

Introduction to Time Series Analysis

Madrid, Spain

10-14 September 2007

Case study: Setting an epidemic threshold for a time series using spectral analysis

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

- **Manipulate Fourier Transform to carry out a spectrum analysis**
- **Decide which Fourier coefficients need to be included in an epidemic threshold**
- **Set up an epidemic threshold**

Denis Coulombier, Fernando Simón, Bruno Coignard

Presentation

This case study includes 4 parts. It illustrates principles of spectral analysis for modelling of food-borne diseases time series in France. Each part requires the participant to perform actions in the spreadsheet.

- **Part 1: performing log transform and removing the trend**
- **Part 2: computation of all Fourier coefficients using Excel**
- **Part 3: application of Fourier transform in surveillance**
- **Part 4: excluding past epidemics from the alert threshold**

Programs needed on the computer:

- Microsoft Excel

Example files used by the case study:

- FOURIER.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

Reference to cells in Excel uses the following syntax:

A1 refers to the first cell of the first row.

When a formula is copied over a range, it is important to keep in mind that Excel considers cells entered in this format as relative references. If the formula:

=A1 + B1

is copied over a range, the output cells will have their references incremented:

=A2 + B2

=A3 + B3

and so on.

Using \$ sign in front of one of the coordinates makes it an absolute reference. If the formula:

=\$A\$1 + \$B1

is copied over, it gives:

=\$A\$1 + \$B2

=\$A\$1 + \$B3

A cell can be named rather than expressed as coordinates. To name a cell, place the cursor in the cell, click on the left cell of the line immediately above the header line and enter the desired name. Subsequently, you can use this name to refer to this cell. This allows for more meaningful names of cells in formulas. In the same way, a range can be named. Select the range to name (it appears against a black background) and indicate the appropriate name in the left cell above the top row. In this case-study, most cells have been renamed.

All formulas and labels have been already entered in the spreadsheet. Results indicated when loading the spreadsheet may be erroneous at this stage because some of the referenced cells are empty. Follow instructions to fill these cells in order to get proper values.

The protection of the spreadsheet is activated in order to avoid erasing a cell. However the protection of cells in which you need to enter, change the value, or copy cells has been disabled.

This case-study uses Excel version 5 or later. Excel uses different names for formulas in the various European languages. This case-study uses English denomination. If you are using non-English version of Excel, you should use the following table to adapt formulas, after having removed the workbook protection.

Lexicon of formulas in European languages

Formula	French	English	German	Italian
P	PI()	PI()	PI()	PI.GRECO()
Sum of a range	SOMME()	SUM()	SUMME()	SOMMA()
Complex modulus	COMPLEXE.MODULUS()	IMABS()	IMABS()	COMP.MODULO()
Variance	VAR()	VAR()	VARIANZ()	VAR()
Covariance	COVARIANCE()	COVAR()	KOVAR()	COVARIANZA()
Standard deviation	STD()	STDEV()	MITELABW()	DEV.ST()
Mean	MOYENNE()	AVERAGE()	MITTELWERT()	MEDIA()

This case-study uses some Excel add-ins. You need to check that these add-ins have been activated before proceeding. In the 'Tools' menu of the main menu bar, you should see options for 'Solver' and 'Analysis tool pack'. If not,

Click on 'add-ins', and activate the 'Solver' and the 'Analysis tool pack'

If you do not see the analysis tool pack option, it means that you did not carry out a complete installation of Excel. When a complete installation is carried out, you should see a 'Analysis' directory under 'EXCEL\LIBRARY' on your hard disk. If not, reinstall Excel.

Intro: setting-up the spreadsheet

Adjust screen display button

The button "Adjust screen display" will optimise the layout of the various spreadsheets for the resolution of your screen.

Load the file **FOURIER.XLS** in Excel.
Click on the "Adjust screen display" button

Reset spreadsheet

The "reset spreadsheet button" copies default values in all the necessary fields. It erases all activities previously carried out by the participant. This button is only activated when the protection has been removed. It is not necessary to execute this procedure upon loading the spreadsheet for the first time.

Remove protection/Protect document

This button removes the protection of the workbook, and allows you to change formulas and values in all the fields. This should be done with caution since the spreadsheet may not work properly if content of cells are altered. It is advised to make a copy of the spreadsheet if you intend to modify its content.

Part 1: stabilizing the variance and assessing the trend

Load the file FOURIER.XLS in Excel.
Activate "DATA" spreadsheet

In this part, we will stabilize the variance and assess the trend. This should be done before using the Fourier transform to build the periodogram, which will be used to study cyclical contributions for each available periods.

Stabilizing the variance

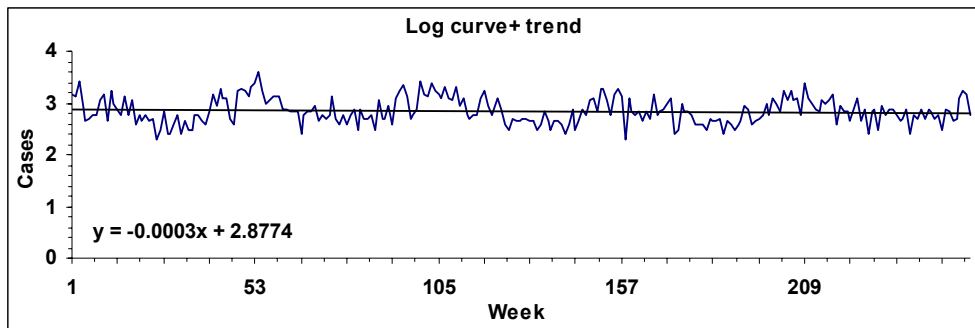
The crude data, as seen on the top graph, visually shows that the variance seems larger during the epidemic peaks than between the peaks. This is an indication to use a log transform to stabilize the variance.

The Neper logarithm of 0 is not defined. Since some values are 0, we will need to displace the entire series by adding a constant term to the series. A value of 10 is usually used.

The value 10 appears in cell I4 for constant term. The cell has been named DECAL
The formula =LN(B3+DECAL) appears in cell C3

Copy this formula over the range C3:C258 by clicking on the "Copy formulas" button (you can copy it manually as well).

The graph should look as follows:



Assessing the trend

The Excel formula to return the characteristics of a regression line has been entered in cells G3 to H6 :

Linear regression				
Slope/Intercept	-0.00025029	2.877382667	Pi	3.141592654
se(m)	0.0002112	0.031307413	Constant	10
R ² /standard error	0.00549889	0.249725891		
F/DDl/p	1.40443991	254	0.23708919	Non significant

It shows that the slope of the regression line is -0.00025 , and the intercept 2.8774 . The F test to test the significance of the slope is not significant. Thus, at this junction, we decide not to remove the trend by regression.


We can visually assess that the variance looks more constant over the peak periods than previously. This series becomes the working series on which we will apply the Fourier transform in next section.

Part 2: Computation of Fourier coefficients using Excel Fast Fourier Transform (FFT) add-in macro

Computation

Load the file **FOURIER.XLS** in Excel.
Activate **PART 4** spreadsheet

The « analysis » add-in macro should be loaded in order to use the FFT (add-ins in the tool directory of Excel 5). When the macro is loaded, an additional option appears on the tool menu, for « other analysis ». When activated, a list of add-ins appears.

Click on the button 

If you want to practice Excel, you can carry-out this operation manually as follows:
Click on the **Fast Fourier Transform**

The Fourier Transform dialog takes only 2 parameters: the input and output range:



Put the cursor in the Input range box, erase its content if any, then click on the spreadsheet with the mouse and select the range B3 B258 with the mouse.

Click on 'Output range', erase its content if any, and then click on cell C3 on the spreadsheet. The dialog box should look as above (but not necessarily in French...)

Click on the 'Ok' button

Sample output

	A	B	C
1			
2	Time	Dat	Fourier coefficients
		a	
3	1	3.18	728.376316055872
4	2	3.14	-0.423732853829142-5.3737340648125i
5	3	3.40	-2.32038287062972+6.27750543598104i

6	4	2.94	-1.60620944527231-2.66814945285176E-003i
---	---	------	--

The output lists the coefficients expressed as complex numbers, starting with the cell indicated for output range. The first coefficient (cell C3) does not have an imaginary component. It is calculated for 0 oscillations and corresponds to the sum of observations over the entire range of values. The next coefficient (cell C4) corresponds to 1 oscillation over the 256 data points (period of 256 weeks); next one (cell C5) corresponds to 2 oscillations (period of 128 weeks). There are 256 coefficients, but the last 128 coefficients are mirrored images of the first 128. We will just look at the first 128.

Building the periodogram

The information we want to get is for which period in weeks (or which frequency) do we get the strongest oscillations in the signal. The periodogram is a graph of the period by the energy of oscillations. In order to get the period and the energy, a couple of additional values and formulas have been entered.

The range F4 to F131 contains numbers from 1 to 128. They correspond to the 128 cosine curves that can fit 256 data points. 1 means one oscillation over the 256 data points. Cell D4 to D131 contains the value for the corresponding period, $=256/F4$ to $=256/F131$.

The energy is the square root of the sum of the square of the imaginary and real component of the complex number. Excel provides a formula (under 'scientific' in the list of available formulas) which extracts it directly: $\text{IMABS}(\text{Complex Number Cell})$. Cell F4 to F131 contains the formula $=\text{LN}(\text{IMABS}(C4))$ to $=\text{LN}(\text{IMABS}(C131))$ which correspond to the modulus of the first sine curve.

Formulas in the example:

	A	B	C	D	E	F
1						
2	Time	Data	Fourier coefficients	Period	Energy	F
3	1	3.18	728.376316055872			
4	2	3.14	-0.423732853829142-5.3737340648125i	$=256/F4$	$=\text{LN}(\text{IMABS}(C4))$	1
5	3	3.40	-2.32038287062972+6.27750543598104i	$=256/F5$	$=\text{LN}(\text{IMABS}(C5))$	2
6	4	2.94	-1.60620944527231-2.66814945285176E-003i	$=256/F6$	$=\text{LN}(\text{IMABS}(C6))$	3

Corresponding values in the spreadsheet:

	A	B	C	D	E	F
1						
2	Time	Data	Fourier coefficients	Period	Energy	F
3	1	3.18	728.376316055872			
4	2	3.14	-0.423732853829142-5.3737340648125i	256.00	1.68	1
5	3	3.40	-2.32038287062972+6.27750543598104i	128.00	1.90	2
6	4	2.94	-1.60620944527231-2.66814945285176E-003i	85.33	0.47	3
7	5	2.64	-2.77868210393391+3.79743136449813i	64.00	1.55	4
8	6	2.71	26.6111729576257-7.77290536824154i	51.20	3.32	5
9	7	2.77	0.99636357282925-3.17974720068084i	42.67	1.20	6
10	8	2.77	-2.5275524066516-3.75324878917275i	36.57	1.51	7

Cells H28 to I30 contain the mean and standard deviation of the "Energy" coefficients. The "Cut off" cell represents the value of the mean + ALPHA standard deviations, which is the cut-off value we will use to assess the significance of the corresponding cyclical contributions. If a coefficient is greater than the cut-off, it is displayed in bold and red, using the Excel conditional formatting that can be set through the "Format" menu.

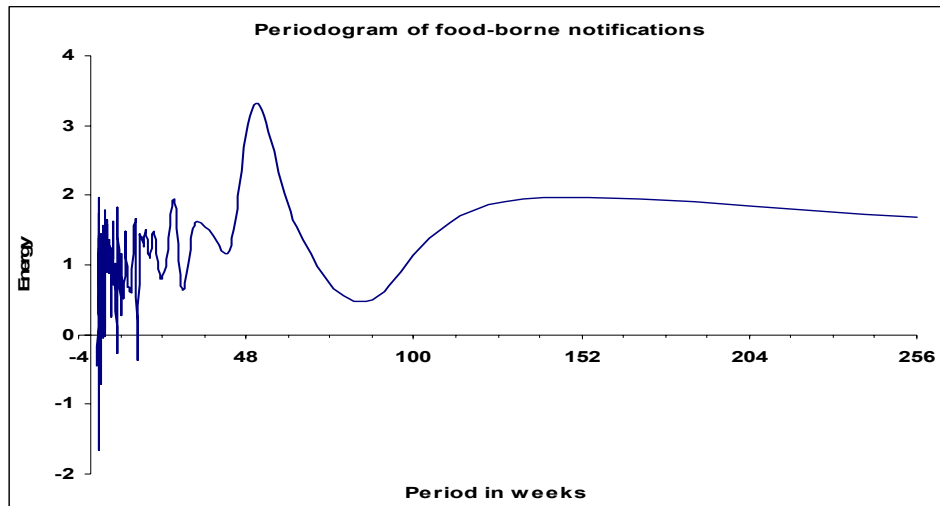
Mean	0.928637
STD	0.60223399

CutOff	1.91922339
P(%)	90
Alpha	1.64

Looking at the E column shows that the energy is maximal for a period of 51.2 weeks. This shows a strong yearly seasonal effect in our data set. The energy is significant for 25.6 weeks as well, which can be considered as an harmonic ($25.6=51.2/2$).

Interpretation of the periodogram

Figure 3.1. Periodogram of food-borne outbreaks



The visual analysis of the periodogram shows a strong contribution for a period of 52 weeks, and a smaller significant peak for 26 weeks, corresponding respectively to half a year.

Summary

In this section, we have used Excel to perform a spectral analysis of our signal. We have built the periodogram, and showed a strong contribution of 52-week oscillation in our data set and a weaker one for 26 weeks. In fact, it was rather obvious from looking at the original time series. However, this is not always the case, especially when there are 2 or 3 years cycle contribution to the series. We will consider 52 weeks and 26 weeks oscillation for modelling the series in next section.

Part 3: applications of Fourier transform in surveillance

In this part, we will apply findings of the periodogram to set up a threshold for food-borne disease time series in France. This requires the following steps:

- Taking into account the trend,
- Taking into account cyclical contributions, and
- Defining the threshold, based on confidence intervals of the residuals between the signal and the model.

Activate "Threshold" spreadsheet

First step: definition of the parameters

The model for the data uses a linear trend component plus 1 to 4 seasonal components. The number of seasonal components to include in the model can be set by clicking on the buttons:

The meaning of other parameters is as follows:

- Cell M4 is the intercept of the linear trend component. Its value will be set by the solver when clicking on "Optimise model"
- Cell M4 is the slope of the linear trend component. Its value will be set by the solver when clicking on "Optimise model"
- Cell M5 is coefficient of determination. Its value gives the proportion of the overall variance accounted for by the model. The formula used by Excel is : =COVAR(MODEL;DATA)/(VAR(MODEL)*VAR(DATA))^0.5.
- Cell M6 corresponds to the sum of the square of the difference between the data and the model. It is used by the solver as the target value to minimize, in order to get the best fit for the model. It contains the formula =SUM(PEAK), PEAK being defined as the range D3 to D258.
- Cell M7 contains the standard deviation of the difference between the data and the model. It corresponds to the square root of the variance, which was calculated in cell M6. This value will be used to define the level of the threshold for epidemics. It will be calculated when optimising the model.
- Cell M8 contains the number of standard deviation that will be used to set the thresholds. The formula used is =NORMSINV(1-(1-M9/100)/2).
- Cell M9 contains the P-level for the threshold. You can use the spin buttons to adjust the confidence interval used for the threshold calculations. As you change the value, the number of standard deviations in cell M8 will adjust accordingly.
- Cell M11 contains the cut-off value that can be used to exclude epidemic periods in the data when optimising the model. It is not indicated to include past epidemics when optimising the model. Any data point above this cut-off value will be excluded of the calculation, and appear in red in column D.
- Cell M12 contains the number of degree of freedom used for the calculations of various parameters. It relates to the number of terms selected and equals 2 * the number of cyclical components (phase + amplitude) + 1 for the slope of the linear trend. It will be adjusted when optimising the model.
- Cells M13, M16, M19, M22 contain the amplitude for each of the four available cyclical components of the model. A value of 0 means that the cyclical component will be ignored.
- Cells M14, M17, M20, M23 contain the phase or lag of the cyclical component. It will be adjusted when optimising the model
- Cells M15, M18, M21, M24 contain the period. These values are artificially fixed to one season (52 weeks) for the first one, 26 for the second one, 17 for the third one (1/3 of a year, or three cycle per year), and 13 (4 cycles per year).

Second step: setting the parameters

1. Number of cyclical terms

In the list of buttons:

Number of sine curves in the model

- 1 cyclic term
 2 cyclic terms
 3 cyclic terms
 4 cyclic terms

Click on 2 cyclic terms since 2 coefficients were significant on the periodogramme.

2. Level of confidence interval

In order to have a 95% confidence interval for the threshold,

Set cell M9 to 95%

3. Mode of exclusion of epidemic periods

Three modes are proposed for excluding epidemic values when optimising the model:

- **On p value.** Any value exceeding the mean + alpha standard deviation will be excluded
- **By exclusion.** This option allows the manual removal of any value in the series when optimising the model. If the value in column K is set to 1, the value in column D is included in the optimisation. If the value in column F is set to 0, then the value in column D is excluded and appears in red in the column. This allows for a much more precise exclusion of historical data, but requires manual inspection of all values, while excluding values on P-value allows the automation of the process.
- **All data points.** This will reset the cells in column F to 1 and include all data points in the optimisation process.

Click on "On P-value"

4. Optimising the model

Click on "Optimise the model"

Resulting output in the spreadsheet:

Time	Working series				Original series				
	Data	Model	Diff^2	95% CI	95% CI	Data	Model	95%CI	95% CI
1	3.18	3.07	0.012	3.43	2.71	14.0	11.52	20.9	5.0
2	3.14	3.06	0.006	3.42	2.69	13.0	11.24	20.5	4.8
3	3.40	3.04	0	3.40	2.68	20.0	10.89	20.0	4.6
4	2.94	3.02	0.006	3.38	2.66	9.0	10.51	19.4	4.3

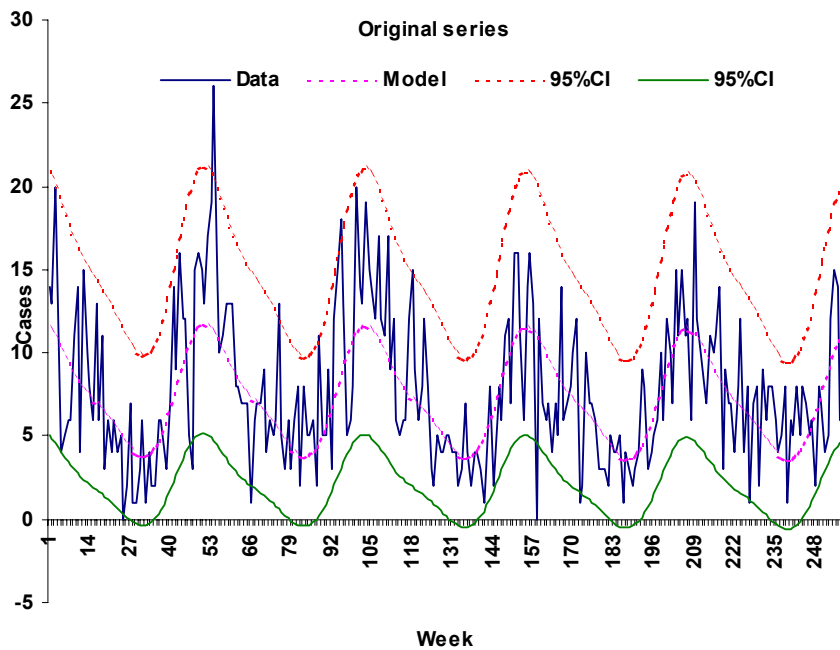
- Column 1 represents time intervals
- Column 2 is the logarithm of the original data
- Column 3 contains the formula for the model, which includes 2 seasonal components plus a linear trend in our example: $=A0+B0*(TIME)+ AMP10*COS(2*PI*((TIME-LAG10)/WEEK10))+ AMP20*COS(2*PI*((TIME-LAG20)/WEEK20))$
- Column 4 contains the sum of the square of the difference between column 2 and 3. However, if the data exceeds the cut-off value in cell M11, a value of 0 appears, meaning that this data point has been ignored.
- Column 5 contains the upper confidence interval of the model, based on the standard deviation and P-value indicated in the list of parameters.
- Column 6 contains the lower confidence interval of the model, based on the standard deviation and P-value indicated in the list of parameters.

- Column 7 contains the original data. In fact, it is the exponent of the value in column 2 minus the constant that had been added to account for nil values.
- Column 8 contains the exponent of the model, plus the constant
- Column 9 contains the exponent of the upper confidence interval
- Column 10 contains the exponent of the lower confidence interval

The graph below shows the working series and its 95% confidence interval.



The graph below shows the original series and its 95% confidence interval:



Summary

We have built a threshold for early detection of epidemic periods in our time series. We should keep in mind that this technique applies for stable series with strong cyclical components. Spectrum analysis gives the same importance to all data points in the data set. It takes the series as a whole. However, when one want to take more into account recent variations than past variation Box & Jenkins modelling is indicated.

Bibliography

- Eickhoff T., Sherman I., Serfling R. Observation on excess mortality associated with epidemic influenza, JAMA, June 3, 1961; 776-779
- Serfling R. Methods for current statistical analysis of excess pneumonia-influenza deaths. Public Health Reports, vol. 78, N°6, June 1963 ; 494-505
- Housworth J. Langmuir A. Excess mortality from epidemic influenza, 1957-1966. American Journal of Epidemiology, Vol. 100, N°1, 1974; 40-48
- Choi K, Thacker S. An evaluation of influenza mortality surveillance, 1962-1979. American journal of epidemiology. Vol 113, N°3, 1981; 215-226
- Chatfield C. Time series analysis, in Problem solving, a statistician guide . London, Chapman & Hall, 1993155-159; 222-229
- Chatfield C. The analysis of time series: an introduction, 2nd edition, Chapman and Hall, London, 1980.
- Bliss C. Periodic regression in biology and climatology. Bulletin of the Connecticut agricultural experiment station, New Haven, June 1958.
- Choi K. Improved accuracy and specificity of forecasting deaths attributed to pneumonia and influenza. The journal of infectious diseases. Vol 144, N°6, December 1981.