Geospatial Data Analysis - Lab Assignment

Student: Lorenzo Carlassara 101724

The provided time series are real data made available by ARPA Veneto, the regional agency for environment protection. The whole dataset is composed by measurements from several boreholes each identified by a numeric ID.

The following project takes in consideration only a borehole, in particular one carrying a large amount of time samples, regardless it can be exentend to all the wells.

Contents

Data Preprocessing

The missing values are removed from the data table and they will be replaced if needed; the features used are:

- epochs 'DATA' converted upstream in integer numbers;
- piezometer measurements 'LIVELLOSTATICO [mslm]' taken between 1999 and 2021 (roughly);
- boreholds ID planar cartographic coordinates in the district of Padua;
- altitude of the boreholds ID 'QUOTA P.R. [mslm]';
- interpolated equidistant epochs;

Smoothing Least Square Interpolation

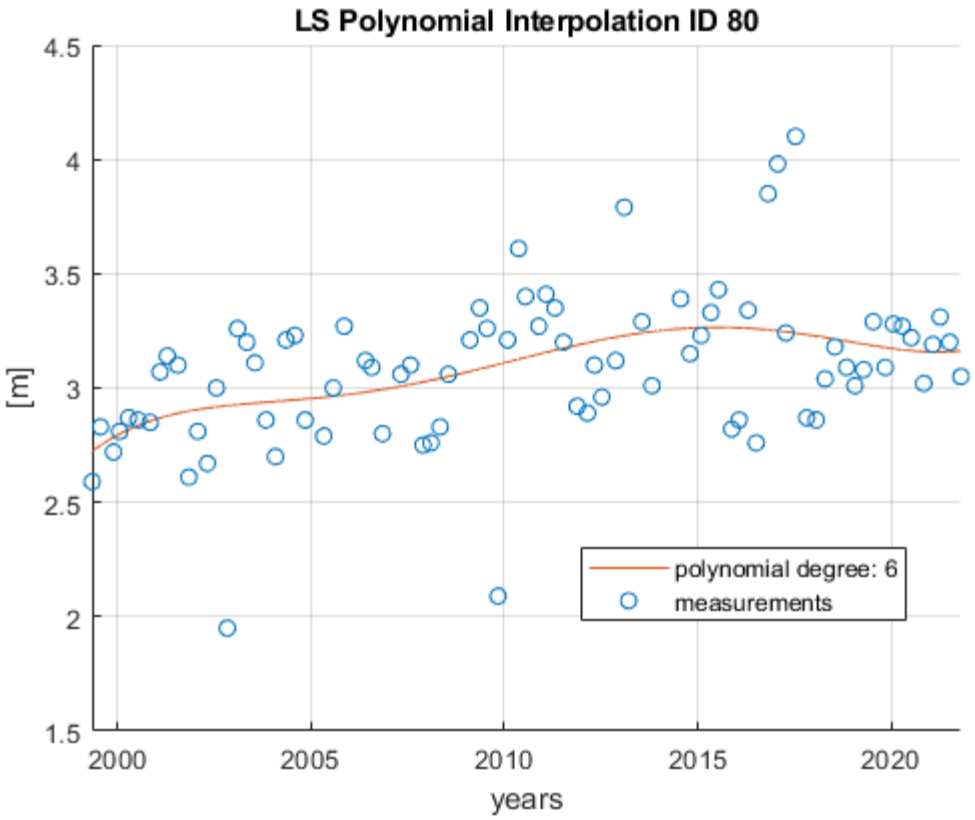
Find the best fitting polynomial with a low degree using a treshold on the ratio between consecutive polynomials of dregree *k* and *k+1* until the improvement is too low.

Once obtained the best number of parameters for the Smoothing Least Square Interpolation, the polynomial model is build computing the optimum solution that minimize the residuals.

The polynomial describes the overall overview of the samples behaviour and, assuming this as the deterministic component of the given samples, it will be deployed later on the further analysis.

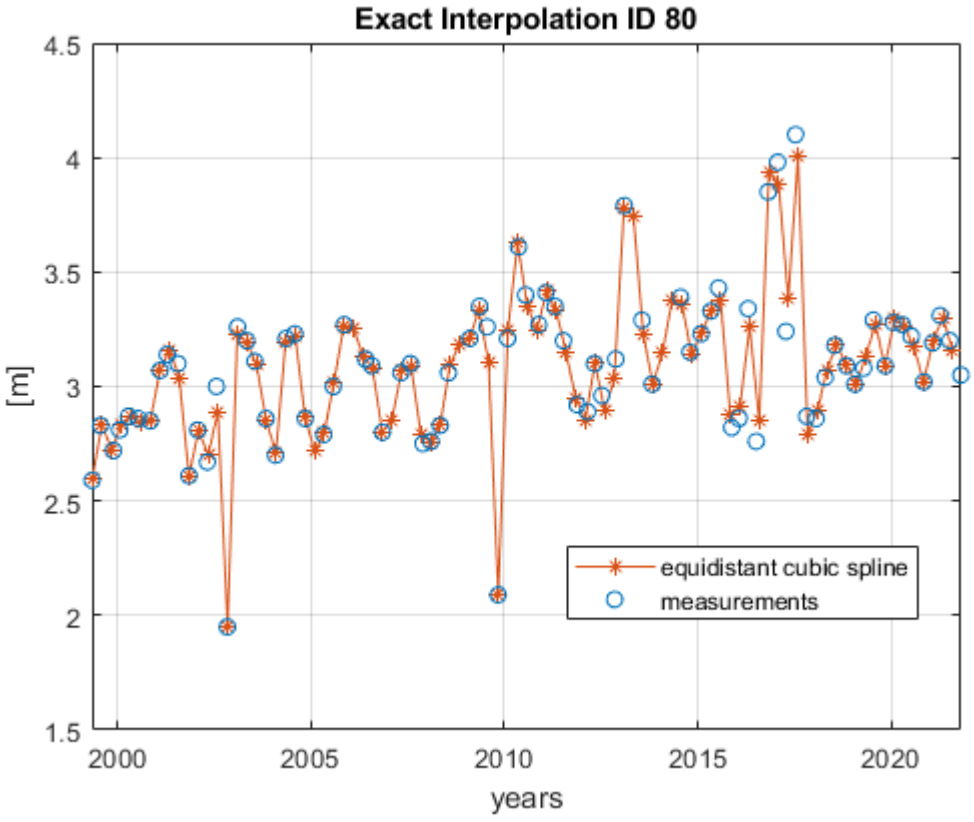
```
Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
RCOND =  1.053039e-16.
Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
RCOND =  1.087710e-24.
Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
RCOND =  1.157125e-32.
Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
RCOND =  1.254279e-40.
Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
RCOND =  1.353331e-48.
```

```
Warning: Matrix is close to singular or badly scaled. Results may be inaccurate.
RCOND =  1.353331e-48.
```



Exact Cubic Spline

Each timeseries covers a different period and although there are 4 times per year the dates of acquisition are different. In order to get an equally distant interval of sampling and fill the holes provided by the missing values, Exact Cubib Spline is a smooth suitable interpolation method based on a polynomial of third order in the Newton form.

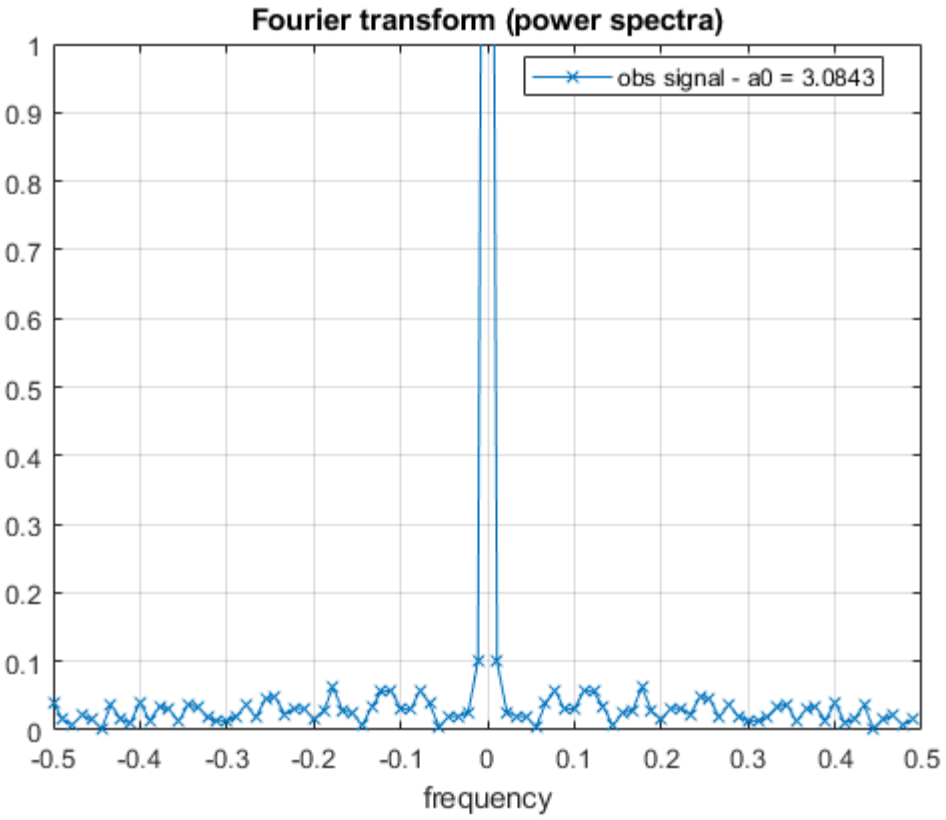


Discrete Fourier Transformation

Assuming the provided piezometer measurements to be samples of a signal that can be decomposed in a finite sum of sin and cosine functions.

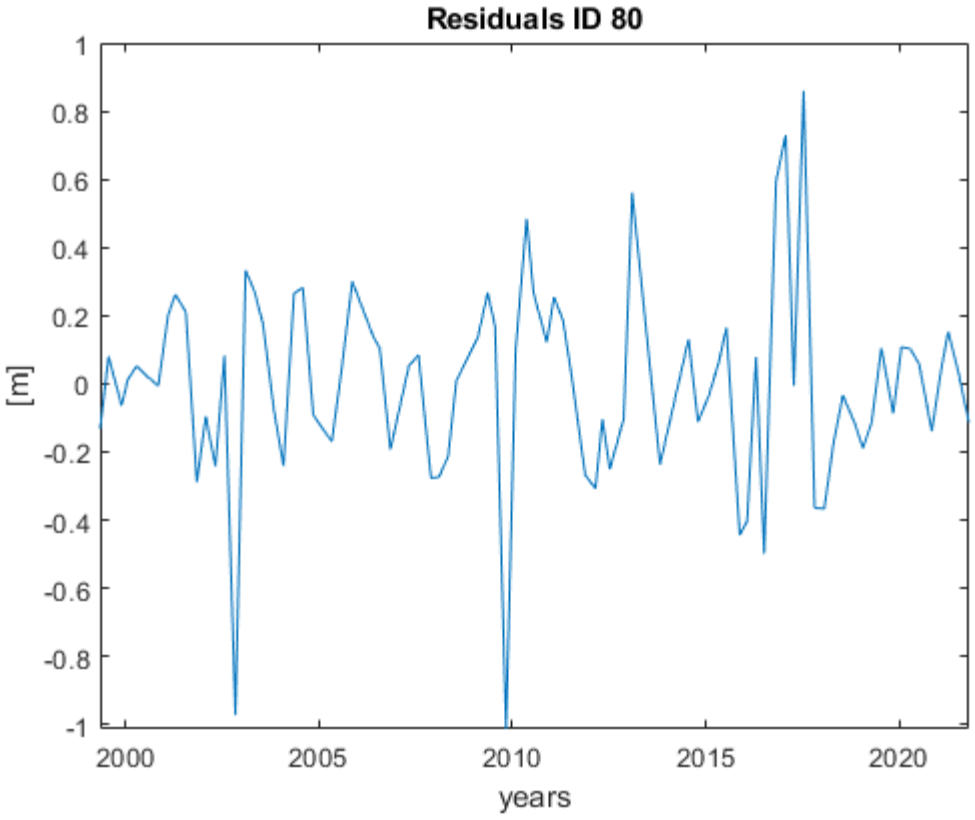
The resulting equidistant sampling got by the Exact Spline is required to build the power spectrum of the signal.

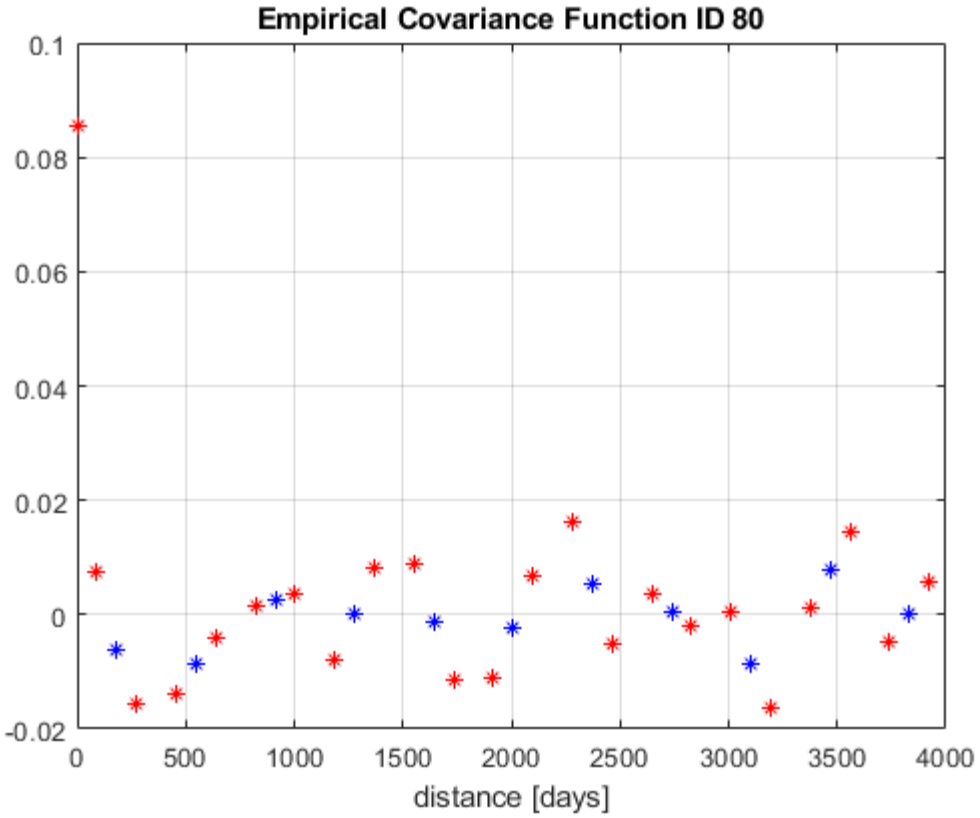
The power spectrum shows that there is no clear evidence to say if the frequencies are noise or harmonic component for example at 6 months.

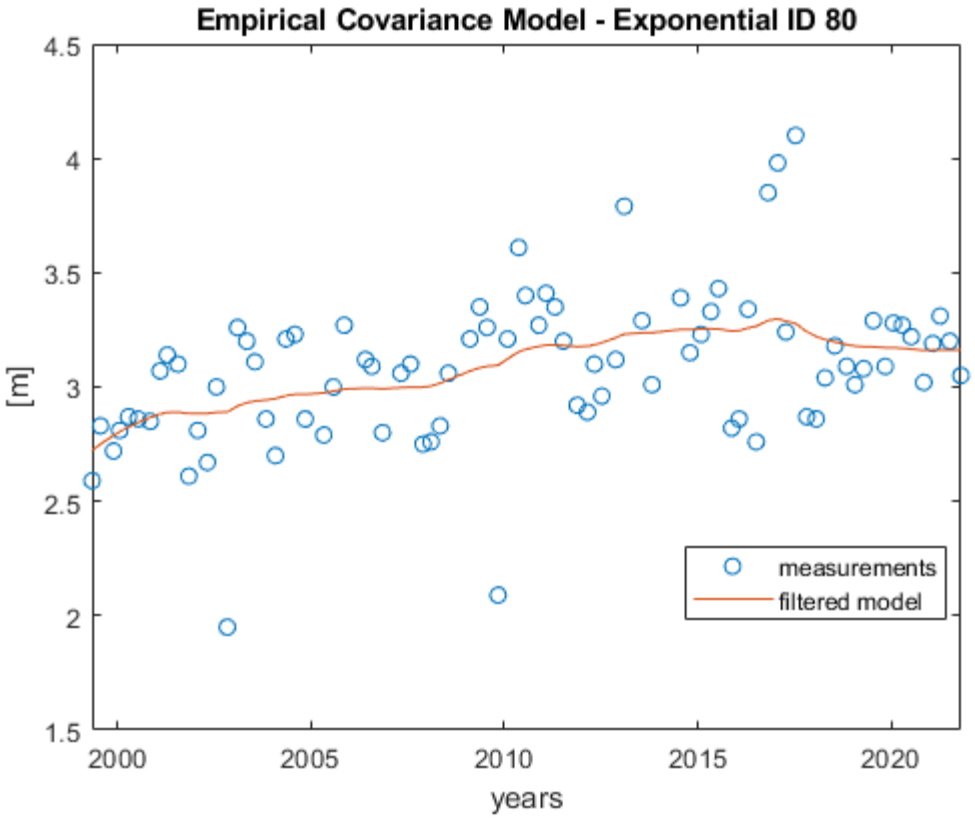


Collocation Technique

1. Empirical covariance function estimation by computing the residuals
2. Covariance model parameters by exact polynomial interpolation
3. Filter samples using the evaluated ECF model

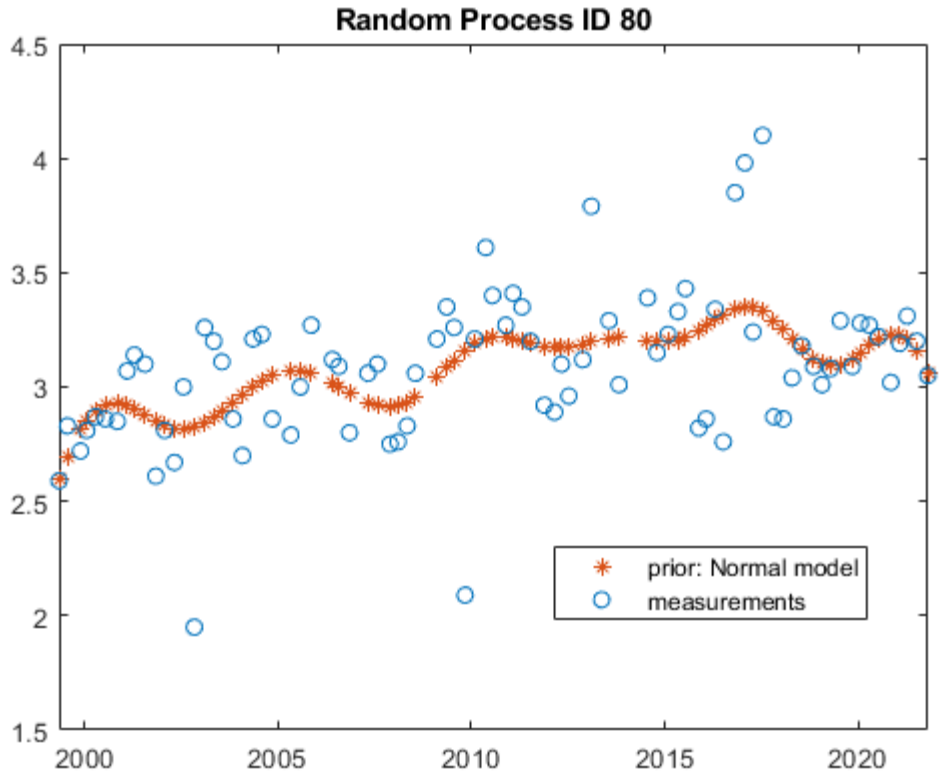


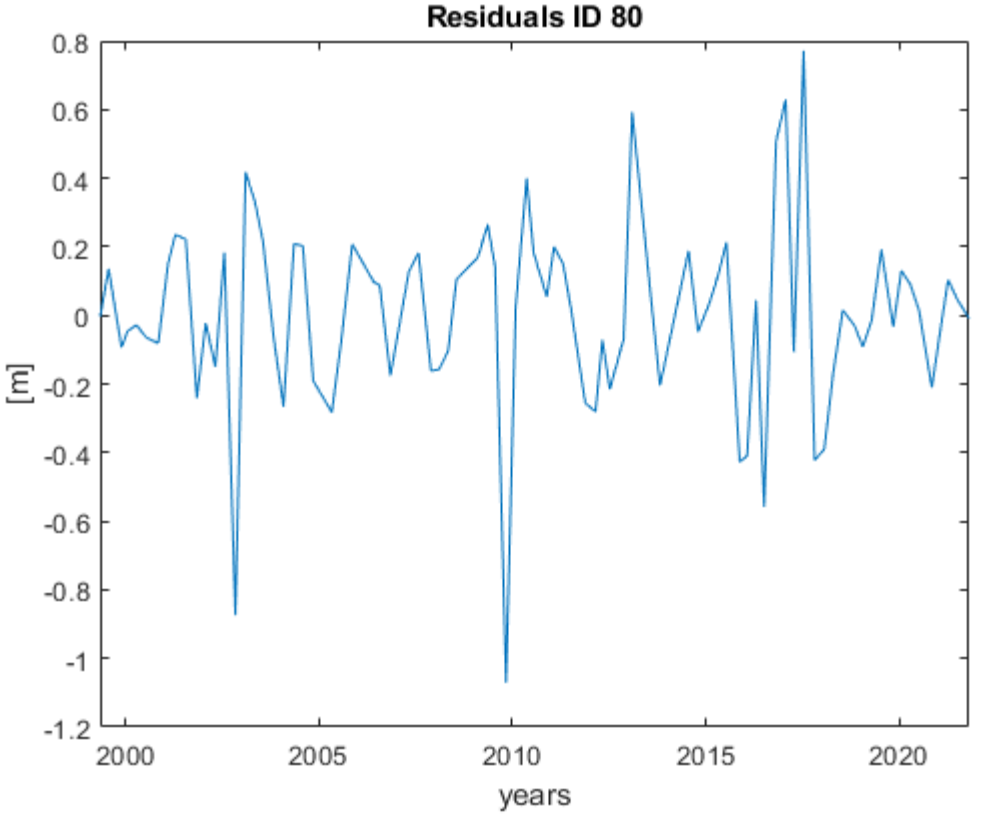


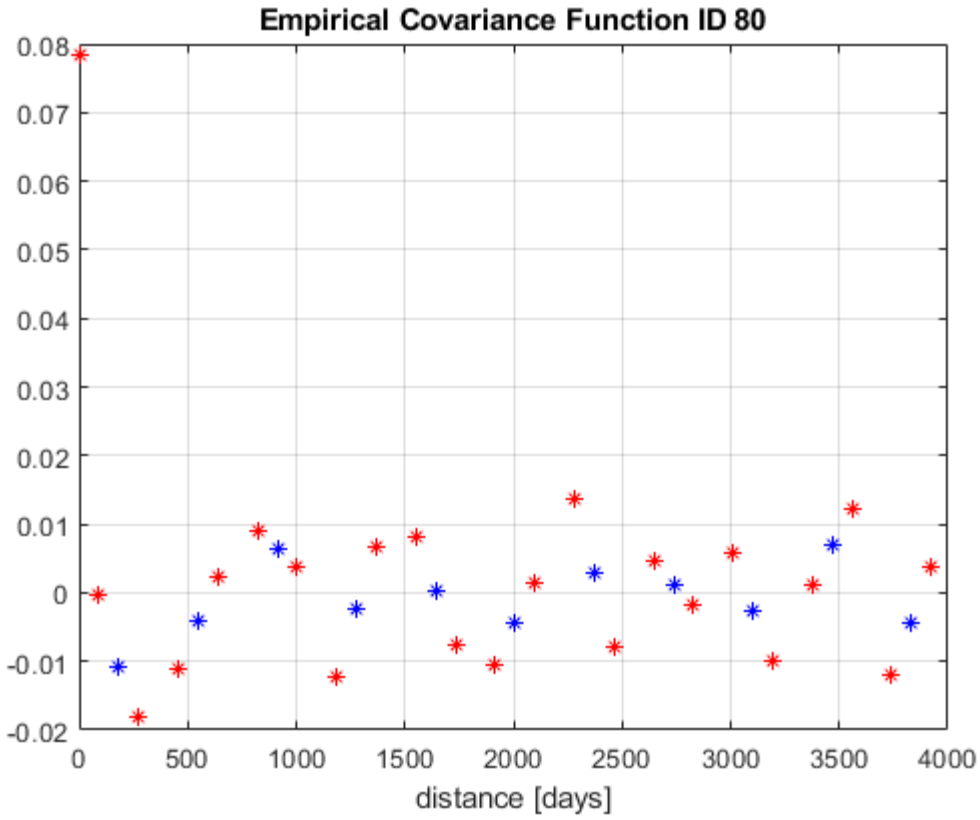


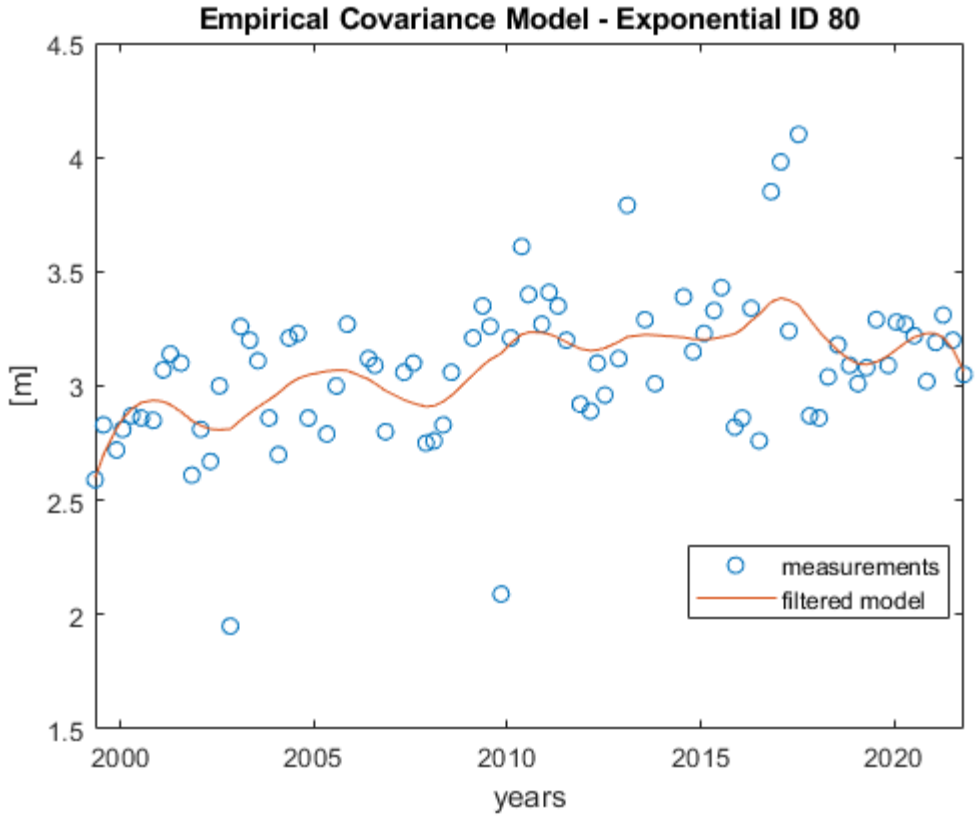
Kernel method for linear regression interpolation

Comparison of the result provided by the Collocation ECF with a priori assumption on the model distribution to be the popular Normal model with hyperparameters provided by maximum likelihood estimator MLE


Random Process ID 80
4.5
4
3.5
3
2.5
2
1.5
2000
2005
2010
2015
2020
prior: Normal model
measurements


Residuals ID 80
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
[m]
2000
2005
2010
2015
2020
years



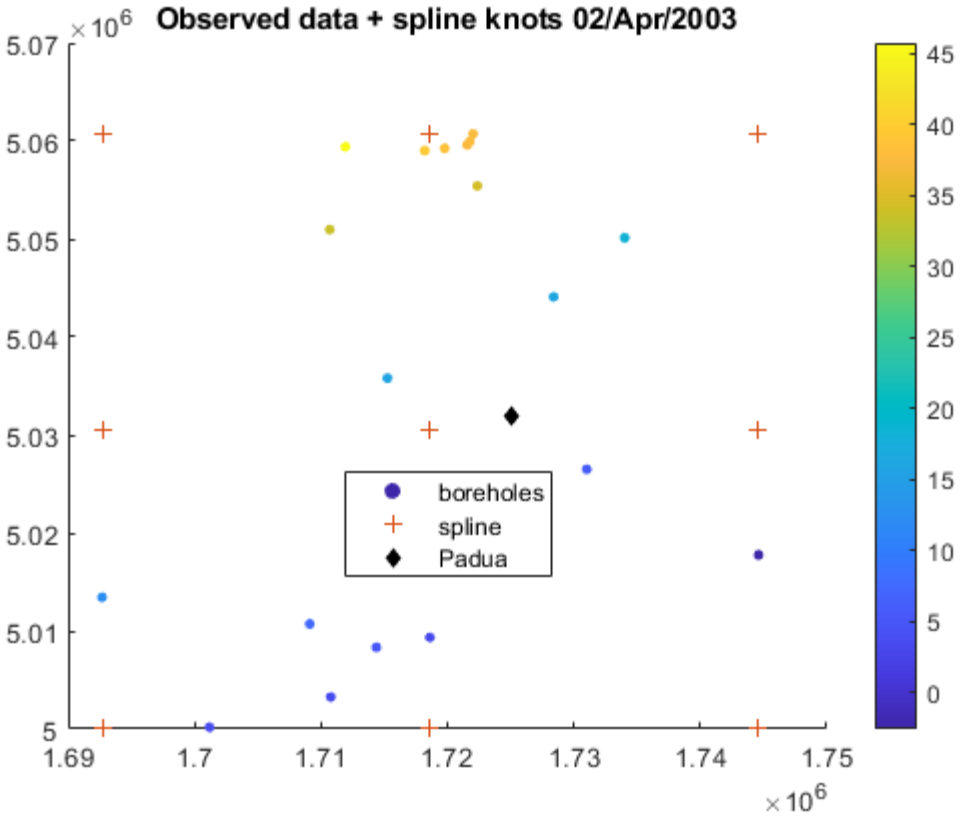


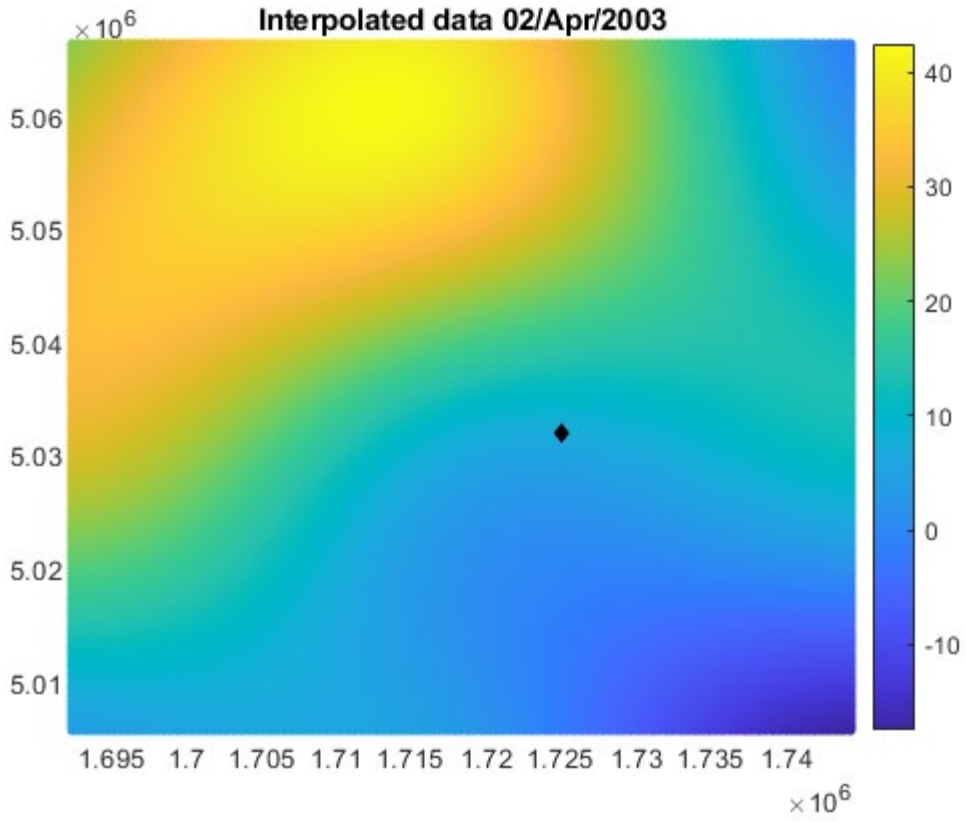
2D LS Cubic Spline (3rd order mother spline)

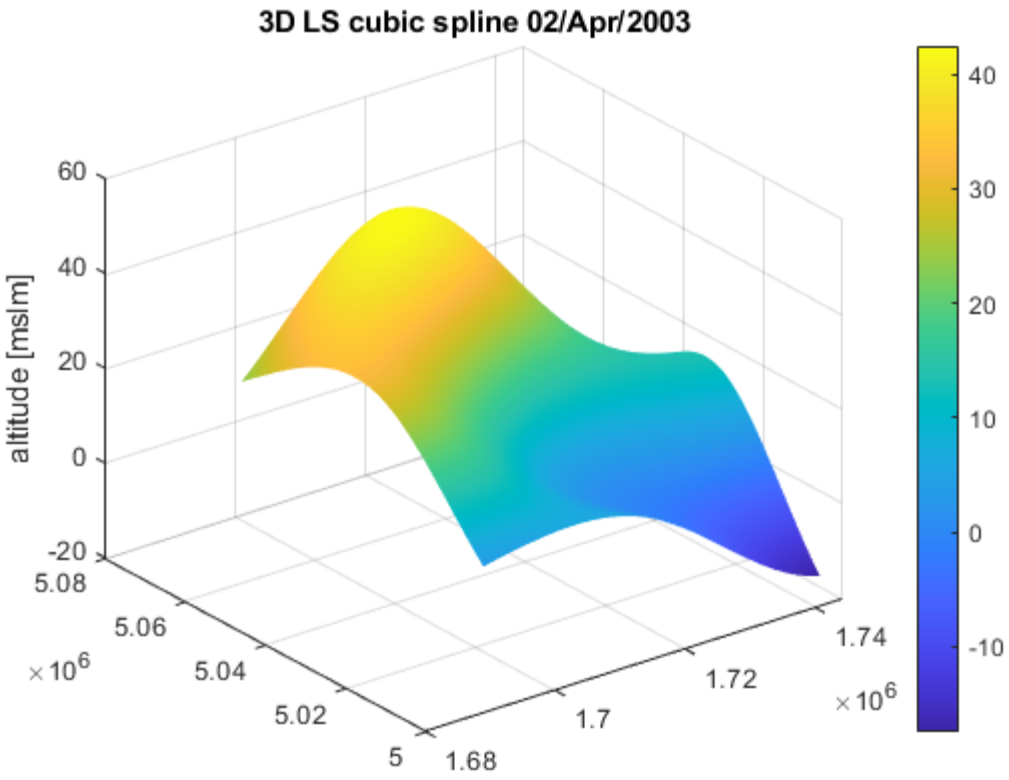
Exploring the data along time shows heterogeneous samples between borehols Exploring the data along space shows that the water levels measurements are uniform with the respective coordinates. Moreover, there isn't a particular evidence of changing in the shape of the levels over time.

If the minimum numbers of boreholds required to describe the surface is not satisfied may lead to twisted results

Anyway the reduction of boreholes seems to not affected the 2D interpolation until the number of samples are heavily decreased.







Brief conclusion

The developped script can be more accurate in some boreholes than other, due to the variety in the samples distribution between different time series. Considering clustering them could help Including more information about the data may be helpful, for example adding some layers could help find a stronger pattern on the behaviour.

A proposal for further analysis may be the implementation of Linearized LS on the hyperparameters and also the implementation of Kriging method. Furthermore, a comparison of the performance of different technique or different boreholes with evaluation metrics (eg R-squared, R-squared adj, AIC, BIC)

Thanks for the attention