GEOSPATIAL DATA ANALYSIS

Lab 1 – Multiple Linear Regression

Roberto Monti – roberto.monti@polimi.it

Alberto Vavassori – alberto.vavassori@polimi.it

A.Y. 2023/2024

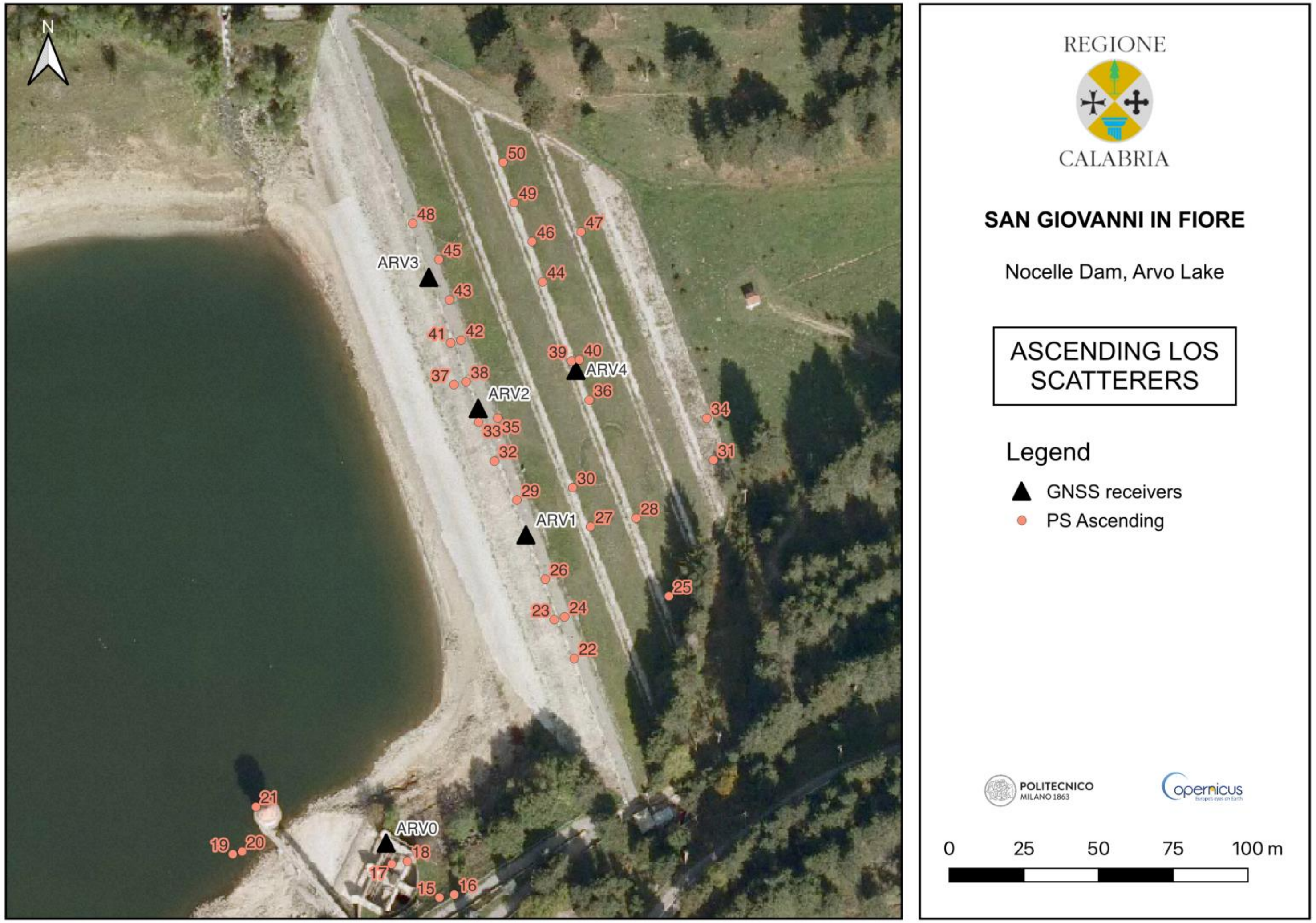
Overview

The scope of the laboratory is to perform a Multiple Linear Regression between the following variables:

- One response variable:
    - **displacement [mm]**, $Y(t)$, of different points (i.e. PSs) of a dam, measured with Synthetic-Aperture Radar (SAR) technology (one measurement per day).

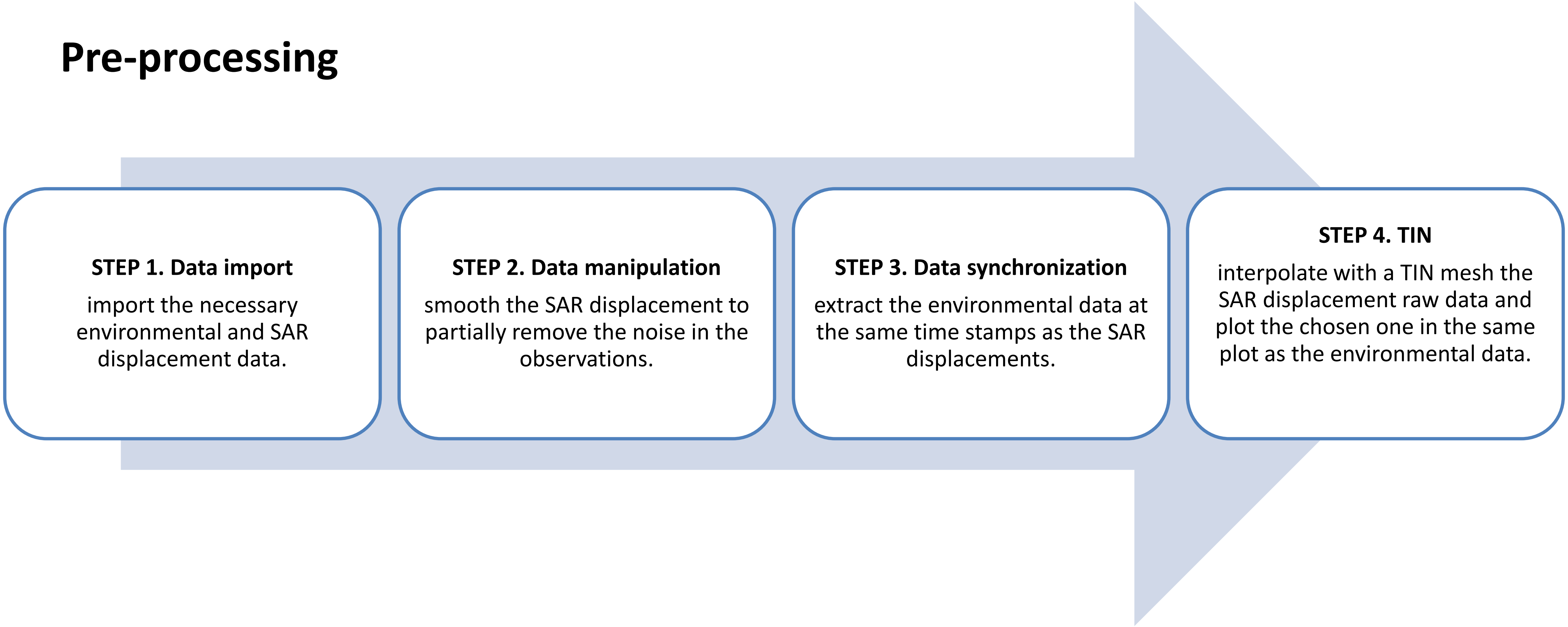
- Two environmental variables, called predictors or independent variables:
    - **water level [m]**, $h(t)$, and **water surface temperature [°C]**, $T(t)$, relative to the reservoir (more than one measurement per day – daily resampling is needed).

Overview


N
50
49
48
47
46
45
ARV3
44
43
41
42
39
40
ARV4
37
38
36
ARV2
34
33
35
32
31
30
29
28
ARV1
27
26
25
23
24
22
21
ARV0
19
20
18
17
15
16
REGIONE
CALABRIA
SAN GIOVANNI IN FIORE
Nocelle Dam, Arvo Lake
ASCENDING LOS SCATTERERS
Legend
▲ GNSS receivers
● PS Ascending
POLITECNICO MILANO 1863
opernicus
0
25
50
75
100 m

Workflow

Pre-processing



STEP 1. Data import

import the necessary environmental and SAR displacement data.

STEP 2. Data manipulation

smooth the SAR displacement to partially remove the noise in the observations.

STEP 3. Data synchronization

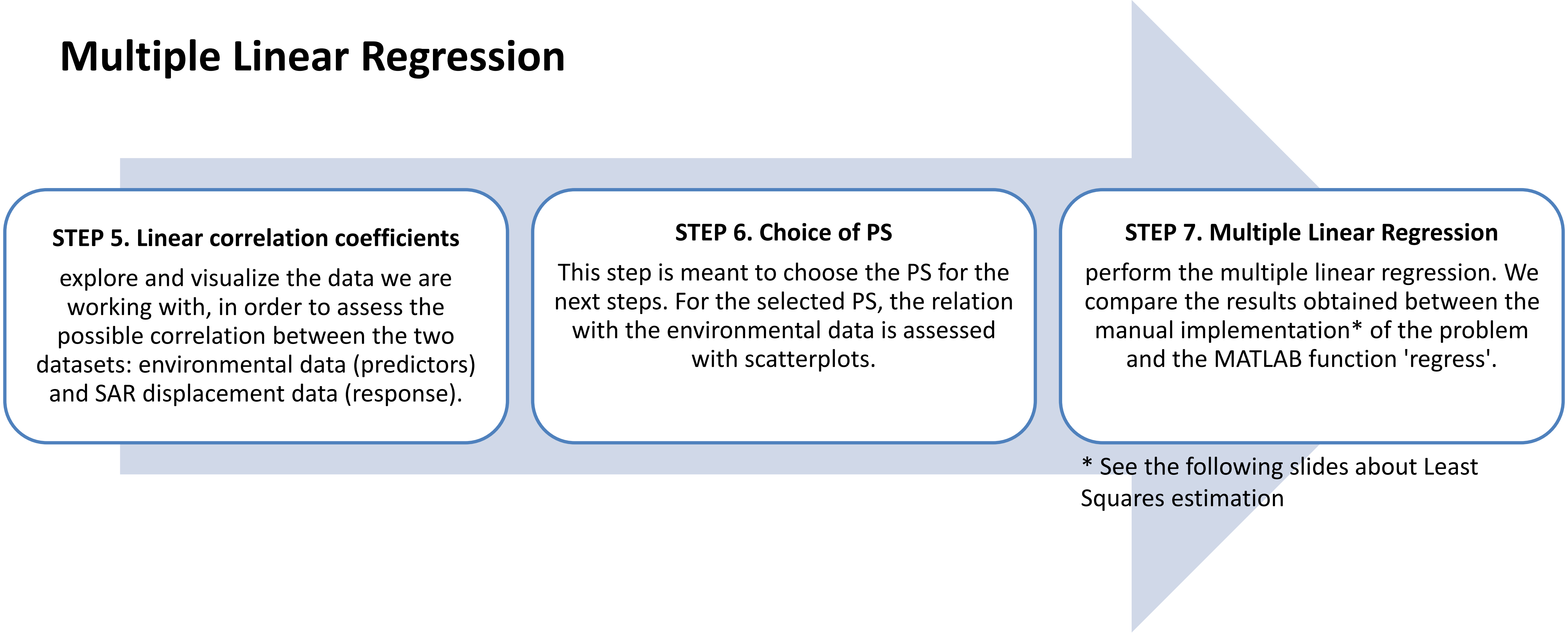
extract the environmental data at the same time stamps as the SAR displacements.

STEP 4. TIN

interpolate with a TIN mesh the SAR displacement raw data and plot the chosen one in the same plot as the environmental data.

Workflow

Multiple Linear Regression



STEP 5. Linear correlation coefficients

explore and visualize the data we are working with, in order to assess the possible correlation between the two datasets: environmental data (predictors) and SAR displacement data (response).

STEP 6. Choice of PS

This step is meant to choose the PS for the next steps. For the selected PS, the relation with the environmental data is assessed with scatterplots.

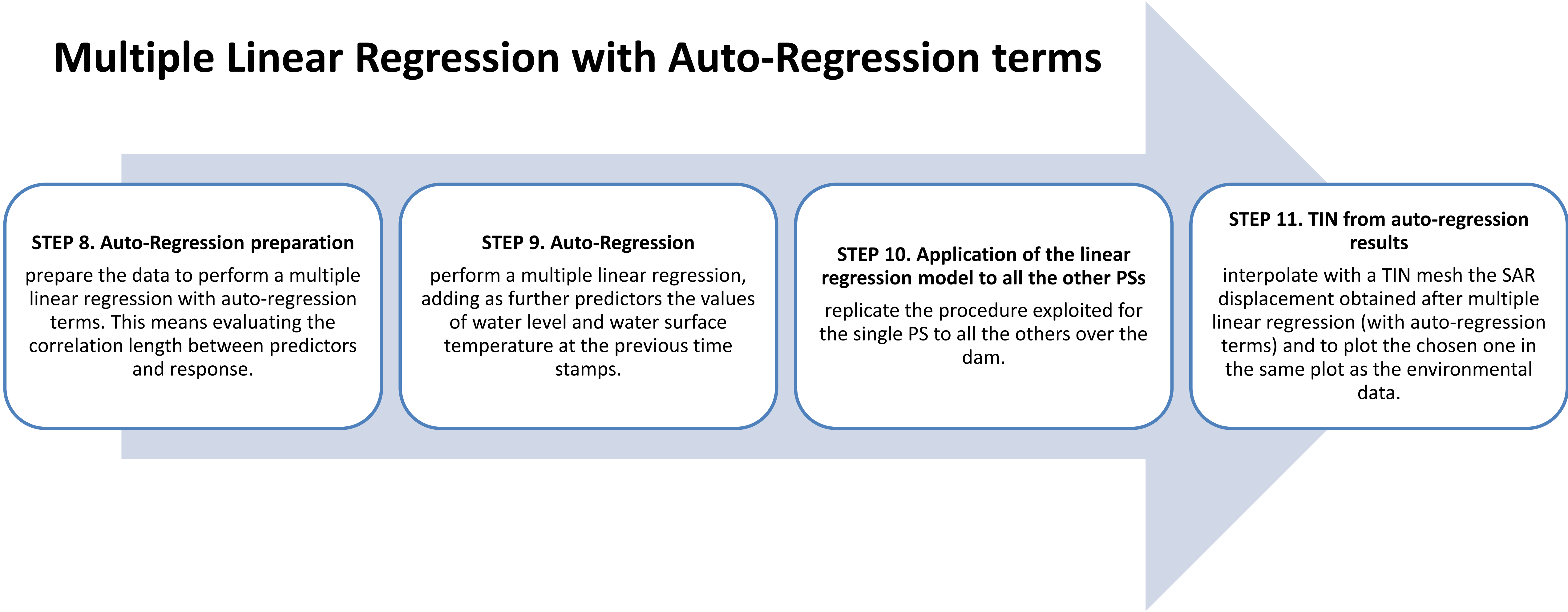
STEP 7. Multiple Linear Regression

perform the multiple linear regression. We compare the results obtained between the manual implementation* of the problem and the MATLAB function 'regress'.

* See the following slides about Least Squares estimation

Workflow

Multiple Linear Regression with Auto-Regression terms



STEP 8. Auto-Regression preparation

prepare the data to perform a multiple linear regression with auto-regression terms. This means evaluating the correlation length between predictors and response.

STEP 9. Auto-Regression

perform a multiple linear regression, adding as further predictors the values of water level and water surface temperature at the previous time stamps.

STEP 10. Application of the linear regression model to all the other PSs

replicate the procedure exploited for the single PS to all the others over the dam.

STEP 11. TIN from auto-regression results

interpolate with a TIN mesh the SAR displacement obtained after multiple linear regression (with auto-regression terms) and to plot the chosen one in the same plot as the environmental data.

Least Squares estimation

Multi linear model between the SAR displacement $Y(t)$, the water level $h(t)$, and the water surface temperature $T(t)$:

$$Y(t) - \bar{Y} = a\ (h(t) - \bar{h}) + b\ (T(t) - \bar{T}) + c$$

Least squares solution:

$$Y_0 = \ Ax$$

$$A = \begin{bmatrix} h_1 - \bar{h} & T_1 - \bar{T} & 1 \\ h_2 - \bar{h} & T_2 - \bar{T} & 1 \\ h_3 - \bar{h} & T_3 - \bar{T} & 1 \\ \dots & \dots & \dots \end{bmatrix} \quad Y_0 = \begin{bmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ Y_3 - \bar{Y} \\ \dots \end{bmatrix} \quad x = [a \quad b \quad c] \quad \bar{h} = \frac{1}{n}\sum_{i=1}^{n} h_i\, , \bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i\, , \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$Q = I \qquad N = A^T A = \begin{bmatrix} \sum_i (h_i - \bar{h})^2 & \sum_i (h_i - \bar{h})(T_i - \bar{T}) & 0 \\ \sum_i (h_i - \bar{h})(T_i - \bar{T}) & \sum_i (T_i - \bar{T})^2 & 0 \\ 0 & 0 & n \end{bmatrix} \qquad A^T Y_0 = \begin{bmatrix} \sum_i (h_i - \bar{h})(Y_i - \bar{Y}) \\ \sum_i (T_i - \bar{T})(Y_i - \bar{Y}) \\ 0 \end{bmatrix}$$

Least Squares estimation

The Least Squares adjustment leads to the following solution in terms of estimate of regression parameters:

$$\hat{x} = \begin{bmatrix} \dfrac{\sum_i (h_i - \bar{h})(Y_i - \bar{Y})}{\sum_i (h_i - \bar{h})^2} \\ \dfrac{\sum_i (T_i - \bar{T})(Y_i - \bar{Y})}{\sum_i (T_i - \bar{T})^2} \\ 0 \end{bmatrix} = \begin{bmatrix} \dfrac{S_{hY}}{S_{hh}^2} \\ \dfrac{S_{TY}}{S_{TT}^2} \\ 0 \end{bmatrix} = \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \end{bmatrix}$$

Therefore, the estimate of the response is:

$$\boxed{\hat{Y}_i = \bar{Y} + \hat{a}(h_i - \bar{h}) + \hat{b}(T_i - \bar{T}) + \hat{c}}$$

Least Squares estimation

NOTE

If in the initial model, we do not subtract the mean of the variables, i.e.

$$Y(t) = ah(t) + bT(t) + c$$

then the estimate of the response will be:

$$\widehat{Y_i} = \hat{a}h_i + \hat{b}T_i + \widehat{c'}$$

Specifically, the estimate of the slope parameters is the same, while the constant term is different.
The relationship between $\hat{c}$ and $\widehat{c'}$ is the following:

$$\widehat{c'} = \hat{c} + \bar{Y} - \hat{a}\bar{h} - \hat{b}\bar{T}$$

POLITECNICO
MILANO 1863

13th Nov. 2023