

Data analysis and applications

Project

The project includes two independent parts. Before implementing the project, you need to understand the related content. The final score of the project depends on the implementation, the experimental results, and the questions I will ask you when you will present the project. You should present a personal work.

Part1

Let us consider the univariate temporal data in the file IBM.txt. It represents IBM's stock prices achieved in the days between 2006-01-03 and 2017-12-29. Write a Python computer programme for

1. The estimation of the autoregressive (AR) model from the data. You need to use the maximum a posterior method.
2. Find the best AR model order by studying the fitting error. Hint: The error is a normal random variable.
3. Plot the data, the data synthesised by using the AR model (having the best order), and the error histogram.

Part2

Let us consider the spatial data in the file Soil.dat. It includes the geographical coordinates, Potassium (K mg/l), Phosphorus (P mg/l), and pH levels in the soil. In this project, we only focus on the study of the pH level. Write a Python computer programme allowing to

1. Estimate the empirical variogram.
2. Plot the empirical variogram.
3. Fit two different variogram functions to the empirical variogram.
4. Predict values at no less than three positions by using each of the variogram function.
5. Compare the predictions by using the prediction errors.

Hint: For the two parts, you need to check whether the data meets the stationarity requirements. If not, you need to add transformations leading to a stationary data.

Deliverable

A report, which includes the description of the preprocessing step (if any), a breve description of the estimators, code sources, experimental results, and the analysis of the results.

Deadline

June 21 between 3pm and 5pm. First arrived, first served. You need to bring the report to my office and to show me a demo on your own laptop.