Time Series Analysis

Identifying possible ARIMA models

Andrés M. Alonso   Carolina García-Martos

Universidad Carlos III de Madrid

Universidad Politécnica de Madrid

June – July, 2012

8. Identifying possible ARIMA models

Outline:

- Introduction
- Variance stabilizing transformation
- Mean stabilizing transformation
- Identifying ARMA structure

Recommended readings:

▷ Chapter 6 of Brockwell and Davis (1996).

▷ Chapter 17 of Hamilton (1994).

▷ Chapter 7 of Peña, Tiao and Tsay (2001).

Introduction

▷ We had studied the theoretical properties of ARIMA processes. Here we are going to analyze how to fit these models to real series. Box and Jenkins (1976) proposed carrying out this fit in three steps.

- The first step consists of identifying the possible ARIMA model that the series follows, which requires: (1) deciding what transformations to apply in order to convert the observed series into a stationary one; (2) determining an ARMA model for the stationary series.
- Once we have provisionally chosen a model for the stationary series we move on to the second step of estimation, where the AR and MA model parameters are estimated by maximum likelihood.
- The third step is that of diagnosis, where we check that the residuals do not have a dependence structure and follow a white noise process.

Introduction

▷ These three steps represented an important advancement in their day, since the estimation of the parameters of an ARMA model required a great deal of calculation time. Nowadays, estimation of an ARIMA model is straightforward, making it much simpler to estimate all the models we consider to be possible in explaining the series and then to choose between them.

▷ This is the philosophy of the automatic selection criteria of ARIMA models, which work well in practice in many cases, and are essential when we wish to model and obtain predictions for a large set of series.

▷ Nevertheless, when the number of series to be modelled is small, it is advisable to carry out the identification step that we will look at next in order to better understand and familiarize ourselves with the dynamic structure of the series of interest.

▷ The objective is not to choose a model now for the series, but rather to identify a set of possible models that are compatible with the series graph and its simple and partial autocorrelation functions.

Introduction

▷ These models will be estimated and in the third phase, diagnosis, we ensure that the model has no detectable deficiencies.

▷ **Identification** of the model requires identifying the non-stationary structure, if it exists, and then the stationary ARMA structure:

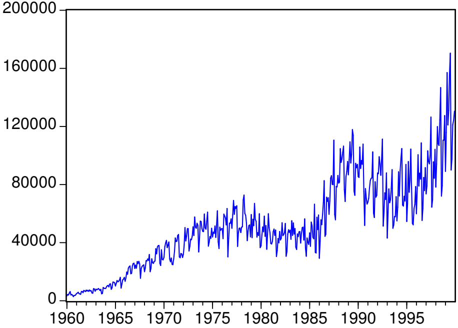
- The identification of the non-stationary structure consists in detecting which transformations must be applied to obtain a stationary ARMA process with constant variance and mean:
  - to transform the series so that it has constant variance;
  - to differentiate the series so that it has constant mean.
- Later we will identify the ARMA model for the stationary series and analyze these aspects.

Variance stabilizing transformation

▷ In many series the variability is often greater when the series takes high values than when it takes low ones.

Example 78

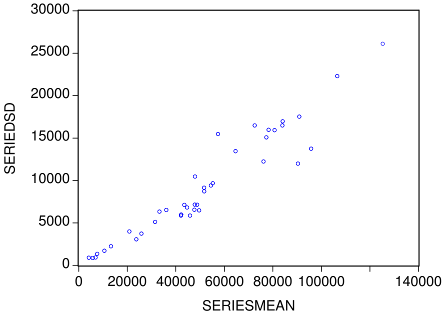
The figure shows the Spanish vehicle registration series and we observe that the variability is much greater when the level of the series is high, which is what happens at the end of the series, than when it is low.



▷ The variability does not increase over time but rather with the level of the series: the variability around 1975-1980 is high and corresponds to a maximum of the level, and this variability decreases around 1980-1985, when the level drops.

▷ We can confirm this visual impression by plotting a graph between a measure of variability, such as the standard deviation, and a measure of level, such as the local mean.

▷ In order to make homogeneous comparisons since the series is seasonal we take the 12 observations from each year and calculate the standard deviation and the mean of the registrations in each year.

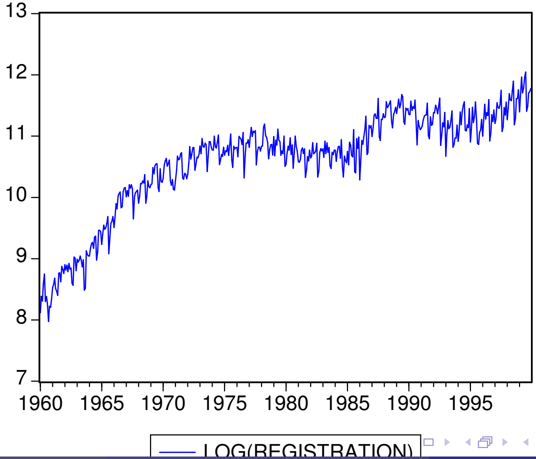


▷ We can see that a clear linear-type dependence exists between both variables.

▷ When the variable of the series increases linearly with the level of the series, as is the case with the vehicle registration, by taking logarithms we get a series with constant variability.

Example 78

The figure shows the series in logarithms which confirms this fact. The series in logs has constant variability.



Variance stabilizing transformation

▷ The dependency of the variability of the level of the series might be a result of the series being generated as a product (instead of sum) of a systematic or predictable component, μ_t, times the innovation term, u_t, that defines the variability. Then:

$$z_t = \mu_t u_t. \tag{156}$$

▷ Let us assume that the expectation of this innovation is one, such that $E(z_t) = \mu_t$. The standard deviation of the series is:

$$\sigma_t = \left[E(z_t - \mu_t)^2\right]^{1/2} = \left[E(\mu_t u_t - \mu_t)^2\right]^{1/2} = |\mu_t| \left[E(u_t - 1)^2\right]^{1/2} = |\mu_t| \, \sigma_u \tag{157}$$

and while the innovation u_t has constant variability, the standard deviation of the observed series z_t is not constant in time and will be proportional to the level of the series.

▷ As we have seen, this problem is solved by taking logarithms, since then, letting $a_t = \ln u_t$:

$$y_t = \ln z_t = \ln \mu_t + a_t$$

and an additive decomposition is obtained for the variable y_t, which will have constant variance.

▷ The above case can be generalized permitting the standard deviation to be a potential function of the local mean, using:

$$\sigma_t = k\mu_t^{\alpha}, \tag{158}$$

and if we transform variables z_t into new variables y_t by means of:

$$y_t = x_t^{1-\alpha}$$

these new variables y_t have the same standard deviation.

▷ A more general way of describing this transformation is:

$$y_t = \frac{x_t^{1-\alpha} - 1}{1-\alpha} \tag{159}$$

which is part of the family of **Box-Cox transformations** and includes the powers and the logarithm of the variable.

▷ In the series, when a relationship is observed between the level and the variability we can estimate the value of α needed to obtain constant variability by making consecutive groups of observations, calculating the standard deviation in each group, s_i, and the mean $\overline{x}_i$ and representing these variables in a graph.

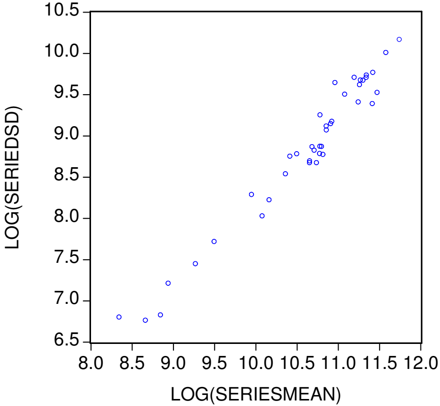
▷ Taking logarithms in (158), the slope of the regression

$$\log s_i = c + \alpha \log \overline{x}_i \tag{160}$$

estimates the value of α, and the transformation of the data by means of (159) will lead to a series where the variability does not depend on the level.

Example 78

The figure gives the relationship between the logarithms of these variables for the registration series. We observe that the slope is close to the unit and if we estimate regression (160) we have $c = -2.17$ *and* $\alpha = 1.04$.



▷ As a result, a transformation using $\alpha = 0$, that is, by means of logarithms, should produce a variance constant with the level.

▷ We had seen that the vehicle registration series expressed in logarithms and its variability is approximately constant.

▷ It is advisable to plot graphs between variability and mean using groups of data that are as homogeneous as possible.

Mean stabilizing transformation

▷ To stabilize the mean of the series we apply regular and seasonal differences. The decision to apply these differences can be based on the graph of the series and on the sample autocorrelation function, but we can also use formal tests.

Determining the order of regular difference

▷ If the series has a trend or shows changes in the level of the mean, we differentiate it in order to make it stationary. The need to differentiate is clearly seen from the graph of the series. For example, the vehicle registration series clearly shows periodic behavior and non-constant mean.

▷ Note that

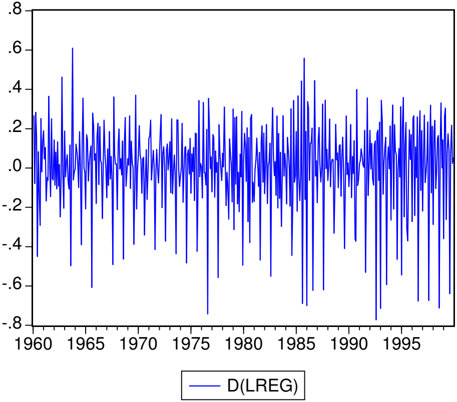
$$(1-B)\ln z_t = \ln z_t - \ln z_{t-1} = \ln\left(1 + \frac{z_t - z_{t-1}}{z_{t-1}}\right) \approx \frac{z_t - z_{t-1}}{z_{t-1}}$$

where we have used $\ln(1+x) \approx x$ if x is small.

▷ Therefore, series $\nabla \ \ln z_t$ is equivalent to the relative growth rates of z_t.

Example 79

The figure shows the first difference of the registration series and we can see that it contains very noticeable variations month to month, up to 0.8, that is 80% of its value.



▷ This is due to the presence of strong seasonality, appearing in the graph as sharp drops, which correspond to the month of August of each year.

▷ As a result of this effect the series does not have a constant mean. Next, we will look at how to eliminate this pattern by means of a seasonal difference.

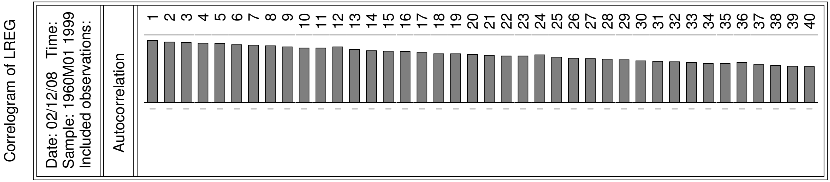
Mean stabilizing transformation

▷ When the decision to differentiate is not clear from analyzing the graph, it is advisable to study its autocorrelation function, *ACF*.

▷ We have seen that a non-stationary series has to show positive autocorrelations, with a slow and linear decay.

Example 80

The figure shows gives the ACF of the vehicle registration series: a slow linear decay of the coefficients can be observed, which indicates the need to differentiate.

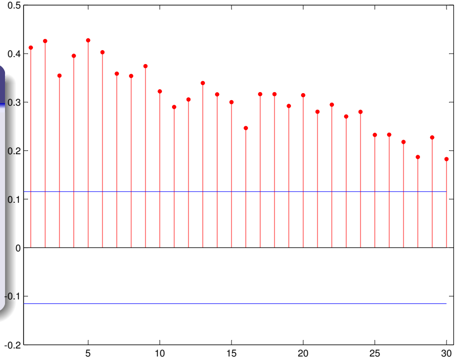


Mean stabilizing transformation

▷ It is important to point out that the characteristic which identifies a non-stationary series in the estimated ACF is not autocorrelation coefficients close to the unit, but rather the slow linear decay.

Example 81

In the IMA(1,1) process if θ is close to one, then the expected value of the sample autocorrelation coefficients is always less than 0.5. Nevertheless, a smooth linear decrease is still to be expected.



▷ To summarize, if the ACF does not fade for high values (more than 15 or 20) it is generally necessary to differentiate in order to obtain a stationary process.

Mean stabilizing transformation

▷ It is possible to carry out a test to decide whether or not the series should be differentiated.

▷ But, if the objective is forecasting, these tests are not very useful because in case of doubt it is always advisable to differentiate, since the negative consequences of overdifferentiating are much less that those of underdifferentiating.

▷ Indeed, if we assume that the series is stationary when it is not, the medium term prediction errors can be enormous because the medium term prediction of a stationary period is its mean, whereas a non-stationary series can move away from this value indefinitely and the prediction error is not bounded.

▷ The undifferentiated model will also be non-robust and with little capacity to adapt in the presence of future values.

▷ However, if we overdifferentiate we always have adaptive predictors and, while we lose accuracy, this loss is limited.

Mean stabilizing transformation

▷ A further reason for taking differences in case of doubt is that it can be proved that although the series is stationary, if the AR parameter is close to the unit we obtain better predictions by overdifferentiating the series than by using the true stationary model.

▷ When the aim is not to forecast, but rather to decide whether a variable is stationary or not, a common situation in economic applications, then we must use the unit-root tests:

- Dickey-Fuller test (with or without intercept, with trend and intercept)
- Augmented Dickey-Fuller test
- Phillips Perron test ...

Mean stabilizing transformation

Unit root tests

▷ These tests tell us whether we have to take an additional difference in a series in order to make it stationary.

▷ We will first present the simplest case of the Dickey-Fuller test. Let us assume that we wish to decide between the non-stationary process:

$$\nabla z_t = a_t, \tag{161}$$

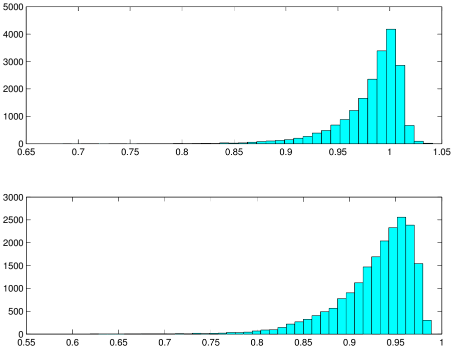
and the stationary:

$$(1 - \rho B)\, z_t = c + a_t. \tag{162}$$

▷ The Dickey-Fuller test was developed because the traditional procedure for estimating both models and choosing the one with least variance is not valid in this case.

▷ In fact, if we generate a series that follows a random walk, (161), and we fit both models to this series, in model (161) we will not estimate any parameter whereas in (162) we estimate parameters ρ and c so that the variance of the residuals is minimum.

▷ To illustrate this aspect, the figure shows the distribution of the least squares estimator of parameter ρ in samples of 100 which have been generated by random walk (161).



▷ The distribution is observed to be very asymmetric and that it can be much smaller than the true value of the parameter which is one, especially with the second estimator, the estimated value, which is the one that provides the best fit.

▷ In conclusion, comparing the variances of both models is not a good method for choosing between them and, especially if we want to protect ourselves with respect to the error of not differentiating.

Unit root tests

▷ One might think that a better way of deciding between two models is to estimate the stationary model (162) by least squares and test whether coefficient ρ is one, comparing the estimator with its estimated standard deviation.

▷ The test would be $H_0 : \rho = 1$ compared to the alternative $H_1 : \rho < 1$, and the statistic for the test:

$$t_\mu = \frac{\widehat{\rho} - 1}{\widehat{s}_\rho} \tag{163}$$

where $\widehat{s}_\rho$ is the estimated standard deviation of the least squares estimator.

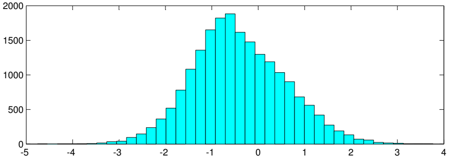
▷ We could carry out the test comparing the value of t_μ using the usual tables of the Student's t with a unilateral test.

▷ Nevertheless, this method is **not correct**, since the distribution in the sample of the estimator $\widehat{\rho}$ is not normal with a mean of one, which is necessary for the statistic (163) to be a Student's t.

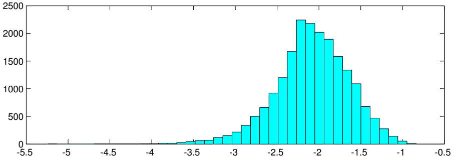
▷ Furthermore, when the process is non-stationary the least squares estimation provides an erroneous value of the variance of estimator $\widehat{\rho}$ and statistic (163) does **not follow** a Student's t-distribution.

▷ The reason is that when the null hypothesis is true the value of the parameter is found in the extreme of the parametric space (0,1) and the conditions of regularity which are necessary for the asymptotic properties of the ML estimator are not satisfied.

▷ To illustrate this fact the figure gives the result of calculating the statistic (163) in 20000 samples of 100 generated by random walks.



▷ Notice that the distribution of the statistic differs greatly from that of the Students's t and does not even have zero mean.



Dickey-Fuller test

▷ In the **Dickey-Fuller test**, the null hypothesis of the test, H_0, is that the series is non-stationary and it is necessary to differentiate, $\rho = 1$, and test whether this hypothesis can be rejected in light of the data.

▷ The statistic for the test is (163), but its value is compared with the true distribution of the test, which is not that of the Student's t

▷ A simple way of obtaining the value of this statistic (163) is to write the model (162) that we want to estimate subtracting z_{t-1} from both members of the equation and write:

$$\nabla z_t = c + \alpha z_{t-1} + a_t \tag{164}$$

where $\alpha = \rho - 1$.

▷ When we write a model including the variable in differences and in levels in the equation, ∇z_t and z_{t-1}, we call it an error correction model.

Dickey-Fuller test

▷ In model (164) if $\alpha = 0$ we have a random walk and if $\alpha \neq 0$ a stationary process.

▷ The null hypothesis which states that the process is non-stationary and $\rho = 1$ in this model becomes the null hypothesis $H_0 : \alpha = 0$, and the alternative, that the process is stationary, becomes $H_1 : \alpha \neq 0$.

▷ The test consists in estimating parameter α in (164) by least squares and rejecting that the process is stationary if the value of t_μ is significantly small.

▷ The statistic (163) is now written:

$$t_\mu = \frac{\widehat{\alpha}}{\widehat{s}_\alpha} \tag{165}$$

where $\widehat{\alpha}$ is the least squares estimation of α in (164) and $\widehat{s}_\alpha$ is its estimated standard deviation calculated in the usual way.

Dickey-Fuller test

- Under the hypothesis that $\alpha = 0$ (which implies that $\rho = 1$) the distribution of t_μ has been tabulated by Dickey and Fuller. An extract from their tables is:

Table: **Critical values for the Dickey-Fuller unit-root test**

T	without constant				with constant			
	.01	.025	.05	.1	.01	025	.05	.1
25	-2.66	-2.26	-1.95	-1.60	-3.75	-3.33	-3.00	-2.63
50	-2.62	-2.25	-1.95	-1.61	-3.58	-3.22	-2.93	-2.60
100	-2.60	-2.24	-1.95	-1.61	-3.51	-3.17	-2.89	-2.58
250	-2.58	-2.23	-1.95	-1.62	-3.46	-3.14	-2.88	-2.57
500	-2.58	-2.23	-1.95	-1.62	-3.44	-3.13	-2.87	-2.57
∞	-2.58	-2.23	-1.95	-1.62	-3.43	-3.12	-2.86	-2.57

- We observe that the estimation of $\widehat{\alpha}$ will be negative since $\rho \leq 1$.

Dickey-Fuller test

▷ The distribution of the statistic is different when we include a constant or not in the model and, for greater security, it is always recommendable to include it, since the alternative is that the process is stationary but does not necessarily have zero mean.

▷ The decision of the test is:

$$\text{Reject non-stationarity if } t_\mu \leq t_c$$

where the value of t_c is obtained from Table 1.

▷ We recommend carrying out the test with a low level of significance, .01, such that there is a low probability of rejecting that the process is non-stationary when this hypothesis is true.

▷ Since the negative consequences of underdifferentiating, we need to protect ourselves against the risk of rejecting that the process is non-stationary when in fact it is.

Augmented Dickey-Fuller test

▷ The test we have studied analyzes whether there is a unit-root in an AR(1). This test is generalized for ARMA processes.

▷ We begin by studying the case in which we have an AR($p+1$) process and we want to know if it has a unit root. That is, we have to choose between two models:

$$H_0 : \phi_p(B)\nabla z_t = a_t, \tag{166}$$

$$H_1 : \phi_{p+1}(B)z_t = c + a_t. \tag{167}$$

▷ The null hypothesis establishes that the largest root of an AR($p+1$) is equal to one, model (166), and the process is non-stationary. The alternative states that it is less than one, as in (167), and we will have a stationary process.

▷ The idea of the Augmented Dickey-Fuller test is to try to test the condition of a unit root directly in the operator $\phi_{p+1}(B)$.

Augmented Dickey-Fuller test

▷ To implement this idea, the operator $\phi_{p+1}(B)$ is decomposed

$$\phi_{p+1}(B) = (1 - \alpha_0 B) - (\alpha_1 B + ... + \alpha_p B^p)\nabla. \tag{168}$$

▷ This decomposition can always be done, because in both sides of the equality we have a polynomial in B of order $p+1$ with $p+1$ coefficients. Thus by identifying powers of B in both members we can obtain the $p+1$ coefficients $\alpha_0, ..., \alpha_p$, given $\phi_1, ..., \phi_{p+1}$.

▷ We will see that this decomposition has the advantage of transferring the condition of the unit root in the first member to a condition on a coefficient that we can estimate in the second member.

▷ Indeed, if $\phi_{p+1}(B)$ has a unit root, then $\phi_{p+1}(1) = 0$, and if we make $B = 1$ in (168) the term $(\alpha_2 B + ... + \alpha_{p+1} B^p)(1 - B)$ is cancelled and it will have to be verified that $(1 - \alpha_0) = 0$, that is $\alpha_0 = 1$.

Augmented Dickey-Fuller test

▷ The unit root condition in $\phi_{p+1}(B)$ implies $\alpha_0 = 1$.

▷ We will see that $\alpha_0 = 1$ also implies a unit root: if $\alpha_0 = 1$, we can write the polynomial on the left as $(1 - \alpha_1 B - ... - \alpha_p B^p)\nabla$, and the polynomial has a unit root.

▷ Therefore, we have proved that the two following statements are equivalent: (1) the polynomial $\phi_{p+1}(B)$ has a unit root and (2) the coefficient α_0 in decomposition (168) is one.

▷ The model (167) can be written, utilizing (168)

$$\phi_{p+1}(B)z_t = (1 - \alpha_0 B)z_t - (\alpha_1 B + ... + \alpha_p B^p)\nabla z_t = c + a_t, \tag{169}$$

that is:

$$z_t = c + \alpha_0 z_{t-1} + \sum_{i=1}^{p} \alpha_i \nabla z_{t-i} + a_t$$

Augmented Dickey-Fuller test

▷ To obtain the statistic of the test directly we can, as in the above case, subtract z_{t-1} from both members and write the model in the error correction form, that is, with levels and differences as regressors.

▷ This model is:

$$\nabla z_t = c + \alpha z_{t-1} + \sum_{i=1}^{p} \alpha_i \nabla z_{t-i} + a_t \tag{170}$$

where $\alpha = (\alpha_0 - 1)$.

▷ Equation (170) can be estimated by least squares and the test for whether the series has unit root, $\alpha_0 = 1$ is equivalent to the test $\alpha = 0$.

▷ The test utilizes the same statistic from above:

$$t_\mu = \frac{\widehat{\alpha}}{\widehat{s}_\alpha}, \tag{171}$$

where $\widehat{\alpha}$ is the least squares estimation of α in (170) and $\widehat{s}_\alpha$ is its estimated standard deviation calculated in the usual way.

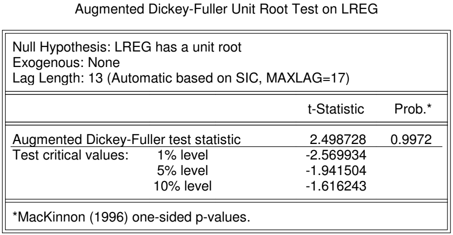
Augmented Dickey-Fuller test

▷ The distribution of t_μ when the hypothesis of a unit root $\rho = 1$ is true has been tabulated by Dickey and Fuller. Again, the distribution depends on whether or not we have a constant in the model.

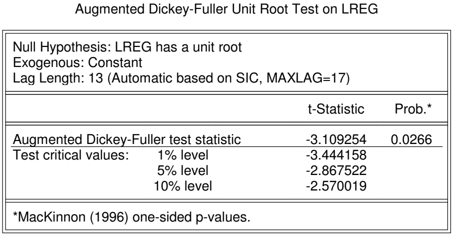
▷ In seasonal series we must be careful to introduce all of the necessary lags.

Example 82

The ADF unit-root test in the vehicle registration series logarithm:



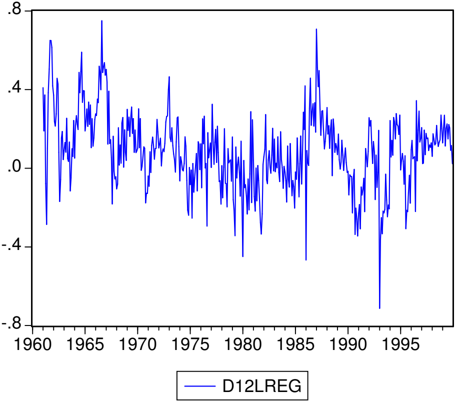
```
Augmented Dickey-Fuller Unit Root Test on LREG

Null Hypothesis: LREG has a unit root
Exogenous: None
Lag Length: 13 (Automatic based on SIC, MAXLAG=17)

                                            t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic        2.498728    0.9972
Test critical values:      1% level          -2.569934
                           5% level          -1.941504
                           10% level         -1.616243

*MacKinnon (1996) one-sided p-values.
```

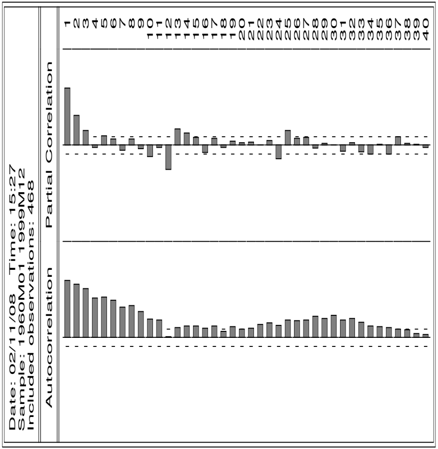


```
Augmented Dickey-Fuller Unit Root Test on LREG

Null Hypothesis: LREG has a unit root
Exogenous: Constant
Lag Length: 13 (Automatic based on SIC, MAXLAG=17)

                                            t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic       -3.109254    0.0266
Test critical values:      1% level          -3.444158
                           5% level          -2.867522
                           10% level         -2.570019

*MacKinnon (1996) one-sided p-values.
```

Example 82

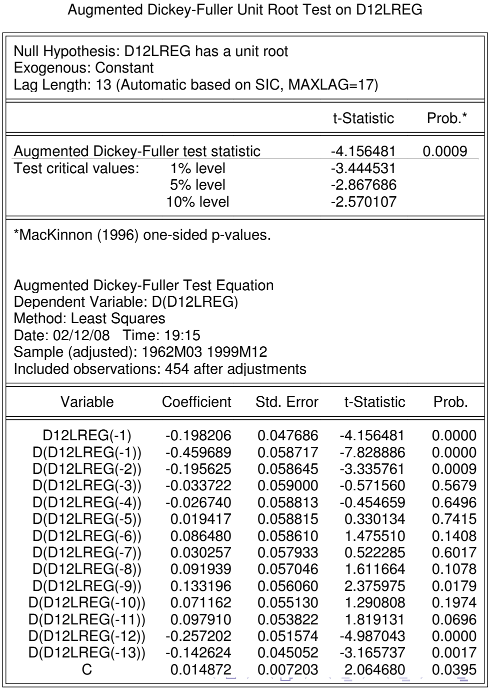
Now, if we consider the seasonal differentiated series, we observe that it not have constant level, but it no longer shows any clear trend. The ACF of this series has a lot of slowly decreasing positive coefficients, suggesting the need for a difference.





Example 82

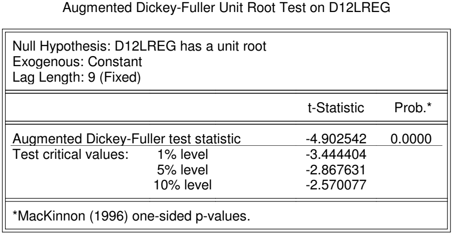
The ADF unit-root test in the seasonal differentiated series:



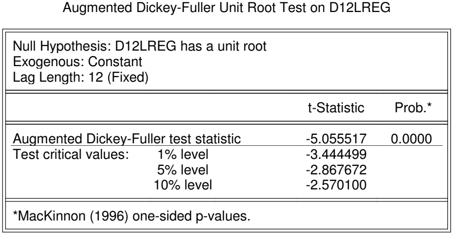
```
Augmented Dickey-Fuller Unit Root Test on D12LREG

Null Hypothesis: D12LREG has a unit root
Exogenous: Constant
Lag Length: 13 (Automatic based on SIC, MAXLAG=17)

                                                  t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic            -4.156481     0.0009
Test critical values:        1% level             -3.444531
                             5% level             -2.867686
                            10% level             -2.570107

*MacKinnon (1996) one-sided p-values.


Augmented Dickey-Fuller Test Equation
Dependent Variable: D(D12LREG)
Method: Least Squares
Date: 02/12/08   Time: 19:15
Sample (adjusted): 1962M03 1999M12
Included observations: 454 after adjustments

      Variable        Coefficient    Std. Error    t-Statistic    Prob.

    D12LREG(-1)        -0.198206     0.047686      -4.156481     0.0000
   D(D12LREG(-1))      -0.459689     0.058717      -7.828886     0.0000
   D(D12LREG(-2))      -0.195625     0.058645      -3.335761     0.0009
   D(D12LREG(-3))      -0.033722     0.059000      -0.571560     0.5679
   D(D12LREG(-4))      -0.026740     0.058813      -0.454659     0.6496
   D(D12LREG(-5))       0.019417     0.058815       0.330134     0.7415
   D(D12LREG(-6))       0.086480     0.058610       1.475510     0.1408
   D(D12LREG(-7))       0.030257     0.057933       0.522285     0.6017
   D(D12LREG(-8))       0.091939     0.057046       1.611664     0.1078
   D(D12LREG(-9))       0.133196     0.056060       2.375975     0.0179
  D(D12LREG(-10))       0.071162     0.055130       1.290808     0.1974
  D(D12LREG(-11))       0.097910     0.053822       1.819131     0.0696
  D(D12LREG(-12))      -0.257202     0.051574      -4.987043     0.0000
  D(D12LREG(-13))      -0.142624     0.045052      -3.165737     0.0017
         C              0.014872     0.007203       2.064680     0.0395
```

Example 82

The ADF unit-root test in the seasonal differentiated series with different number of lags



```
Augmented Dickey-Fuller Unit Root Test on D12LREG

Null Hypothesis: D12LREG has a unit root
Exogenous: Constant
Lag Length: 9 (Fixed)

                                              t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic        -4.902542     0.0000
Test critical values:       1% level          -3.444404
                            5% level          -2.867631
                           10% level          -2.570077

*MacKinnon (1996) one-sided p-values.
```



```
Augmented Dickey-Fuller Unit Root Test on D12LREG

Null Hypothesis: D12LREG has a unit root
Exogenous: Constant
Lag Length: 12 (Fixed)

                                              t-Statistic    Prob.*

Augmented Dickey-Fuller test statistic        -5.055517     0.0000
Test critical values:       1% level          -3.444499
                            5% level          -2.867672
                           10% level          -2.570100

*MacKinnon (1996) one-sided p-values.
```

▷ This example illustrates the importance of including the right lags in the test.

Mean stabilizing transformation

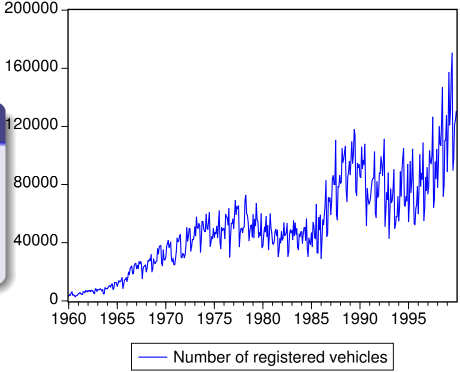
Determining the order of seasonal differencing

▷ If the series has seasonality a seasonal difference, $\nabla_s = 1 - B^s$, will have to be applied in order to make the series stationary. Seasonality is shown:

- In the series graph, which will show a repeating pattern of period s.
- In the autocorrelation function, which shows positive coefficients that slowly decrease in the lags s, $2s$, $3s$...

Example 83

The figure suggests a seasonal pattern because the value of the series is systematically lower in August.



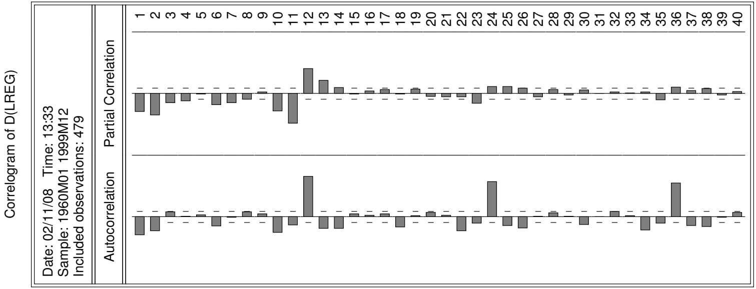
Determining the order of seasonal differencing

▷ Apart from the series graph it is advisable to look at the ACF, since a series with seasonality will show high and positive autocorrelation in seasonal lags.

Example 83

The figure shows the sample ACF of the regular differentiated series.

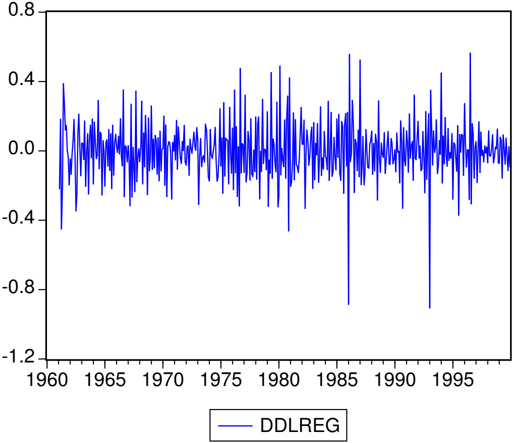
▷ *Notice that there are high and persistent correlations in seasonal lags 12, 24 and 36.*



Determining the order of seasonal differencing

▷ This suggests the need for taking a seasonal difference to obtain a stationary series.

▷ The figure gives the series with two differences, one regular and the other seasonal.



Identifying the ARMA structure

▷ Once we had determined the order of regular and seasonal differences, the next step is to identify the ARMA structure.

▷ Identification of the orders of p and q is carried out by comparing the estimated partial and simple autocorrelation functions with the theoretical functions of the ARMA process.

▷ Letting ω_t denote the stationary series, $\omega_t = \nabla^d \nabla_s^D z_t$, where in practice d takes values in $(0, 1, 2)$ and D in $(0, 1)$, the autocorrelations are calculated by:

$$r_k = \frac{\sum_{t=d+sD+1}^{T-k} (\omega_t - \overline{\omega}) (\omega_{t+k} - \overline{\omega})}{\sum_{t=d+sD+1}^{T} (\omega_t - \overline{\omega})^2}, \qquad k = 1, 2, ... \tag{172}$$

▷ In order to judge when a coefficient r_k is different from zero we need its standard error, whose determination depends on the structure of the process.

Identifying the ARMA structure

▷ One simple solution is to take $1/\sqrt{T}$ as the standard error, which is approximately the standard error of a correlation coefficient between independent variables. If *all* the theoretical autocorrelation coefficients were null, the standard deviations of estimation would be approximately $1/\sqrt{T}$.

▷ Therefore, we can place confidence bands at $\pm 2/\sqrt{T}$ and consider as significant, in the first approximation, the coefficients which lie outside those bands.

▷ The partial autocorrelations are obtained with the regressions:

$$\widetilde{\omega}_t = \alpha_{k1}\widetilde{\omega}_{t-1} + ... + \alpha_{kk}\widetilde{\omega}_{t-k},$$

where $\widetilde{\omega}_t = \omega_t - \overline{\omega}$. The sequence $\widehat{\alpha}_{kk}$ $(k = 1, 2, ...)$ of least squares coefficients estimated in these regressions is the partial autocorrelation function.

▷ In the graphs of the *PACF* we will always use the asymptotic limits $\pm 2/\sqrt{T}$, and we will considered as approximate limits of reference.

Identifying the ARMA structure

▷ If the process is seasonal, we study the coefficients of the sample *ACF* and *PACF* in lags $s, 2s, 3s, ...$, in order to determine the seasonal ARMA structure.

▷ Identifying an ARMA model can be a difficult task. With large sample sizes and pure AR or MA processes, the structure of the sample *ACF* and *PACF* usually indicates the required order.

▷ Nevertheless, in general, the interpretation of the sample *ACF* and *PACF* is complex for three main reasons:

- when autocorrelation exists the estimations of the autocorrelations are also correlated, which introduces a pattern of random variation in the sample *ACF* and *PACF* that is superposed on the true existing pattern;
- the limits of confidence that we use, $2/\sqrt{T}$, are asymptotic and not very precise for the first autocorrelations;
- for mixed ARMA processes it can be extremely difficult to estimate the order of the process, even when the theoretical values of the autocorrelations are known.

Identifying the ARMA structure

▷ Fortunately, it is not necessary in the identification step to decide what the order of the model is, but only to choose a set of ARMA models that seem suitable for representing the main characteristics of the series.

▷ Later we will estimate this set of selected models and choose the most suitable.

▷ Identification with the simple and partial autocorrelation function of the possible models can be done using the following rules:

1. Decide what the maximum order of the AR and MA part is from the obvious features of the *ACF* and *PACF*.
2. Avoid the initial identification of mixed ARMA models and start with AR or MA models, preferably of low order.
3. Utilize the interactions around the seasonal lags, especially in the *ACF*, in order to confirm the concordance between the regular part and the seasonal.

Identifying the ARMA structure

▷ In practice, most real series can be approximated well using ARMA models with p and q less than three, for non-seasonal series, and with P and Q less than two for seasonal series.

▷ In addition to selecting the orders $(p,q)(P,Q)$ of the model we have to decide in this step if the stationary series, ω_t, has a mean different from zero:

$$\overline{\omega} = \frac{\sum \omega_t}{T_c},$$

where T_c is the number of summands (normally $T_c = T - d - sD$). Its standard deviation can be approximated by:

$$s\left(\overline{\omega}\right) \simeq \frac{s_\omega}{\sqrt{T}} \left(1 + 2r_1 + ... + 2r_k\right)^{1/2}$$

where s_ω is the standard deviation of the stationary series, ω_t, and r_i the estimated autocorrelation coefficients.

Identifying the ARMA structure

▷ In this formula we are assuming that the first k autocorrelation coefficients are significant and that k/T is unimportant.

▷ If $\overline{\omega} \geq 2s\left(\overline{\omega}\right)$ we accept that the mean of the stationary process is different from zero and we will include it as a parameter to estimate; in the opposite case we assume that $E(\omega_t) = 0$.

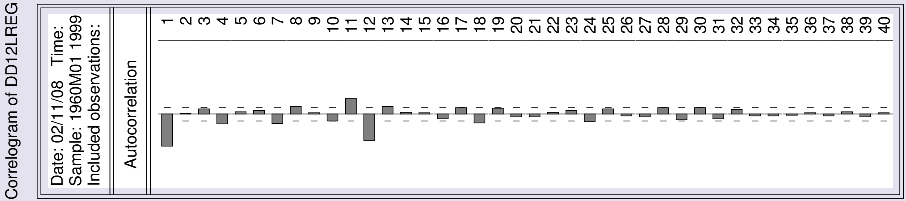
Additionally

▷ There are automatic selection procedures, such as the one installed in the TRAMO program, which avoid the identification step and estimate all the possible models within a subset, which is usually taken as $p \leq 3, q \leq 2, \ P \leq 2, Q \leq 1$.

▷ TRAMO also identify the mean, the log transformations, etcetera.

Example 84

Lets consider the regular and seasonal differentiated vehicle registration series. The most notable features of its ACF are: (i) a significant r_1 coefficient; (ii) significant coefficients in the seasonal lags, r_{12}, r_{24} and r_{36}; (ii) interaction around the seasonal lags, as shown by the positive and symmetric values of the coefficients r_{11} and r_{13} as well as r_{23} and r_{25}.

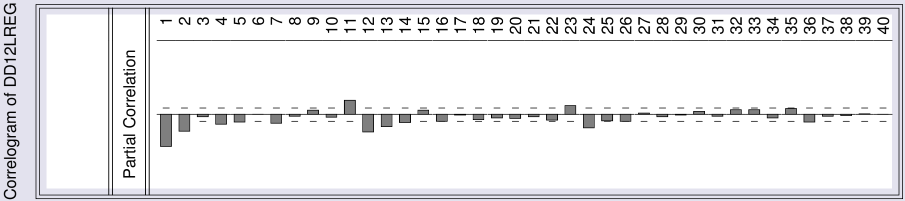


▷ *The regular part suggests an MA(1) model.*

▷ *The seasonal part is more complicated, since the observed structure is compatible with that of an AR(1)$_{12}$ with negative coefficient and with longer AR or ARMA(1,1) $_{12}$ models as well.*

Example 84

The PACF of this series confirms the MA(1) structure for the regular part: a geometric decay is observed in the first lags and, by the interaction as well, which repeats after the seasonal lags.



▷ *The two significant coefficients in the seasonal lags cause us to reject the hypothesis of an AR(1)$_{12}$, but they are compatible with an AR(2)$_{12}$ or with an ARMA(1,1)$_{12}$.*

▷ *Therefore, we move on to estimating models with MA(1) for the regular part and AR(2) or ARMA(1,1) for the seasonal part.*

And TRAMO selects ...

TRANSFORMATION: Z $->$ LOG Z

NONSEASONAL DIFFERENCING D= 1

SEASONAL DIFFERENCING BD= 1

MEAN OF DIFFERENCED SERIES -0.8281D-03

MEAN SET EQUAL TO ZERO

MODEL FITTED

NONSEASONAL P= 0 D= 1 Q= 1

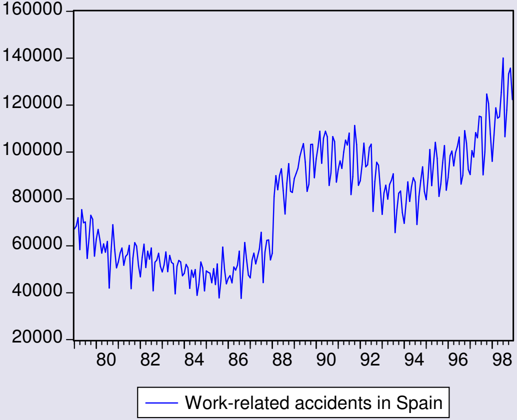
SEASONAL BP= 0 BD= 1 BQ= 1

PERIODICITY MQ= 12

TRAMO selects an $ARIMA(0,1,1)\times(0,1,1)_{12}$**.**

Example 85

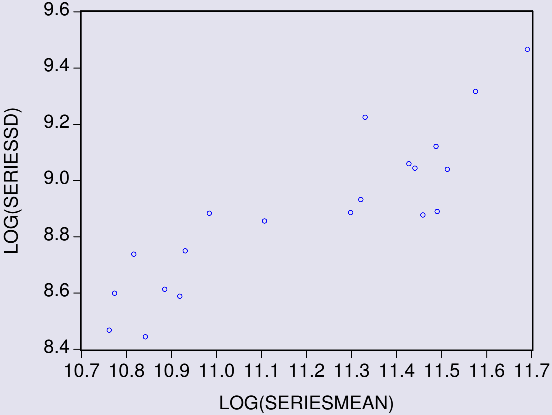
We are going to identify a model for the Spanish work related accidents series found in the accidentes.dat file. This file contains 20 years of monthly data from January 1979 to December 1998. The figure gives the graph of this series.



▷ *The series seems to show an increase in variability with level.*

Example 85

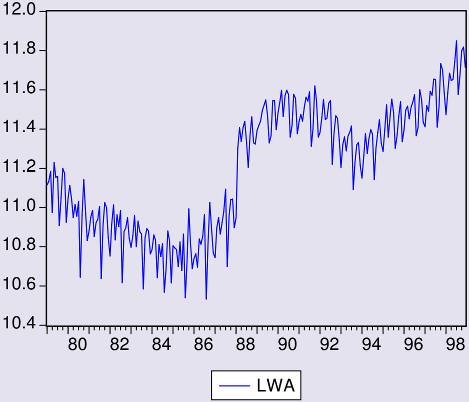
The figure gives the relationship between the logarithm of the standard deviation each year and the logarithm of the mean for the year.



▷ *A linear relationship is observed with a slope that is slightly less than the unit, thus we will take logarithms as a first approximation.*

Example 85

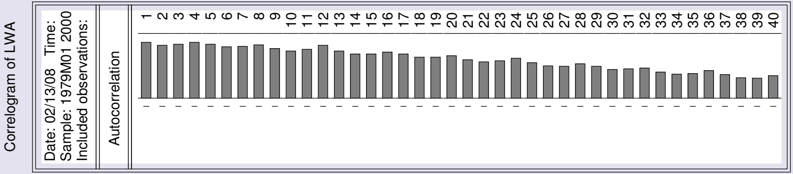
The graph of the series in logarithms is shown in the figure. The log transformation may be too strong because the variability of the first two years now seems slightly greater than that of the last.

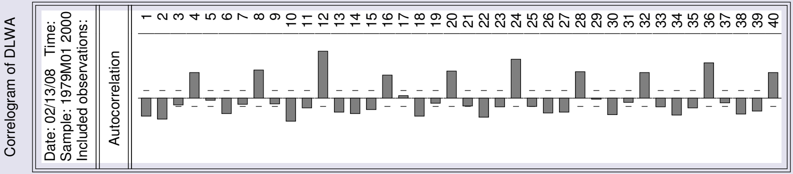


▷ *We have tried using the square root as well, and the result is a little better, but for ease of interpretation we will work with the series in logarithms.*

Example 85

The graph of the series indicates the need for at least one regular difference for the series to be stationary and the estimated ACF of the $\nabla \log z_t$ transformation shows high coefficients and decays slowly in lags 12, 24, 36.

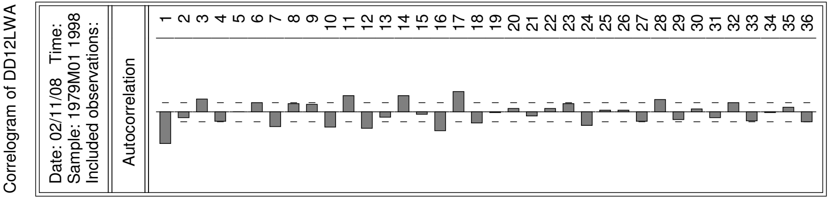


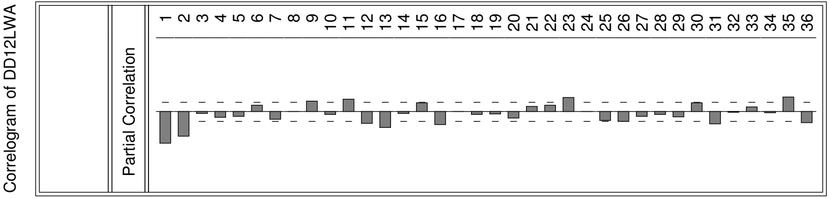


▷ *So we take a regular and a seasonal difference.*

Example 85

The figures give the simple and partial correlogram of the series $\nabla\nabla_{12}\log z_t$. In the ACF we see significant coefficients in the regular lags 1 and 3, and in the seasonal lags 12, 24 and 36. Furthermore, several significant coefficients appear around the seasonal lags.





Example 85

- *Starting with the regular part, the ACF suggests an AR process for the regular part, which is supported by the numerous coefficients of interaction around the seasonal lags.*
- *As far as seasonality, there are significant lags in 12 and 24 and in the limit for 36. The simplest hypothesis is an MA(2)$_{12}$, but it could also be AR or ARMA.*
- *Regarding the PACF, two significant lags appear in the regular part, which suggests an AR(2) for this part.*
- *In the seasonal lags there are significant coefficients in lags 12 and 36, which suggests that the seasonal structure might be either an MA or an AR greater than two, or, alternatively, ARMA.*

▷ *As a conclusion to this analysis, in the next section we will estimate an* $AR(2) \times MA(2)_{12}$ *as well as more complex models of type* $ARMA(2,1) \times ARMA(2,1)_{12}$.

And TRAMO selects ...

TRANSFORMATION: Z $->$ LOG Z

NONSEASONAL DIFFERENCING D$=1$

SEASONAL DIFFERENCING BD$=1$

MEAN OF DIFFERENCED SERIES 0.5834D-03

MEAN SET EQUAL TO ZERO

MODEL FITTED

NONSEASONAL P$=2$ D$=1$ Q$=0$

SEASONAL BP$=0$ BD$=1$ BQ$=1$

PERIODICITY MQ$=12$

TRAMO selects an $ARIMA(2,1,0)\times(0,1,1)_{12}$.