CSC 425 – Time Series Analysis

Retail Sales Data

Final Project

Bill Qualls
Laura Floyd
3/13/2013

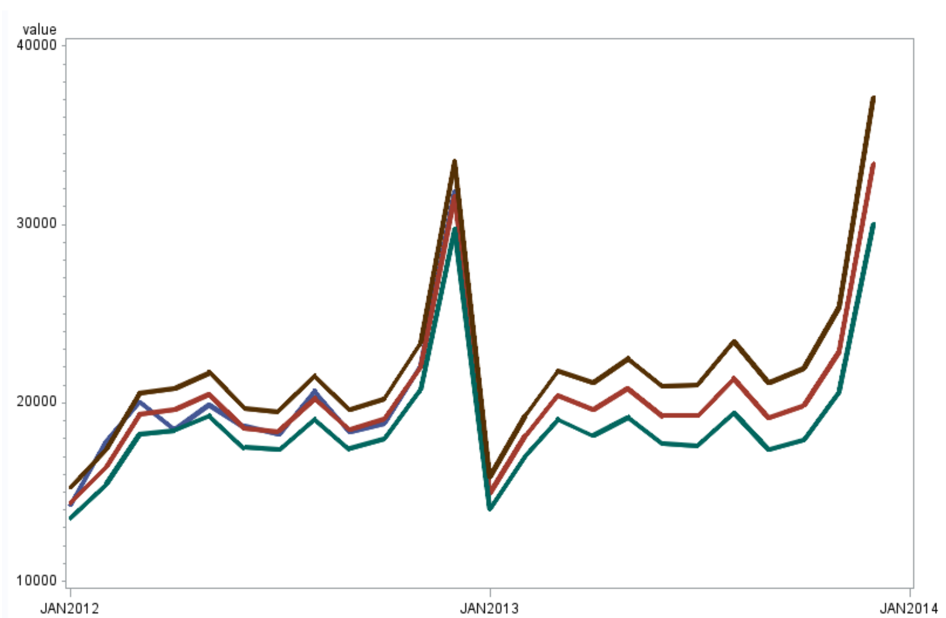
NON-TECHNICAL SUMMARY

The goal of the analysis was to model retail sales data for clothing & accessories stores, given a dataset of monthly sales data from January 1992 through December 2012. The data set contains 252 observations, which include clothing and accessory store monthly sales data (in millions), as reported by retailers selected by a rigorous selection process by the US Government, ensuring sample representivity for the US. (More detail: http://www.census.gov/retail/mrts/how_surveys_are_collected.html)

The data appears to have an increasing average value over time, which is to be expected given inflation, at the very least. Additionally, there's a sharp spike in sales data around the holiday season every year in December. In order to properly predict future sales data, this increasing mean and sharp spike in December needs to be removed. A few transformations were made to the data to make it suitable for modeling. Initially, the log is taken for each datapoint. This will (with later transforms) serve to stabilze the variation in the data. Next, by taking the change in log value month-to-month (rather than the actual value every month), we can eliminate the upward trend of the data over time. Finally, if we take the difference between the current month and a month twelve months out, then the spike in December is accounted for.

The final model fitted to our data takes these transformations into account and fits a model to the dataset that predicts the values well. Our residuals, or the difference between the actual data point and our predicted data point, show a relatively consistent pattern across all points in our data; our model does not favor any set of values in the data. We've predicted the sales data for 2013, and in early returns for January that the government has begun to collect, the value they have reported is within our range of values.

In the chart below, the purple solid line in the first half of the chart are our actual values. The red dotted line is the predicted value, and the the brown dotted line (the "top" line) is the max prediction and the blue dotted line (the "bottom" line) is the minimum prediction. Our model follows the trend of the actual 2012 data well, and will ideally follow 2013 beyond January, as well.



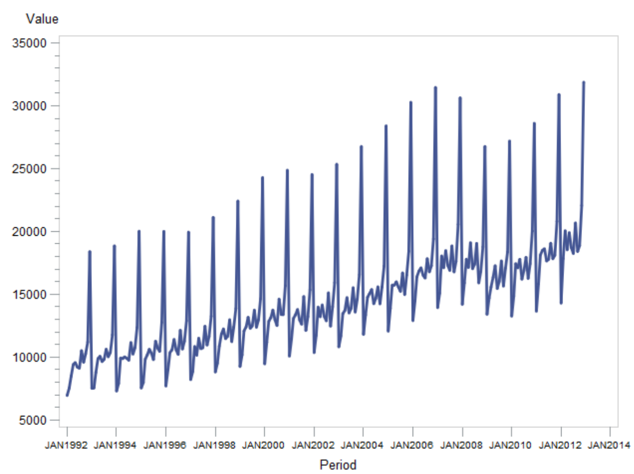
TECHNICAL SUMMARY

DATASET

The dataset used for this analysis includes monthly sales data for clothing & accessories stores in the US from January 1992 through December 2012. There are 252 observations on sales data (in millions), as reported by retailers selected by a rigorous selection process by the US Government, ensuring sample representivity for the US. (More detail is available at http://www.census.gov/retail/mrts/how_surveys_are_collected.html). The variables present in the dataset are period (month) and value (sales data, in millions).

EXPLORATORY ANALYSIS

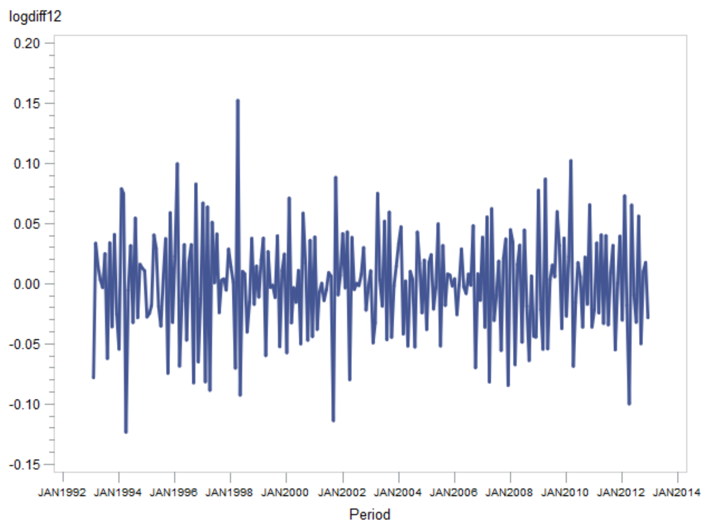
Evaluating a plot of values over time, the data shows strong evidence of seasonality, with a sharp jump at the 12th month in every year (understood to be a jump for holiday sales). Additionally, there appears to be a positive trend in the data, as well, with the mean increasing year over year.



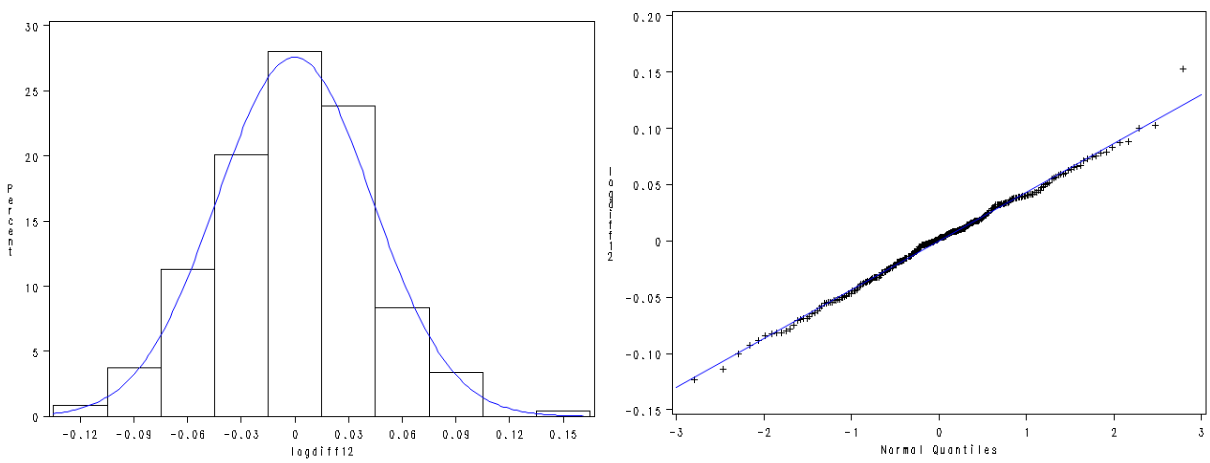
Given the data is not adequate for analysis as-is, several steps are taken to clean and prepare the data for fitting a time series model:

- Given increasing average, the difference of the data is taken (value minus the first lag). The trend is gone, but there is still evidence of seasonality. (See Appendix, figure 2)
- Next, the lag12 difference is taken to account for seasonality. The trend and seasonality are both gone, but the variance is not constant. (See Appendix, figure 3)
- The log transform is computed on the original data in order to stabilize the variance. Trend and seasonality are still evident on the log transform. (See Appendix, figure 4)
- The difference is taken to account for the trend. (See Appendix, figure 5)
- Removed seasonality by taking the lag12 difference. Data now shows constant mean and variance, with no obvious trends or seasonality present. (See Appendix, figure 6)

As noted, taking the log transform, first difference in the data, followed by the 12th difference in the data removes all evidence of a trend as well as seasonality. The visual representation of this transformed data shows all evidence of seasonality and trend removed, and it also appears to have a constant mean and variance, as well.



Visual tests of normality show that the data appears to be normally distributed, and more detailed tests evaluating normality, skewness, and kurtosis all show our data is normal. (See Appendix, section 1.8 for detailed tests of normality.)



Once the appropriate transformations of the data were identified and the resulting values on which to fit a time series model were normal, the model fitting process could begin.

MODEL FITTING

Given the type of seasonality seen in the data, the first model used as an attempt to fit the data is the Airline Model, or a seasonal model. The correlations of the differenced values are analyzed, and assumptions needed to use the Airline Model appear to be fulfilled: the data is highly correlated at lags 1, 11, 12, and 13.

Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.083639	1.00000	********************	0
1	-0.033811	-.40425	******** .	0.063119
2	-0.013499	-.16139	*** .	0.072706
3	-0.0000274	-.00033	. .	0.074120
4	0.0093467	0.11175	. **.	0.074120
5	-0.0012606	-.01507	. .	0.074788
6	-0.0036160	-.04323	. * .	0.074800
7	-0.0015019	-.01796	. .	0.074900
8	0.010023	0.11983	. **.	0.074917
9	-0.0003364	-.00402	. .	0.075676
10	-0.013927	-.16651	*** .	0.075677
11	-0.032247	-.38555	******** .	0.077123
12	0.078780	0.94190	. *******************	0.084454
13	-0.032525	-.38887	******** .	0.119170
14	-0.012285	-.14688	. *** .	0.124123
15	0.00001757	0.00021	. .	0.124813
16	0.0088407	0.10570	**	0.124813

All other autocorrelations in the data are zero, or in other words, are not statistically significant from zero.

The first attempt to fit the data is using an additive MA(1, 12, 13) on the differenced data. However, the residuals show evidence of serial correlations, which means the model is not an adequate fit.

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	29.50	4	<.0001	-0.101	0.017	0.236	-0.146	0.052	0.175
12	60.67	10	<.0001	-0.162	0.084	0.153	-0.201	0.144	0.082
18	77.04	16	<.0001	-0.151	0.154	0.016	-0.101	0.064	0.052
24	107.66	22	<.0001	-0.175	0.099	-0.042	-0.111	0.228	-0.095
30	143.06	28	<.0001	-0.084	0.104	-0.212	-0.066	0.027	-0.248
36	162.05	34	<.0001	-0.026	0.097	-0.171	0.023	0.069	-0.152
42	186.28	40	<.0001	0.149	-0.067	-0.158	0.108	-0.116	-0.085

Additionally, the residuals of this model have a high negative correlation at the 10th lag, which may be something to incorporate into the final model.

Several models were attempted via trial-and-error. Ultimately, the best fit model using our log transform variable is an ARIMA(1,2)x(10,12)12, which takes into the first and second difference as well as a 10th and 12th seasonal lag. All coefficients of this model are significant (values are significantly different from zero), which means all parameters used in the model are necessary and fit our dataset appropriately.

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.16051	0.06596	2.43	0.0150	10
MA2,1	0.47026	0.06215	7.57	<.0001	12
AR1,1	-0.72471	0.06066	-11.95	<.0001	1
AR1,2	-0.40097	0.06020	-6.66	<.0001	2

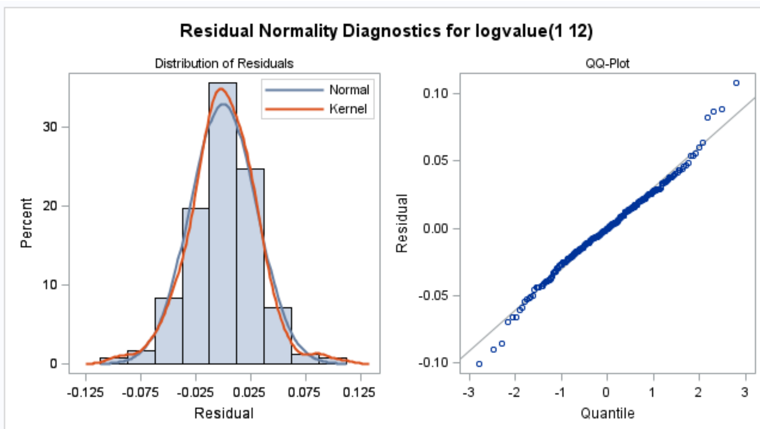
The final model can be written as:

$$(1 - 0.725B + 0.401B^2)X_t = (1 - 0.161B^{10})(1 - 0.470B^{12})a_t$$

RESIDUAL ANALYSIS AND MODEL DIAGNOSTICS

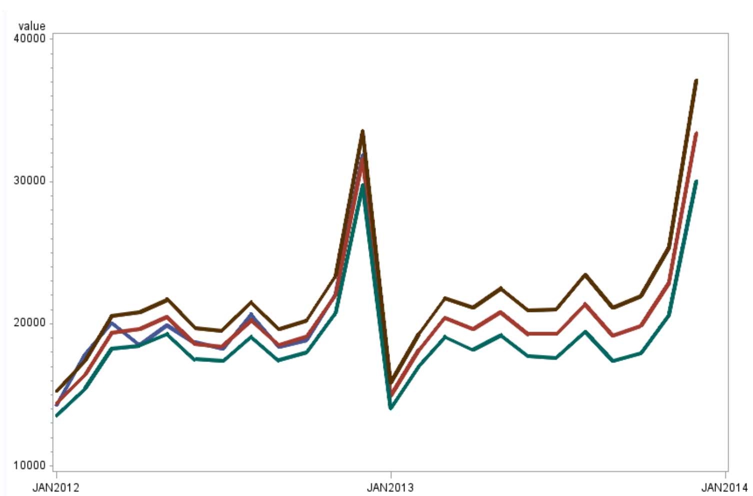
Using the fitted model, our residuals are white noise (autocorrelations are not significantly different from zero), which shows no evidence of serial correlation. Additionally, the residuals appear to be relatively normally distributed, as well.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.35	2	0.0690	0.010	-0.035	-0.097	-0.078	0.034	0.063
12	11.90	8	0.1555	0.101	0.043	0.092	0.008	0.075	0.012
18	16.49	14	0.2844	0.071	0.091	-0.003	-0.031	-0.012	0.059
24	30.75	20	0.0586	0.130	-0.013	-0.035	-0.021	0.186	-0.026
30	46.73	26	0.0075	0.035	0.028	-0.115	-0.050	-0.064	-0.191
36	50.41	32	0.0203	0.035	0.078	-0.045	-0.016	0.054	-0.027
42	62.75	38	0.0070	0.103	-0.087	-0.118	0.024	-0.100	-0.005



These results further confirm that the fitted model adequately explains the time series found in the data.

FORECAST ANALYSIS



Evaluating the trend of our data in 2012, our model appears to forecast at least the trend of the actual values rather well. Projecting ahead through 2013, the results our model generates are:

Period	Forecast	Lower 95% Conf Limit	Upper 95% Conf Limit
JAN2013	14936.14	14063.29	15848.42
FEB2013	18155.60	17055.57	19307.25
MAR2013	20399.92	19077.51	21788.94
APR2013	19623.76	18183.13	21147.11
MAY2013	20806.31	19204.16	22505.25
JUN2013	19269.43	17699.50	20940.04
JUL2013	19290.11	17625.27	21068.51
AUG2013	21379.10	19454.26	23411.30
SEP2013	19188.63	17386.08	21125.80
OCT2013	19864.61	17923.25	21957.27
NOV2013	22939.91	20665.29	25394.60
DEC2013	33410.02	30001.74	37096.55

An advanced estimate for clothing & accessory store retail sales data for January 2013 (based on early reports from a small sampling of firms) was $15,119 million (or $15 billion). The actual value is well within the confidence interval for our forecast. At least for that one new data point, our model appears to predict the sales data well.
(http://content.govdelivery.com/attachments/USESAEI/2013/02/13/file_attachments/190485/Advance%2BMonthly%2BSales%2Bfor%2BRetail%2Band%2BFood%2BServices%2B%2528January%2B2013%2529.pdf)

Additionally, looking at the backtesting results of the selected model versus the simpler airline model, we see a marginally better MAFE, MSFE, and RMSFE.
Backtest of accepted model:

Backtest results for sales
Model: VAR=logvalue DIFF=(1,12) p=(1,2) q=(10)(12) DATE=period TRAINPCT=80

Obs	_TYPE_	_FREQ_	mafe	msfe	rmsfe
1	0	50	0.025288	.001134414	0.033681

Backtest of rejected model:

Backtest results for sales
Model: VAR=logvalue DIFF=(1,12) q=(1)(12) DATE=period TRAINPCT=80

Obs	_TYPE_	_FREQ_	mafe	msfe	rmsfe
1	0	50	0.028404	.001469803	0.038338

ANALYSIS OF RESULTS AND DISCUSSION

As noted, our model assumptions appear to fit the sales data well. The first data point beyond our dataset for January 2013 (again, though not the true value, this is the predicted value based on actual results reported early) is well within our confidence range for predictions. Our parameters are all significant and our residuals appear to be white noise.

It should be noted that the first model (a simple, additive airline model MA(1,12,13)) was rejected because the residuals showed evidence of autocorrelation. Our selected model also has a lower MAFE, MSFE, and RMSFE than this rejected model. Despite our selected model being relatively simple, these results showed that the data needed more parameters than the most simple airline model provided. Additionally, the MAPE for our accepted model is 2.3%, whereas the MAPE of our rejected model is 2.4%.

Another rejected model had many more parameters and seemed to provide a better fit on the data, but the model failed to converge in the estimation process (likely due to being overparameterized). As seen in many cases of time series modeling, there may be multiple models that fit the data well (or some may have some diagnostics better than other models), which means there may be several models that might fit the data reasonably well, just as we've seen with this data set, as well.

APPENDIX

Figure 1: Plot of input values
Evidence of trend and seasonality

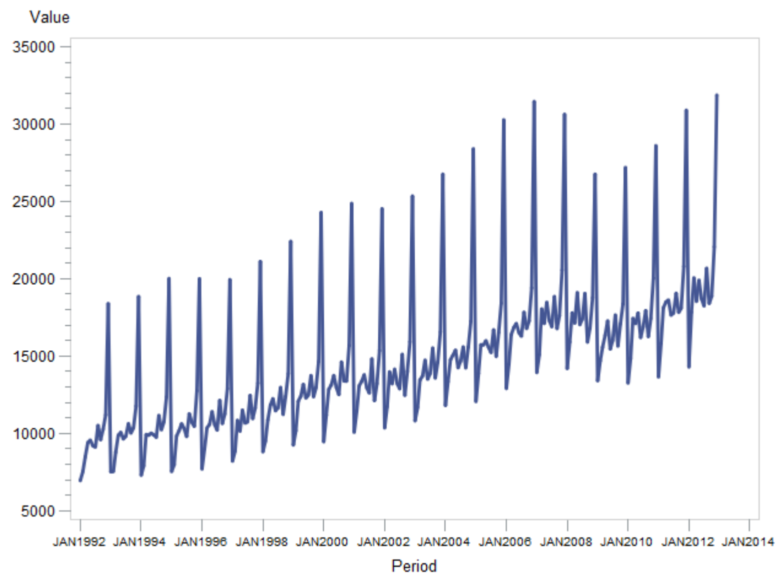


Figure 2: Plot of differenced input values
Evidence of seasonality; trend has been removed

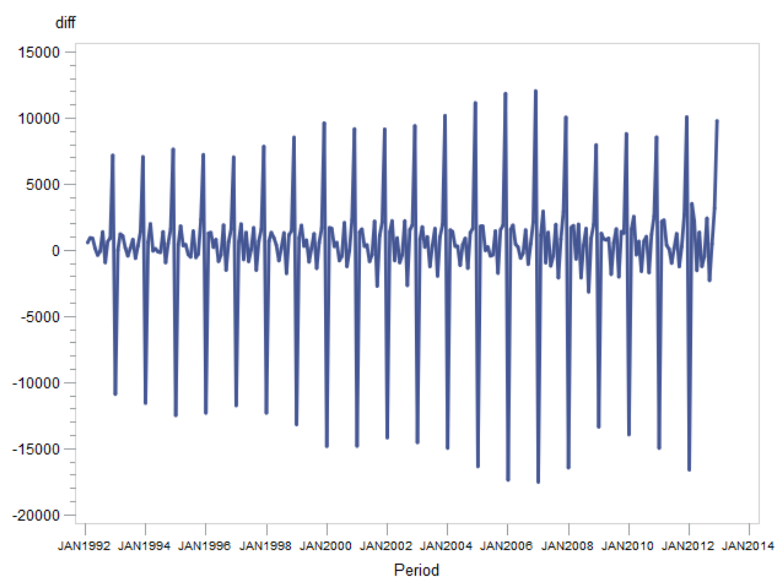


Figure 3: Plot of lag1 and lag12 differenced input values
Trend and seasonality are gone; data does not have a constant variance

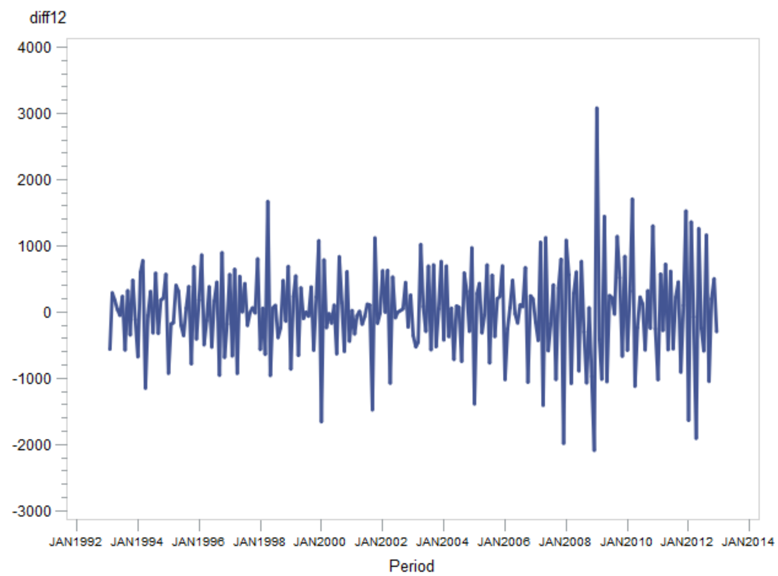


Figure 4: Plot of log transform (to stabilize variance)
Evidence of trend and seasonality remains

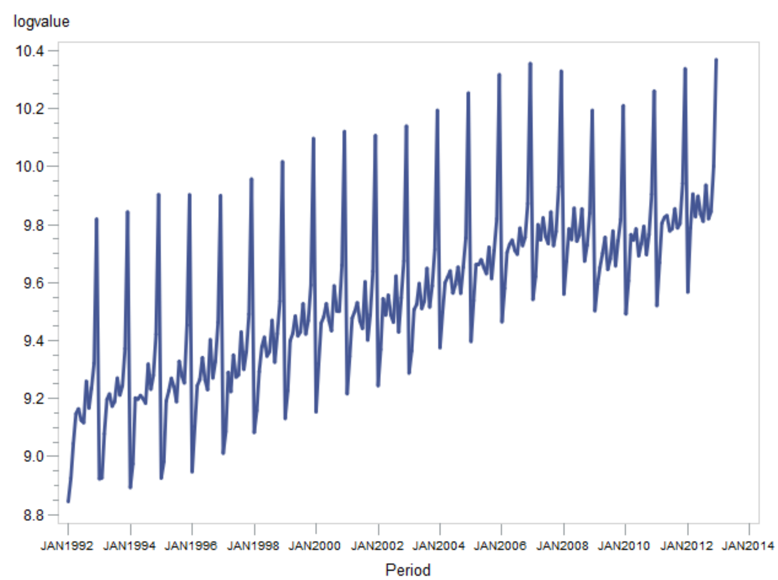


Figure 5: Plot of differenced log transform
Evidence of seasonality; trend has been removed

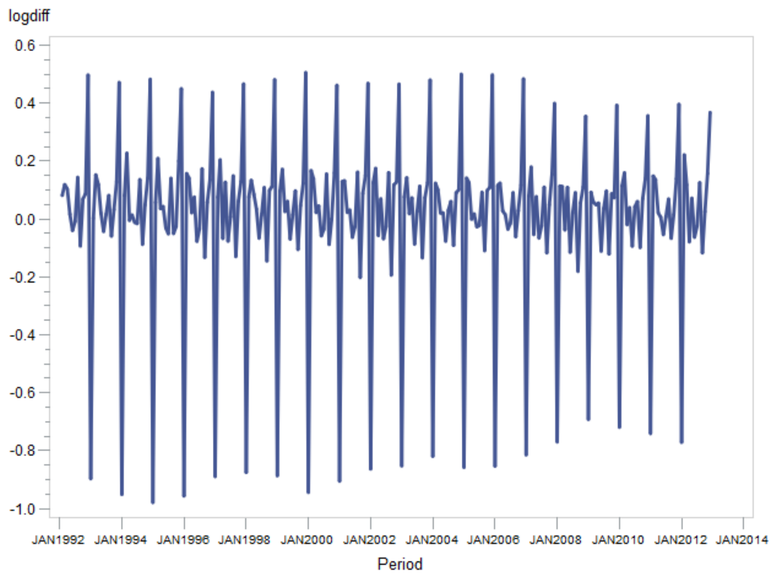
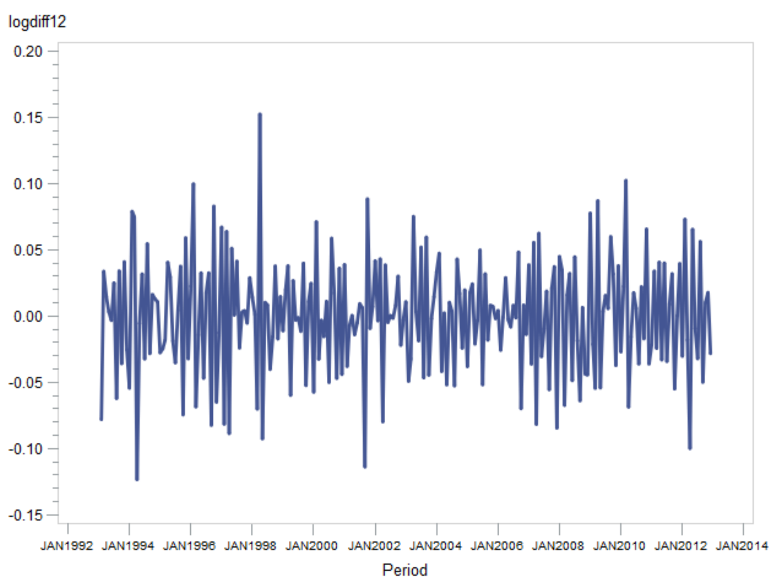
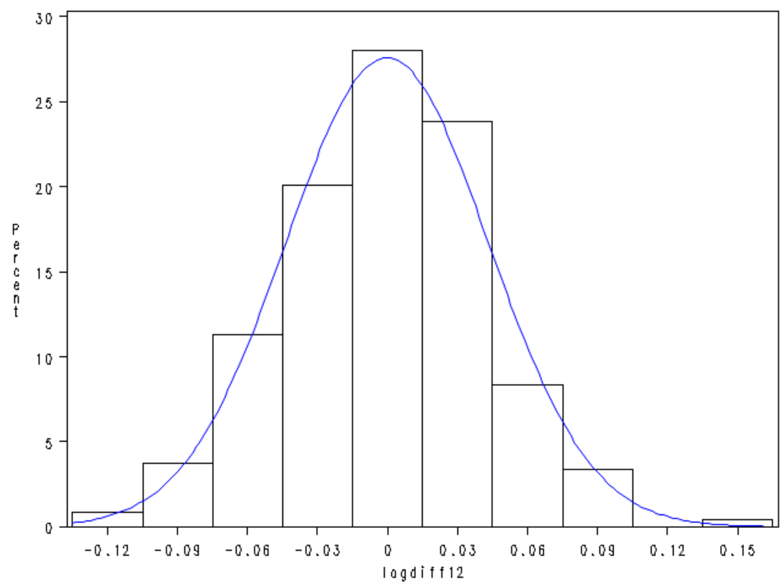
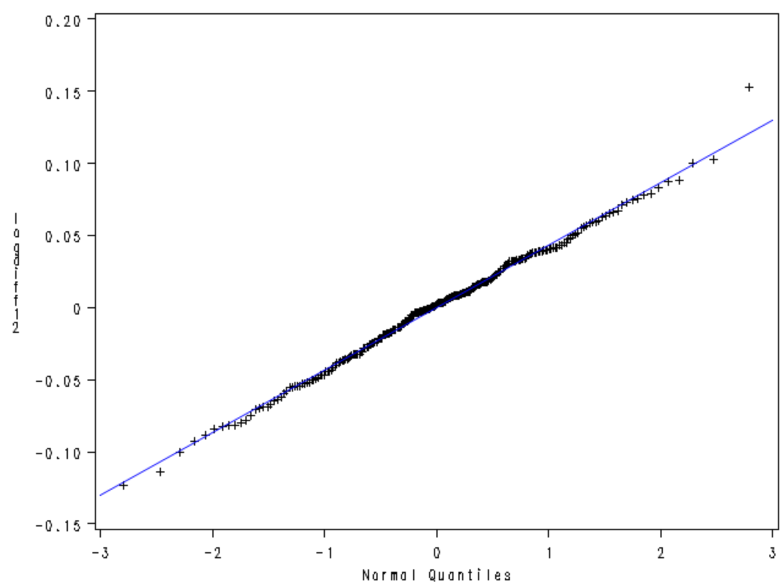


Figure 6: Plot of lag1 and lag12 differenced log transform
Trend and seasonality are gone; variance is more constant



Section/Figure 7: Tests of Normality





Results of test on skewness

Obs	skewness, logdiff12	skew_test	P-value for skewness test
1	-0.013194	-0.083269	1.06636

Results of test on kurtosis

Obs	kurtosis, logdiff12	kurt_test	P-value for kurtosis test
1	0.20296	0.64046	0.52187

Results of Jacque and Bera test on normality

Obs	skewness, logdiff12	kurtosis, logdiff12	Jarque & Bera statistic	P-value for Jarque & Bera test
1	-0.013194	0.20296	0.41713	0.81175

Figure/Section 8: Analysis of autocorrelations for applicability of airline model.

Autocorrelations				
Lag	Covariance	Correlation	-1 9 8 7 6 5 4 3 2 1 0 1 2 3 4 5 6 7 8 9 1	Std Error
0	0.083639	1.00000	|********************|	0
1	-0.033811	-.40425	********| .	0.063119
2	-0.013499	-.16139	***| .	0.072706
3	-0.0000274	-.00033	. | .	0.074120
4	0.0093467	0.11175	. |**.	0.074120
5	-0.0012606	-.01507	. | .	0.074788
6	-0.0036160	-.04323	. *| .	0.074800
7	-0.0015019	-.01796	. | .	0.074900
8	0.010023	0.11983	. |**.	0.074917
9	-0.0003364	-.00402	. | .	0.075676
10	-0.013927	-.16651	***| .	0.075677
11	-0.032247	-.38555	********| .	0.077123
12	0.078780	0.94190	. |*******************	0.084454
13	-0.032525	-.38887	********| .	0.119170
14	-0.012285	-.14688	. ***| .	0.124123
15	0.00001757	0.00021	. | .	0.124813
16	0.0088407	0.10570	. |** .	0.124813
17	-0.0011815	-.01413	. | .	0.125169
18	-0.0035643	-.04262	. *| .	0.125176
19	-0.0014437	-.01726	. | .	0.125234
20	0.0099040	0.11841	. |** .	0.125243
21	-0.0007335	-.00877	. | .	0.125688
22	-0.013261	-.15855	. ***| .	0.125691
23	-0.030325	-.36257	*******| .	0.126485
24	0.074509	0.89083	. |******************	0.130560
25	-0.031058	-.37134	*******| .	0.152870
26	-0.011128	-.13304	. ***| .	0.156423
27	-0.0003212	-.00384	|	0.156873

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	51.91	6	<.0001	-0.404	-0.161	-0.000	0.112	-0.015	-0.043
12	338.11	12	<.0001	-0.018	0.120	-0.004	-0.167	-0.386	0.942
18	387.81	18	<.0001	-0.389	-0.147	0.000	0.106	-0.014	-0.043
24	657.35	24	<.0001	-0.017	0.118	-0.009	-0.159	-0.363	0.891
30	704.57	30	<.0001	-0.371	-0.133	-0.004	0.101	-0.011	-0.044
36	958.45	36	<.0001	-0.014	0.114	-0.013	-0.148	-0.344	0.841

Figure/Section 9: Results of additive MA(1,12,13) airline model

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	98.75	6	<.0001	-0.533	0.005	0.206	-0.245	0.105	0.095
12	164.85	12	<.0001	-0.189	0.104	0.119	-0.244	0.276	-0.257
18	174.18	18	<.0001	-0.006	0.126	-0.013	-0.055	0.005	0.130
24	199.23	24	<.0001	-0.170	0.098	-0.027	-0.095	0.195	-0.092

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	1	-836.228	0.0001	-20.44	<.0001		
	3	7703.597	0.9999	-10.94	<.0001		
	5	6489.729	0.9999	-7.63	<.0001		
Single Mean	1	-836.231	0.0001	-20.40	<.0001	208.04	0.0010
	3	7705.760	0.9999	-10.92	<.0001	59.59	0.0010
	5	6512.051	0.9999	-7.61	<.0001	28.97	0.0010
Trend	1	-836.253	0.0001	-20.35	<.0001	207.15	0.0010
	3	7634.667	0.9999	-10.89	<.0001	59.35	0.0010
	5	6222.075	0.9999	-7.60	<.0001	28.88	0.0010

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.66912	0.04858	13.77	<.0001	1
MA2,1	0.50583	0.05884	8.60	<.0001	12

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	29.50	4	<.0001	-0.101	0.017	0.236	-0.146	0.052	0.175
12	60.67	10	<.0001	-0.162	0.084	0.153	-0.201	0.144	0.082
18	77.04	16	<.0001	-0.151	0.154	0.016	-0.101	0.064	0.052
24	107.66	22	<.0001	-0.175	0.099	-0.042	-0.111	0.228	-0.095
30	143.06	28	<.0001	-0.084	0.104	-0.212	-0.066	0.027	-0.248
36	162.05	34	<.0001	-0.026	0.097	-0.171	0.023	0.069	-0.152
42	186.28	40	<.0001	0.149	-0.067	-0.158	0.108	-0.116	-0.085

Model for variable logvalue	
Period(s) of Differencing	1,12

No mean term in this model.

Moving Average Factors	
Factor 1:	1 - 0.66912 B**(1)
Factor 2:	1 - 0.50583 B**(12)

Figure/Section 10: Results of multiplicative ARIMA(1,2)x(10,12)12 airline model

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	98.75	6	<.0001	-0.533	0.005	0.206	-0.245	0.105	0.095
12	164.85	12	<.0001	-0.189	0.104	0.119	-0.244	0.276	-0.257
18	174.18	18	<.0001	-0.006	0.126	-0.013	-0.055	0.005	0.130
24	199.23	24	<.0001	-0.170	0.098	-0.027	-0.095	0.195	-0.092

Augmented Dickey-Fuller Unit Root Tests							
Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	1	-836.228	0.0001	-20.44	<.0001		
	3	7703.597	0.9999	-10.94	<.0001		
Single Mean	1	-836.231	0.0001	-20.40	<.0001	208.04	0.0010
	3	7705.760	0.9999	-10.92	<.0001	59.59	0.0010
Trend	1	-836.253	0.0001	-20.35	<.0001	207.15	0.0010
	3	7634.667	0.9999	-10.89	<.0001	59.35	0.0010

Maximum Likelihood Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.16051	0.06596	2.43	0.0150	10
MA2,1	0.47026	0.06215	7.57	<.0001	12
AR1,1	-0.72471	0.06066	-11.95	<.0001	1
AR1,2	-0.40097	0.06020	-6.66	<.0001	2

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.35	2	0.0690	0.010	-0.035	-0.097	-0.078	0.034	0.063
12	11.90	8	0.1555	-0.101	0.043	0.092	0.008	0.075	0.012
18	16.49	14	0.2844	-0.071	0.091	-0.003	-0.031	-0.012	0.059
24	30.75	20	0.0586	-0.130	-0.013	-0.035	-0.021	0.186	-0.026
30	46.73	26	0.0075	-0.035	0.028	-0.115	-0.050	-0.064	-0.191
36	50.41	32	0.0203	-0.035	0.078	-0.045	-0.016	0.054	-0.027
42	62.75	38	0.0070	0.103	-0.087	-0.118	0.024	-0.100	-0.005

Model for variable logvalue	
Period(s) of Differencing	1,12

No mean term in this model.

Autoregressive Factors	
Factor 1:	1 + 0.72471 B**(1) + 0.40097 B**(2)

Moving Average Factors	
Factor 1:	1 - 0.16051 B**(10)
Factor 2:	1 - 0.47026 B**(12)

Figure/Section 11: Backtest of accepted model

Backtest results for sales
Model: VAR=logvalue DIFF=(1,12) p=(1,2) q=(10)(12) DATE=period TRAINPCT=80

Obs	_TYPE_	_FREQ_	mafe	msfe	rmsfe
1	0	50	0.025288	.001134414	0.033681

Figure/Section 12: Backtest of rejected model

Backtest results for sales
Model: VAR=logvalue DIFF=(1,12) q=(1)(12) DATE=period TRAINPCT=80

Obs	_TYPE_	_FREQ_	mafe	msfe	rmsfe
1	0	50	0.028404	.001469803	0.038338

Figure 13: Plot of forecast and confidence limits

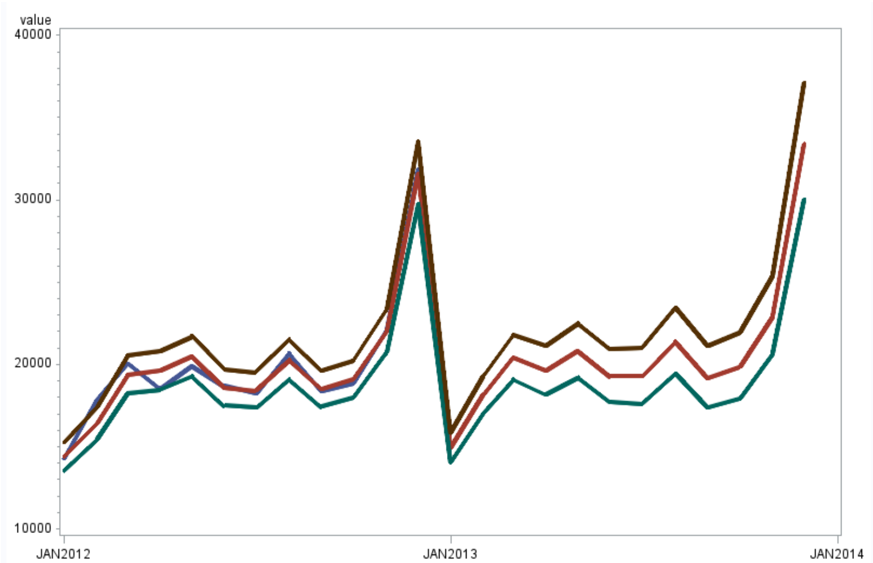


Figure 14: MAPE for accepted model

MAPE
2.3%

SAS SOURCE CODE:

```
options nodate;
ods listing close;

%let MY_FOLDER = D:\Users\lfloyd\Documents\School Misc;
%include "&MY_FOLDER\backtest_macro.sas";

title "Step 1: Import";
proc import datafile="&MY_FOLDER\RetailTimeSeries - not adj.csv" out=sales
replace;
delimiter = ",";
getnames = yes;
run;
title;

title "Step 2: A quick plot of the input values.";
title2 "(Analysis: Evidence of trend and seasonality.)";
symbol interpol = join;  * gives connected lines ;
proc gplot data=sales;
plot value * period;
run;
title;

title "Step 3: Plot shows positive trend. Take difference.";
data sales;
set sales;
diff = value - lag(value);
run;
title;

title "Step 4: Plot after taking difference.";
title2 "(Analysis: Trend gone, but still seasonality.)";
symbol interpol = join;  * gives connected lines ;
proc gplot data=sales;
plot diff * period;
run;
title;

title "Step 5: Remove seasonality by taking lag12 difference.";
data sales;
set sales;
diff12 = diff - lag12(diff);  * remove seasonailty ;
run;
title;

title "Step 6: Plot after taking lag12 difference.";
title2 "(Analysis: Now trend *and* seasonality are gone, but";
title3 "variance is not constant.)";
symbol interpol = join;  * gives connected lines ;
proc gplot data=sales;
```

```
plot diff12 * period;
run;
title;


title "Step 7: Stabilize variance by using log transform.";
data sales;
set sales;
logvalue = log(value);
run;
title;


title "Step 8: Plot after log transform.";
title2 "(Analysis: Need to once again remove trend and seasonality.)";
symbol interpol = join;  * gives connected lines ;
proc gplot data=sales;
plot logvalue * period;
run;
title;


title "Step 9: Plot still shows positive trend, so take the difference.";
data sales;
set sales;
logdiff = logvalue - lag(logvalue);
run;
title;


title "Step 10: Plot after taking difference.";
title2 "(Analysis: Trend gone, but still seasonality.)";
symbol interpol = join;  * gives connected lines ;
proc gplot data=sales;
plot logdiff * period;
run;
title;


title "Step 11: Remove seasonality by taking lag12 difference.";
data sales;
set sales;
logdiff12 = logdiff - lag12(logdiff);  * remove seasonailty ;
run;
title;


title "Step 12: Plot after taking lag12 difference.";
title2 "(Analysis: Now trend *and* seasonality are gone";
title3 "and variance is more constant.)";
symbol interpol = join;  * gives connected lines ;
proc gplot data=sales;
plot logdiff12 * period;
run;
title;



title "Step 13: Analyze data, check for normality.";
```

```
title2 "(Analysis: Yes, histogram and qq-plot indicate normality.)";
goption reset=global;
proc univariate data=sales;
var logdiff12;
histogram/normal;
qqplot / normal (mu=est sigma=est);
* output out=stats kurtosis=kurtosis skewness=skewness N=ntot;
run;
title;


goptions reset=global;
proc univariate;
var logdiff12;
histogram/normal;
probplot / normal(mu=est sigma=est);
output out=stats kurtosis=kurtosis skewness=skewness N=ntot;
run;


/* steps to compute skewness, kurtosis and Jarque-Bera tests*/
data computation;
set stats;
label pv_kur = "P-value for kurtosis test";
skew_test = skewness/sqrt(6/Ntot);
kurt_test = kurtosis/sqrt(24/Ntot);
jb = skew_test*skew_test+kurt_test*kurt_test;
pv_skew = 2* (1-cdf('NORMAL', skew_test));
pv_kur = 2*(1-cdf('NORMAL', kurt_test));
pv_jb = 1-cdf('CHISQUARE', jb,2);
label pv_kur = "P-value for kurtosis test"
pv_skew= "P-value for skewness test"
pv_jb = "P-value for Jarque & Bera test"
jb = "Jarque & Bera statistic";

/* Print out results of tests*/
Title " Results of test on skewness";
proc print data= computation label;
var skewness skew_test pv_skew;
run;
Title " Results of test on kurtosis";
proc print data= computation label;
var kurtosis kurt_test pv_kur;
run;
Title " Results of Jacque and Bera test on normality";
proc print data= computation label;
var skewness kurtosis jb pv_jb;
run;


title "Step 14: Check for applicability of airline model";
title2 "by analyzing correlations of differenced values.";
title3 "(Analysis: Highly correlated at lags 1, 11, 12, and 13.)";
proc arima data=sales;
identify var=logdiff nlag=36;
run;
title;
```

```
* Multiplicative models (see notes, Week 6, slide 15).
* For monthly time series, annual seasonality has s=12 and the ACF
* is not zero at lag 1,11,12 and 13 only.  Furthermore, we expect
* (lag1 coeff) * (lag12 coeff) approx = (-1) * (lag13 coeff).
* Check: (-0.40425) * (0.94190) approx = (-1) * (-0.38555) ==>
* -0.38076 approx = -0.38555, so YES! ;


title "Step 15: Try a simple airline model by fitting";
title2 "additive MA(1,12,13) model on differenced data.";
title3 "(Analysis: Residuals are correlated.  Try another model.)";
proc arima data=sales;
identify var=logvalue(1,12) stationarity=(adf=(1 3 5));
estimate q=(1)(12) noconstant;
run;
title;


title "Step 16: Through trial and error, we settled on this.";
title2 "(Analysis: All coefficients are significant and residuals are white
noise.)";
proc arima data=sales;
identify var=logvalue(1,12) nlag=24 stationarity=(adf=(1 3));
estimate p=(1,2) q=(10)(12) noconstant method=ml plot;
run;
title;


title "Step 17: Backtest of accepted model." ;
%backtest(trainpct=80, dataset=sales, date=period, var=logvalue,
      diff=(1,12), p=(1,2) q=(10)(12), interval=month, noconstant=Y);
run;


title "Step 18: Backtest of rejected model for comparison purposes." ;
%backtest(trainpct=80, dataset=sales, date=period, var=logvalue,
      diff=(1,12), q=(1)(12), interval=month, noconstant=Y);
run;


title "Step 19: Run accepted model again, this time writing forecasts to file.";
proc arima data=sales;
identify var=logvalue(1,12) nlag=24 stationarity=(adf=(1 3));
* estimate q=(1)(12) noconstant method=uls plot;
estimate p=(1,2) q=(10)(12) noconstant method=uls plot;
forecast out=forecasts lead=12 id=period interval=month noprint;
run;


title "Step 20: Retransform the forecast values to get forecasts in the original
scales.";
data retransform;
set forecasts;
value    = exp( logvalue );
forecast = exp( forecast + std*std/2 );
l95      = exp( l95 );
u95      = exp( u95 );
run;
```

```
title "Step 21: Plot the forecasts and their confidence limits.";
title2 "(Showing last two years only for readability.)";
goption reset=symbol;
symbol1 i=join width=3;
symbol2 i=join width=3;
symbol3 i=join width=3;
symbol4 i=join width=3;
proc gplot data=retransform;
where period >= '01Jan2012'd;
plot value * period
     forecast * period
     l95 * period
     u95 * period      /
     overlay haxis= '01Jan2012'd to '01Jan2014'd by year;
run;


title "Step 22: Compute Mean Absolute Percent Error (MAPE).";
data mape (keep=mape);
retain sum 0;
retain count 0;
set retransform end=eof;
where value ne . and forecast ne . ;
ape = abs(forecast - value) / value;
sum = sum + ape;
count = count + 1;
if (eof) then do;
   MAPE = sum / count;
   format MAPE percent7.1;
   output;
end;
run;

* Note the MAPE for the accepted model is 2.3%
* while the MAPE for the rejected model is 2.4% ;

proc print data=mape noobs;
run;
```

SAS BACKTEST MACRO:

```
%macro backtest(TRAINPCT=80, DATASET=, VAR=, DIFF=, P=, Q=, DATE=date,
INTERVAL=month, NOCONSTANT=N);

* ------------------------------------------------------------------------- ;
*   T I M E   S E R I E S   B A C K T E S T   M A C R O
* ------------------------------------------------------------------------- ;
*   DePaul CSC425, Winter 2013, Dr. Raffaella Settimi
*   Macro written by Bill Qualls, First Analytics
* ------------------------------------------------------------------------- ;
*  E X P L A N A T I O N   O F   P A R A M E T E R S
*  (Order of variables is insignificant)
* ------------------------------------------------------------------------- ;
*   TRAINPCT   .. Percent of dataset to be used for training.
*                 So, (100 - TRAINPCT)% will be used for evaluation.
*                 Example: TRAINPCT=80
*   DATASET    .. Time series dataset. Libname optional, defaults to Work.
*                 Example: DATASET=Work.Unemp
*   VAR        .. Name of time series variable.
*                 Example: VAR=ratechg
*   P          .. Specified for AR models. Omit otherwise.
*                 Example: P=(1 3 6)
*   Q          .. Specified for MA models. Omit otherwise.
*                 Example: Q=(1 3 6)
*   DATE       .. Name of date variable. Defaults to date.
*                 Example: DATE=date
*   INTERVAL   .. Date interval. Defaults to month.
*                 Example: INTERVAL=day
* ------------------------------------------------------------------------- ;
*   Additional parameters added 20130309 for final project
*   DIFF       .. Differencing. Default to none.
*                 Example: DIFF=(1,12)
*   NOCONSTANT .. Add NOCONSTANT if Y, otherwise omit.
*                 Example: NOCONSTANT=Y
* ------------------------------------------------------------------------- ;
*  S A M P L E   U S A G E
*  %backtest(trainpct=80, dataset=work.unemp, var=ratedif, p=(1), interval=day);
* ------------------------------------------------------------------------- ;


%put TRAINPCT=&TRAINPCT;
%put DATASET=&DATASET;
%put VAR=&VAR;
%put DATE=&DATE;
%put P=&P;
%put Q=&Q;
%put DIFF=&DIFF;
%put INTERVAL=&INTERVAL;
%put NOCONSTANT=&NOCONSTANT;


* How many records are in the dataset? ;
data _null_;
call symput('NRECS', trim(left(nrecs)));
set &DATASET nobs=nrecs;
stop;
run;
```

```
* Determine which ones are exclusively for training based on TRAINPCT ;
%let SIZE_OF_ROLLING_WINDOW = %sysfunc(round(&NRECS * &TRAINPCT / 100));

* create a working copy of dataset with observation number  ;
* as a variable for use in a where clause with proc arima.  ;
* also add a placeholder for the predicted value.           ;

data Work._MY_COPY_ (keep = &VAR &DATE _OBS_ _PRED_);
set &DATASET;
    _OBS_ = _N_;
    _PRED_ = .;
run;

* turn off log -- too lengthy ;
filename junk dummy;
proc printto log=junk print=junk;
run;

* Will build the model once for each record used in evaluation. ;
* Each time I will predict one record forward.                   ;

%let MODELS_TO_BE_BUILT = %sysevalf(&NRECS - &SIZE_OF_ROLLING_WINDOW);

%do i = 1 %to &MODELS_TO_BE_BUILT;

    * Model using SIZE_OF_ROLLING_WINDOW records, and make one prediction ;
    proc arima data=Work._MY_COPY_ plots=none;
    where _OBS_ ge &i
        and _OBS_ le (&i + &SIZE_OF_ROLLING_WINDOW - 1);
    identify var=&VAR &DIFF noprint;
    estimate
        %if ("&P" ne "") %then %do; p=&P %end;
        %if ("&Q" ne "") %then %do; q=&Q %end;
            %if ("&NOCONSTANT" eq "Y") %then %do; NOCONSTANT %end;
        method=ml noprint;
    forecast lead=1 id=&DATE interval=&INTERVAL out=Work._MY_RESULTS_ noprint;
    run;

    * get the predicted value (in the last record) as a macro variable ;
    data _null_;
    p = nrecs;
    set Work._MY_RESULTS_ point=p nobs=nrecs;
    call symput("FORECAST", forecast);
    stop;
    run;

    * move that prediction to its place in the output file ;
    proc sql noprint;
    update Work._MY_COPY_
        set _PRED_ = &FORECAST
        where _OBS_ = &i + &SIZE_OF_ROLLING_WINDOW;
    quit;
    run;

    * show progress so far ;
    %if (%sysfunc(mod(&i, 20)) = 0) %then %do;

            * print on;
```

```
        proc printto log=log print=print;
        run;

        %put Finished &i iterations;

        * print off again;
        proc printto log=junk print=junk;
        run;

    %end;

%end;

* turn print back on ;
proc printto log=log print=print;
run;

* calculate prediction error;
data Work._MY_COPY_;
set Work._MY_COPY_;
Predicted_Error_Squared = (&VAR - _PRED_) ** 2;
run;

* turn print back on ;
* proc printto;
* run;

* calculate prediction error;
data Work._MY_COPY_;
set Work._MY_COPY_;
Predicted_Error = (&VAR - _PRED_) ;
Predicted_Error_Squared = (&VAR - _PRED_)**2;
absresidual = abs(Predicted_Error);
run;

* compute and report the mean square forecast error;
%if ("&P" eq "") %then %let PP = ; %else %let PP = p=&P;
%if ("&Q" eq "") %then %let QQ = ; %else %let QQ = q=&Q;

title "Backtest results for &DATASET";
title2 "Model: VAR=&VAR DIFF=&DIFF &PP &QQ DATE=&DATE TRAINPCT=&TRAINPCT";
proc summary data=Work._MY_COPY_;
where _OBS_ > &SIZE_OF_ROLLING_WINDOW;
var Predicted_Error absresidual;
output out=outm mean(absresidual)=mafe mean(Predicted_Error_Squared)=msfe;
run;

data outm;
set outm;
rmsfe=sqrt(msfe);
run;

proc print data=outm;
run;

%mend backtest;
```