}\left(\frac{2\pi t}{4}\right)\right\} + \underset{[17.40]}{0.85} Y_{t-1} \quad (6.1)$$

$R^2 = .75 \quad \bar{R}^2 = .74 \qquad F = 108.35 \qquad DW = 1.75 \qquad h = 1.57$

The above model yields a surprisingly high adjusted $\bar{R}^2$, with the F statistic indicating that the overall relationship is highly significant. The bracketed t-statistics indicate that the intercept and slope coefficients of all explanatory variables differ significantly from zero at the 1% significance level. Finally, Durbin's h statistic indicates the absence of significant first order serial autocorrelation in the disturbance term at any reasonable significance level.

Comparative Performance of Simple Time Series Models

In Table 1 below, find summary diagnostics for each of the models estimated in Section 1 over the sample period. Examination of the mean percentage error (MPE) diagnostic provides an appreciation of the degree of forecasting bias associated with each of the models. The least biased model appears to be the **dxsm** model whose MPE is .05 of one percent ! However, three other models - **sxsm**, **wxsm1** and **wxsm2** - enjoy MPE values that are very close to 0%.

The mean squared error (MSE) diagnostic which effectively penalises models with large forecasting errors, suggests that the best performers are the **wxsm1** and **wxsm2** models. These two models also perform best according to the two remaining diagnostics in Table 1 : the mean absolute deviation (MAD) and the mean absolute percentage error (MAPE). The former diagnostic is a useful measure when the analyst's intent is to measure the magnitude of average forecasting error in much the same units as the initial series. On the other hand, if the focus is on the average *relative* magnitude of the forecasting error in relation to the original values, the MAPE would be the appropriate diagnostic by which to gauge the performance of competing forecasting models. The two poorest performing models - **cdtsm** and **sdtm** - have almost identical diagnostics.

	Model	MAD	MAPE	MSE	MPE
Single exponential smoothing	sxsm	1932	8.09%	5,742,855	-0.30%

Brown's double exponential smoothing	dxsm	2132	8.97%	7,072,559	0.05%
Winter's exponential smoothing	wxsm1	143	0.58%	30,537	-0.47%
Winter's exponential smoothing	wxsm2	98	0.40%	14,266	-0.34%
Holt's exponential smoothing	hxsm	1927	8.12%	5,804,286	-1.48%
Classical decomposition	cdtsm	2908	12.32%	12,705,235	-2.19%
Seasonal dummy variables	sdtm	2909	12.33%	12,727,721	-2.19%
Trigonometric seasonal forecasting	tsfm	1449	6.09%	3,642,779	-0.62%

MAD refers to the mean absolute deviation
MAPE to the mean absolute percentage error
MSE to the mean square error
MPE to the mean percentage error

Table1: Performance of Models over Sample Period

In Table 2, diagnostics have been computed for each of the models in the *out of sample* period. Since, this only comprises a total of four quarters, this is clearly too short a span of time to arrive at a definitive conclusion as to which of the models performs best beyond the sample period. However, the following tentative observations may be made.

Model	MAD	MAPE	MSE	MPE
sxsm	2011	8.18%	5604328	-0.48%
dxsm	2424	10.25%	9449662	-0.42%
wxsm1	818	3.53%	1651037	-3.39%
wxsm2	1561	6.53%	3765044	-6.53%
hxsm	1664	7.13%	5229824	-6.85%
cdtsm	837	3.54%	1148899	-3.54%
sdtm	866	3.68%	1278807	-3.65%
tsfm	2561	10.29%	7110835	-6.92%

Table 2 : Performance Beyond Sample Period

Firstly, and perhaps surprisingly, the two poorest performing models over the sample period - **cdtsm** and **sdtm** - enjoy the best two MSE figures and the second and third best MAD and MAPE figures. Secondly , whilst **wxsm1** and **wxsm2** were the best performing models over the sample period, only **wxsm1** reigns supreme beyond the sample period as far as the MAD and MAPE criteria are concerned. Finally, it is the **sxsm** and **dxsm** models that seem to provide the least biased forecasts beyond the sample period. The authors await further observations, before they can reach a definitive conclusion regarding the relative performance of the studied models.

The entries in Table 3, are p-values for the Ljung-Box-Pierce test statistics[14] determined for the residual autocorrelations at successive lags for each model listed in Table 1. These statistics suggest that there is significant serial correlation amongst the residuals of each model beyond a lag of 4. What is particularly surprising is that two of the models that purport to model seasonality - **wxsm1** and **tsfm** - yield a significant residual autocorrelation coefficient at the fourth lag. This suggests that neither of these so called seasonal models are properly accounting for the impact that seasonality has on the original series.

	Model							
	sxsm	dxsm	wxsm1	wxsm2	hxsm	cdtsm	sdtm	tsfm
Lag								
1	0.693	0.272	0.000	0.000	0.691	0.000	0.000	0.201
2	0.424	0.022	0.000	0.000	0.423	0.000	0.000	0.001
3	0.556	0.043	0.000	0.000	0.553	0.000	0.000	0.004
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001
5 to 24	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The entries in this table indicate p-values for the Ljung-Box-Pierce test statistics determined for the residual autocorrelations at successive lags for each model estimated in Section 1.

Table 3 : Box-Ljung-Pierce Analysis of Model Residuals

Section 2 Arima Modelling

This section is devoted to the development of an appropriate seasonal autoregressive integrated moving average (ARIMA) model[15] for private new housing starts.

Employing the same notation as Bowerman and O'Connell(1999, p.570), the general form of a seasonal ARIMA model is given by :

$$\phi_p(B)\phi_P(B^L)Z_t = \delta + \theta_q(B)\theta_Q(B^L)a_t \qquad (7.0)$$

where:

B denotes the backshift operator.[16]

Z_t denotes a *stationary* variable[17] that is obtained through an appropriate transformation of the original series of interest; this being private new housing starts in the case of the present paper.

$\phi_p(B)$ denotes the nonseasonal autoregressive operator of order p.[18]

$\phi_P(B^L)$ denotes the seasonal autoregressive operator of order P.[19]

$\theta_q(B)$ denotes the nonseasonal moving average operator of order q.[20]

$\theta_Q(B^L)$ denotes the seasonal moving average operator of order Q.[21]

δ denotes a constant term.[22]

a_t denotes a classically well behaved disturbance term.[23]

The Shazam software program was used to implement all four iterative stages of the Box Jenkins methodology [24] for arriving at an appropriate ARIMA model including forecasts beyond the sample period.

The original data series was found to be non-stationary. Consequently, a search was made for a plausibly stationary series for which an adequate ARIMA model could be fitted. A total of 36 different historical series - including the original series - were

investigated for stationarity. The original series was subjected to all combinations of three levels of regular (i.e. non-seasonal) differencing (d = 0, 1 and 2) and three levels of seasonal differencing (D = 0, 1 and 2). This procedure was repeated for three separate pre-differencing transformations of the original series; the square root of the original series, its quartic root and finally its logarithm. A total of 12 plausibly stationary series were identified. However, in only two cases was a reasonably adequate model fitted. In both cases the modelled series had not been subjected to a pre-differencing transformation.

The first of these estimated models - ARIMA1 - took the form :

$$(1 - \hat{\phi}_1 B - \hat{\phi}_2 B^2)(1 - \hat{\phi}_{1,4} B^4 - \hat{\phi}_{2,4} B^8) Z_t = e_t$$

where : $Z_t = (1 - B^4)(1 - B)^2 Y_t$ and e_t denotes the residual

The second of these estimated models - ARIMA2 - assumed the form :

$$(1 - \hat{\phi}_1 B - \hat{\phi}_2 B^2)(1 - \hat{\phi}_{1,4} B^4 - \hat{\phi}_{2,4} B^8 - \hat{\phi}_{3,4} B^{12} - \hat{\phi}_{4,4} B^{16}) Z_t = e_t$$

where : $Z_t = (1 - B^4)^2(1 - B)^2 Y_t$

Table 4 below provides some useful diagnostics by which to compare the adequacy of both models over the sample period.

	Diagnostic			
Model	$\bar{R}^2$	AIC	SC	Se
ARIMA1	.4187	15.327	15.426	2051.3
ARIMA2	.7459	15.565	15.717	2263.6

In this table $\bar{R}^2$ denotes the adjusted R^2, AIC (SC) denotes the logarithm of the Akaike Information Criterion Statistic (Schwartz Information Criterion Statistic) and S_e denotes the standard error of estimate.

Table 4 : Diagnostic Comparison of ARIMA Models

The estimated parameters and their associated t-statistics are presented in Table 5. For both models, all estimated parameters appear to be significant at all reasonable significance levels.

Estimated Parameter	ARIMA1	ARIMA2
$\hat{\phi}_1$	-.38812 [-4.311]	-.40477 [-4.382]
$\hat{\phi}_2$	-.41046 [-4.559]	-.42244 [-4.582]
$\hat{\phi}_{1,4}$	-.63190 [-6.767]	-1.4442 [-16.27]
$\hat{\phi}_{2,4}$	-.41693 [-4.680]	-.41693 [-11.78]
$\hat{\phi}_{3,4}$	-	-1.0977 [-8.66]
$\hat{\phi}_{4,4}$	-	-.52065 [-6.607]

The bracketed entries in this table refer to the calculated t-statistics for the estimated parameters of the tentatively identified ARIMA model. The latter are the un-bracketed figures in this table.

Table 5 : t-Values for the Estimated Parameters

Table 6 below confirms that the residual autocorrelations of either fitted ARIMA model are not significantly different from zero at the 1% significance level.

Residual Auto-Correlations for the ARIMA1 Model													
Lags													SE
01-12	.04	.02	-.01	-.09	-.19	-.04	-.18	-.10	-.07	.02	-.08	-.12	.10
13-24	.06	.07	.05	.10	.09	.00	.10	.05	.01	.08	-.08	-.13	.11
25-36	-.07	-.07	-.16	-.01	-.06	.09	.11	.07	.03	.12	.06	.02	.11
37-48	-.05	-.06	-.06	-.01	.02	.00	-.13	.02	-.04	-.04	.03	-.01	.12
49-60	.03	.09	.09	.07	.03	-.06	-.02	-.09	-.12	-.03	-.02	0.00	.12
Residual Auto-Correlations for the ARIMA2 Model													
Lags													SE
01-12	-.01	-.01	-.08	-.07	-.15	-.05	-.12	-.11	.03	.02	.02	-.14	.10
13-24	.11	.00	.10	-.12	.08	-.07	.10	-.08	.04	.08	.02	.07	.11
25-36	.02	-.07	-.10	-.02	-.12	.04	.04	-.04	-.04	.13	.05	.00	.11
37-48	.03	-.02	-.02	.05	.01	-.03	-.07	.05	-.05	-.09	.02	-.07	.12

49-60	.01	.06	.07	.05	.07	-.06	-.02	-.06	-.11	-.01	.05	.00	.12

Table 6 : Residual Autocorrelations for Both ARIMA Models

Again, for both models, the p-values of the Ljung Box Pierce Statistics for residual autocorrelations suggest that there is no pattern in the residuals at the 1% significance level.

P-Values for ARIMA2 Model

Lags												
05-16	.019	.056	.023	.030	.044	.076	.092	.082	.107	.127	.161	.154
17-28	.162	.211	.204	.239	.294	.305	.310	.243	.263	.282	.173	.211
29-40	.232	.224	.199	.209	.244	.200	.219	.254	.276	.291	.309	.351
41-52	.393	.438	.355	.394	.425	.458	.494	.536	.570	.547	.514	.514
53-60	.545	.551	.587	.553	.467	.497	.531	.569				

P-Values for ARIMA2 Model

Lags												
07-18	.017	.030	.068	.126	.202	.139	.129	.189	.189	.167	.184	.208
19-24	.206	.222.	.270	.289	.348	.370	.430	.454	.427	.486	.418	.459
31-36	.505	.492	.528	.580	.621	.666	.708	.731	.770	.801	.801	.817
43-48	.828	.811	.839	.840	.866	.868	.866	.873	.868	.868	.887	.889
55-60	.844	.866	.874	.894								

Table 7 : Box-Ljung-Pierce Analysis of Model Residuals

Although the ARIMA1 model has a lower $\bar{R}^2$ than ARIMA2, in some other respects it is the preferred model. Firstly, its AIC, SC and Se statistics are lower. Secondly it is a far less complex model; less parameter estimates are involved and the level of differencing is somewhat simpler. Finally, as will be seen subsequently in Table 8, the ARIMA1 model appears to be a much better forecasting model beyond the sample period. A comparison with the results in Table 2 reveal that ARIMA1 is the least biased in percentage terms of all non-causal models. On the other hand, its MAD, MAPE and MSE values are not as favourable as those pertaining to some of the other models considered earlier in Section 1.

Model	MAD	MAPE	MSE	MPE
ARIMA1	1197	4.81%	1617046	-0.17%
ARIMA2	5949	24.38%	41949677	-24.38%

MAD refers to the mean absolute deviation, MAPE to the mean absolute percentage error, MSE to the mean square error and MPE to the mean percentage error.

Table 8 : Performance Beyond Sample Period

Section 3 Combining Forecasts

Combining forecasts typically improves the error statistics such as MAD and MSE, the criteria used to select a model. The data series in question does not display a long run secular trend, seasonality is present but cyclical variations, while present, are subject to phase shifts. The ARIMA modelling approach described above comes closest to satisfying the intuitive perception of the underlying data generating process. Due to the nature of the data, as well as the results reported in table 1, it was decided to incorporate some of the simpler time series models to form the combined model.[25]

The weights were determined using the Solver add-in in Excel and the error statistic minimised is MSE. Weights are in the range, $0 \leq w_i \leq 1$ and sum to unity. The results are provided in table 9. The relative magnitude of the weights indicate the bias for a particular model. For example, the largest weight is assigned to **wxsm2**, this results in a proportionate contribution of 48.08% by this model to the combined model. This is consistent with the error statistics, given in table 1, for **wxsm2** which produced the lowest MSE statistics. The classical decomposition model is assigned the lowest weight, almost zero, suggesting that it has negligible influence in the combined model.

The error statistics indicate an improvement over all models with the exception of Winter's exponential smoothing. Purely statistical criteria is a poor basis upon which to make a judgement. The algorithm employed to carry out the analysis will attempt to optimally meet the criteria imposed. On the basis of the error statistics, Winter's model is optimal. However, there may be additional information, or the informed view of the forecaster, that recognises the value of alternative models in the environment being forecasted. A combined model may be employed to capture elements of the objective statistical criteria and the knowledge of the forecaster. For example, a condition may be imposed which places an upper and/or lower bound on the contribution of a particular model to the combined model. Such a restriction may be easily implemented in Excel.

	Model		**Weight**	
$(1 - B^4)(1 - B)^2Y_t$	ARIMA1	w_1	0.1808	**Error Statistics**
Winter's exponential smoothing$_1$	WXSM1	w_2	0.2578	MSE 363,268
Winter's exponential smoothing$_2$	WXSM2	w_3	0.4808	MAD 1070
Single exponential smoothing	SXSM	w_4	0.0808	MAPE 2.043%
Classical decomposition	CDTSM	w_5	0.0000	MPE -0.210%
	Sum of weights			1.0000

Table 9 : Weights for the Combined Model and Error Statistics

Section 4 Causal Economic Modelling

Housing starts are a leading economic indicator and are an important source of information when it comes to the general economic outlook. Changes in starts directly impact on the housing and construction industry and indirectly impact on those industries that depend upon it. Successive governments in the past have looked to the housing sector as an engine of economic growth and have introduced policy measures to take advantage of this sector's ability to stimulate the economy. A key policy instrument traditionally has been the control of housing interest rates. With the deregulation of the financial markets these controls have been removed and replaced with competitive market forces.

Throughout the decade of the nineties several new mortgage providers have entered the market. This has lead to greater competition among the established providers of mortgage finance, the banks, as well as in the industry as a whole. This has resulted in historically low interest rates over most of this period. The low inflation rates, and relatively highly levels of unemployment, during the nineties have also meant a downward pressure on interest rates. Market expectations have come to reflect this environment and the outlook for the foreseeable future is for a moderate increase in interest rates.

Despite the obvious theoretical link between interest rates and housing starts, attempts to model the relationship are difficult by the unsuitability of the data to test this relationship. The controls over interest rates has meant that a historical series is not truly reflective of market behaviour and tests incorporating this information are inappropriate.

The decision to acquire a mortgage is based partly on the interest cost and also on the current and expected future wealth of the borrower. A positive outlook for the future is likely to weigh much more heavily on the decision to take on a mortgage. Thus the general state of the economy should influence changes in the housing sector. A simple regression relationship between housing starts is provided in table 10.

Regression 1 — Period Sept. 1959 to June 1999.			Number of Obs = 160		
	Coefficients	*Standard Error*	*t Stat*	*F*	95.10
Intercept	15195.26	782.21	19.43	Adj. R Sq	0.37
GDP	0.09	0.01	9.75	Std Error	3613.44

Table 10. Regression relationship Between GDP and Housing Starts

The information contained in table 10 is overly simplistic and provides very little in the way of useful information. In subsequent analysis the first 49 observations, the period 1959 to 1969 inclusive, were removed for reasons discussed in the introduction. A time trend was included in a modified regression and various lags of GDP. The time trend was not significant, this is consistent with the previous ARIMA analysis, and lags of GDP other than 4, 12 and 16 were not significant. Table 11 contains the results for GDP and lags of 4, 8 and 12.

	Coefficients	*Standard Error*	*t Stat*		
Intercept	16835.32	820.63	20.52	Obs.	148
GDP	0.46	0.14	3.20	Std Error	3297.5
GDP_{-4}	-0.74	0.20	-3.72	R Sq	0.426
GDP_{-8}	-0.31	0.20	-1.58	Adj R Sq	0.410
GDP_{-12}	0.70	0.15	4.68	F	26.539

Table 11. Regression relationship with lagged values of GDP

Using past values of GDP to explain housing starts has to be interpreted with care, the causal relationship is likely to be the other way around. Our interest is in discovering whether the general state of the economy influences housing starts. The results in table 10 confirm that this is indeed the case; developers or builders take account of current and past GDP information. The seasonal and cyclical nature of the data primarily account for the significance of the various lags. This is apparent from table 12 when lagged values of the dependent variable (housing starts) are included in the model. The previous quarter has the greatest impact, the twelfth quarter (3 years back) is also significant.

	Coefficients	*Standard Error*	*t Stat*		
Intercept	2540.9	2119.00	1.199		
GDP	0.22529	0.1073	2.099		
GDP4	-0.33226	0.1489	-2.232	R Sq	0.692
GDP8	-0.10751	0.1452	-0.741	Adj R Sq	0.675
GDP12	0.20933	0.1080	1.939		
$Starts_{-1}$	0.74973	0.0644	11.650		
$Starts_{-12}$	0.17802	0.0633	2.811		

Table 12. Regression relationship with lagged values of GDP and Housing Starts

Section 5 Conclusion and Prospects for Future Research

The data is quarterly, which dampens the variability in the data, and a larger number of observations would be desirable for both the ARIMA and regression analysis. The results reflect the quality of the data and a larger sample with higher frequency would be preferred. Higher frequency data would also allow more extensive testing to be carried out to establish causal relationships.

The analysis has been carried on national data. While there are many similarities across the different regions of Australia, it may prove useful to conduct a separate analysis for the different regions. Housing cycles tend to differ between states and interstate migration impacts on the underlying trend. This latter effect is difficult to identify in national data which is an aggregation of the states.

Some consideration should perhaps be given to identifying the nature of housing demand. In the short run demand is determined by expectations about the state of the economy and the major variable of influence in the long term is population. Medium term demand is more difficult to identify. One approach that may be worthy of consideration is to use a stock adjustment model, however, at present a suitable data series to quantify stock does not exist.

Bibliography

Bowermann B. L., O'Connell R. T. 1993 ***Forecasting and Time Series An Applied Approach*** 3rd ed. Wadsworth Inc.

Diebold, F.X., 1998 ***Elements of Forecasting*** South-Western College Publishing.

Flaherty J. et al. 1999 ***A Spreadsheet Approach to Business Quantitative Methods*** ISBN 0 86444 790 6

Hanke J. E., Reitsch A. G. 1998 ***Business Forecasting*** 6th ed. Prentice Hall International Inc.

Kacapyr, E., 1996 ***Economic Forecasting, The State of the Art***, M.E. Sharpe. ISBN 1-56324-765-8

Levin R. I. 1987 ***Statistics for Management*** 4th ed. Prentice Hall, Inc. Englewood Cliffs, New Jersey.

White K. J. 1997 ***Shazam User's Reference Manual Version 8.0*** McGraw-Hill ISBN 0-07-069870-8.

Notes

[1] The specific data on quarterly private new housing starts that were obtained through AUSSTATS were sourced from Table 1 of the Australian Bureau Statistics Catalogue Number 8750 (Building Activity - Dwelling Unit Commencements - Preliminary - Australia - Quarterly).

[2] A detailed explanation of how to use SOLVER to obtain the parameters of all the smoothing models considered in this section is to be found in Flaherty et. al. (1999, Ch. 13).

[3] A detailed explanation of how to use EXCEL and its regression analysis tool to estimate cdtsm, sdtm and tsfm is to be found in Flaherty et. al. (1999, Chs. 13).

[4] Both a_t and b_t are functionally related to two variables S_t and SS_t Their respective formulae are given by:

$$a_t = 2S_t - SS_t \qquad b_t = \left(\frac{\alpha}{1-\alpha}\right)[S_t - SS_t]$$

where S_t is the exponentially smoothed value of Y_t at time t and SS_t - known as the double exponentially smoothed value - is the smoothed value of S_t at time t. Formulae for S_t and SS_t are reproduced below:

$$S_t = \alpha Y_t + (1-\alpha)S_{t-1} \qquad SS_t = \alpha S_t + (1-\alpha)SS_{t-1}$$

It is observed that both S_t and SS_t employ the same smoothing constant $0 \leq \alpha \leq 1$ and that this same constant appears in the formula for b_t. The value of α is chosen in such a way as to minimise MSE over the sample period.

[5] In the hxsm model, the adjustment equations for a_t and b_t at time period t are given by the following two smoothing equations :

$$a_t = \alpha Y_t + (1-\alpha)(a_{t-1} + b_{t-1}) \qquad b_t = \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1}$$

where $0 \leq \alpha, \beta \leq 1$ are smoothing constants that are chosen in such a way as to minimise MSE over the sample period.

[6] An alternative approach suggested by Hanke and Reitsche (1997, p162 - 163) in the absence of prior historical data, is to set $a_1 = Y_1$ and $b_1 = 0$. This approach was in fact attempted but the results were discarded. This is because the approach generated a rather trivial forecasting model - with a zero slope coefficient. By the last quarter of the sample period the resultant forecasting model was given by :

$$\hat{Y}_{t+\tau} = 25838.0 = Y_{114} \qquad t=114$$

which is equivalent to the constant term in expression (3.0)

The above result essentially arises because the optimal smoothing constants α and β (see footnote 5) took on the values of 1 and 0 respectively. This meant that $a_t = Y_t$ and $b_t = 0$ at each quarter t of the sample period. Another interesting remark that should be made about this result, is that this particular hxsm model, when used for one period ahead forecasting, is exactly equivalent to the sxsm model estimated earlier with a smoothing constant $\alpha = 1$ (see expression 1.1). For this reason, all diagnostics, relating to the residuals of either estimated model will be identical over the sample period.

[7] In the wxsm model, terms : a_t , b_t and S_t are exponentially adjusted according to the following three smoothing equations :

$$a_t = \alpha\left(\frac{Y_t}{S_{t-L}}\right) + (1 - \alpha)(a_{t-1} + b_{t-1}) \qquad b_t = \beta(a_t - a_{t-1}) + (1 - \beta)b_{t-1} \; S_t = \gamma\left(\frac{Y_t}{a_{t-1}}\right) + (1 - \gamma)S_{t-L}$$

where $0 \leq \alpha, \beta, \gamma \leq 1$ are smoothing constants that are chosen in such a way as to minimise MSE over the sample period. The terms a_t and b_t have much the same interpretation as in the dxsm and hxsm models. On the other hand, an additional term : S_t denotes the smoothed estimate of the seasonal multiplicative index at time t.

As with most smoothing models that are quite sensitive to initial settings, several different approaches are available for providing the commencing values required for the model estimation process. The two approaches adopted in this paper are discussed below.

The first approach - referred to as wxsm1 - is a slight variation of one suggested by Flaherty et al (1999, pp. 454-455). More specifically, use is made of the eight quarters of data preceding the sample period. The initial values a_o and b_o were arrived at as follows :

$$a_o = \left(\frac{\sum_{t=1}^{4} Y_t}{4}\right) \qquad b_o = \frac{1}{4}\left(\frac{\sum_{t=1}^{4}\{ Y_{t+4} - Y_t \}}{4}\right)$$

with the initial seasonal index for the qth quarter given by :

$$S_{qo} = \frac{4 \infty Y_q}{\sum_{n=1}^{4} Y_i} \quad q = 1 \ldots 4$$

The second approach - referred to as wxsm2 - is one suggested by Hanke and Reitsche (1997, p.166). Here, a_o is to be estimated by averaging a few values prior to the sample period and b_o is to be estimated by using the slope of the trend equation fitted to prior data. Finally, the initial seasonal indices may be generated by using the method of Time Series Decomposition. In this paper, data for the eight quarters preceding the sample period were averaged to obtain a_o. The estimate of the initial seasonal indices as well as b_o were obtained by estimating a Classical Decomposition of Time Series Model applied to the forty quarters preceding the sample period.

[8] The set of smoothing constants that minimised MSE over the sample period were given by $(\alpha, \beta, \gamma) = (.99, .03, 0)$. Hence, with γ set to zero, the seasonal indices remained unchanged throughout the sample period.

[9] The set of smoothing constants that minimised MSE over the sample period were given by $(\alpha, \beta, \gamma) = (1, .01, 0)$. As in the case of wxsm1, a γ-value of zero, renders the seasonal indices invariant throughout the sample period.

[10] The trend component of the estimated model in expression (5.1) was estimated by fitting a least squares linear trend model to the *deseasonalised* values of $\{Y_t\}$ over the sample period. Diagnostics for this trend model were as follows :

$$T = \underset{[33.32]}{22526.58} + \underset{[2.68]}{27.37t} \qquad R^2 = .06 \qquad \bar{R}^2 = .05 \; F = 7.19$$

where the bracketed terms underneath the intercept and slope coefficients indicate t-values. The seasonal indices used for the deseasonalisation process described previously, were obtained using the Ratio to Moving Average Technique for measuring seasonal variation. This technique is described in most elementary business statistics textbooks. See for example, Levin (1987, pp.707).

[11] The dummy variables $\{S_q\}$ appearing in expression (5.2) are zero-one variables. If time period t happens to be the q^{th} quarter, then the seasonal dummy variable $S_q = 1$; and 0 otherwise. It should

therefore follow that the estimated seasonal influence that the q^{th} quarter exerts on housing starts is given by the slope coefficient b_q of S_q.

[12] There is a very important technical reason for the deliberate omission of a dummy variable for the fourth quarter, this being that any one of the four dummy variables would be a linear combination of the other three. In multiple regression analysis, if any one explanatory variable is a linear combination of the others, then it is impossible to obtain the least squares estimating hyper plane. This is a problem referred to in the econometric literature as perfect multi-collinearity.

[13] In the right hand side of expression (6.0)

$a + b_1 t$	captures the long term linear trend of the series
$b_2 Cos\left(\frac{2\pi t}{4}\right) + b_3 Sin\left(\frac{2\pi t}{4}\right)$	captures the seasonality of the time series
$b_4\left\{tCos\left(\frac{2\pi t}{4}\right)\right\} + b_5\left\{tSin\left(\frac{2\pi t}{4}\right)\right\}$	captures the change in seasonality across time
$b_6 Y_{t-1}$	captures the influence that the previous quarterly starts might have on current quarterly starts

[14] The p-values appearing in Table 3 were obtain using the SHAZAM econometric package (version 8.0).

[15] ARIMA models attempt to reproduce the historical behaviour of a stationary time series. Such a time series is one whose mean and variance remain unchanged through time.

[16] The backshift operator B when applied to a time series observation Z_t shifts its subscript one period backwards in time. For instance:

$BZ_t = Z_{t-1}$

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

$B^L Z_t = Z_{t-L}$

[17] If the time series of interest $\{Y_t\}$ is non-stationary, then it must be transformed into a stationary one denoted $\{Z_t\}$. The general form of the transformation is presented below using the Bowermann and Cooper (1993, pp. 568-570) notation.

$Z_t = (1 - B^L)^D (1-B)^d Y_t^*$

where :

- B denotes the backshift operator (described in Footnote 17)
- L denotes the length of seasonality (with quarterly data L = 4)
- $\{Y_t^*\}$ denotes a suitable *pre-differencing* transformation of the series $\{Y_t\}$ designed to ensure that the transformed series enjoys constant variance over time. Commonly applied pre-differencing transformations involve, the extraction of the square root, the quartic root and the logarithm of the original series $\{Y_t\}$. Note that if the variability in the original series appears to remain constant over time, no pre-differencing transformation is required so that $\{Y_t^*\}$ is equivalent to $\{Y_t\}$.
- d (D) denotes the level of regular (seasonal) differencing applied to the series $\{Y_t^*\}$ designed to ensure that the resultant series $\{Z_t\}$ fluctuates with constant variation about a constant mean.

For example, suppose $\{ Z_t \}$ is stationary when $\{ Y_t^* \} = \{Y_t\}$, d = 1, D=1 and L = 4, then :

$Z_t = (1 - B^4)^1 (1-B)^1 Y_t = (1 - B^4)(1-B)Y_t = (1 - B - B^4 + B^5)Y_t = Y_t - BY_t - B^4 Y_t + B^5 Y_t =$

$= Y_t - Y_{t-1} - Y_{t-4} + Y_{t-5}$

[18] The nonseasonal autoregressive operator of order p is given by : $\phi_p(B) = 1 - \phi_1 B^1 - \phi_2 B^2 - .. - \phi_p B^p$ where: $\phi_1, \phi_2, \ldots \phi_p$ are parameters that must be estimated from sample data.

[19] The seasonal autoregressive operator of order P is given by: $\phi_P(B) = 1 - \phi_{1,L} B^{BL} - \phi_{2,L} B^{2L} - .. - \phi_{P,L} B^{PL}$ where: $\phi_{1,L}, \phi_{2,L}, \ldots \phi_{P,L}$ are parameters that must be estimated from sample data.

[20] The nonseasonal moving average operator of order q is given by : $\theta_q(B) = 1 - \theta_1 B^1 - \theta_2 B^2 - ... - \theta_q B^q$ where: $\theta_1, \theta_2, \ldots \theta_q$ are parameters that must be estimated from sample data.

[21] The seasonal moving average operator of order Q is given by: $\theta_Q(B) = 1 - \theta_{1,L} B^{BL} - \theta_{2,L} B^{2L} - .. - \theta_{Q,L} B^{QL}$ where $\theta_{1,L}, \theta_{2,L}, \ldots \theta_{Q,L}$ are parameters that must be estimated from sample data.

[22] The constant term is given by: $\delta = \mu\phi_p(B)\phi_P(BL)$ where: μ is the true mean of the modelled stationary series.

[23] This is a disturbance term that is normally distributed with a zero mean and constant variance. It is also serially uncorrelated with constant variance across time.

[24] The Box Jenkins methodology is a four step iterative procedure that may be summarised as follows :

Stage 1 : Tentative Model Identification. The first part of this stage establishes whether or not the historical series is stationary. As described in Bowermann and O'Connell (1993, pp. 521 - 528), temporal stability of the variance may be gauged by inspecting a scattergram of the series against time. Moreover, the absence of trend, at both the seasonal and nonseasonal level, may be confirmed by examining the sample autocorrelation function of the series. Note that if the original series of interest is deemed to be non-stationary, it must be *transformed* into one that is (see Footnote 18 for details). Having identified a stationary series, the analyst is in a position to undertake the second part of stage 1. This involves the tentative identification of a plausible ARIMA model that would generate the type of behaviour exhibited by the series' sample autocorrelation and partial autocorrelation functions. In undertaking this part of Stage 1, the authors chose to abide by the guidelines of model identification provided by Bowermann and O'Connell (1993, pp. 572-574).

Stage 2 : Estimation. At this stage, the stationary historical data referred to at Stage 1 are used to estimate the parameters of the tentatively identified model. The algorithm used to estimate this model is described in White (1997, pp. 126 - 128).

Stage 3 : Diagnostic Checking. Here the analyst makes use of several diagnostics to judge the adequacy of the tentatively identified model that was estimated at Stage 2. These involve an inspection of the adjusted R^2 , the t-statistics of each of the estimated parameters, conducting at various lags, a Box-Ljung-Pierce Test of serial autocorrelation of the residuals, generation of the sample autocorrelation function for the residuals to investigate whether the autocorrelations of residuals at low order or seasonal lags differ significantly from zero. If the estimated model is deemed to be inadequate, then the diagnostics may suggest a new improved model.

Stage 4 : Forecasting. Once an appropriate model has been identified and estimated it may be employed to forecast future values of the time series.

[25] Combining forecasting models using Excel is described in Flaherty et. al pp. 474-475.