|  |
| --- |
| STAT 6820 Time Series Analysis |
| Monthly Carbon Dioxide Level Analysis |
| Project 2 due 11/18/2019 |

|  |
| --- |
| Minjiao Yang  11-18-2019 |

**Table of Contents**

1. Introduction
2. Stationary and Transformation of Data
3. Model Selection by AIC/BIC/RMSE/LBP-test
4. Candidate 1: ARIMA(0,1,1)x(0,1,1)12 Model
5. Candidate 2&3: ARIMA(0,1,(1,9,18))x(0,1,1)12 Model & ARIMA(0,1,(1,18))x(0,1,1)12 Model
6. Best model: ARIMA(0,1,(1,9,18))x(0,1,1)12 Model
7. Model Selection by MAPE value
8. Forecast
9. Conclusion
10. Reference
11. Appendix
12. Summary of 25 Possible Models
13. Summary of ARIMA(0,1,1)x(0,1,1)12 Model
14. Summary of ARIMA(0,1,(1,9,18))x(0,1,1)12 Model
15. Summary of ARIMA(0,1,(1,18))x(0,1,1)12 Model
16. SAS code

**Introduction**

Carbon dioxide is a greenhouse gas, which absorbs heat. Warmed by sunlight, Earth’s land and ocean surface continuously radiate thermal infrared energy (heat). Unlike oxygen or nitrogen (which make up most of our atmosphere), greenhouse gases absorb that heat and release it gradually over time. Increases in greenhouse gases have tipped the Earth’s energy budget out of balance, trapping additional heat and raising Earth’s average temperature. According to the State of the Climate in 2018 report from NOAA and the American Meteorological Society, global atmospheric carbon dioxide was 408.53 0.1 ppm in 2019. Carbon-dioxide levels today are higher than at any point in at least the past 800,000 years.

Levels of carbon dioxide (CO2) are monitored at several sites around the world to investigate atmospheric changes. One of the sites is at Alert, Northwest Territories, Canada, near the arctic circle. A record from carbon dioxide level monitoring site shown the Monthly CO2 levels from January 1994 to December 2004. We want to identify trends and seasonality of this data set and fit a time series model to forecast the future carbon dioxide level.

**Stationary and Transformation of Data**

The data set contains 132 records of monthly carbon dioxide levels from January 1994 to December 2004.



A screenshot of a cell phone

Description automatically generated

CO2 Level



Figure 1. Times series graph of monthly CO2 level from Jan.1994 to Dec.2004

In figure 1, the level of CO2 increases over the year. The overall maximum Co2 level is 383.58ppm in April 2004. And the CO2 level reaches a peak in May every year. So, the time series appears to have an overall trend and seasonal effects.

A close up of a logo

Description automatically generated

Figure 2. Trend and Seasonality decomposition of Time Series

In figure 2, the time series is strongly seasonal with an overall increasing trend and obviously non-stationary. We know that stationary is important because, in its absence, a model describing the data will vary inaccuracy at different time points. In order to achieve stationary, we performed box-cox transformation for the data.

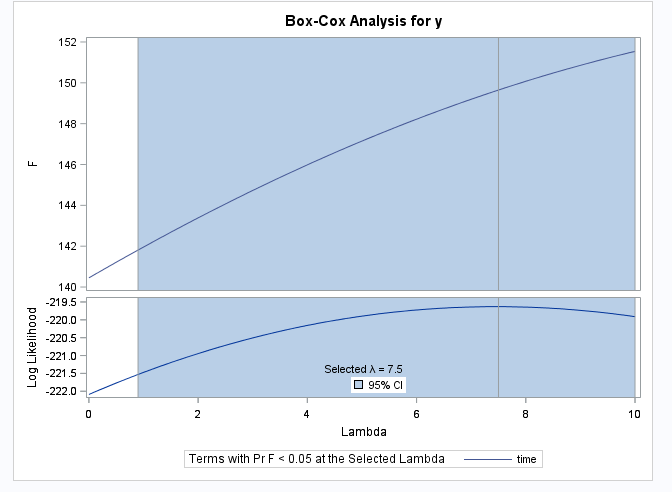


Figure 3. Box-Cox Transformation for y

However, BOX-COX transformation suggested the best lambda is 7.5. lambda=1 is included in a 95% Confidence Interval of log-likelihood indicating transformation on data is unnecessary. Then, we did simple differencing on data to remove overall trend and seasonal differencing to remove seasonality, the resulting time series finally achieved stationary in both mean and variance (as shown in Figure 5).

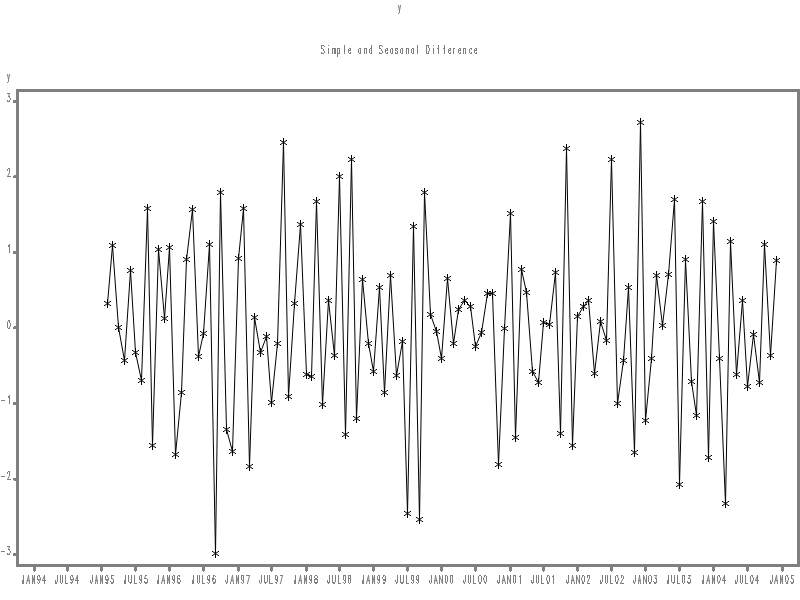


Figure 5. time series graph after adjustment

**Model Selection By AIC/BIC/LBP Test**

1. Candidate 1: ARIMA(0,1,1)x(0,1,1)12

A screenshot of a video game

Description automatically generated

Figure 6. Autocorrelation plots of simple and seasonally differenced data (lag<12)

In the autocorrelation plots of simple and seasonally differenced data, In the non-seasonal lags (lag<12), there are 2 significant spikes in PACF at lag1& 2, while ACF decays exponentially. This suggests a possible AR(2) term. On the other hand, there is 1 significant spike in ACF at lag 1, while PACF decays exponentially. This suggests a possible MA(1) term.

A screenshot of a computer

Description automatically generated

Figure 7. Autocorrelation plots of simple and seasonally differenced data (more lags)

In addition, there are spikes in the PACF at lag 12 and 24, while ACF decays exponentially at lag 12, 24, etc. This may be suggestive of a seasonal term AR(2) term. On the other hand, there is a spike in the ACF at lag 12, while PACF decays exponentially at lag 12, 24. This may be suggestive of a seasonal term MA(1).

Consequently, this initial analysis suggests that possible models for these data are ARIMA(2,1,0)x(2,1,0)12, ARIMA(2,1,0)x(2,1,0)12, ARIMA(0,1,1)x(2,1,0)12, and ARIMA(0,1,1)x(0,1,1)12. We fit these models, along with some variations on them (in total 25 models), compute the Akaike information criterion (AIC=2k-2ln(loglikelihood)), Bayesian information criterion (BIC=2nk-2ln(loglikelihood) and look at the Ljung–Box test (LBP-test) (details shown in Appendix). The models ARIMA(0,1,1)x(0,1,2)12, ARIMA(0,1,1)x(1,1,1)12, and ARIMA(0,1,1)x(0,1,1)12 are selected as the best three models based on AIC/BIC criteria (shown in table.1).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AIC | BIC | LBP test(lag>0.05) | Significant vbles |
| ARIMA(0,1,1)x(0,1,2)12 | 254.182 | 262.201 | Lag 6, 12, 54, 60 pass | MA at lag 24 is not |
| ARIMA(0,1,1)x(1,1,1)12 | 254.184 | 262.202 | Lag 6, 12, 54, 60 pass | AR at lag 12 is not |
| ARIMA(0,1,1)x(0,1,1)12 | 252.248 | 257.593 | Lag 6, 12, 48,54, 60 pass | All variables are significant |

Table 1. best 3 Models among 25 Comparison (all models without intercept)

Among three models, ARIMA(0,1,1)x(0,1,1)12 is selected as the first candidate model because:

* A preferred model should have the smallest AIC and BIC. ARIMA(0,1,1)x(0,1,1)12 has the least AIC/BIC among 3 models.
* Residuals of a model should behave like a white noise series. LBP-test is to check if residuals are correlated with each other with null hypothesis “data are independently distributed”. If the p-value of lags is all greater than 0.05 then it passes the LBP-test (accept null hypothesis), the residuals are not correlated with each other.

None of the 3 models passed LBP Test at all lags. While, ARIMA(0,1,1)x(0,1,1)12 passed LBP test at more lags.

* It’s the only one with all variables that are significant. (p-value < 0.05)
* If we take out the insignificant variables from the other two models, both of them will end up with ARIMA(0,1,1)x(0,1,1)12 model.

So, the first candidate model is:

And, the fitted model is:

A detailed summary of the model can be found in the Appendix.

Then we look at the residual diagnose of the ARIMA (0,1,1)x(0,1,1)12.

A screenshot of a social media post

Description automatically generatedA close up of a map

Description automatically generated

Figure 8. Residual Diagnose for ARIMA(0,1,1)x(0,1,1)\_12

Residuals of ARIMA(0,1,1)x(0,1,1)12 seems normally distributed in histogram and QQ-plot. But if we take a closer look at residual autocorrelation plots (ACF&PACF), not all spikes stay within the significance limits, so the residuals do not appear to be white noise. The Ljung-Box test also shows that the residuals have remaining autocorrelations.

1. Candidate 2&3: ARIMA(0,1,(1,9,18))x(0,1,1)12 & ARIMA(0,1,(1,18))x(0,1,1)12

Both the ACF and PACF show significant spikes at lag 18, indicating that some additional non-seasonal terms need to be included in the models. So model ARIMA(0,1,18)x(0,1,1)12 was fitted and backward model reduction was perform (take out insignificant variables), the resulting models ARIMA(0,1,(1,9,18)x(0,1,1)12 and ARIMA(0,1,(1,18)x(0,1,1)12 are shown in table 2.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | AIC | BIC | LBP test(lag>0.05) | Significant vbles |
| ARIMA(0,1,(1,9,18))x(0,1,1)12 | 247.314 | 258.000 | All pass | All significant |
| ARIMA(0,1,(1,18))x(1,1,1)12 | 248.792 | 256.810 | Lag 24 not pass | All significant |

Table 2. Developed models from ARIMA(0,1,1)x(0,1,1)\_12 (all models without intercept)

So, the second candidate model is:

And, the fitted model is:

A detailed summary of the model can be found in the Appendix.

Then we look at the residual diagnose of the ARIMA (0,1,(1,9,18))x(0,1,1)12.

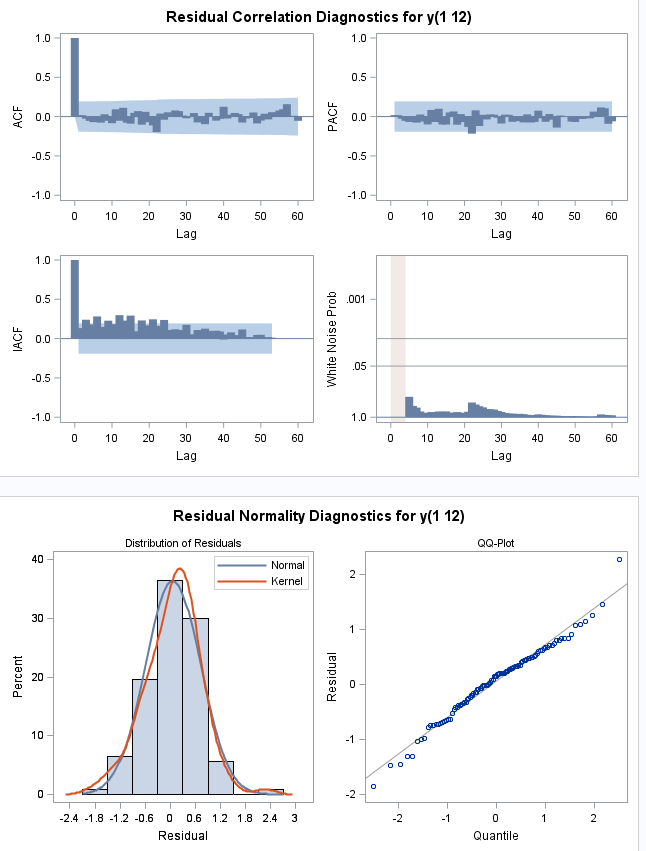
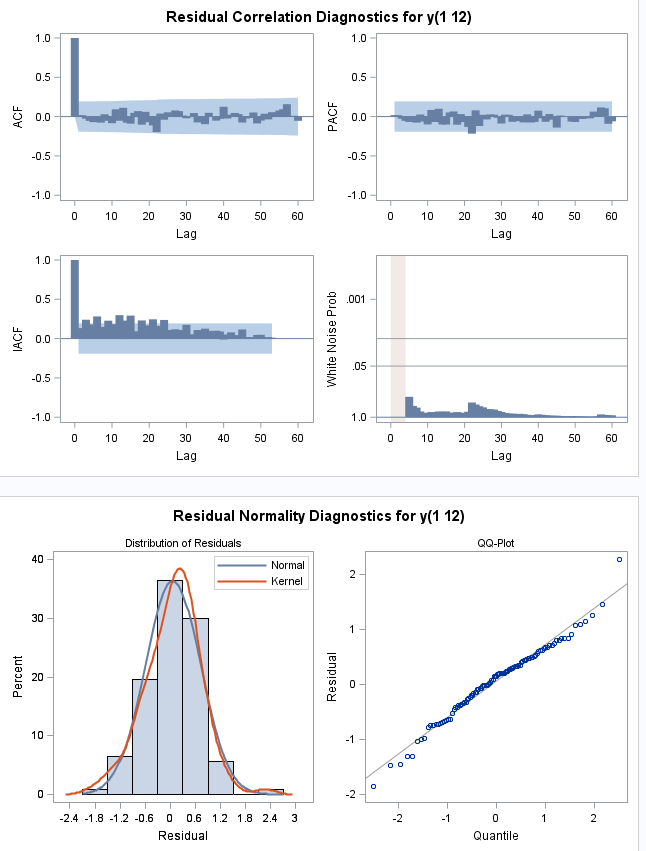


Figure 9. Residual Diagnose for ARIMA(0,1,(1,9,18))x(0,1,1)\_12

Residuals of ARIMA(0,1,(1,9,18))x(0,1,1)12 seems normally distributed in histogram and QQ-plot (figure 9). All spikes are within the significance limits in residual autocorrelation plots, so the residuals appear to be white noise. The Ljung-Box test also shows that the residuals have no remaining autocorrelations.

The third candidate model is:

And, the fitted model is:

A detailed summary of the model can be found in the Appendix.

Then we look at the residual diagnose of the ARIMA (0,1,(1,18))x(0,1,1).

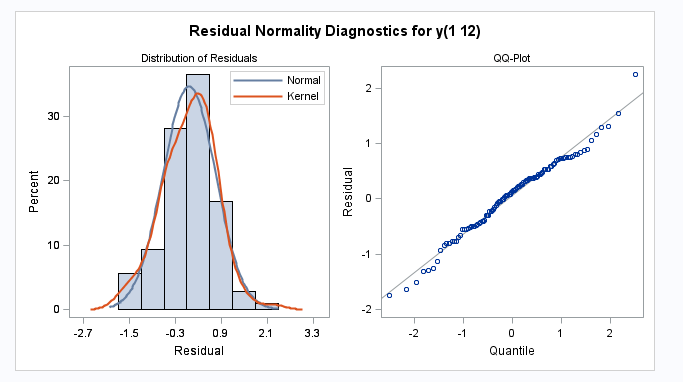
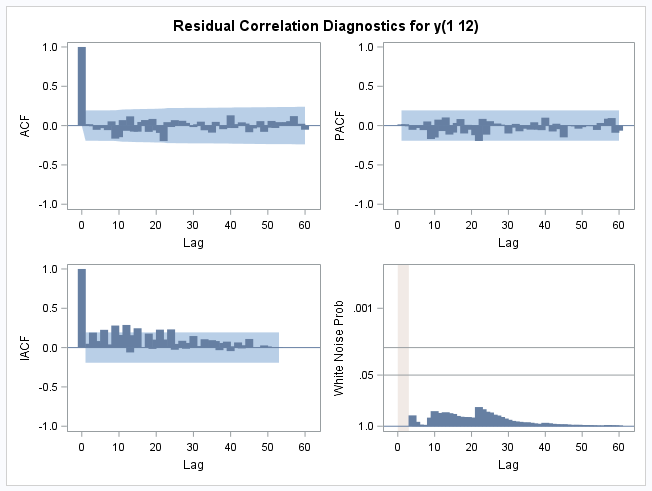


Figure 10. Residual Diagnose for ARIMA(0,1,(1,18))x(0,1,1)\_12

Residuals of ARIMA(0,1,(1,18))x(0,1,1)12 seems normally distributed in histogram and QQ-plot. Most spikes are within the significance limits in residual autocorrelation plots. It failed the LBP-test at lag 24, but the model can still be the candidate forecasting model, just the prediction intervals may not be accurate due to the correlated residuals.

1. Select the best model among three candidates

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | AIC | BIC | RMSE | LBP test(lag>0.05) | Significant vbles (P<0.05) |
| ARIMA(0,1,1)x(0,1,1)12 | 252.248 | 257.593 | 0.71 | Lag 6, 12, 48,54, 60 pass | MA at lag 24 is not |
| SARIMA(0,1,(1,9,18))x(0,1,1)12 | 247.314 | 258.000 | 0.67 | All pass | AR at lag 12 is not |
| SARIMA(0,1,(1,18))x(0,1,1)12 | 248.792 | 256.810 | 0.70 | Lag 24 not pass | All variables are significant |

Table 3. Candidates models comparison (all models without intercept)

Among three candidate models, ARIMA(0,1,(1,9,18))x(0,1,1)12 is selected as the best model because:

* It has the least AIC (248.79) among 3 models. Although its BIC (258.00) is slightly larger than the other two models.
* It has a lower standard error (0.56).
* It’s the only one passed the LBP Test at all lags.
* All variables are significant.
* The residuals are white noise. No ACF/PACF is significantly different from zero.
* QQ-plot and histogram of residual distribution indicate the residuals are asymptotically normally distributed.

**Model Selection By MAPE**

The mean absolute *percentage* error (MAPE) is also often useful for comparing the model, which expresses the prediction accuracy of the forecasting model as a percentage of the error. By computing the MAPE value for the two best models from the previous step using the data from January 2004 to December 2004, ARIMA(0,1,(1,9,18)x(0,1,1)12 is selected as the best model with the least MAPE value. Since the percentage error is smaller, the prediction accuracy of ARIMA(0,1,(1,9,18)x(0,1,1)12 model is higher than the other one.

|  |  |
| --- | --- |
| Model | MAPE |
| ARIMA(0,1,(1,9,18))x(1,1,1)12 | 0.01695 |
| ARIMA(0,1,(1,18))x(0,1,1)­12 | 0.01952 |

Table 4. MAPE values of competing models

**Forecast**

Forecasts from the best model ARIMA(0,1,(1,9,18)x(1,1,1)12 for next year each month (year 2004) are shown in Figure.11. Forecasts have shown in red dashed lines, and blues lines represent upper and lower 95% confidence interval for forecast values.

A close up of a map

Description automatically generated

Figure 11. MAPE values of competing models

The forecasts follow the recent trend in the data, reach a peak in May 2004 at carbon dioxide level 383 ppm, and drop rapidly till 368ppm in August then raise again.

**Conclusion**

The time-series data of carbon dioxide is simple differenced and seasonal differenced for stationary. Under model selection criteria, ARIMA(0,1,(1,9,18)x(1,1,1)12 is selected as the best model with least AIC, RMSE, MAPE and the only one passed LBP-test at all lags. According to the selected best model, the next peak of carbon dioxide level next year (in 2004) will be in May again with 384ppm.

**Reference**

Eleanor Imster and Deborah Byrd. (2019. June 17). Atmospheric CO2 hits record high in May 2019. Retrieved from <https://earthsky.org/earth/atmospheric-co2-record-high-may-2019>.

Rebecca Lindsey. (2019. Sep 19). Climate Change: Atmospheric Carbon Dioxide. Retrieved from <https://earthsky.org/earth/atmospheric-co2-record-high-may-2019>.

**Appendix**

1. Summary of 25 possible models

A screenshot of a cell phone

Description automatically generated

1. Summary of ARIMA (0,1,1)x(0,1,1)12

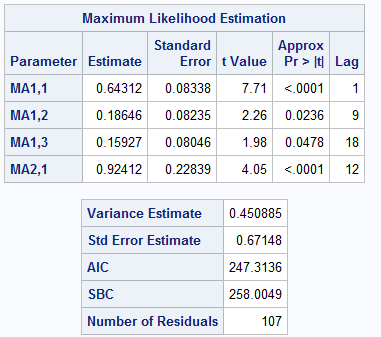
**A screenshot of a cell phone

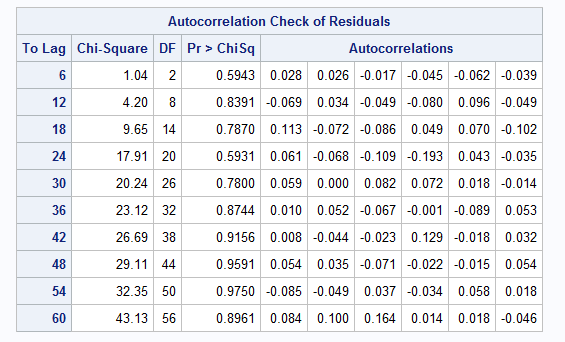
Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

1. Summary of ARIMA (0,1,(1,9,18))x(0,1,1)12

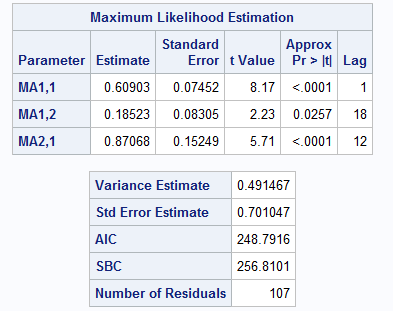
****

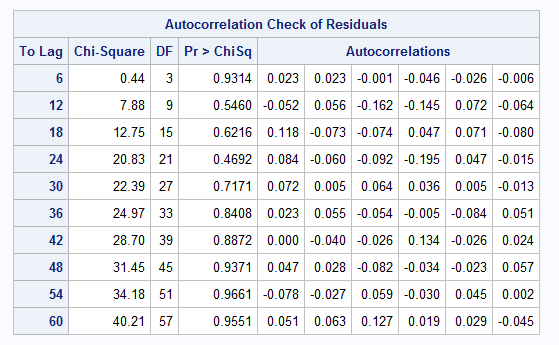
****

**A screenshot of a cell phone

Description automatically generated**

1. Summary of ARIMA (0,1,(1,18))x(0,1,1)12

****

****

A screenshot of a cell phone

Description automatically generated

**SAS Code**

data work.co2;

infile 'C:\Users\my57617\Desktop\Project2\co2.dat' firstobs=2;

input y;

run;

data work.co2;

set work.co2;

year=1994+int((\_n\_+1-1)/12-0.001);

month=MOD(\_n\_+1-1,12);

if month=0 then month=12;

time=MDY(month,1,year);

format time monyy.;

drop year month;

n=\_n\_;

run;

quit;

data work.co2new;

set work.co2;

if n > 120 then delete;

run;

quit;

proc gplot data=work.co2;

plot y\*time/ vaxis=axis1 haxis=axis2 frame grid;

title2 "co2 vs time";

axis1 label = (a=90 'co2');

axis2 label=('time');

symbol1 v=dat h=.1 i=join ci=blue ;

run;

quit;

proc transreg data=work.co2;

title2 'BOX-COX';

model boxcox(y/lambda=0to 10 by 0.1) = identity(time);

output out=trans;

run;

\*1 is in 95CI no transformation;

title2 "SARIMA models\_1 for Y without Intercept";

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate p=(1,2)(12,24) noint printall grid method=ml plot;

estimate p=(1,2) q=(0)(12,24) noint printall grid method=ml plot;

estimate p=(1,2)(12) q=(0)(12) noint printall grid method=ml plot;

estimate p=(1,2) q=(0)(12) noint printall grid method=ml plot;

estimate p=(1,2)(12) noint printall grid method=ml plot;

forecast out=out1 id=time alpha=0.05 lead=0;

run;

quit;

title2 "SARIMA models\_2 for Y without Intercept";

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate p=(0)(12,24) q=(1,2) noint printall grid method=ml plot;

estimate q=(1,2)(12,24) noint printall grid method=ml plot;

estimate p=(0)(12) q=(1,2)(12) noint printall grid method=ml plot;

estimate q=(1,2)(12) noint printall grid method=ml plot;

estimate p=(0)(12) q=(1,2) noint printall grid method=ml plot;

forecast out=out2 id=time alpha=0.05 lead=0;

run;

quit;

title2 "SARIMA models\_3 for Y without Intercept";

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate p=(1)(12,24)q=(1) noint printall grid method=ml plot;

estimate p=(1) q=(1)(12,24) noint printall grid method=ml plot;

estimate p=(1)(12) q=(1)(12) noint printall grid method=ml plot;

estimate p=(1) q=(1)(12) noint printall grid method=ml plot;

estimate p=(1)(12) q=(1)noint printall grid method=ml plot;

forecast out=out3 id=time alpha=0.05 lead=0;

run;

quit;

title2 "SARIMA models\_4 for Y without Intercept";

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate p=(1)(12,24)noint printall grid method=ml plot;

estimate p=(1) q=(0)(12,24) noint printall grid method=ml plot;

estimate p=(1)(12) q=(0)(12) noint printall grid method=ml plot;

estimate p=(1) q=(0)(12) noint printall grid method=ml plot;

estimate p=(1)(12)noint printall grid method=ml plot;

forecast out=out4 id=time alpha=0.05 lead=0;

run;

quit;

title2 "SARIMA models\_5 for Y without Intercept";

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate p=(0)(12,24)q=(1) noint printall grid method=ml plot;

estimate q=(1)(12,24) noint printall grid method=ml plot;

estimate p=(0)(12) q=(1)(12) noint printall grid method=ml plot;

estimate q=(1)(12) noint printall grid method=ml plot;

estimate p=(0)(12) q=(1)noint printall grid method=ml plot;

forecast out=out5 id=time alpha=0.05 lead=0;

run;

quit;

\*best 3 models (0,1,1)x(0,1,2),(1,1,1),(0,1,1);

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate q=(1)(12,24) noint printall grid method=ml plot;

\*drop significant lag will be last model;

forecast out=best1 id=time alpha=0.05 lead=0;

run;

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate p=(0)(12) q=(1)(12) noint printall grid method=ml plot;

\*drop significant lag will be last model;

forecast out=best2 id=time alpha=0.05 lead=0;

run;

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate q=(1)(12) noint printall grid method=ml plot;

forecast out=best3 id=time alpha=0.05 lead=12;

run;\*best;

data best3;

merge best3 work.co2;

run;

quit;

proc gplot data=best3;

where n >120;

plot y\*time forecast\*time L95\*time U95\*time / overlay haxis=axis1

vaxis=axis2 frame legend=legend1 grid;

axis1 label = ("time");

axis2 label = (a=90 "y\_t");

title2 "Forecast plot for y\_t: SARIMA(0,1,1)x(0,1,1)\_12";

symbol1 i=join h=.1 v=dot l=1 ci=black;

symbol2 i=join h=.1 v=square l=2 ci=red;

symbol3 i=join l=1 r=2 ci=blue;

legend1 label = none

position = (bottom center outside)

across = 4

down = 1

mode = reserve

frame

offset =(0.5, 0.5)

value = (j=l h=0.3 'Observed' 'Forecast' 'C.I.' 'C.I.');

run;

quit;

title2 "MAPE Computations for SARIMA (0,1,1)x(0,1,1)\_12 without Intercept";

data forecastbest;

set best3;

where n> 120;

keep forecast n;

run;

quit;

data PartCompbest;

set work.co2;

where n > 120;

keep y n;

run;

quit;

data MAPEbest;

merge forecastbest PartCompbest;

ratio = (abs(y-forecast))/y;

proc print;

run;

quit;

proc means data=MAPEbest noprint;

var ratio;

output out=mape mean=mape;

run;

proc print;

title2 "MAPE value for SARIMA (0,1,1)x(0,1,1)\_12 for logY without Intercept";

run;

quit;

\*develop model;

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate q=(1,2,3,4,5,6,7,8,9,10,11,18)(12) noint printall grid method=ml plot;

estimate q=(1,9,18)(12) noint printall grid method=ml plot;

estimate q=(1,18)(12) noint printall grid method=ml plot;

forecast out=deve id=time alpha=0.05 lead=12;

run;\*develope model

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate q=(1,18)(12) noint printall grid method=ml plot;

forecast out=best31 id=time alpha=0.05 lead=12;

run;\*best of best;

data best31;

merge best31 work.co2;

run;

quit;

data best31;

set best31;

forecast2 = forecast;

L952 = L95;

U952 = U95;

run;

quit;

proc gplot data=best31;

where n >120;

plot y\*time forecast2\*time L952\*time U952\*time / overlay haxis=axis1

vaxis=axis2 frame legend=legend1 grid;

axis1 label = ("time");

axis2 label = (a=90 "y\_t");

title2 "Forecast plot for y\_t: SARIMA(0,1,(1,18))x(0,1,1)\_12";

symbol1 i=join h=.1 v=dot l=1 ci=black;

symbol2 i=join h=.1 v=square l=2 ci=red;

symbol3 i=join l=1 r=2 ci=blue;

legend1 label = none

position = (bottom center outside)

across = 4

down = 1

mode = reserve

frame

offset =(0.5, 0.5)

value = (j=l h=0.3 'Observed' 'Forecast' 'C.I.' 'C.I.');

run;

quit;

title2 "MAPE Computations for SARIMA (0,1,(1,18))x(0,1,1)\_12 without Intercept";

data forecastbest2;

set best31;

where n> 120;

keep forecast2 n;

run;

quit;

data PartCompbest2;

set work.co2;

where n > 120;

keep y n;

run;

quit;

data MAPEbest2;

merge forecastbest2 PartCompbest2;

ratio2 = (abs(y-forecast2))/y;

proc print;

run;

quit;

proc means data=MAPEbest2 noprint;

var ratio2;

output out=mape2 mean=mape2;

run;

proc print;

title2 "MAPE value for SARIMA (0,1,(1,18))x(0,1,1)\_12 for logY without Intercept";

run;

quit;

proc arima data=work.co2new;

identify var = y(1,12) nlag=60;

estimate q=(1,9,18)(12) noint printall grid method=ml plot;

forecast out=best32 id=time alpha=0.05 lead=12;

run;

data best32;

merge best32 work.co2;

run;

quit;

data best32;

set best32;

forecast3 = forecast;

L953 = L95;

U953 = U95;

run;

quit;

proc gplot data=best32;

where n >120;

plot y\*time forecast3\*time L953\*time U953\*time / overlay haxis=axis1

vaxis=axis2 frame legend=legend1 grid;

axis1 label = ("time");

axis2 label = (a=90 "y\_t");

title2 "Forecast plot for y\_t: SARIMA(0,1,(1,18))x(0,1,1)\_12";

symbol1 i=join h=.1 v=dot l=1 ci=black;

symbol2 i=join h=.1 v=square l=2 ci=red;

symbol3 i=join l=1 r=2 ci=blue;

legend1 label = none

position = (bottom center outside)

across = 4

down = 1

mode = reserve

frame

offset =(0.5, 0.5)

value = (j=l h=0.3 'Observed' 'Forecast' 'C.I.' 'C.I.');

run;

quit;

title2 "MAPE Computations for SARIMA (0,1,(1,18))x(0,1,1)\_12 without Intercept";

data forecastbest3;

set best32;

where n> 120;

keep forecast3 n;

run;

quit;

data PartCompbest3;

set work.co2;

where n > 120;

keep y n;

run;

quit;

data MAPEbest3;

merge forecastbest3 PartCompbest3;

ratio3 = (abs(y-forecast3))/y;

proc print;

run;

quit;

proc means data=MAPEbest3 noprint;

var ratio3;

output out=mape3 mean=mape3;

run;

proc print;

title2 "MAPE value for SARIMA (0,1,(1,9,18))x(0,1,1)\_12 for logY without Intercept";

run;

quit;