EPIET TIME SERIES MODULE

ARIMA MODELLING

Bordeaux

10-12 May 2004

Case study: Forecasting a time series using Box-Jenkins modeling

Objectives: at the end of the case-study, the participant should:

- **Understand principles of Box-Jenkins forecasting**
- **Use the autocorrelogramme and the periodogramme to describe a time series**
- **Manipulate CDC Statistical Software for Surveillance (SSS1)**
- **Perform modeling of a time series**
- **Understand the model**

Denis Coulombier*, Philippe Quenel, Bruno Coignard****
*** WHO/CDS/CSR/LYO/EPS, Lyon, France**
**** Institut de Veille Sanitaire, Paris, France**

Presentation

This case study includes 4 parts. The first part illustrates principles of Box & Jenkins analysis while the last 3 parts are meant to be done by the participant using his/her computer.

- **Part 1: definition and principle**
- **Part 2: plotting the time series and achieving stationarity**
- **Part 3: identifying the model and running diagnostics**
- **Part 4: forecasting values**

Programs needed on the computer:

- SSS1: Statistical Software for Surveillance

Files used by the case study:

- CTYPH.DAT
- JENKINS.XLS

Text style used in the case-study

Commands to type in the computer. The text between ' and ' is the text you actually need to type

Additional information about the programs

Output of SSS1

Part 1: definitions and principles

Introduction

The analysis of time series in public health contributes to improve knowledge about diseases. Usual statistical methods assume that the observed data are realization of independent random variables. Analysis of time series requires specific techniques to account for lack of independence of values in the series. Box & Jenkins modeling was developed in the 70s and first applied to business. Armitage applied it successfully to disease surveillance later. Surveillance data has the reputation of not being suitable for sophisticated analysis because of its poor quality. However, we use these data routinely to make decision that impact on the health of people.

With these techniques, it is possible to

- forecast future values of health indicators, and define thresholds for change in disease pattern
- evaluate the impact of interventions
- study correlation between several series such as atmospheric pollution and health indicators

This case study represents a short introduction to Box & Jenkins modeling.

Definition

Time series: a time series is a series of health indicators recorded at regular intervals over time.

Stationary series: a time series is stationary if it has a constant mean, variance, and autocorrelation through time (e. g. seasonal dependencies have been removed). It means that the probability structure of the series does not change with time.

Variance: it is the sum of the square of the difference between individual values and the mean, divided by the number of observations:

$$\text{Variance} = \frac{\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2}{N}$$

Where :

$$\bar{x} = \frac{\sum_{t=1}^{n} x_t}{N}$$

Covariance: for 2 series, the covariance is the sum of the product of deviations from the mean from each series. For 2 sets of data, X and Y, the covariance is calculated by:

$$\text{Covariance} = \frac{\sum\left(x_i - \bar{x}\right) \times \left(y_i - \bar{y}\right)}{N}$$

Correlation coefficient: it is the ratio of the covariance of 2 series over the square root of the product of the variances in each series. It indicates how much of the variation in one series is explained by variation in the other.

Autocorrelation coefficient: same as the correlation coefficient, but for 2 series extracted from the same signal, at a certain interval apart. Since data is gotten from one unique series, it is called **AUTO**correlation.

The autocorrelation coefficient for a lag = k is given by the formula:

$$r_k = \sum_{t=1}^{N-k}(x_t-\bar{x})\times(x_{t+k}-\bar{x}') \Bigg/ \sqrt{\sum_{t=1}^{N-k}(x_t-\bar{x})^2 \sum_{t=1}^{N-k}(x_{t+k}-\bar{x}')^2}$$

which can be approximated by:

$$r_k \approx \sum_{t=1}^{N-k}(x_t-\bar{x})\times(x_{t+k}-\bar{x}') \Bigg/ \sum_{t=1}^{N}(x_t-\bar{x}'')^2$$

Where $\bar{x}$: mean of (N-k) observations (lag 0)
$\bar{x}'$: mean of (N-k) observations (lag k)
$\bar{x}''$: mean of the entire series

Partial autocorrelation: measures the correlation between observations at lag k, taking in account correlation at lag k-1. It corresponds to the relative autocorrelation.

ACF: autocorrelation function, the set of autocorrelations coefficient for different lags.

PACF: partial autocorrelation function, the set of partial autocorrelation coefficients for different lags.

Principles

Studying time series consists of studying the probability laws which generates the series. Classical analytic methods are not appropriate for time series because observations are not independent as they are in a random sample. The chronology of observations is usually correlated to their value. This is particularly true when looking at factors such as pollution. If the pollution indicator is high a given day, chances are that this would influence the level of pollution for the next day. The smaller the interval between 2 measurements, the higher should be the correlation. When the series appear to be stable, we can use linear models, such as ARIMA models developed by Box & Jenkins in 1970. Such models are used to forecast series and develop thresholds, to study correlation between 2 series, or to evaluate impact of interventions (comparing 'before' with 'after'). Because these models imply a 'stable' series, we need to transform the series to achieve stationarity before modeling.

Stationarity

A time series is stationary if it has a constant mean, variance, and autocorrelation through time (e. g. seasonal dependencies have been removed). In other words, their value is independent of the time of observation. This is achieved through differencing. The observation at time t is replaced by the difference between observation at time (t-1) and t (differencing of order 1)

Autocorrelation

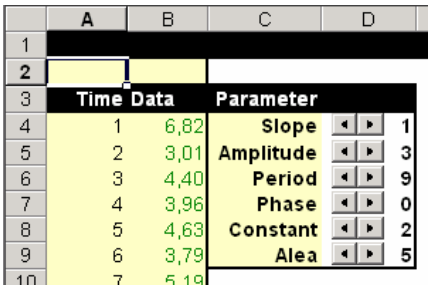
The autocorrelation measures correlation between observations at different distances apart. In other words, how much an observation at a certain time is dependent of earlier observations. In a random "white noise" signal, such as the one in figure 1, the correlogram shows rapid decay to 0 since random values are not correlated to each other. We plot the ACF and PACF functions to study the original series first. This will indicate how much correlation appears in the time series, and guide us through identifying the best suited transform.

The ACF correlogram plots the value of the autocorrelation coefficient for different values of lag. The pattern of the ACF function gives us indication about the underlying process. If it does not decay to 0 with increasing lags, it indicates a strong trend in the series (figure 2). In other words, a value at a given time is highly dependent of the value at previous times. In a random time series (figure 1), the ACF function decays to 0 rapidly, meaning values are not correlated, and seem to appear randomly distributed.

You can "play" with correlograms using the JENKINS.XLS file provided.

```
Start EXCEL
Load the file JENKINS.XLS
```

Time units appear in cells A4:A67, cases in cells B4:B67. Rows 68 and 69 include information but are not displayed on the Time series graphic.



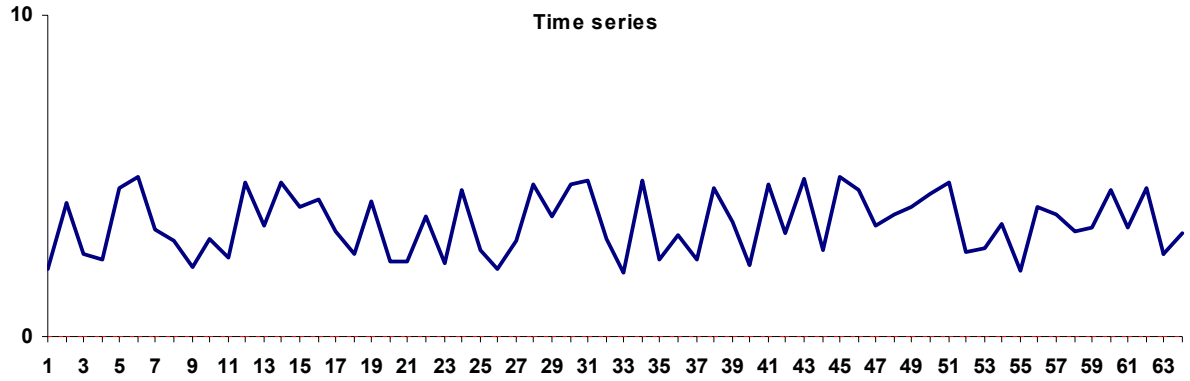
	A	B	C	D
1				
2				
3	Time	Data	Parameter	
4	1	6,82	Slope	1
5	2	3,01	Amplitude	3
6	3	4,40	Period	9
7	4	3,96	Phase	0
8	5	4,63	Constant	2
9	6	3,79	Alea	5
10	7	5,19		

The characteristics of the time series can be altered in several ways:

- A linear slope, positive or negative can be set by clicking on the spin button next to it
- A sine curve, in which the amplitude, the period of the phase can be adjusted using the corresponding spin buttons
- A constant and a random factor(Alea) can be added

Assessing randomness

The default time series present with neither trend, nor cycles.

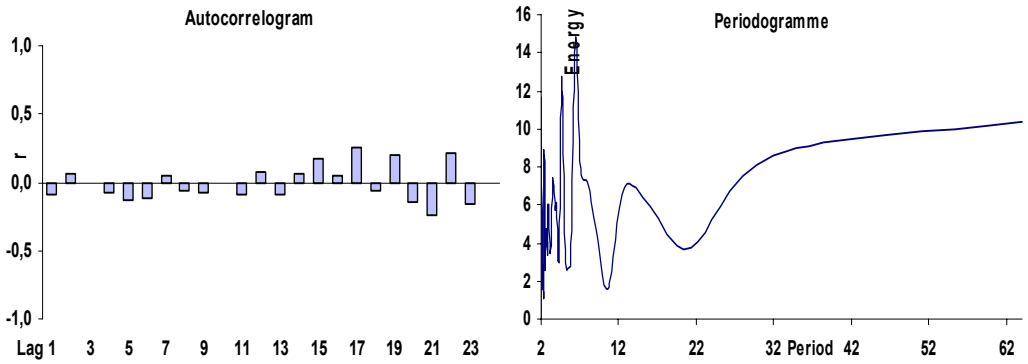


Values randomly vary from 2 (the constant), to 5 (the constant + the random factor).

On such a random signal, the auto-correlogramme will show spikes of low amplitude, and not presenting with any pattern. Similarly, the periodogramme will not show frequency for which there would be significant cyclical contributions. However, just by chance, there may be some level of cycles in the data generated randomly. In such case, what you get is a snapshot of randomness which would not persist if new random values are generated.

You can assess "the stability" of the auto-correlogramme and of the periodogramme by using the Recalculate button, which recalculates the spreadsheet and updates the graphs.

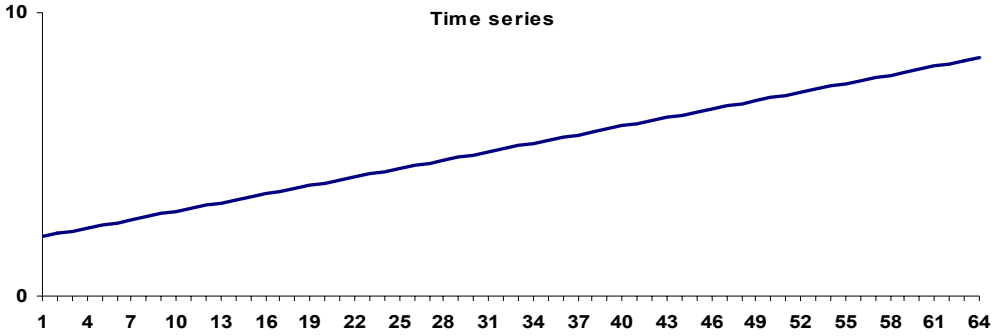
Example of correlogramme and periodogramme of a random function:



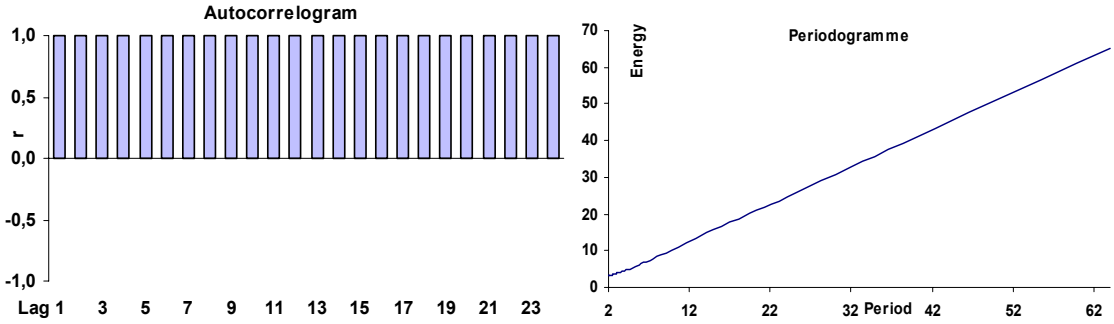
In time series analysis, you cannot use such features to assess the stability of the figures. However, you can display them for subsets of data and check if you get a consistent pattern on various areas of the dataset.

Assessing trend

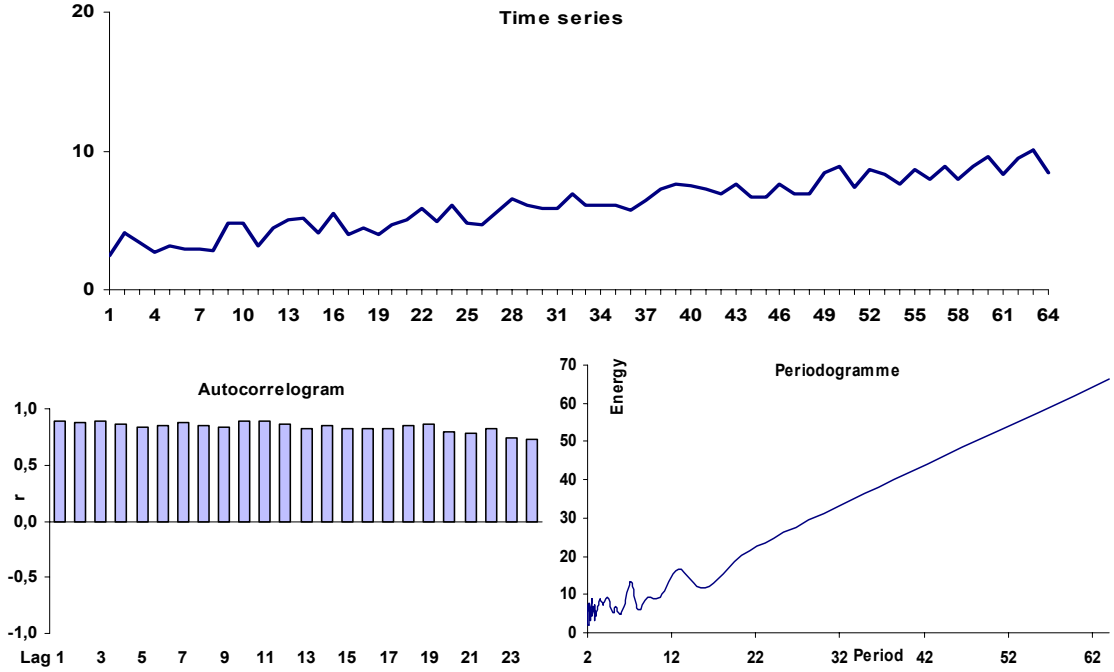
A pure linear positive trend will be reflected on the periodogramme by autocorrelation coefficients of 1 at all lags. A pure linear trend will generate a periodogramme with increasing coefficients for increasing periods being tested.



Example of correlogramme and periodogramme of a pure linear trend

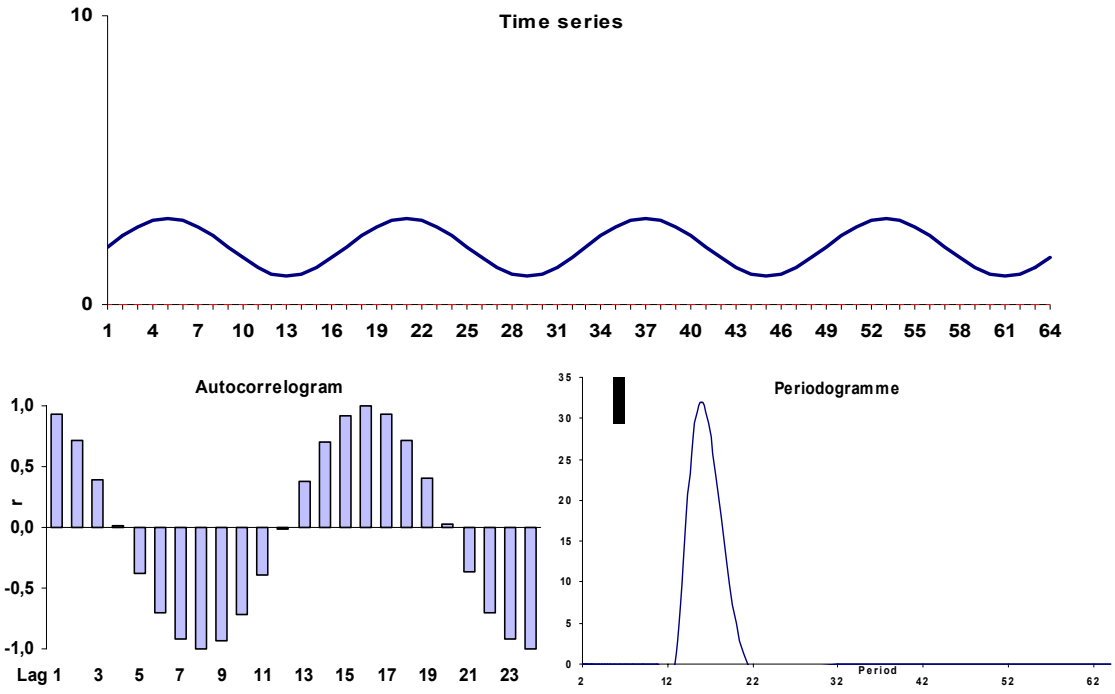


If a random factor is added, the correlation coefficients will remain positive, but decrease in magnitude, and some fluctuations will be added to the short period on the periodogramme.

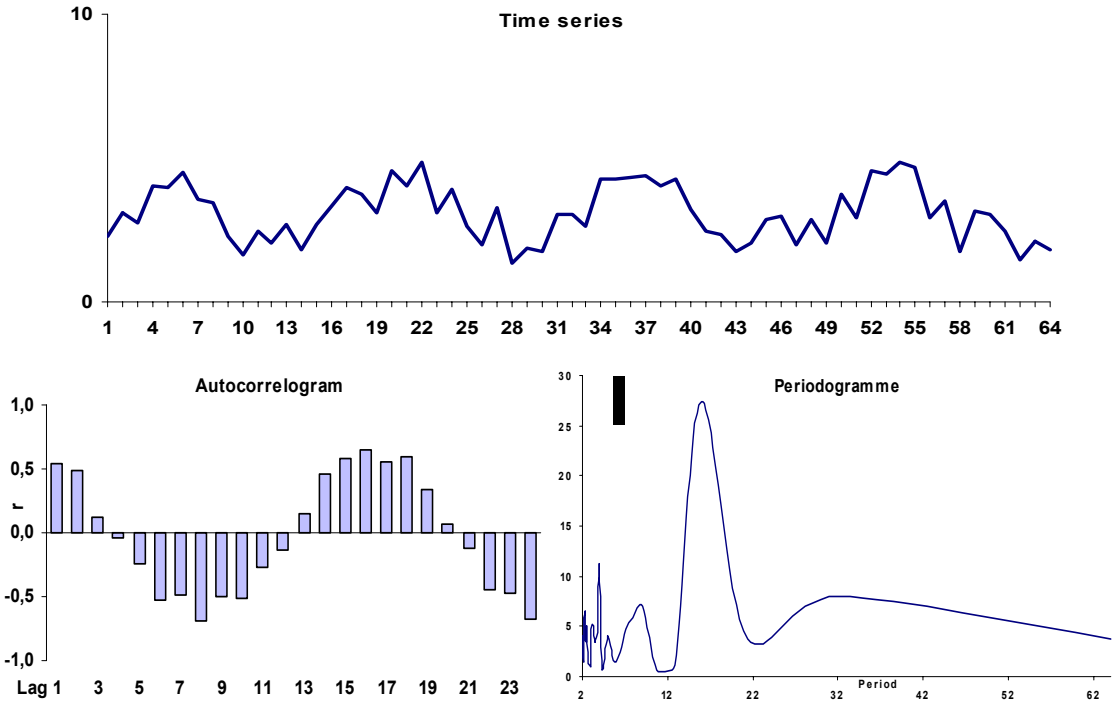


Assessing cycles

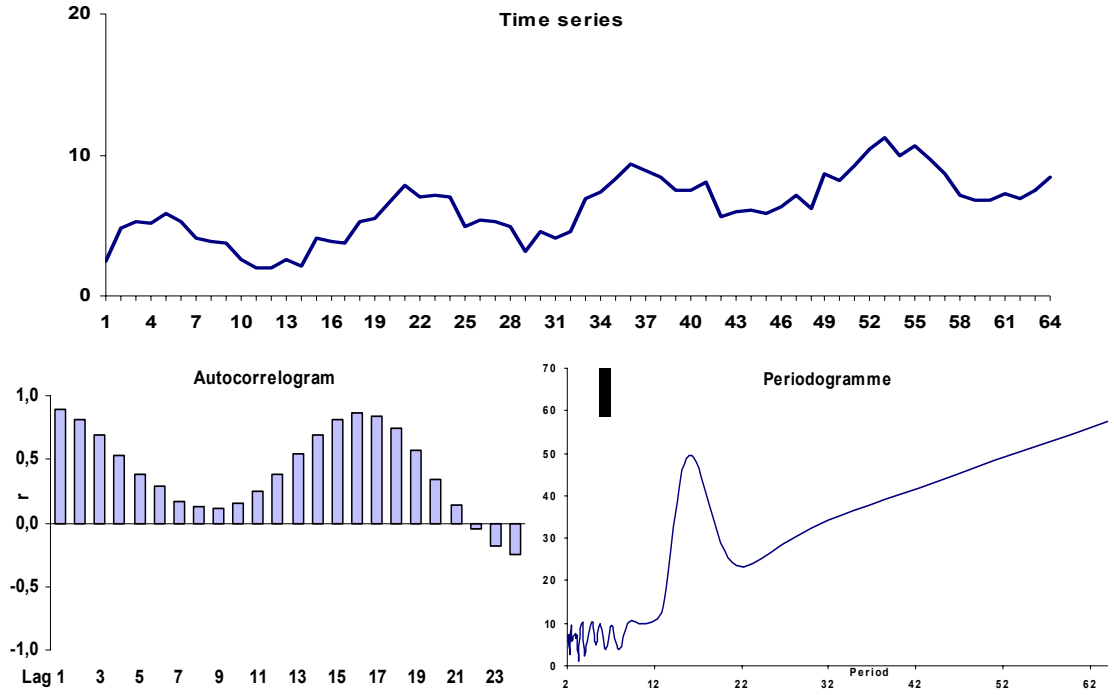
If the signal corresponds to a a pure sine curve, the autocorrelogramme will show sinusoid variations of the autocorrelation coefficients, returning to the value of 1 after 1 cycle (16 time units in the example below). The periodogramme will show one peak for the corresponding period.



If a random factor is added, the correlation coefficients will remain sinusoidal, but will decrease in magnitude, and some fluctuations will be added to the short period on the periodogramme.



If the signal is a combination of a trend, a cyclical component, and some random variation, the autocorrelogramme will show a decaying sine curve and the periodogramme will show a peak at the seasonal lag (cycle), increasing values at longer lags (trend), and some fluctuations for short periods.



Explore the behavior of the autocorrelogramme and the periodogramme for various combinations of trend, cycles and random factor

Note that the phase does not affect the autocorrelogramme or the periodogramme.

Steps involved in modeling a signal

- Transform the series to make it stationary
- Identify the model
- Estimate model parameters
- Check the model
- Forecast values

First step: transform the series to make it stationary

In real world, most series are not stationary. The first step is to observe the series, and detect visually any trend or cyclical component in the series. The variance may increase over time, requiring stabilizing it. A log transform is usually the first one to try. Other transforms may be required. A linear trend can be removed by a difference of order 1. In this case, the value at time $Y_{t+1} = X_{t+1}-X_t$. To remove a seasonal factor, we can use a difference of order corresponding to seasonality. For example for monthly values, we would use a difference of order 12.

Choice of the appropriate transformation

Log transform:

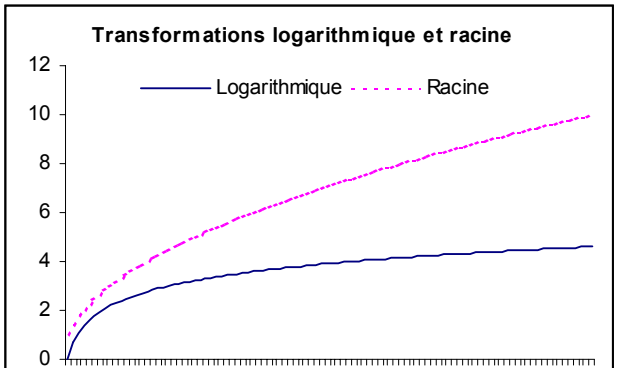
- Indicated when the variance increases with the mean
- to remove an exponential slope

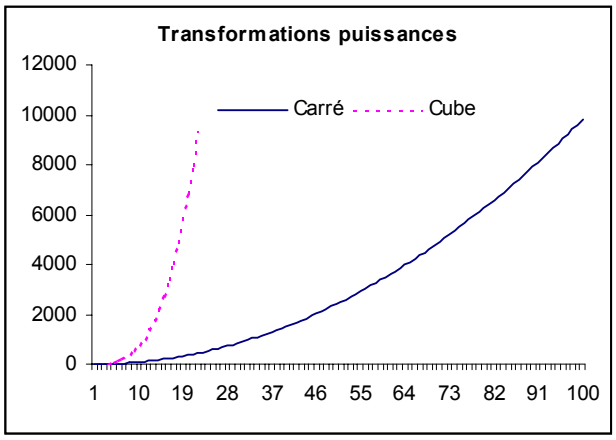
Square root transform :

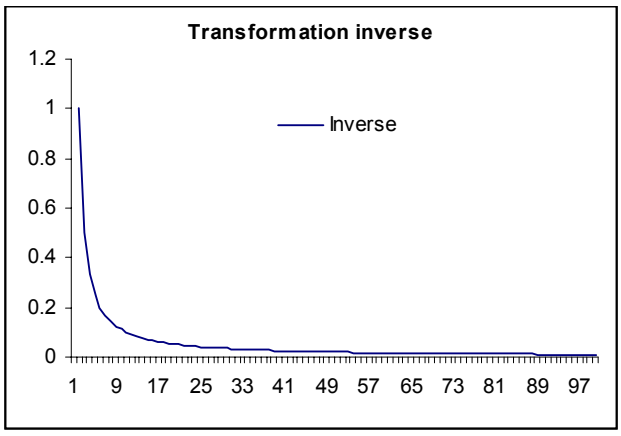
- Same indication as for log, but less marked

Reciprocal transform

- Indicated when the variance is proportional to the mean to a power 4
-







Differencing

A linear trend can be removed by a differencing of order 1. In such case, value at time (t) is replaced by the difference between value at time (t-1) and t.. To remove a seasonal component, a differencing of seasonal order is used. For example, for a series showing a 52 week seasonal component, a differencing of order 52 is recommended.

Second step: Identify the model

Box & Jenkins have proposed several theoretical models. These models involve 2 different processes:

Autoregressive process AR(p):
A process is said autoregressive if

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_k X_{t-p} + \varepsilon_t$$

where ϕ are constants, the parameters of the model, and ε_t the remaining "white noise". This process shows dependency between observations at certain lags in time. In a pure autoregressive model of order p, the ACF decays exponentially to 0, and the PACF is null after p.

Moving average process MA(q):
A process is said moving average if:

$$X_t = \theta_0 \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_q \varepsilon_{t-q}$$

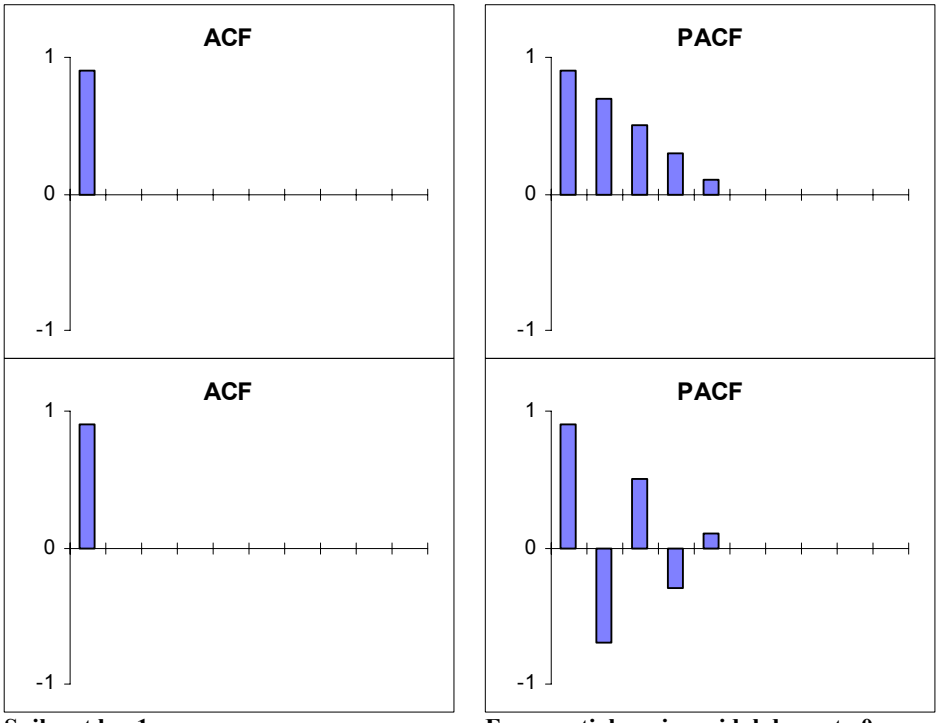
where θ are constants, the parameters of the model, and ε_t a "white noise". In other words, a value at time t depends on past and present variations in the series, and not values as in the autoregressive process. In a pure moving average model of order q, the PACF decays exponentially to 0, and the ACF is null after q.

Mixed process or autoregressive moving average process ARMA(p,q):
This is a combination of the 2 above processes. In this case, both the ACF and PACF decay exponentially to 0.

ARIMA process ARIMA(p,d,q):
It is an ARMA process in which differencing has been performed (d) to remove trend and stabilize the mean. This is the process used by SSS1.

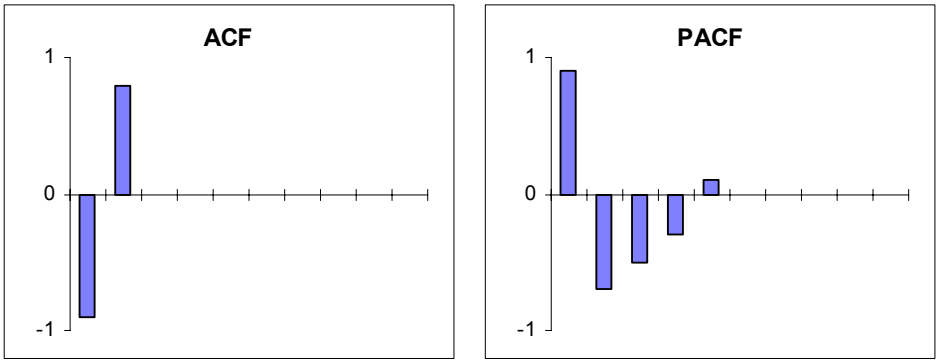
Theoretical models:

They appear in the manual pages 50-52.



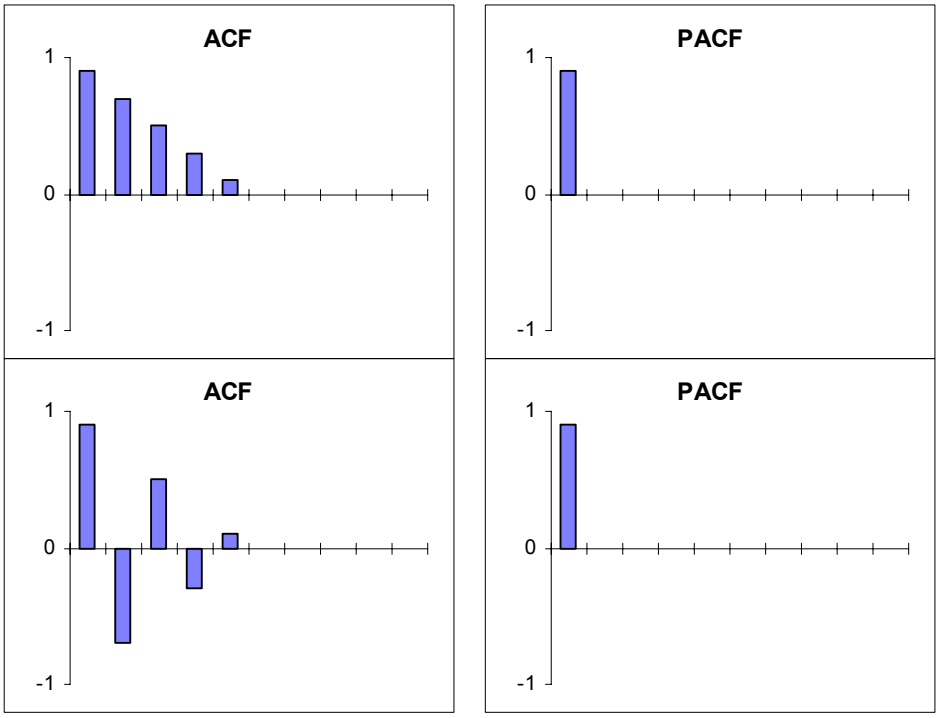
Spike at lag 1 **Exponential or sinusoidal decay to 0**

MA(1), ARIMA (0,d,1)



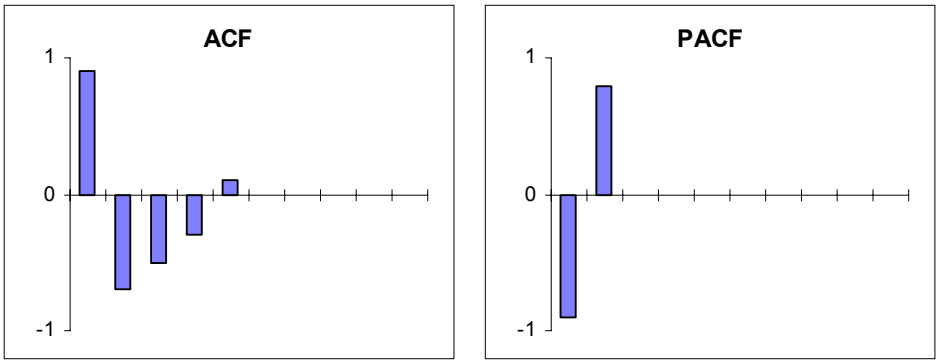
Spike at lag 2 **Tails off to 0**

MA(2), ARIMA (0,d,2)



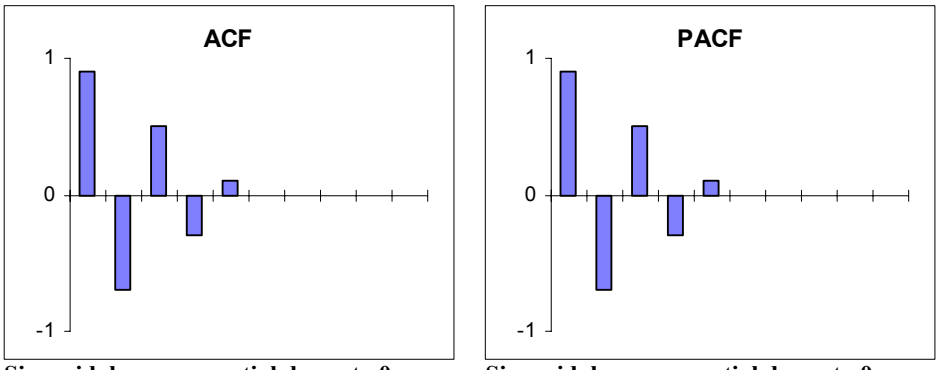
Sinusoidal or exponential decay to 0 **Spike at lag 1**

AR(1), ARIMA(1,d,0)



Tails off to 0 **Spike at lag 2**

AR(2), ARIMA (2,d,0)



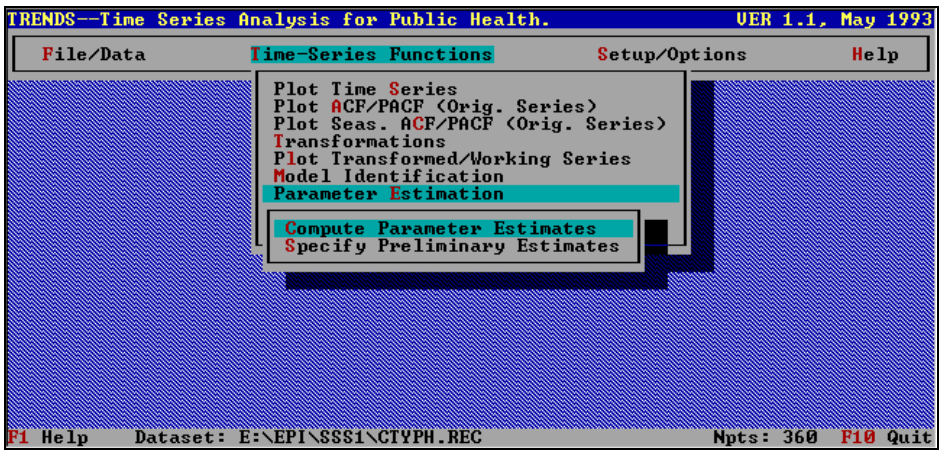
Sinusoidal or exponential decay to 0 **Sinusoidal or exponential decay to 0**

AR(1), MA(1), ARIMA (1,d,1)

The selection of the proper model is done by carefully looking at the ACF and PACF correlograms. Furthermore, in Box & Jenkins modeling, this process has to be done for both trend modeling and seasonal modeling. For a given time series there may be several models fitting the data. In this case, we choose the simplest, most economical model, the fewest terms explaining most variation. It is the principle of parsimony. In fact if one includes all autoregressive lags in a model, the series can be totally reproduced. But this would be meaningless since the residual white noise would be in the model. The behavior of this residual white noise can be modeled in the observed series, but cannot be predicted in the future.

Estimate model parameters

This is done (fortunately) by the software! Least square and maximum likelihood methods are used.



Check adequacy of the model

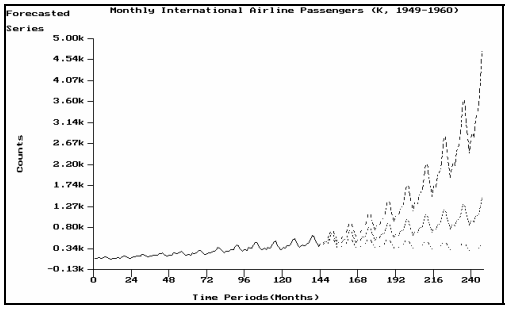
Checking the adequacy of the model requires looking at :

- parameters by a Student test on each
- randomness of residuals by residual ACF and PACF
- autocorrelation of residuals by Box-Ljung statistical test

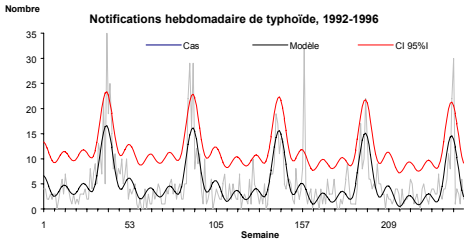
If the model is not adequate, then we change it with guidance from the residual ACF/PACF. Any parameter not significant is removed, and new one can be added. This is a recursive process until the "most adequate" model is identified.

Forecast values

This is the final step. Once the model is adequate, we use it to derive future values. As a difference with spectral analysis, the interval of confidence around the prediction increases with time.



Diverging interval of confidence for Box & Jenkins



Non diverging confidence interval for spectral analysis

Summary

When building a model, it may be advisable to remove outliers, or to remove known epidemics in the time series. This is done according to the objective of the modeling. If epidemic periods are included, they will be forecasted, and thus the model will not permit early detection of epidemics, but rather detection of epidemics of unusual magnitude.

We have looked at general principles in modeling time series using Box & Jenkins method. Further reading are needed in order to understand all underlying statistics. Among other:

- Ulrich Helfenstein, The use of Transfer Function Models, Intervention Analysis and Related Time Series Methods in Epidemiology. *International Journal of Epidemiology*, 1991, **20**, 808-815.

- Laurence Watier. Revue méthodologique de quelques techniques spécifiques à l'analyse des séries temporelles en épidémiologie et santé publique. *Revue d'Epidémiologie et de Santé Publique*, 1995, **43**, 162-172.

- SSS1 manual! Pages 27 to 62 and 105 à 128.

- C. Chatfield, *The analysis of time series, an introduction*. Chapman and Hall, 1975, ISBN: 0-412-26030-1

- Daniel Schnell, Akbar Zaidi,Gladys Reynolds. A time series analysis of gonorrhea surveillance data. *Statistics in Medicine*, 1989, **8**, 343-352.
- Akbar Zaidi, Daniel Schnell, Gladys Reynolds. Time series analysis of syphilis surveillance data. *Statistics in Medicine*, 1989, **8**, 353-362.
- Stephen B. Thacker. Commentary. *Statistics in Medicine*, 1989, **8**, 363.

- Ulrich Helfenstein, Ursula Ackermann-Liebrich, Charlotte Braun-Fahrländer. The environmental accident at 'Schweizerhalle' and respiratory diseases in children : a time series analysis. *Statistics in Medicine*, 1991, **10**, 1481-1492.

- Laurence Watier. Analyse des séries temporelles des infections a Salmonelles non typhiques. *Revue d'Epidémiologie et de Santé Publique*, 1995, **43**, 173-185.

- Ulrich Helfenstein. Box-Jenkins modelling of some viral infectious diseases. *Statistics in Medicine*, 1986, **5**, 37-47.

- Laurence Watier, Sylvia Richardson. A time series construction of an alert threshold with application to *S. Bovismorbificans* in France. *Statistics in Medicine*, 1991, **10**, 1493-1509.

Part 2: Plotting the time series and achieving stationarity

In this part, we start building a model for typhoid cases in France. The first steps are:

- Plotting the time series
- Achieving stationarity

First step: plotting the time series

```
Change directory to C:\SSS1: 'CD \SSS1'
Load the Time Series module: 'TS'
```

SSS1 (SSS1.EXE) includes several modules useful for epidemiologist involved in surveillance:
Time series (TS.EXE)
Capture-recapture (CR.EXE)
Figure 1 (FIG1.EXE)
Robust trend analysis (RR.EXE)

The main menu of SSS1 gives access to all modules. However each module can be called separately by typing the name of the executable file. This case study will use only the Time Series module.

From the main 'Time series menu', use the 'File/Data' option to load the file CTYHP.REC

```
Click on 'File/Data'
Select 'Open/Read file'
Choose '*.REC'
Pick 'CTYPH.REC'
Select the field for disease count: 'COUNT'
Enter series seasonality: '52' since we are dealing with weekly
notifications of typhoid in France
```

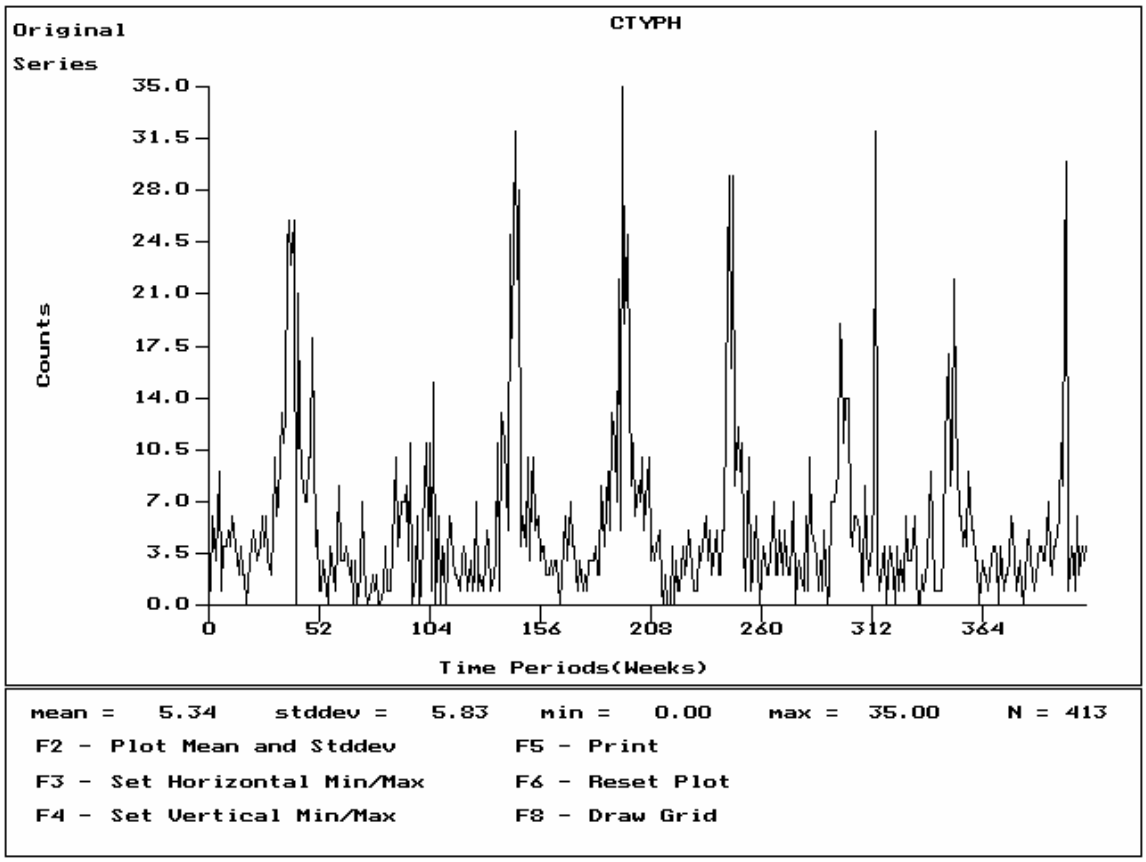
The time series shows on screen for visual inspection.

```
Press 'F2' to show mean and standard deviation
```

Visual inspection involves identifying the 2 basic components: trend and seasonality.

This series does not seem to show obvious trend over the data points. However, this is rather difficult to assess precisely given the strong seasonality. It shows obvious seasonality at 52 week lags. The epidemics are not every year of the same magnitude. Around week 104, the epidemic is much smaller than for other years (this seems to have to do with a strike of the public health officers at this time!). Furthermore, around week 302, there is a pick that does not correspond to the seasonality noticed in previous years.

Figure 1: Typhoid cases in France by week, 1989-1996



Second step: achieving stationarity

A series that varies uniformly about a constant mean is considered stationary. Achieving stationarity involves stabilizing the variance through power transform, and stabilizing the mean through differencing. Variance should be always stabilized first. Statistical tools are used to guide this process.

Stabilizing the variance

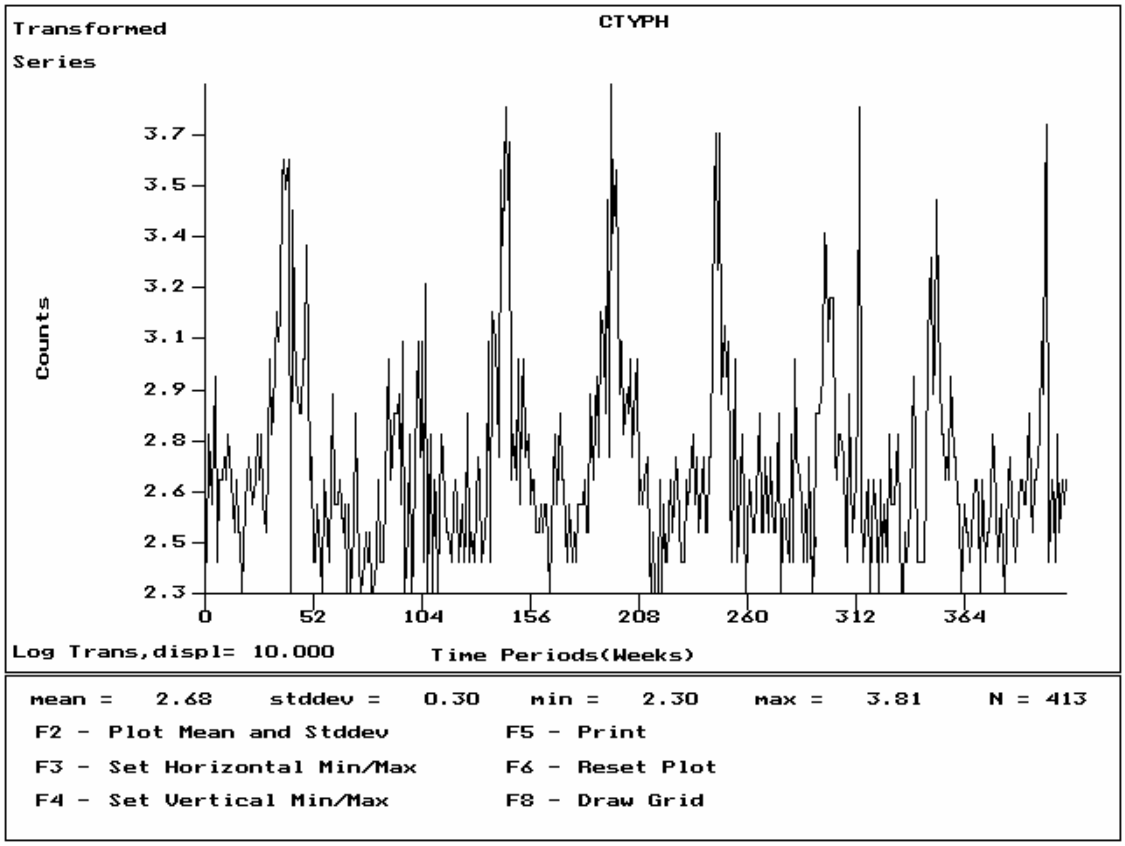
A stabilization of the variance whenever there is a visual impression that the variance differs according to periods or level of mean. To stabilize the variance we use a transform. SSS1 includes a diagnosis tool to identify the best suited transform for the series. In fact, the log transform should be the first tried since it usually stabilize the variance. In our case, SSS1 recommend the reciprocal transform, but it does not yield a better stabilization of the variance, giving a standard deviation of 0.30.

```
Press 'ESC' to go back to the menu
Select 'Transformation' in the 'Time series functions'
Choose 'Stabilize variance'
Pick 'Log' transform
Confirm a displacement of 10 to perform the transform
```

A displacement is required since we have a couple of data points with 0 cases and log(0) does not exist. The displacement adds 10 to each data point, thus allowing a log transform. However this has no effect on the modeling process.

The transformed series shows:

Figure 2: Typhoid cases in France by week, 1989-1996, after LOG transform and displacement



The series ranges between 2.30 and 3.81, with a standard deviation of 0.30.

Stabilizing the mean: differencing

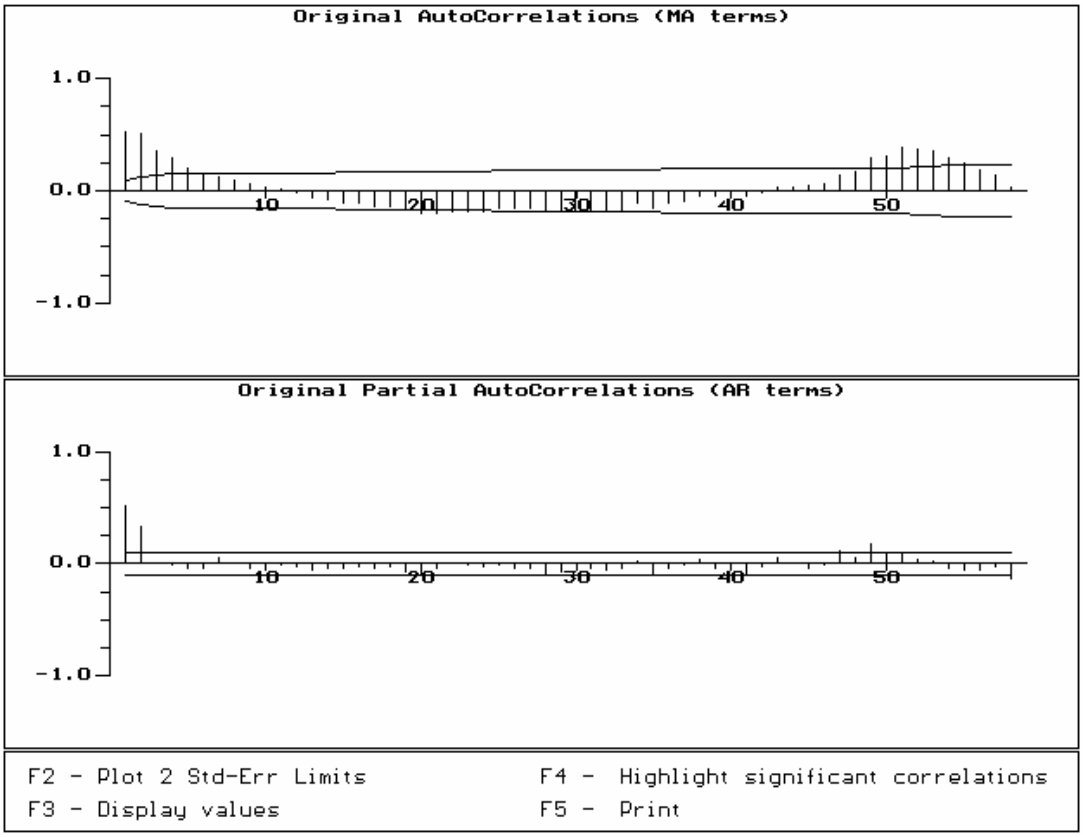
At this junction we need to stabilize the mean to achieve stationarity. We use the Autocorrelation Function to guide us through this process. Remember, an ACF not decaying rapidly to 0 indicates a trend, because values at time t are related to values at previous times.

```
Press 'ESC' to go back to the menu
Select 'Plot ACF/PACF (Orig. series)' to display the ACF function
```

The ACF functions shows on screen (figure 3).

The ACF is the top graphic on figure 3. It moves to 0 after 10 lags, and then peaks again around 52 weeks. This indicates some level of linear trend, and a seasonal component of 52 weeks. The PACF shows the contribution taking into account contributions of previous lags. The ACF shows absolute contributions of each lag. At this stage, we use only the ACF plot to find the best suited differencing. From figure 3 we choose to use a differencing of lag 1 and 52 to account to autocorrelations in the series.

Figure 3: Typhoid cases in France by week, 1989-1996, ACF/PACF plots.



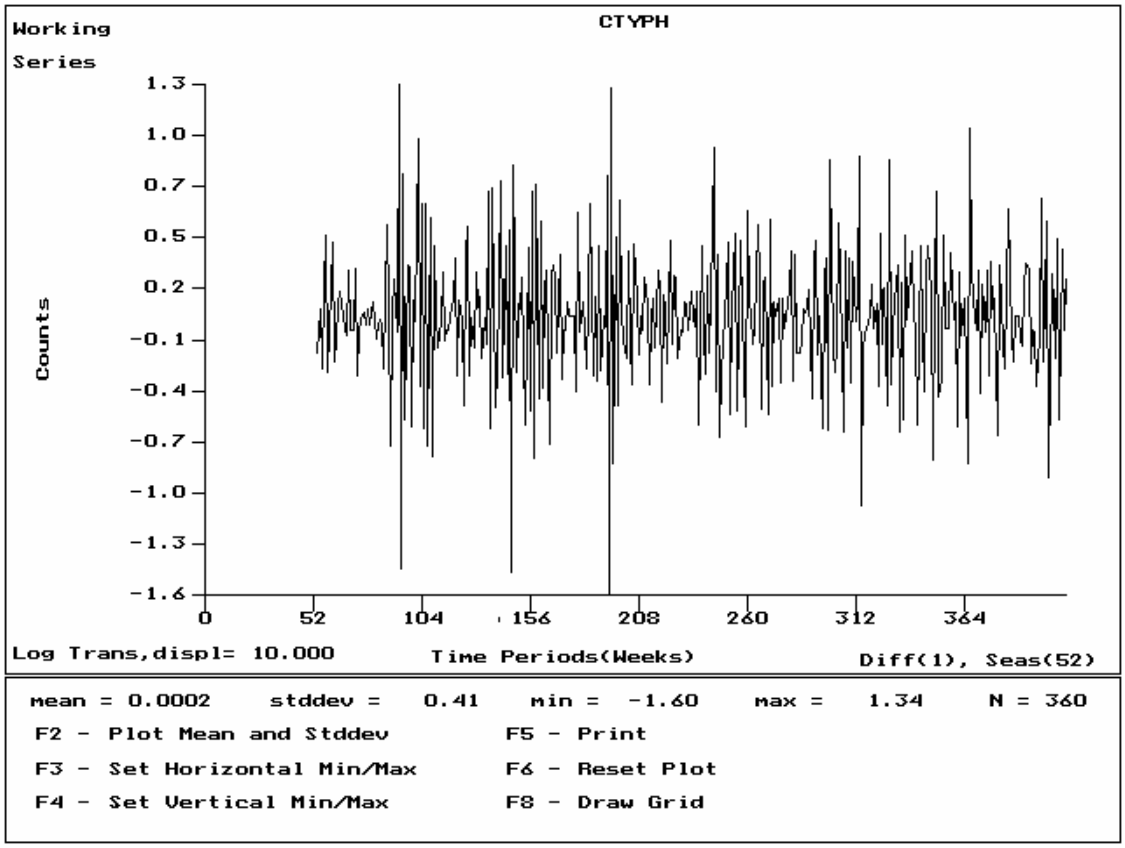

```
Press 'ESC'
Select 'Transformation'
Choose 'Stabilize mean'
Pick 'Regular differencing (lag=1)'
```

The working series after differencing of lag 1 shows.

```
Press 'ESC'
Select 'Transformation'
Choose 'Stabilize mean'
Pick 'Seasonal differencing (lag=52)'
```

The working series after both differencing shows (figure 4).

Figure 4: Typhoid cases in France by week, 1989-1996, after differencing.



Since we have used differencing at lag 52, the first 52 points in the data set are removed. The series now is centered around 0, and seems to oscillate randomly around this value. The computer stores in memory the first 52 values. With these values, it will be able to construct the original signal by summing them up, plotting a cumulative curve.

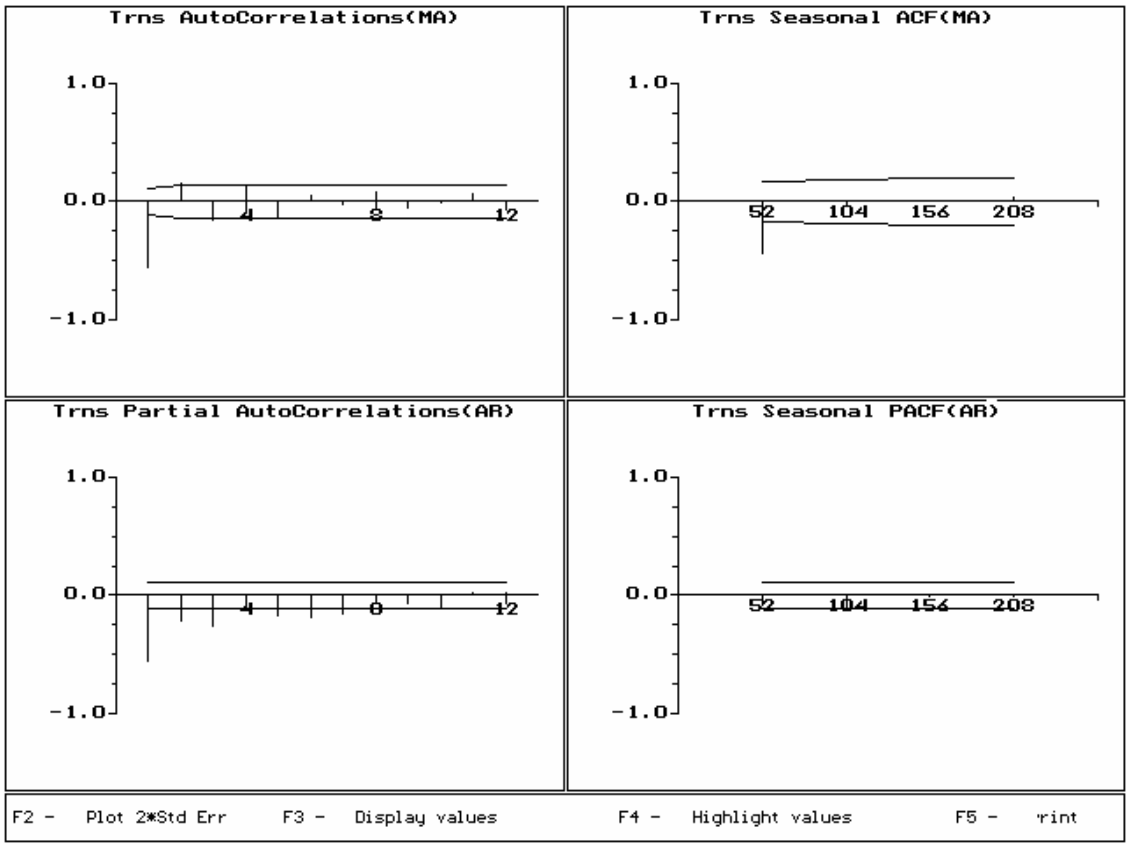
We check the ACF/PACF on the transformed series to check for stationarity (figure 5).

```
Press 'ESC'
Select 'Plot transformed/working series'
Pick 'Plot ACF/PACF'
```

The decay to 0 in the ACF is reached after 5 lag (figure 5). We consider the series stationary, compared with the original ACF (figure 3) where the decay to 0 was taking longer.

At the time of forecasting values, the transformation and differencing will be removed by integration, and the original series will be reconstructed without any loss of information.

Figure 5: Typhoid cases in France by week, 1989-1996, transformed series ACF/PACF.



The 2 left correlograms refer to trend correlations. The 2 right correlogams refer to seasonal correlations. ACF appears on top, PACF on bottom.

Summary

As in most epidemiological study, the first step in modeling data is to look at the series and identify its component:

- trend
- seasonality
- random noise

Statistical tools such as correlogram are used to identify these components and assess stationarity.

Part 3: Identifying the model and running diagnostics

In this part, we will:

- Identify the underlying model
- estimate the autoregressive and moving average terms of the model
- run diagnostics to check the validity of the model
- adjust the model accordingly

First step: model identification

Model identification consists in identification of the AR and MA terms of the model. We use the theoretical patterns of autocorrelation in the manual page 50-52 for guidance. Negative and positive spikes have the same absolute meaning, but in opposed directions. We first look at the 2 left plots for trend terms (figure 5). The ACF has a spike at 1 and then small spikes at 2, 3, and 4. The PACF presents an exponential decay to 0. This fits the bottom theoretical model of page 51 of the SSS1 manual, indicating a need for both AR and MA terms of order 1.

The 2 right plots (figure 5) show the seasonal components of the model. It shows a spike at 52 in the ACF, indicating a need for a 52 MA term.

Second step: parameter estimation

Once the relevant terms have been identified, we use SSS1 to build the model.

```
Goto 'Model identification'
'Build an ARIMA model'
'Select Auto Regular Autoregressive Term'
Select 'Order 1'
A tick mark appears in front of the option.
'ESC' to return to previous menu
'Select Auto Regular Moving Average Term'
Select 'Order 1'
A tick mark appears in front of the option.
'ESC' to return to previous menu
'Select Auto Seasonal Moving Average Term'
Select 'Order 52'
A tick mark appears in front of the option.
'ESC' to return to previous menu
'ESC' to return to previous menu
'ESC' to return to previous menu
Select 'Parameter estimation'
Choose 'Compute parameter estimates'
```

The parameter estimation dialog box shows the following output:

```
PARAMETER ESTIMATION
Itr   Sum of Squares    Parameters
 0  5.550275E+0001   0.100000   0.100000   0.100000
 1  5.091191E+0001  -0.400000  -0.326598   0.100004
...
13  1.912976E+0001   0.048087   0.909883   0.869455
14  1.912976E+0001   0.047503   0.909712   0.869455
15  1.912976E+0001   0.047503   0.909712   0.869455

Estimated residual variance= 0.048192
                                          Press any key to continue
```

Third step: diagnostic checking

```
Press 'ESC' to see the residual ACF/PACF
```

At this junction, we look at the ACF/PACF for the residuals (figure 6), values not 'explained' by the model we have selected.

58 data points are displayed in the ACF/PACF correlogram. We notice a few significant spikes on both ACF and PACF at lag 2 and 43, and one at lag 8 on the PACF. These spikes are not highly significant, and the residuals look pretty much uncorrelated.

Let's look at the model terms.

```
Goto 'Model Identification'
Select 'Display model term'
```

```
  Term#      Type       Order
    1        REG AR       1          (p,d,q) * (P,D,Q), Constant term
    2        REG MA       1           1 1 1     0 1 1          N
    3        SEA MA      52          Log Transformation with
             <end of list>                 10.000 displacement
                                    Maximum Order of Model 54
                                      Npts = 262; Back-Cast ON

                                                        Press any key to continue
```

The syntax (p,d,q) * (P,D,Q) is used to characterize the model. In SSS1, the p,d,q terms refer to trend modeling while the P,D,Q refer to seasonality modeling.

- p is the term for autoregressive process
- d is the term for differencing
- q is the term for moving average process

In this example, we have an ARIMA model:

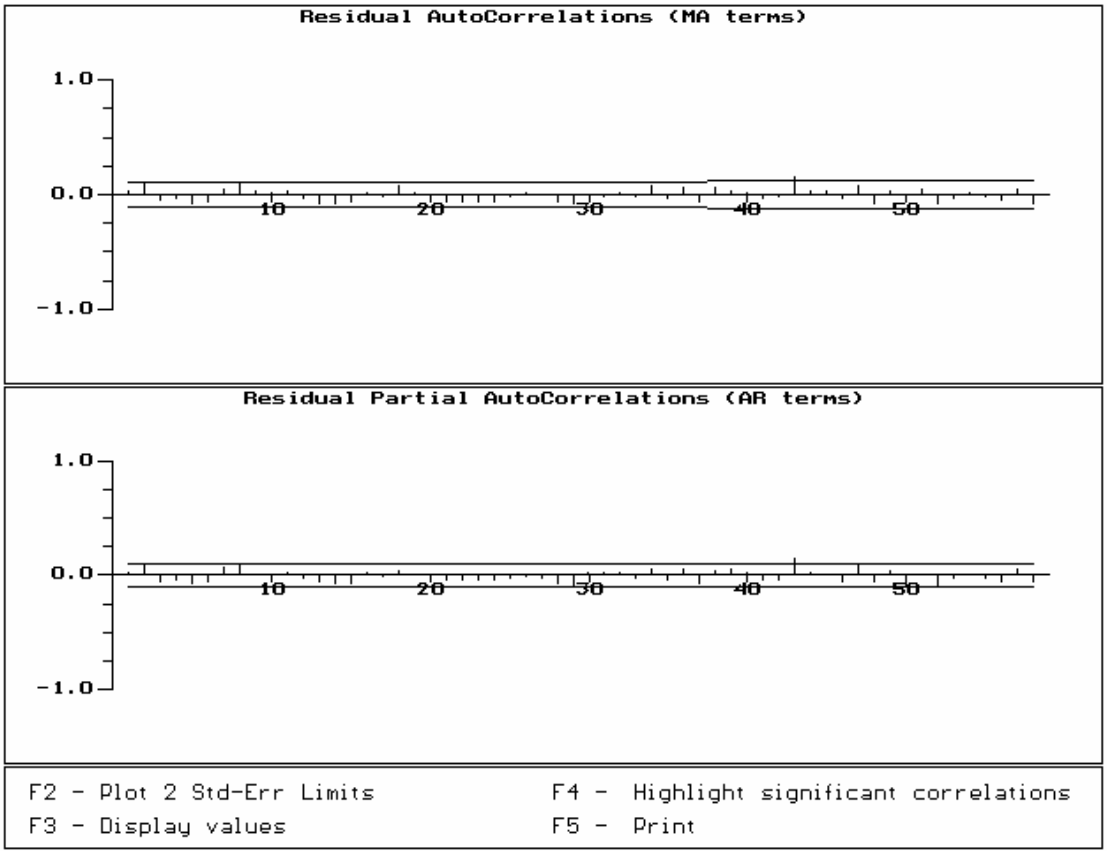
ARIMA(1,1,1) * (0,1,1)

meaning that we have included:

- one AR term (lag 1)
- one differencing (lag 1 used to stabilize the mean)
- one MA term (lag 1)
- no seasonal AR term
- one seasonal differencing (lag 52 used to stabilize the mean)
- one seasonal MA term (lag 52)

SSS1 syntax differs somehow of the usual mathematical syntax. The real syntax shows the actual order of the lags for p,d,q, while SSS1 shows the numbers of terms, and the lags are shown in the order column.

Figure 6: Typhoid cases in France by week, 1989-1996, residual ACF/PACF.



Once parameters have been estimated, we should check their adequacy in describing the series.

```
Press 'ESC' twice to go back to the 'Time series functions'
Goto the 'Residual analysis/diagnostics'
Pick 'Display model statistics'
```

The following file shows on screen.

```
                    BOX-JENKINS TIME SERIES ANALYSIS
                       MODEL IDENTIFICATION
File : ctyph.rec
CTYPH  Npts = 413
(p,d,q) * (P,D,Q), Constant term   Back-Casting
 1 1 1     0 1 1          N              Y
Data Seasonality  52      Maximum Order of Model  54
Log Transformation  with   10.000  displacement

                    PARAMETER ESTIMATION
Itr   Sum of Squares    Parameters
 0  5.550275E+0001   0.100000   0.100000   0.100000
 1  5.091191E+0001  -0.400000  -0.326598   0.100004
...
14  1.912976E+0001   0.047503   0.909712   0.869455
15  1.912976E+0001   0.047503   0.909712   0.869455
Convergence achieved to 3 digits
Relative change in each parameter is <=  1.000E-0003
```

This top section refers to the model and parameter calculations.

The next section relates to parameter diagnostics. Each term in the model is evaluated for its contribution to the fit. An appropriate parameter will show a T-value > 2, and confidence limits not including 0. In our example, the AR(1) term is not significant, and thus should be removed from the model. In fact, the ACF (top left) in figure 5 was showing a spike at lag 1 with borderline spikes at other lags, indicating a MA(1) unique term. Since other spikes were borderline, and for pedagogical reasons, we introduced a AR(1) term in the model.

```
                    PARAMETER DIAGNOSTICS
                                                                95% CONFIDENCE
LIMITS
TERM#   TYPE ORDER   ESTIMATE    STD.ERROR      T-VALUE        LOWER        UPPER
   1  REG AR    1   0.047503     0.055700     0.852837    -0.061691     0.156698
   2  REG MA    1   0.909712     0.023197    39.216563     0.864236     0.955187
   3  SEA MA   52   0.869455     0.041329    21.037571     0.788435     0.950476
```

Since we did not pass this diagnostic checking, we need to change the model.

Fourth step : adjusting the model

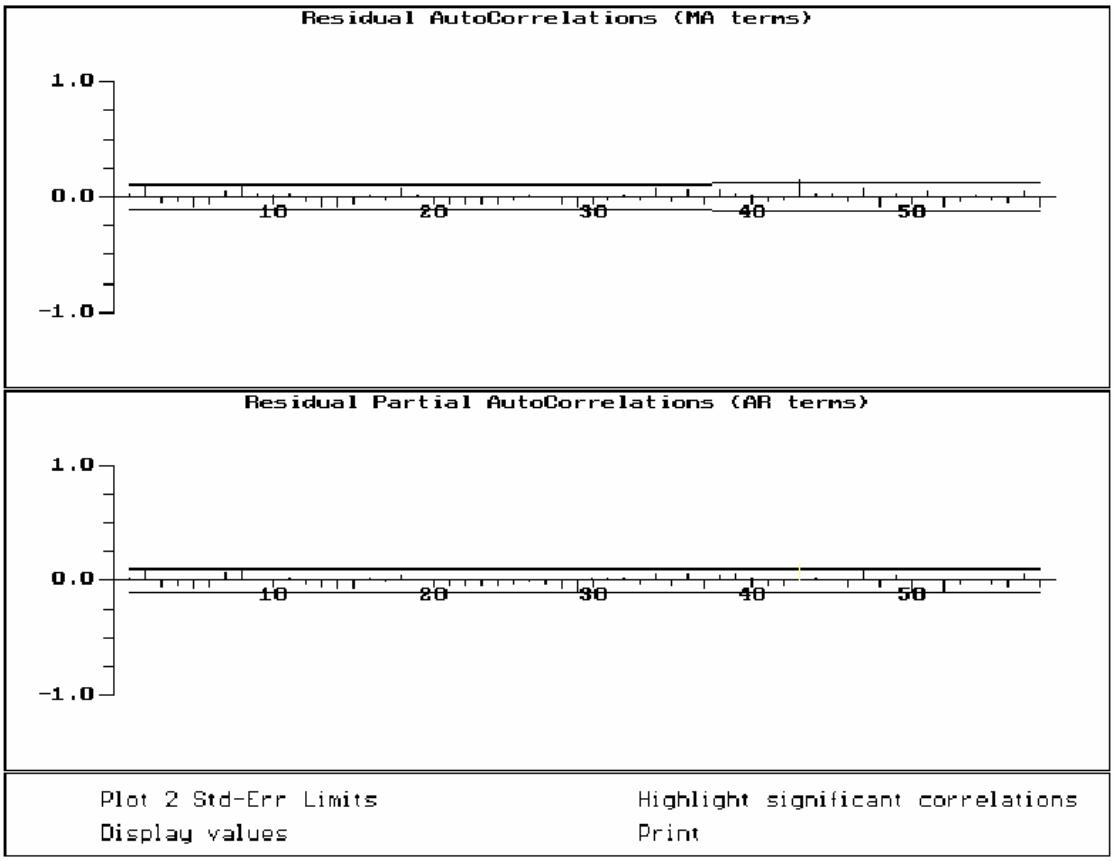
```
Press 'ESC'
Select 'Model identification'
Choose 'Build an ARIMA model'
Pick 'Select regular autoregressive term'
Press 'enter' while the cursor is on 'Regular autoregressive term of order
1' to remove the tick mark on the left of the option indicating it is
included in the model.
The mark should disappear
Press 3 times 'ESC'
```

Now that the model has been changed, parameters should be estimated again.

```
Goto 'Parameter estimation'
Choose 'Compute parameter estimates'
```

The residual variance is now = 0.047941, actually better than the previous residual variance with 3 terms in the model. The residual ACF/PACF shows only one significant spike at lag 43 on both ACF/PACF (Figure 7).

Figure 7: Typhoid cases in France by week, 1989-1996, residual ACF/PACF.



Fifth step : looking at diagnostics on the new model

We need to check the model again:

```
Press 'ESC' twice to go back to the 'Time series functions'
Goto the 'Residual analysis/diagnostics'
Pick 'Display model statistics'
```

The resulting file is displayed:

Model identification

```
                    BOX-JENKINS TIME SERIES ANALYSIS

                   MODEL IDENTIFICATION

File : ctyph.rec
CTYPH  Npts = 413
(p,d,q) * (P,D,Q), Constant term   Back-Casting
 0 1 1     0 1 1         N              Y
Data Seasonality  52      Maximum Order of Model  53
Log Transformation  with   10.000  displacement


                   PARAMETER ESTIMATION
Itr   Sum of Squares    Parameters
 0  5.002814E+0001   0.100000   0.100000
 1  3.465315E+0001   0.600000   0.100032
 2  3.146229E+0001   0.846127   0.100489
 3  2.348084E+0001   0.861368   0.511915
 4  1.927776E+0001   0.876794   0.841669
 5  1.917938E+0001   0.886686   0.878725
 6  1.916380E+0001   0.889804   0.868723
 7  1.916345E+0001   0.891220   0.869588
Convergence achieved to 3 digits
Relative change in each parameter is <=  1.000E-0003
```

The new model includes only 2 moving average terms (lag 1 and 52).

Parameter diagnostics

The 2 parameters are significant (T value = 39 and 21) and thus, should be kept in the model. The correlation matrix indicates whether the 2 parameters included in the model are correlated, meaning whether the effect of one is in fact due to the effect of the other. -2.62E-0002 means the 2 terms are independent, and thus should be kept both. If the correlation coefficient was greater than 0.9, then the 2 parameters would have been dependent of each other and one should have been a candidate for deletion.

```
                   PARAMETER DIAGNOSTICS
                                                          95% CONFIDENCE LIMITS
TERM#  TYPE  ORDER   ESTIMATE   STD.ERROR     T-VALUE     LOWER        UPPER
   1  REG MA    1   0.891220    0.022763   39.152248  0.846595    0.935844
   2  SEA MA   52   0.869588    0.041162   21.125883  0.788894    0.950282

Covariance Matrix
          1           2
  1    5.18E-0004
  2   -2.45E-0005  1.69E-0003
Correlation Matrix
          1           2
  1    1.00E+0000
  2   -2.62E-0002  1.00E+0000
```

Residual diagnostics

The residual mean is close to 0, and varies little around this value (std=0.22). This is satisfactory.

```
                    RESIDUAL DIAGNOSTICS
The Residual Mean                                  =        -0.007123
The Standard Error of the Res Mean                 =         0.011540
The T-value of the Residual Mean                   =        -0.617256
The Residual Variance                              =         0.047941
The Residual Std Deviation                         =         0.218953

The Number of Negative Residuals                   =              200
The Number of Positive Residuals                   =              160
The Number of Zero Crossings                       =              175
The Minimum Residual                               =        -0.626701
The Maximum Residual                               =         1.052013
The Range of the Residuals                         =         1.678714
```

Box-Pierce and Ljung-Box residual correlation are not significant, indicating a random distribution of residuals.

```
 The Box-Pierce and Ljung-Box statistics measure the residual
 correlation as a whole to test for randomness and independence.

 ChiSquare Test    Lags      Q-Statistic      X²(0.05)       DF
 Box-Pierce          6          7.59590        9.48773        4      0.107554
                    12         12.59573       18.30700       10      0.247161
                    18         20.57265       26.29620       16      0.195528
                    24         27.01980       33.92440       22      0.210475
 Ljung-Box           6          7.71566        9.48773        4      0.102567
                    12         12.86228       18.30700       10      0.231474
                    18         21.22920       26.29620       16      0.169868
                    24         28.13631       33.92440       22      0.171208
  A ChiSquare probability (last column on the right) of 0.05 or greater
  (on average) indicates that the model residuals are random (uncorrelated)
  and therefore we are 95% sure that the model is adequate, or conversely,
  the chance of rejecting an adequate model is about 5%. The larger the
  ChiSquare probability the more confidence we have in accepting the
  hypothesis of model adequacy.
         Closeness-of-fit statistics
The Mean Absolute Percent Error                    =         5.258177
     Coefficient of Determination--R Square
R Square of the original series is 0.998773
  The Coefficient of Determination measures the amount of variation in
  the original series that has been accounted for in the fit. A value
  of 0.9 or greater implies that the model has accounted for most of
  the variation in the original data.
```

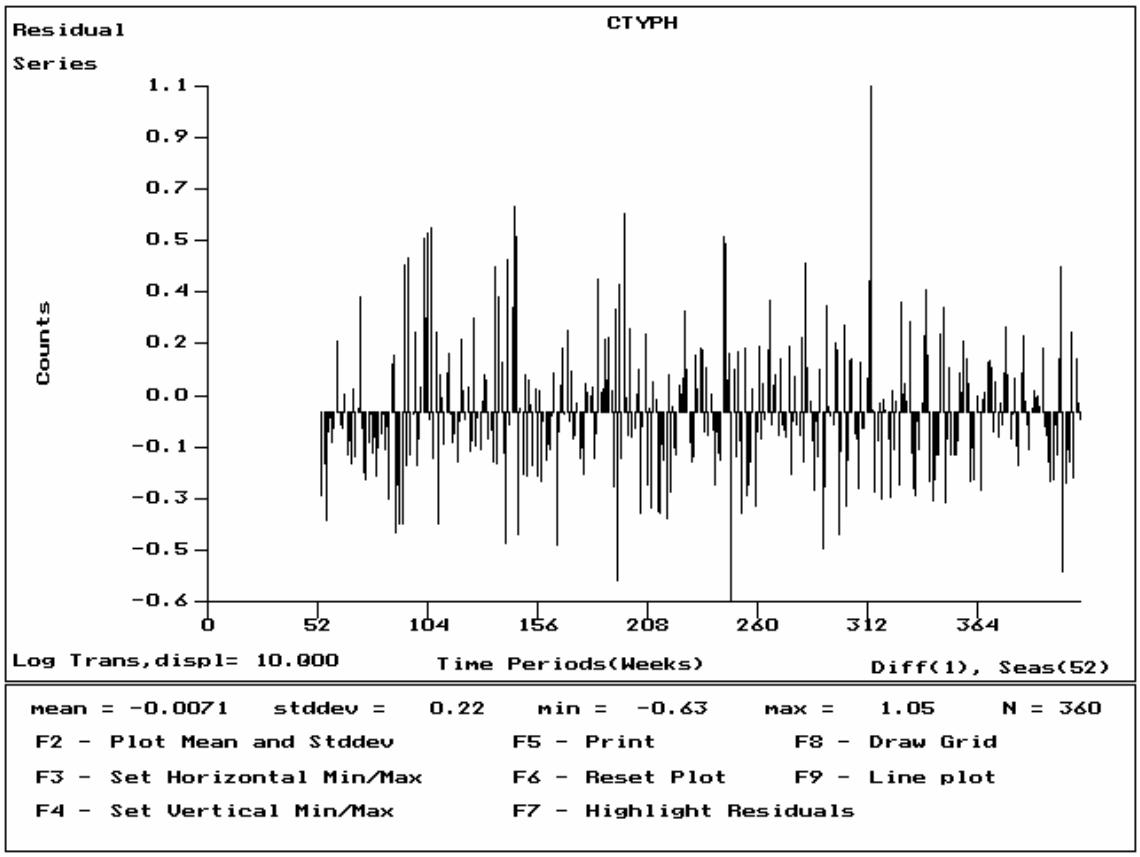
99.9% of the variation in the original series has been accounted for by the model. Here we are !

In SSS1, the file shows on screen. In fact it is created on disk, under the following name: DIAGCHEC.RPT. If you want to keep it, it needs to be saved. Quit SSS1 and from any text editor, open the file and rename it with a different name.

Looking at residuals

```
Goto 'Residual analysis'
Choose 'Plot residuals'
```

Figure 7: Typhoid cases in France by week, 1989-1996, residual plot.



Summary

In this part, we used Box & Jenkins theoretical correlations to pick the best suited model for our series. Then we estimated the value of the parameters, and run checks on the resulting model.

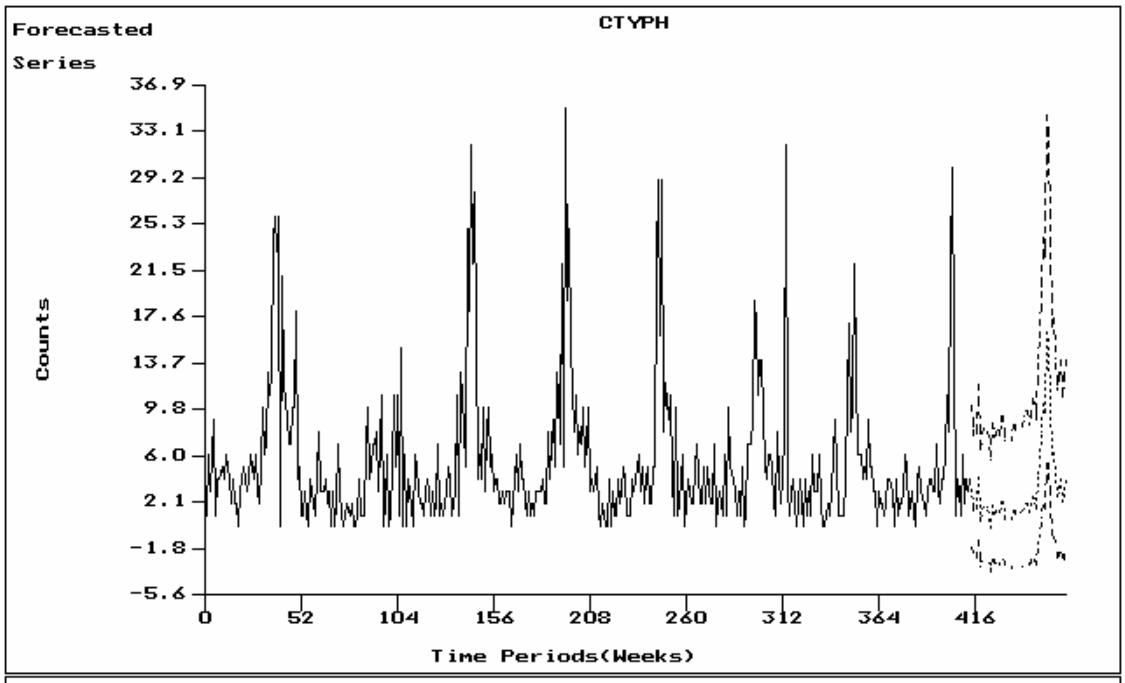
We did that process recursively until we got a suitable model showing randomly distributed residuals.

Part 4: Forecasting values and confidence intervals

Having achieved a satisfactory model, we can now forecast the time series.

```
Goto 'Forecast' in the Time series functions'
Choose 'Specify Forecasting origin and length'
Confirm '413' for Forecast origin
and '52' for length.
```

The original series, with one season forecast, shows on screen.



We could use F4 do set the minimum value to 0 since negative estimates do not make much sense. As a rule, Box & Jenkins models should not be used to forecast too far ahead. One cycle is the maximum. Theoretically only the next point could be forecast for!

You can print the display from here, but you can also save the forecast data which is stored in the file FRCST.RPT. Exit SSS1 and open the file in an editor and rename it. By carefully removing the header of the file, you can export the data to an Excel spreadsheet in order to get better control of the graphical display.

Sample forecast file after removing the header:

PERIOD	FORECAST	LOWER	UPPER
414	3.422	-1.262	10.617
415	1.436	-2.574	7.611
416	2.030	-2.208	8.573
417	1.528	-2.552	7.842

Part 5: Keys for a good model

Do you need to stationnarise the series?

Apply a transform (variance stabilization)?

The variance needs to be stabilized first. Look at the variations in the series. The variance stabilization is indicated when the amplitude of variation varies according to period or level of mean. The log transform should be tested first even though SSS1 can recommend another transform.
If doubtful, compare the residual variance after applying a model with a constant term, and choose the transform yielding the smallest residual variance.

Differentiate (remove a trend)?

Some series do not present an obvious linear trend (indicating a differencing of order 1) or a seasonal component (indicating a differencing of seasonal order). Several tools can help:

- Look at the ACF of the original series:
  the lack of a rapid decay of peaks is in favor of a linear trend
  Peaks reappearing at seasonal lag indicates a seasonal component.
- Apply a constant term model on the original series and the differentiated series. If the differencing yields a smaller residual variance, it is indicated.

Which term to include?

Terms are selected according to ACF and PACF functions and by review of the parameters, keeping in mind their "epidemiological" signification. Some terms can be included "a priori", even if they do not meet the criteria for a good model. Fox example, a 52 week seasonal term can be kept, even if not significant, if the disease is known to have a seasonal behavior (influenza). This term is called a forced term.

Characteristics of a good model are summarized in the following table:

Convergence of the parameter estimation	Achieved
Significance of parameters	Value of t test > 2
Parameter confidence interval	Does not include 0
Parameter correlation	Parameters non correlated
Residual mean	Close to 0
Correlation of residuals	Non significant (white noise)
Residual series	Same number of positive and negative values
Q statistic	Probability (Chi^2) > 0,05
Adjustment of the model	R-square + Mean Absolute Percent Error (MAPE)

Characteristics of a good model

Is the model predictive?

The last step consists to check if the model is predictive. A model can be adequate (well fitted to original data), but not predictive. The random (white noise) component of the prediction is too important. The FORECAST option of SSS1 can be run against the last season on the original series. The correlation between the forecast and the last season can be then visually reviewed.

Summary

Box & Jenkins modeling is not a simple process. Tools and good introductory books exist, and should be used by epidemiologist involved in surveillance activities to acquire better knowledge of the diseases. It requires some experience during model identification. However, statistical tools such as ACF and PACF are well implemented in SSS1, making this process easier than with other packages.