Time Series Analysis and Forecasting Labs

Hugues Garnier

November 2021

Contents

Introduction

At this point of your study at Polytech Nancy, you should have some experience with MATLAB, for MATrix LABoratory, which is a powerful software that can deal with complex problems in various scientific fields, exploit and analyse data in an interactive environment.

It includes the *Econometrics Toolbox* which is particularly useful to data scientists, econometricians and data engineers to perform various time series analysis, modeling and forecasting.

It offers a wide range of visualizations and diagnostics for model selection, including tests for autocorrelation and heteroscedasticity, unit roots and stationarity, causality, and structural change.
The *Econometrics Toolbox* toolbox includes functions and a graphical user interface (or Apps) that we will discover and use during the lab sessions for generating, viewing, analysis and forecasting time series. With the help of this toolbox, you will be able to estimate, simulate, and forecast economic systems using a variety of modeling frameworks. These frameworks include:

- regression models;
- ARIMA models;
- state-space models;
- GARCH models;
- multivariate VAR and VEC models;
- switching models.

The toolbox also provides Bayesian tools for developing time-varying models that learn from new data.
We will use the *Econometrics Toolbox* during the lab sessions and so make sure it is installed on Matlab 2020b or install it before the beginning of the labs.

Lab 1

Time series analysis

Downloading of the data needed for the tutorial

1. Download the zipped file **Lab1_TSAF.zip** from the course website and save it in your Matlab working directory.
2. Start Matlab.
3. By clicking on the browser folder icon, **change the current folder of Matlab so that it becomes your Lab1_TSAF folder** that contains the files needed for this lab.
4. It is highly recommended to use the Matlab Live Editor. In the Live Editor, you can create live scripts that show output together with the code that produced it.

Layout of the tutorial

1. A tutorial introduction to show how to perform exploratory time series analysis and decomposition in the case of the monthly average temperature in Atlanta.
2. Several assignments where you have to perform real-life time series data analysis with the methods and theory that you have learned in Lecture 1.

1.1 Tutorial introduction of time series analysis and decomposition - Monthly average temperature in Atlanta

1. Open and run the tutorial file `Decompose_AtlantaTemperature_data_tutorial.mlx` that illustrates how to perform time series analysis and decomposition on the monthly average temperature in Atlanta.
2. Understand every step of the general procedure and compare the results with those given in the lecture slides.

1.2 Assignments - Time series analysis of data coming from various fields

1.2.1 Forecasting of the USA population in 2040

The file `pop_usa.mat` includes 10-yearly data of the total population of the United States for the years 1920 to 2020. The units are millions of people.

The goal is to model the population growth and predict the population in 2040.

1. Load the dataset in Matlab
```
load pop_usa;
```

2. Plot the time series and observe the population increase over time.
```
plot(t,pop,'r','linewidth',2)
xlabel('Years')
ylabel('Millions')
title('Total population in the USA')
grid
```

3. Partition the time series into training and validation datasets. Select, for example, the first 80% of the time series as the training dataset and the remaining 20% part for the validation dataset.

4. We choose to postulate a polynomial model to capture the population growth as

$$p(t_k) = \beta_0 + \beta_1 t_k + \ldots + \beta_p t_k^p \tag{1.1}$$

5. For $p = 1$ to $p = 2$, determine by simple least squares the parameters of the polynomial models on the training dataset.

6. For each trained model, compute the residuals along with the RMSE and MAPE performance indices on both training and validation datasets.

7. Select the best model and use it to forecast the total population of the United States in 2040.

1.2.2 Airline Passengers

Monthly total international airline passengers (in thousands of passengers), January 1949–December 1960, is available in Matlab.

1. Load the data available in Matlab:
```
load Data_Airline;
```

2. Plot the data and describe the behavior of the series.
```
plot(dates,Data)
shg
datetick
```

3. Does the time series appear to be stationary?

4. Perform a log transformation of the data and plot the resulting series. Compare the behavior of the original and log-transformed series.

5. Do you see an advantage in using a log transformation for modeling purposes?

1.2.3 Sunspot

This assignment aims at analyzing sunspot data which is available from NASA for the years 1749-2012.

Sunspots are temporary phenomena on the Sun's photosphere that appear as spots darker than the surrounding areas on the side of the Sun visible to an Earth observer, as shown in Figure 1.1.

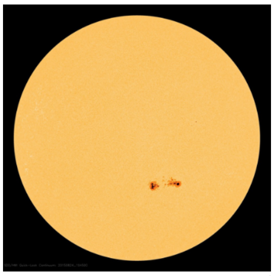

Figure 1.1: Sunspots observed on the Sun - August 24, 2015. NASA.

The Sun is the source of the solar wind: a flow of gases from the Sun that streams past the Earth at speeds of more than 500 km/s. Disturbances in the solar wind shake the Earth's magnetic field and pump energy into the radiation belts. Regions on the surface of the Sun often flare and give off ultraviolet light and x-rays that heat up the Earth's upper atmosphere.

This **"Space Weather"** can change the orbits of satellites and shorten mission lifetimes. The excess radiation can physically damage satellites and pose a threat to astronauts. Shaking the Earth's magnetic field can also cause voltage surges in power lines that destroy equipment and knock out power over large areas on the planet. As we become more dependent upon satellites in space and electricity power on Earth, we will increasingly feel the effects of space weather and therefore we need to better understand and predict it.
To learn more about space weather, browse:
`https://www.futura-sciences.com/sciences/actualites/...`
`soleil-nouveau-cycle-solaire-pourrait-etre-plus-intense-jamais-observe-84601`
`https://en.wikipedia.org/wiki/Space_weather`

The nature and causes of the sunspot cycle constitute one of the great mysteries of solar astronomy. While we now know many details about the sunspot cycle, we are still unable to produce a model that will allow us to reliably predict future sunspot numbers using basic physical principles. This problem is a little like trying to predict the severity of next year's winter or summer weather.
For almost 300 years, astronomers have tabulated the number and size of sunspots every year. Load the sunspot data available in the folder that you have downloaded:

```
load sunspot;
```

Convert the signal to a timetable.

```
sunspot = timetable(years(year),y);
```

1. Examine the properties of your series using plots or other appropriate tools.
2. Determine the main cycle/period in years of the sunspot activity.
3. Predict in which year we should reach a new maximum sunspot number.

1.2.4 Dataset of your choice

Download a time series of your choice from the Internet. Note that financial and economic time series are available from sources such as Google Finance and the Federal Reserve Economic Data (FRED) of Federal Reserve Bank in St. Louis, Missouri, while climate data is available from NOAAs National Climatic Data Center (NCDC).

1. Store the data in a text file or a .csv file and read the data in Matlab.
2. Examine the properties of your series using plots or other appropriate tools.
3. Does your time series appear to be stationary? If not, apply any appropriate transformation and or decomposition to make the time series stationary?

Lab 2

Time series long term trend forecasting by smoothing and polynomial modelling

Downloading of the data needed for the tutorial/lab

1. Download the zipped file **Lab2_TSAF.zip** from the course website and save it in your Matlab working directory.
2. Start Matlab.
3. By clicking on the browser folder icon , **change the current folder of Matlab so that it becomes your Lab2_TSAF folder** that contains the files needed for this lab.
4. It is highly recommended to use the Matlab Live Editor. In the Live Editor, you can create live scripts that show output together with the code that produced it.

Layout of the tutorial

1. Assignments where you have to perform time series smoothing and forecasting with the methods and theory that you have learned in Lecture 2.
2. Assignments where you have to perform time series data polynomial trend estimation by least squares and forecasting with the methods and theory that you have learned in Lecture 2.
3. Several assignments where you have to perform further time series data decomposition and forecasting with the methods and theory that you have learned in Lecture 1.

2.1 Assignments - Smoothing and forecasting of global temperature anomalies

1. Load the `Temperature_anomalies.mat`. It includes the temperature anomalies used to illustrate the moving average smoothing concept in lecture 2.
2. Plot and observe temperature anomaly increase over time.
3. Apply a trailing moving average of window 2 and 10 to the time series data and observe the latest trend of the temperature anomalies. See the `movmean` function in Matlab.
4. Select the best model and use it to forecast the temperature anomalies in the upcoming decade.

2.2 Assignments - Polynomial trend estimation and forecasting of global temperature anomalies

1. Load the `Temperature_anomalies.mat`. It includes the temperature anomalies used to illustrate the trend estimation by using least squares concept in lecture 2.

2. Plot and observe temperature anomaly increase over time.

3. We choose to postulate a polynomial model to capture the temperature anomalies increase as

$$p(t_k) = \beta_0 + \beta_1 t_k + \ldots + \beta_p t_k^p \tag{2.1}$$

4. For $p = 1$ to $p = 2$, determine by simple least squares the parameters of the polynomial models.

5. For each trained model, compute the residuals along with the RMSE and MAPE performance indices.

6. Select the best model and use it to forecast the temperature anomalies in the upcoming decade.

2.3 Assignments - Time series analysis and decomposition of data coming from various fields

2.3.1 Atmospheric CO2 Levels at Mauna Loa

The data for this analysis were collected from the Mauna Loa CO_2 monitoring site in Hawaii. Monthly observations started in March 1958 and ended with the most recent recorded month, November 2020. The variables include the monthly CO_2 concentrations in part per milliions (ppm); the month and year of every measurement were also recorded.

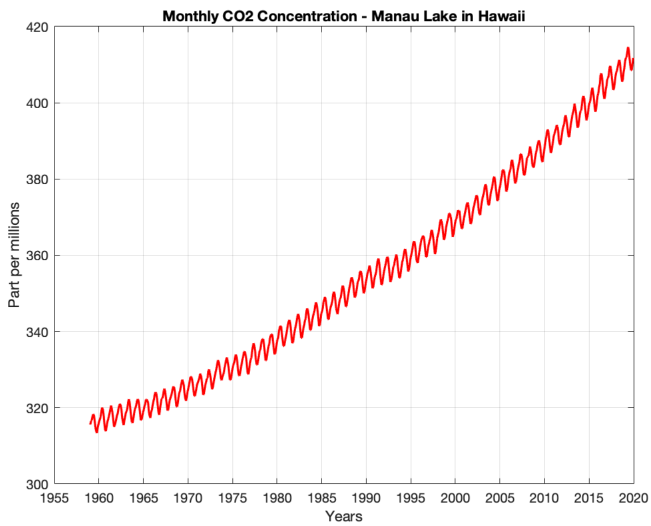

Figure 2.1: Carbon dioxide levels at Mauna Loa, Hawaii

As seen in Figure 4.2, the carbon dioxide levels at Mauna Loa, Hawaii increased above the 400 parts per million threshold. This threshold was first crossed as a global average in 15 March

2015. Why is this figure important? Scientists have warned that crossing this threshold could result in more global warming and the disasters associated with it, like sea-level rise and ocean acidification.
The notion of CO_2 concentration is to be distinguished from CO_2 emissions. Emissions represent what enters the atmosphere as a result of human activities, including the combustion of fossil resources and cement production. The concentration indicates what remains in the atmosphere at the end of the interactions between the air, the biosphere and the oceans. About a quarter of total CO_2 emissions are absorbed by the oceans and another quarter by the biosphere, tempering the impact of human activities.

1. Open and run the file `Decompose_CO2_data.mlx`. It loads the dataset stored in the file `co2_annmean_mlo.csv` and plot the CO_2 concentrations over time.
2. Observe the CO_2 levels increase over time.
3. Investigate the structure of the dataset. Is an additive model acceptable ?
4. Perform a decomposition of the time series.

2.3.2 Airline Passengers

The data for this analysis has been classically used and analyzed to illustrate time series analysis and forecasting. The data represent monthly total passengers of an international airline (in thousands of passengers), January 1949–December 1960. The time series is plotted in Figure 3.3. Does the time series appear to be stationary?

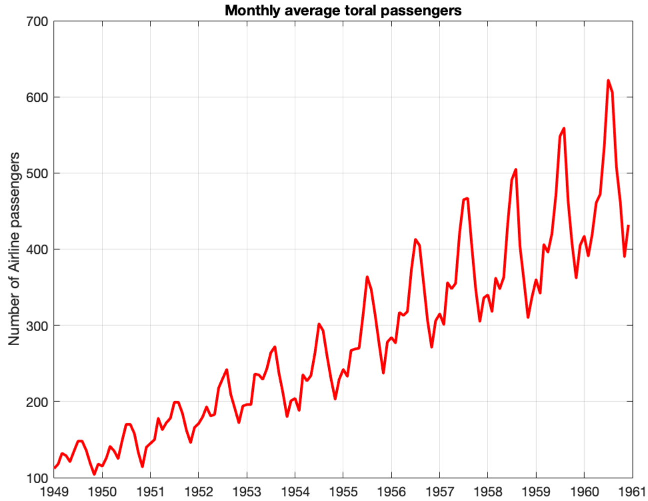

Figure 2.2: Total number of an international airline passengers

1. Open and run the file `Decompose_AirlinePassengers_data.mlx`. It loads the dataset available in Matlab and plot the Monthly total passengers over time.
2. Observe the time series increase over time.
3. Investigate the structure of the dataset. Is an additive model acceptable ?
4. Perform a log transformation of the data and plot the resulting series. Compare the behavior of the original and log-transformed series.

5. Do you see an advantage in using a log transformation for modeling purposes?
6. Perform a decomposition on the log-transformed time series.

Lab 3

The Box-Jenkins forecasting methodology

Downloading of the data needed for the tutorial/lab

1. Download the zipped file **Lab3_TSAF.zip** from the course website and save it in your Matlab working directory.

2. Start Matlab.

3. By clicking on the browse for folder icon , **change the current folder of Matlab so that it becomes your Lab3_TSAF folder** that contains the files needed for this lab.

4. It is highly recommended to use the Matlab Live Editor. In the Live Editor, you can create live scripts that show output together with the code that produced it.

Layout of the tutorial

1. A tutorial introduction to show how to use the recent Econometrics Modeller App.

2. Several assignments where you have to perform time series data forecasting with the ARIMA-based Box-Jenkins methodology and theory that you have learned in Lecture 3.

3.1 Tutorial introduction of the recent Econometrics Modeller App - Airline Passengers time series

The Econometric Modeler App provides a flexible interface for interactive exploratory data analysis of time series and conditional mean (for example, ARIMA), conditional variance (for example, GARCH), and time series regression model estimation.

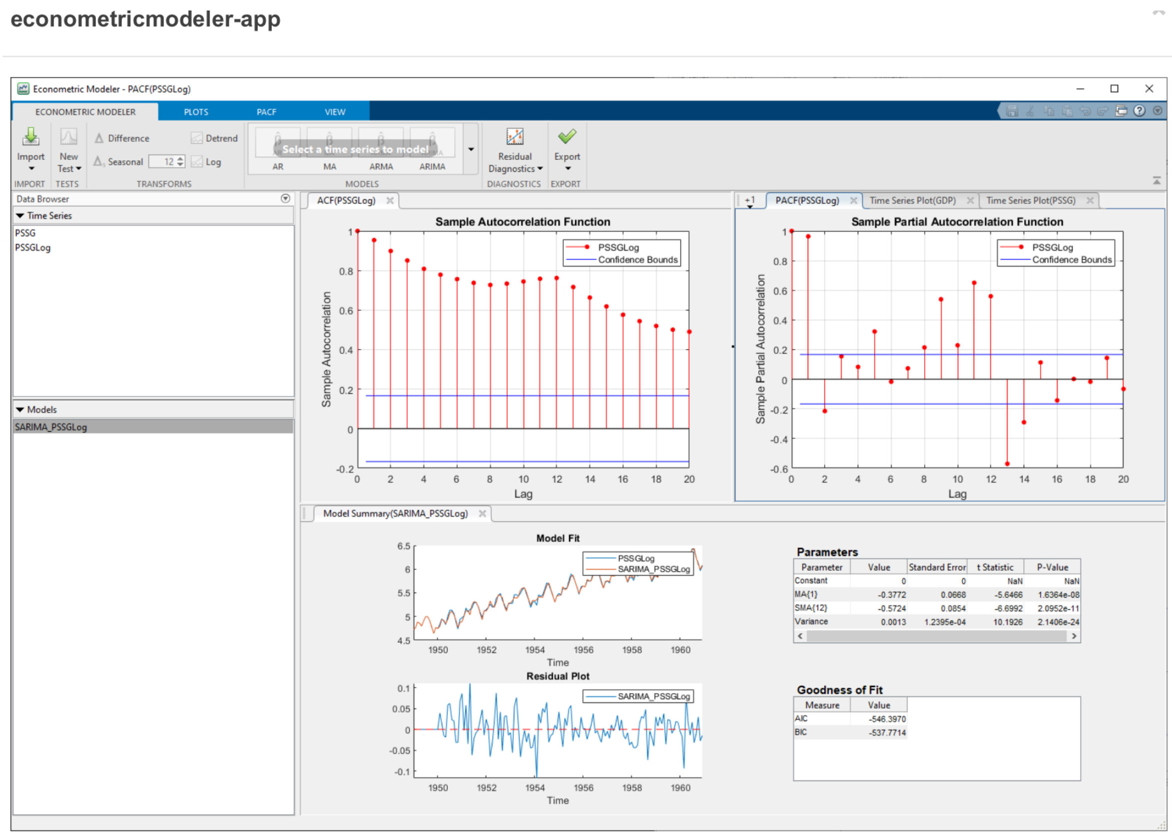

Figure 3.1: Main window of the Econometrics Modeller App

The main window of the App is shown in Figure 3.1. Using the App, you can:

1. Visualize and transform time series data.
2. Perform statistical specification and model identification tests.
3. Estimate candidate models and compare fits.
4. Perform post-fit assessments and residual diagnostics.
5. Automatically generate code or a report from a session.

This quick introduction will show you how to use the Econometric Modeler App for time-series analysis, including data transformation, visualization, statistical tests, and model fitting. The featured example is based on the Airline passengers data which has been classically used in the past to illustrate time-series analysis. This data is available in the Econometric Toolbox. You will learn how to:

1. Visualize data using traditional plots, including ACF and PACF plots.
2. Transform data using log transformation.
3. Create seasonal ARIMA models for time-series analysis.
4. Perform statistical diagnostics on residuals.

Watch the 5mn introductory tutorial video available at:
`fr.mathworks.com/videos/creating-arima-models-using-econometric-modeler-app-1515711526970.html`

More examples can be found from the following link: `https://www.mathworks.com/help/econ/econometricmodeler-app.h`
Note in particular that the results of an Econometric Modeler app session can be easily shared by generating a live functions to use outside the app.

3.2 Assignments - Forecasting of simulated time series data and of the classical international airline passengers

3.2.1 Simulated time series

The time series from a simulated ARIMA model are stored in the file `Simulated_time_series.mat`.
The simulated ARIMA time series is plotted in Figure 4.1.

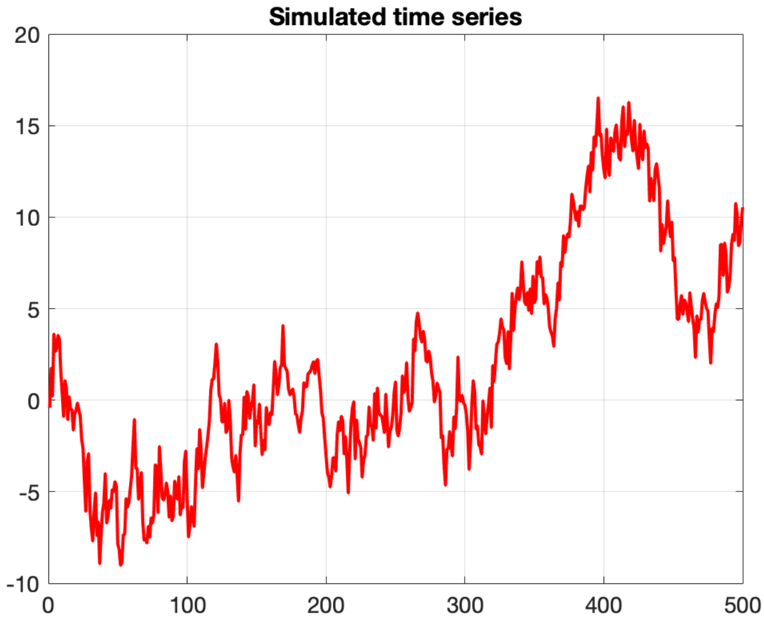

Figure 3.2: Time series data of a simulated ARIMA model

Does the time series appear to be stationary ?
By using the Econometric Modeler App, determine the best ARIMA model from this time series. Explain your choices.

3.2.2 International airline passengers

The data for this analysis has been classically used and analyzed to illustrate time series analysis and forecasting. The data represent monthly total passengers of an international airline (in thousands of passengers), January 1949–December 1960. The time series is plotted in Figure 3.3. Does the time series appear to be stationary ?

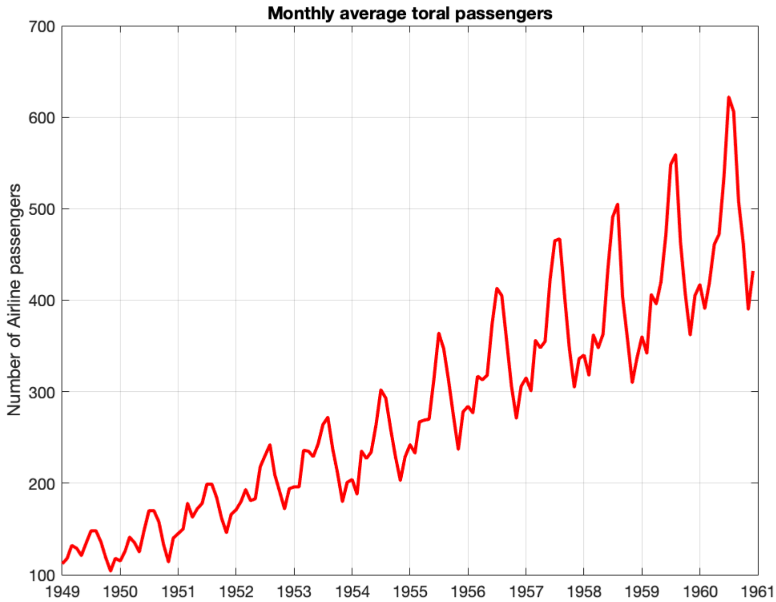

Figure 3.3: Total number of an international airline passengers

The Box-Jenkins methodology of ARIMA modelling and forecast has been applied to this time series by using the Matlab routines in the Econometrics toolbox.

1. Open the file `Estimate_and_forecast_Airline_Passengers.mlx`.
2. Execute every section and understand the different steps of the Box-Jenkins methodology that lead to a forecast of the Airline passengers over the next 5 years. Use the slides of the lectures to understand the full procedure.

Lab 4

Application of the ARIMA-based Box-Jenkins methodology

The work done during this 4h00 lab will be noted in a report and marked. You are asked to follow the instructions given below to write your report. You will also have to prepare a short oral presentation of 10 mn showing your time series analysis and forecasting methodology and results.

Instructions for writing your lab report

A lab report is a scientific document. It should be self-contained: that is, someone who has never seen the lab instructions should be able to understand the problems that you are solving and how you are solving them. All the choices you have made should be clearly motivated. Comment and explain all your plots and your Matlab-code, so as to make it easy to follow your way of thinking.
An example of a report summarizing the use of ARIMA model in Matlab to forecast the stock price of Air Asia Airlines is available at:
`https://www.mathworks.com/matlabcentral/fileexchange/68576-stock-prediction-using-arima?s_tid=srchtitle`

Downloading of the time series data

1. Download the zipped file **Lab4_TSAF.zip** from the course website and save it in your Matlab working directory.

2. Start Matlab.

3. By clicking on the browse for folder icon , **change the current folder of Matlab so that it becomes your Lab4_TSAF folder** that contains the files needed for this lab.

4. It is highly recommended to use the Matlab `Live Editor`. In the Live Editor, you can create live scripts that show output together with the code that produced it. Moreover, from the .mlx file, you can create a pdf file very easily to generate your report.

Layout of the Lab

1. A first assignment where you have to perform time series data modelling and forecasting with the ARIMA-based Box-Jenkins methodology on a simulated time-series data.

2. A second assignment where you have to perform time series data modelling and forecasting on a practical real-life time series with the methodology of your choice.

4.1 Analysis and modelling of a simulated time series

The time series from a simulated ARIMA model are stored in the file `simulated_ARIMA.mat`. The simulated time series is plotted in Figure 4.1.

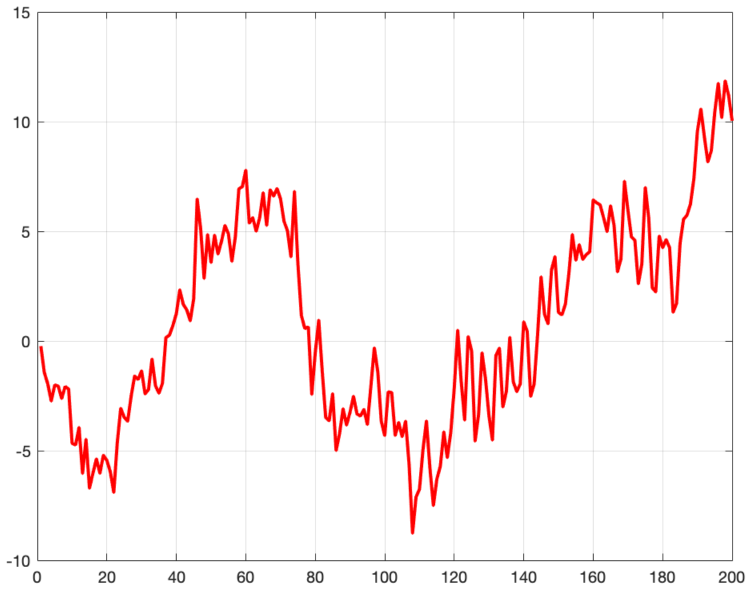

Figure 4.1: Time series data of a simulated ARIMA model

1. Does the time series appear to be stationary ?
2. By using the Econometric Modeler App, determine the best ARIMA model for this time series. Explain all your choices.
3. Export your analysis in a pdf file.

If necessary, you can watch again the 5mn introductory tutorial video available at:
`fr.mathworks.com/videos/creating-arima-models-using-econometric-modeler-app-1515711526970.html`

More examples of the use of the Econometric Modeler app can be found from the following link:
`https://www.mathworks.com/help/econ/econometricmodeler-app.html`
It is reminded that the results of an Econometric Modeler app session can be easily shared by generating a live function to use outside the app.

4.2 Modelling and forecasting of real-life time series

You will have to study one of the real-life time series data that are described in the sub-sections below.

You can adapt the file `Estimate_and_forecast_Airline_Passengers.mlx` that was given in the previous lab or use the Econometrics Modeller App to determine the best SARIMA model that will be used to provide 12-months ahead forecast of the time series. Split the dataset to create a training and test set.
Define in your Matlab code the last 12 months as the test set and use the remaining as the training set by using the code below.

```
y_train = y(1:end-12);
Months_train = Months(1:end-12);
y_test = y(end-11:end);
Months_test = Months(end-11:end);
```

4.2.1 Atmospheric CO2 concentration forecasting

The data for this analysis were collected from the Mauna Loa CO_2 monitoring site in Hawaii. Monthly observations started in March 1958 and ended with the most recent recorded month, November 2020. The variables include the monthly CO_2 concentrations in part per millions (ppm); the month and year of every measurement were also recorded.

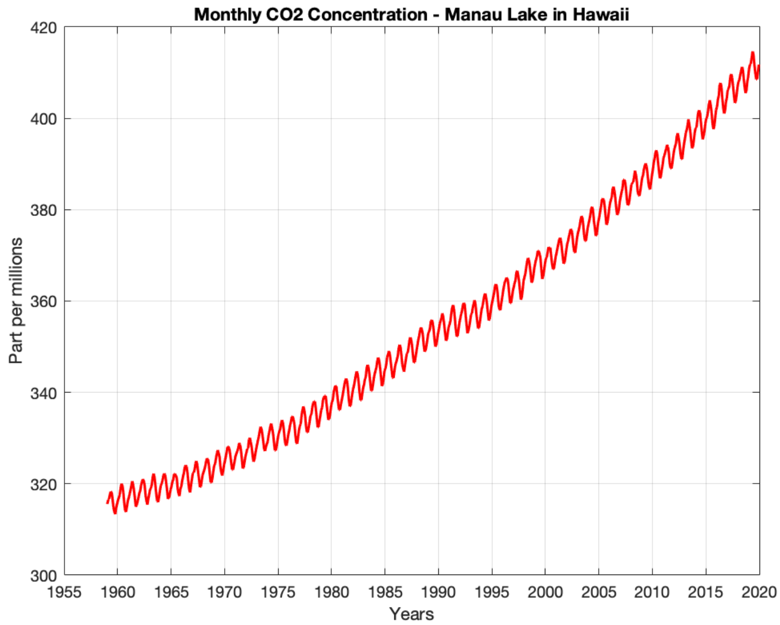

Figure 4.2: Carbon dioxide levels at Mauna Loa, Hawaii

As seen in Figure 4.2, the carbon dioxide levels at Mauna Loa, Hawaii increased above the 400 parts per million threshold. This threshold was first crossed as a global average in 15 March 2015. Why is this figure important ? Scientists have warned that crossing this threshold could result in more global warming and the disasters associated with it, like sea-level rise and ocean acidification.

CO_2 concentrations from 1958 to 2020 are stored in the file `CO2_data.mat`.

1. Split the dataset to create a training and test set. The training set will be used for building the model (estimating its parameters). The test set will be used only for forecasting, i.e. to test the performance of your estimated model.

2. Estimate a model for the CO_2 level given the training set. Explain your choices.

3. Use your best model to forecast the CO_2 concentration over a 12-months horizon on the independent test set.

4. Create a plot to compare the actual level and the predicted level as well as a plot for the computed forecast error. In addition to the visualization, quantify the performance of your forecast by computing two classical metrics: the root mean square error (RMSE) and the mean average percent error (MAPE).

5. Generate a PDF report of all your analysis.

You can get some helpful information from the paper available at:
`https://www.researchgate.net/publication/343449582_A_Statistical_Analysis_of_Atmospheric_CO_2_Levels_at_Mauna_Loa`

4.2.2 Amtrak passenger forecasting

Amtrak is a passenger railroad service that provides medium and long-distance intercity service in the contiguous United States and to nine Canadian cities. The data for this analysis represent the monthly total passengers of Amtrak (in thousands of riderships), January 1991–December 2004. The time series is plotted in Figure 4.3.

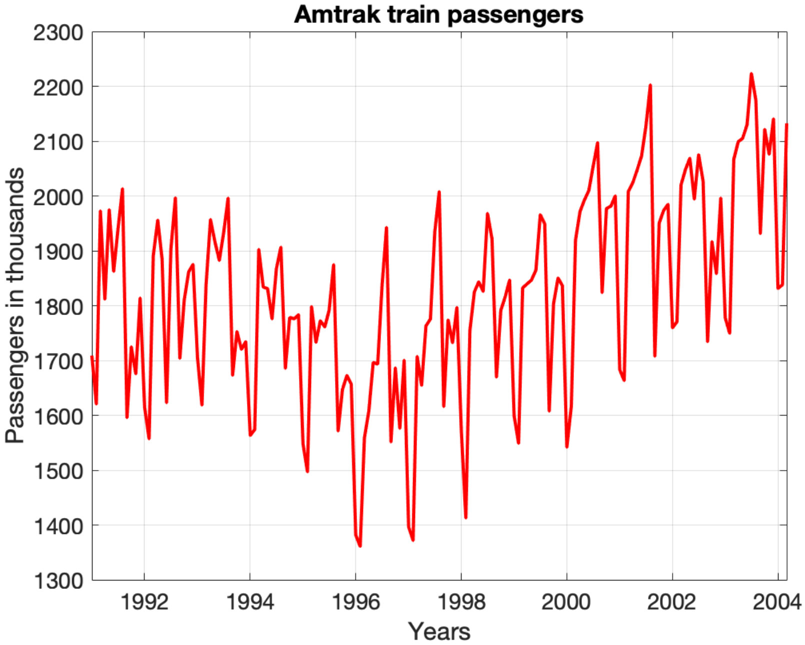

Figure 4.3: Total number of Amtrak passengers

Monthly riderships data from 1991 to 2004 are stored in the file `Amtrak_data.mat`.

1. Split the dataset to create a training and test set. The training set will be used for building the model (estimating its parameters). The test set will be used only for forecasting, i.e. to test the performance of your estimated model. Define the last 12 months as the test set and use the remaining as the training set.

2. Estimate a model for the Amtrak passenger number given the training set. Explain your choices.

3. Use your best model to forecast the ridership number over a 12-months horizon on the independent test set.

4. Create a plot to compare the actual level and the predicted Amtrak riderships as well as a plot for the computed forecast error. In addition to the visualization, quantify the performance of your forecast by computing two classical metrics: the root mean square error (RMSE) and the mean average percent error (MAPE).

5. Generate a PDF report of all your analysis.

The data for this analysis has been used by Galit Schmueli to illustrate time series analysis and forecasting in her video series available on youtube.
You can get some helpful information from the following video:
`https://www.youtube.com/watch?v=0xHf-SJ9Z9U`

4.2.3 Energy demand forecasting

Storing electricity for future use is extremely costly and inefficient. Therefore, it is important for electricity producers to generate only as much as is demanded/needed by consumers. Electricity consumption varies throughout the day. It has peaks and dips depending on the time of day, the day of the week, bank holidays and vacations, but also the seasons and weather conditions. It reflects the daily life of the people in a given country and its economic activity. You can, for example, see the electricity consumption and one-day ahead forecast from:
`https://www.rte-france.com/eco2mix/la-consommation-delectricite-en-france`

The data for this analysis represent monthly energy demand in the State of New South Wales in Australia, from 1999 to mid 2014. The variables are stored in the file `EnergyDemand.mat`. The time series is plotted in Figure 4.4.

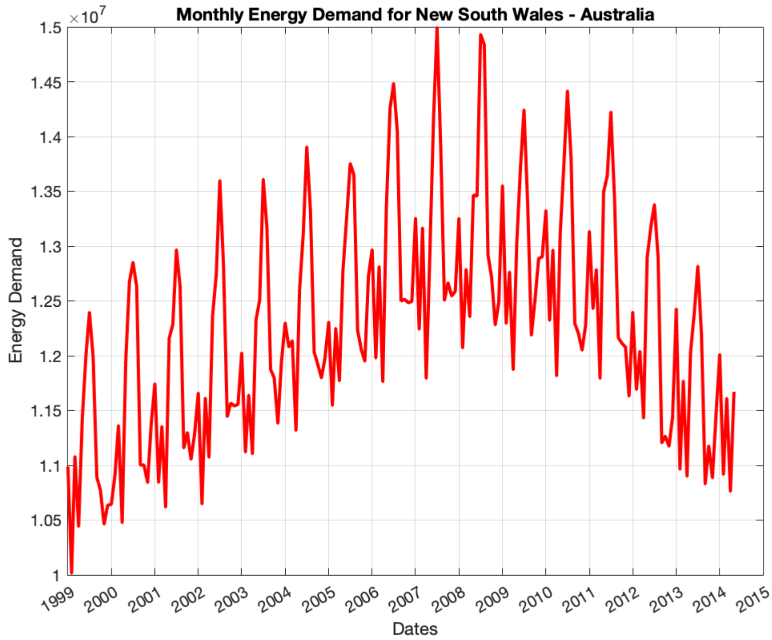

Figure 4.4: Monthly energy demand

As it can be noticed, energy demand in Australia has started to decline since 2008. As a result, traditional regression based models that were being used to forecast long term energy load were highly inaccurate in their predictions. Energy demand has since 2008 continued to fall, however will it increase again? If so, how can this be predicted? A dynamic model to forecast long term energy demand is needed.

1. Split the dataset to create a training and test set. The training set will be used for building the model (estimating its parameters). The test set will be used only for forecasting, i.e. to test the performance of your estimated model. Define the last 12 months as the test set and use the remaining as the training set.
2. Estimate a model for the energy demand given the training set. Explain your choices.
3. Use your best model to forecast the energy demand over a 12-months horizon on the independent test set.
4. Create a plot to compare the actual level and the predicted energy demand as well as a plot for the computed forecast error. In addition to the visualization, quantify the performance of your forecast by computing two classical metrics: the root mean square error (RMSE) and the mean average percent error (MAPE).
5. Generate a PDF report of all your analysis.

The data for this analysis has been used by Mathworks to illustrate time series analysis and forecasting with Matlab.
You can get some helpful information from the video available at (you can start watching at 22:00mn):
`http://mathworks.com/videos/long-term-energy-forecasting-with-econometrics-in-matlab-99301.html`
and associated code:
`http://au.mathworks.com/matlabcentral/fileexchange/49279-long-term-energy-forecasting-with-econometrics-in-matlab`