Federico Cantarelli

Black-box Model for Seismic Activity

Kobe Earthquake, 1995

Project Work Data Driven Modelling of Dynamical System and Optimal Control A.Y. 2022/2023

Context

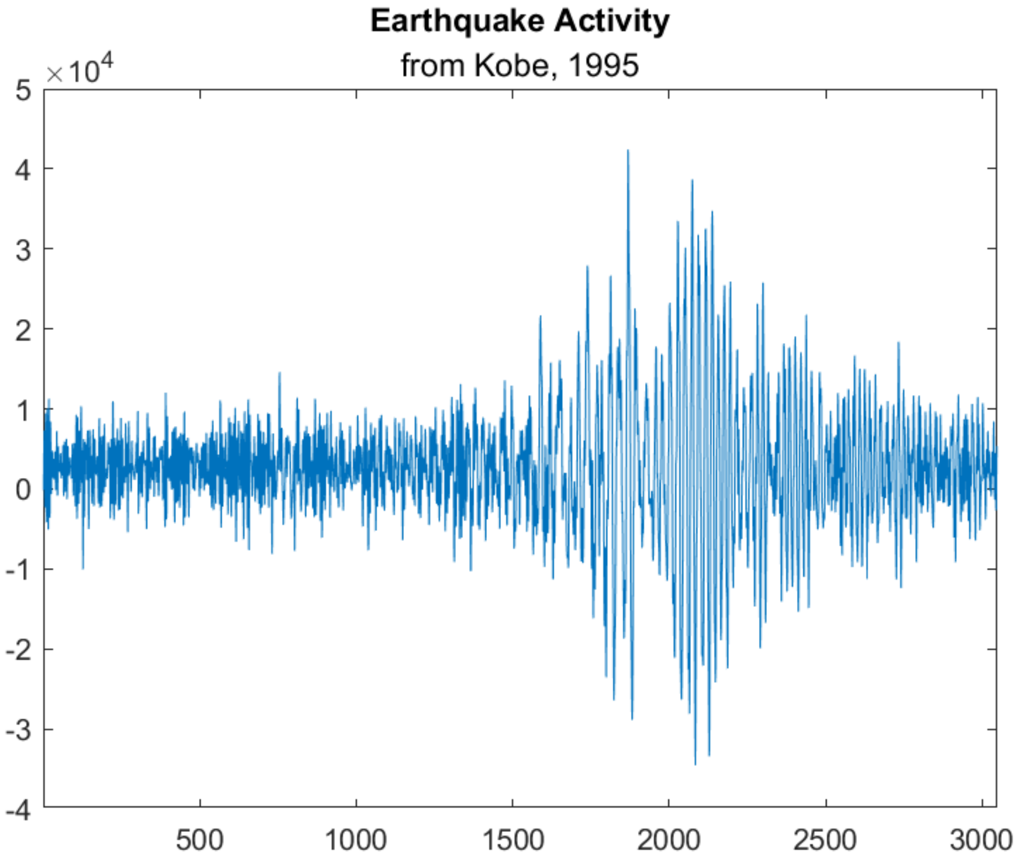
The dataset contains seismic activity from Kobe earthquake happened on 17th January 1995. It had a magnitude of 7.3 MW (USGS) and tremors lasted for about 20 seconds. There are 3048 total samples taken at discrete time periods.

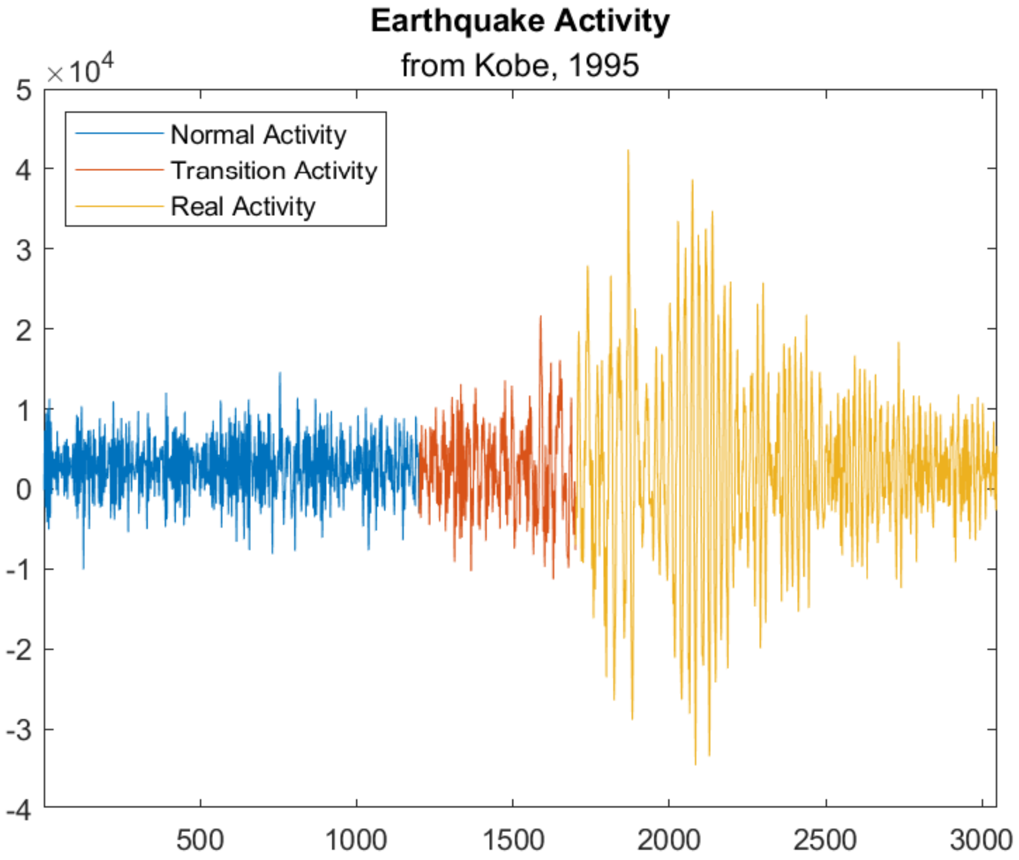
The **goal** of the project is to investigate whether it is possible, from the model built on normal activity data, to predict the earthquake measurements.

01 – Preliminary analysis

- Different activities
- Summary statistics
- Covariance plot and spectrum
- ACF and PACF
- Model Structure

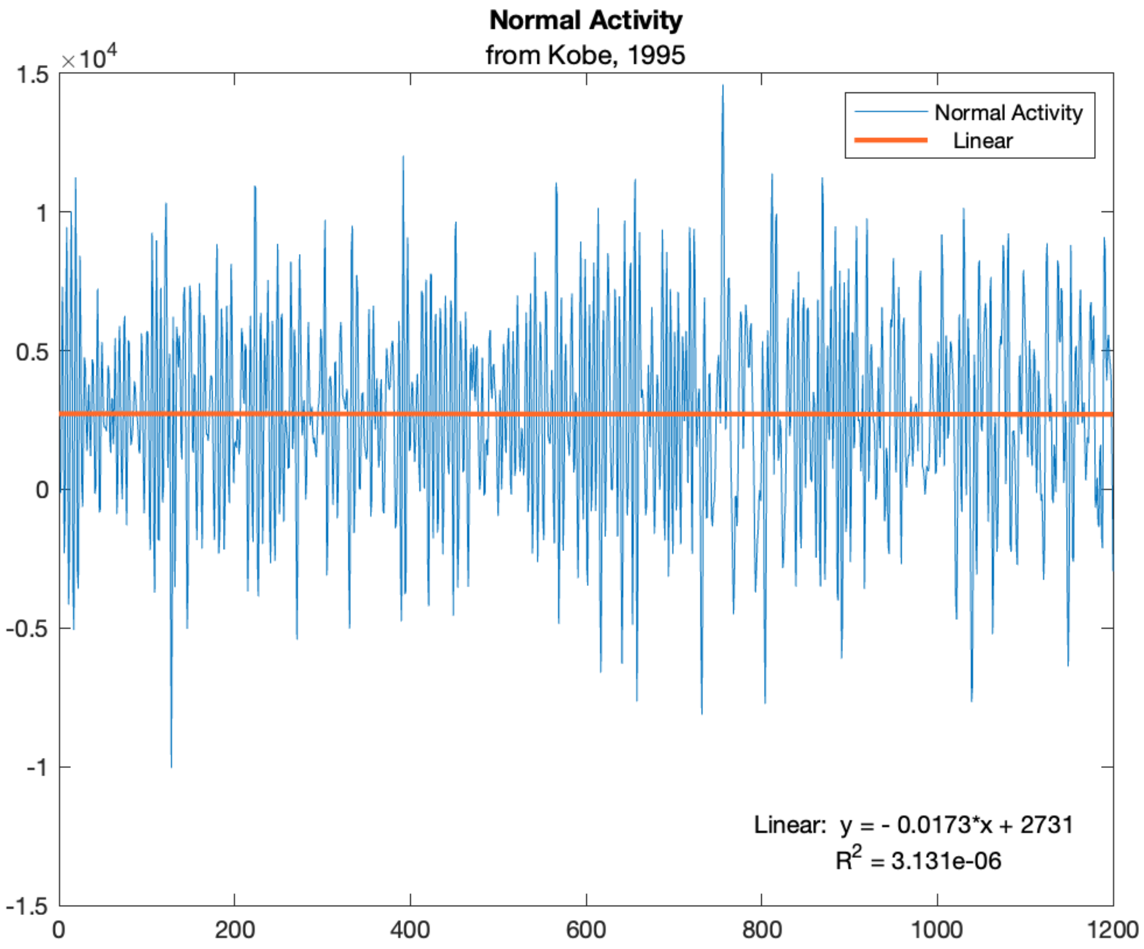
01 – Preliminary analysis





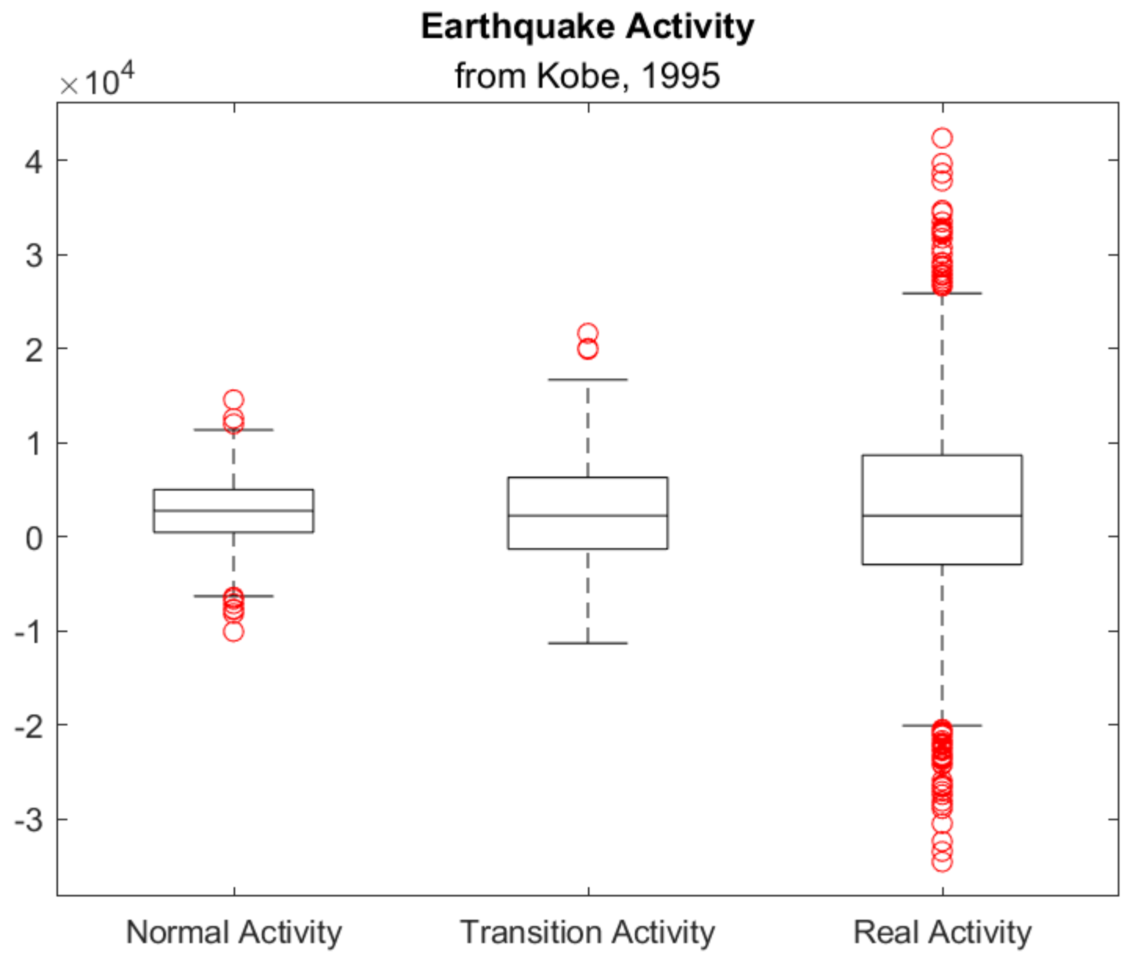
01 – Trend analysis

Time series **doesn't exhibit any trend**, it's **biased** though. Indeed, the mean of the normal activity data is about 2,700.



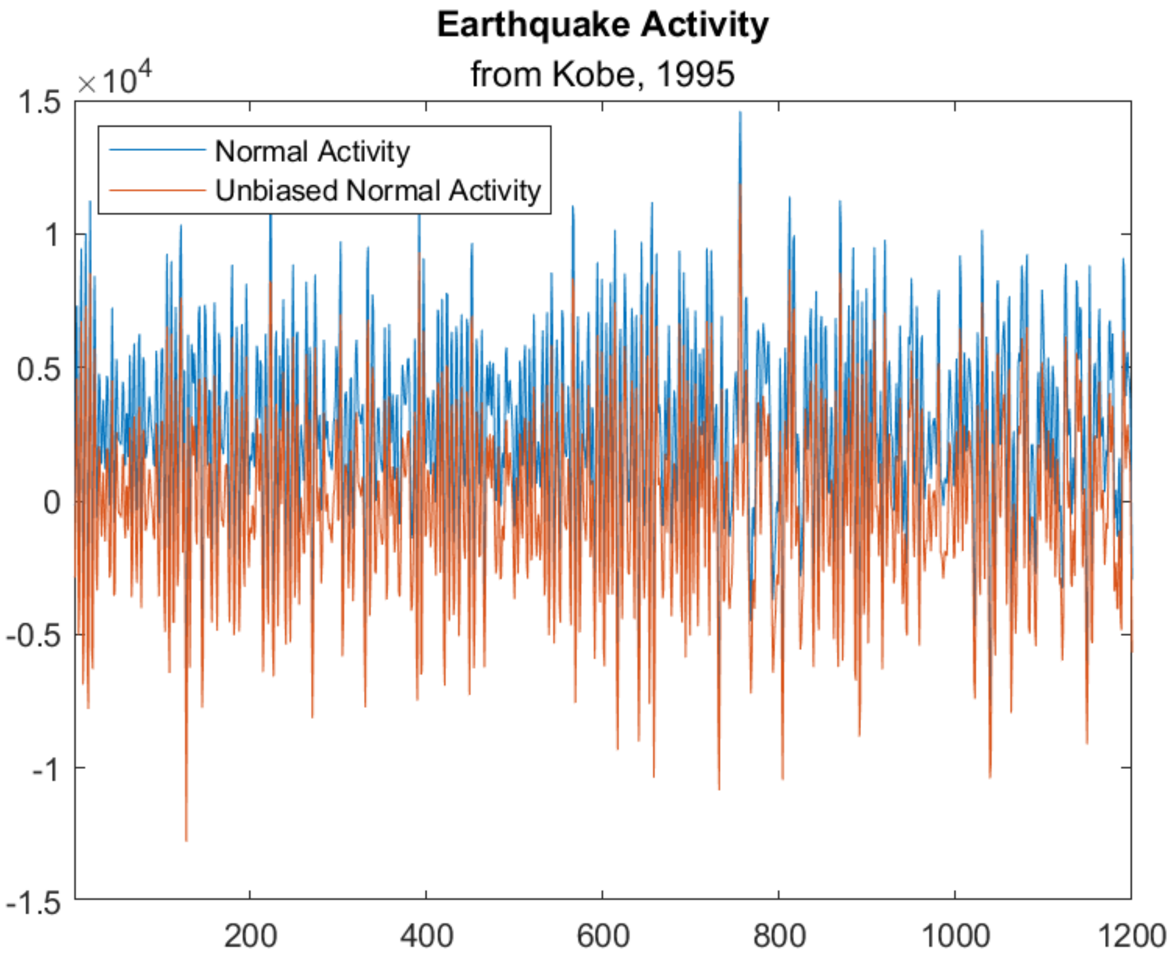
01 – Summary statistics

Activity	Mean	StDev	Range	Min	Median	Max
Normal	2,720.3	3,389.1	24,642	-10,045	27,890	14,597
Transition	2,503.9	5,488.6	32,955	-11,289	22,770	21,666
Real	2,631.6	10,613.58	76,950	-34,522	22,665	42,428



01 – Unbiased signal

As seen in the previous slides, normal activity is not zero-centred. So we can compute un **unbiased version** of the signal by removing the constant bias of the original signal.



01 – Preliminary analysis

Let’s search for a **statistical evidence** that **variance is different** among different phases of the earthquake.

01 – Anderson-Darling test

$H_0: Data\ are\ normal$

Activity	P-value	AD statistics	Result
Normal	0.3204	0.4243	Fail to reject H0
Transition	0.6615	0.2808	Fail to reject H0
Real	0.0005	6.5065	Reject H0

Since real activity observations are not from a normal distribution, we can use Levene's test, which is less sensible to the normality assumption of Bartlett's test, to asses if the variance is the same for all the different phases of the earthquake.

01 – Levene's test for variance

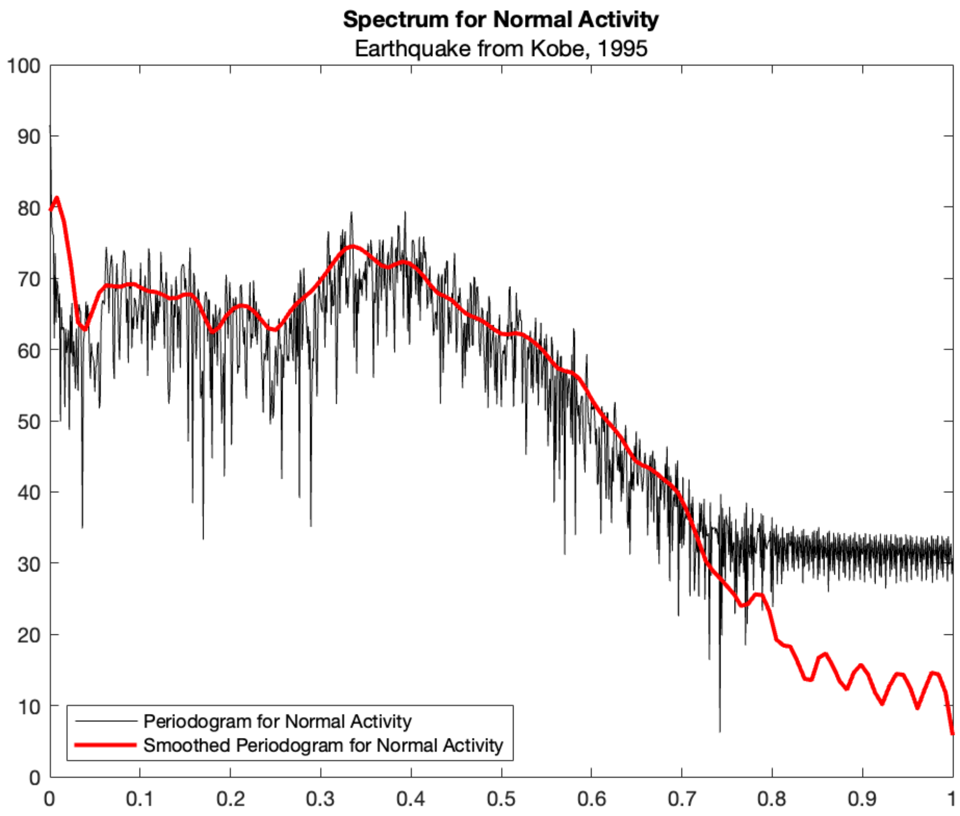
$H_0: \sigma_i = \sigma, \qquad \forall i, i = 1, 2, 3$

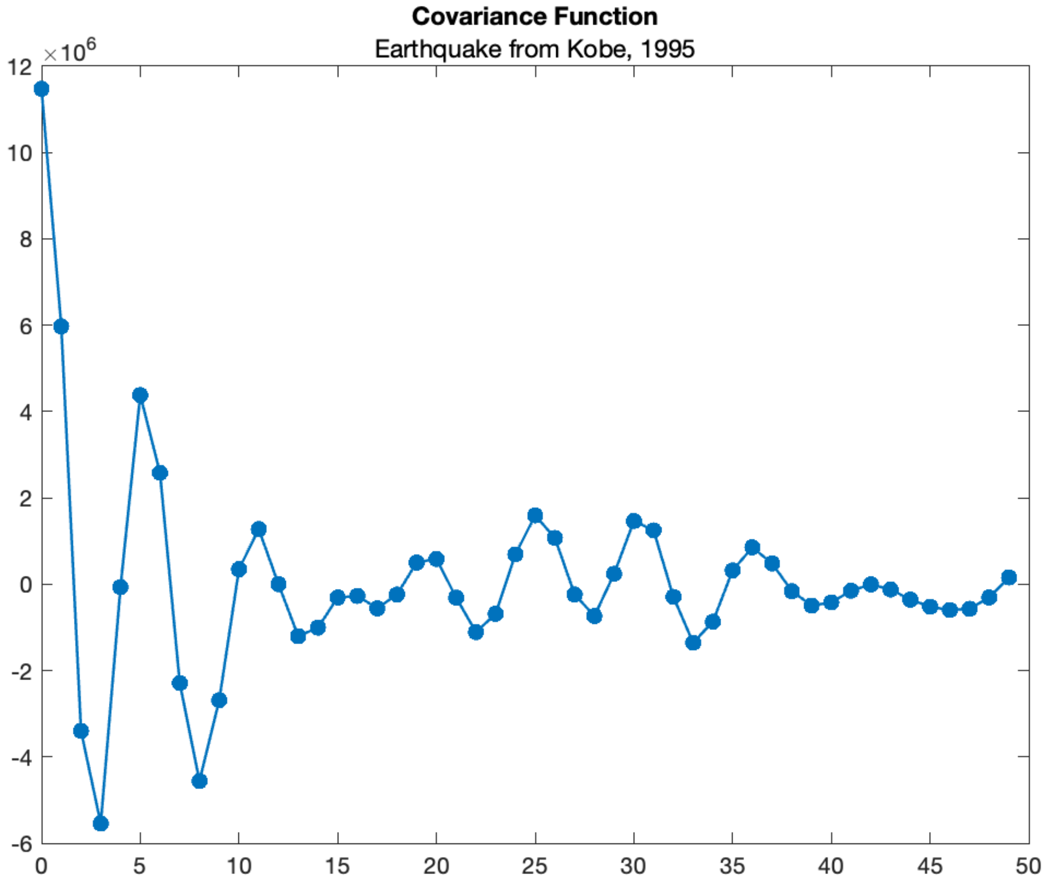
Group	Count	Mean	Std Dev
Normal Activity	1200	2720.34	3389.1
Transition Activity	500	2503.86	5488.6
Real Activity	1348	2631.65	10613.6
Pooled	3048	2645.61	7700.1

Levene's statistic (absolute)	350.078
Degrees of freedom	2, 3045
p-value	0

As we could have been expected from the graphical interpretation of the time series, the variance is different considering earthquake phases.

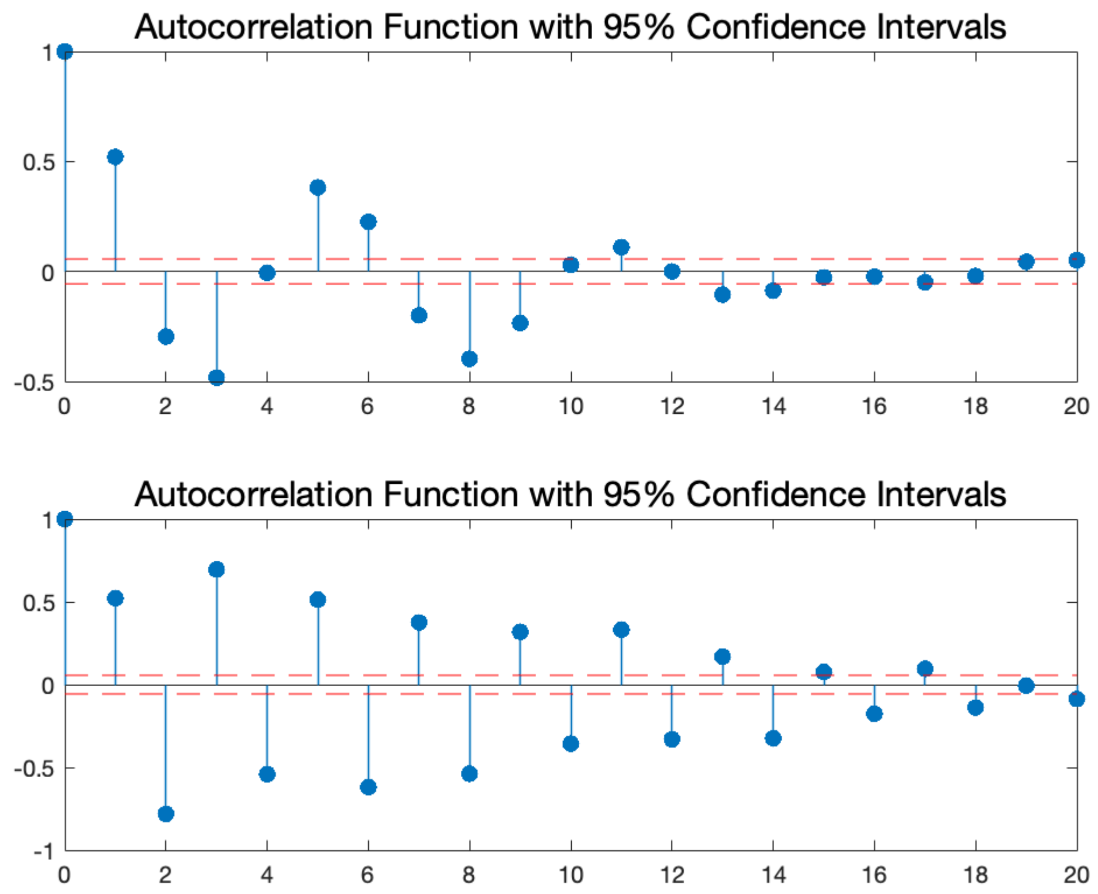
01 – Spectrum and covariance function





01 – ACF & PACF plots

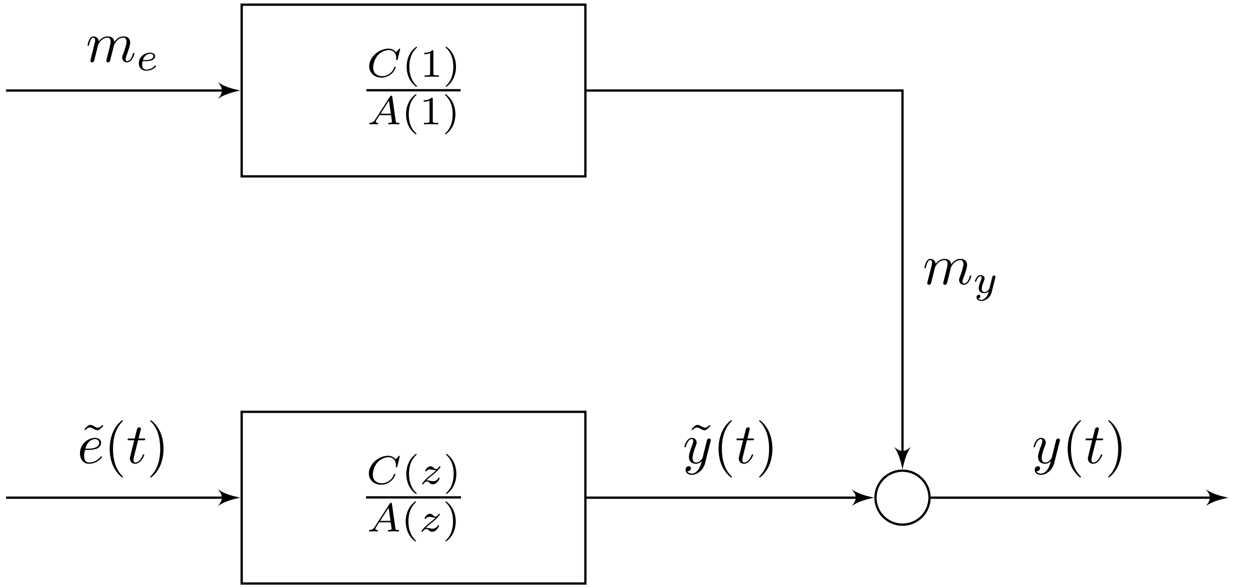
Both the autocorrelation function and partial autocorrelation plots exhibit some criticalities. In particular we can see high values for lag up to 18. The oscillatory behaviour of the PAC function suggest us to search also for an MA part of the model. So we can expect to find a high order ARMA model.



01 – Model structure

Where m_y is process bias, $e(t) \sim WN(m_e, \lambda^2)$

and $\tilde{e}(t)$ is the unbiased version of $e(t)$.



02 – Model Training

- Train & validation split
- Model training
- Model selection
- Analysis of model residuals

02 – Train & validation split

1200

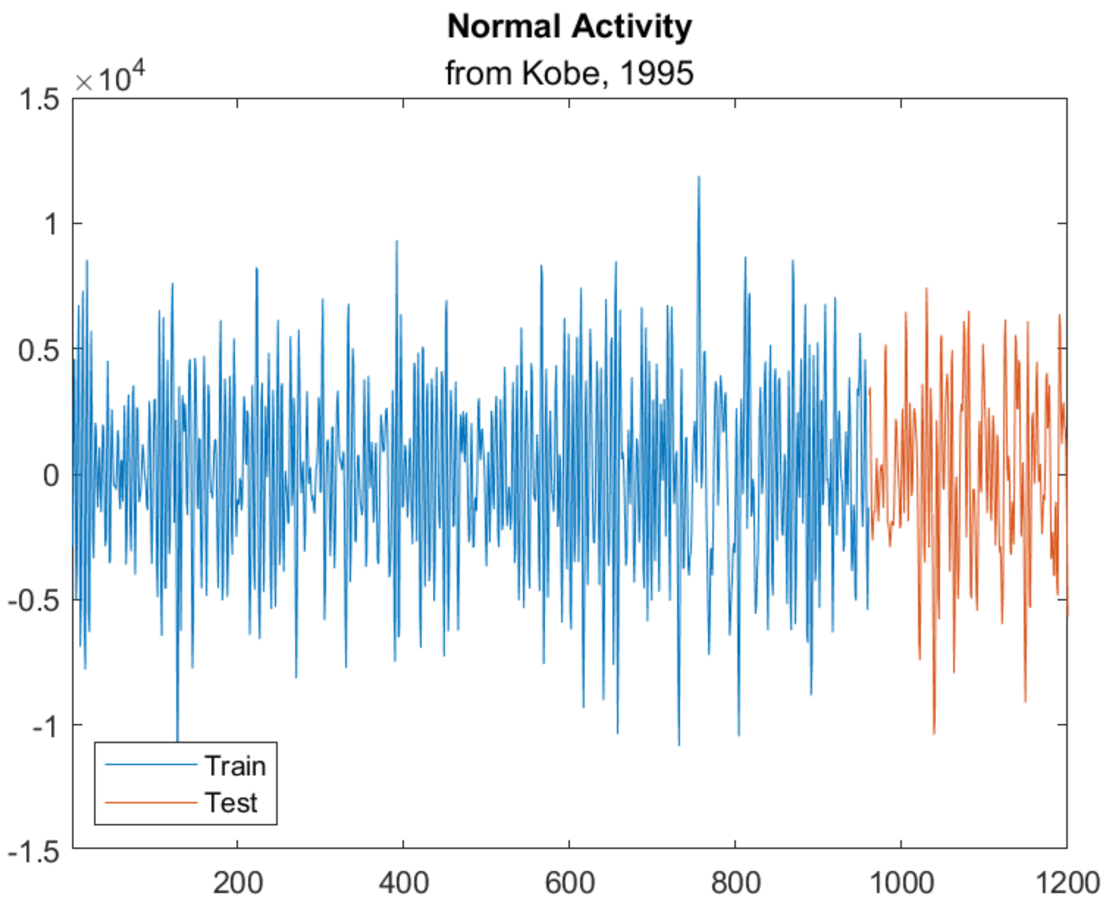
Samples for Normal Activity

960

Training Data

240

Validation Data



02 – Model training

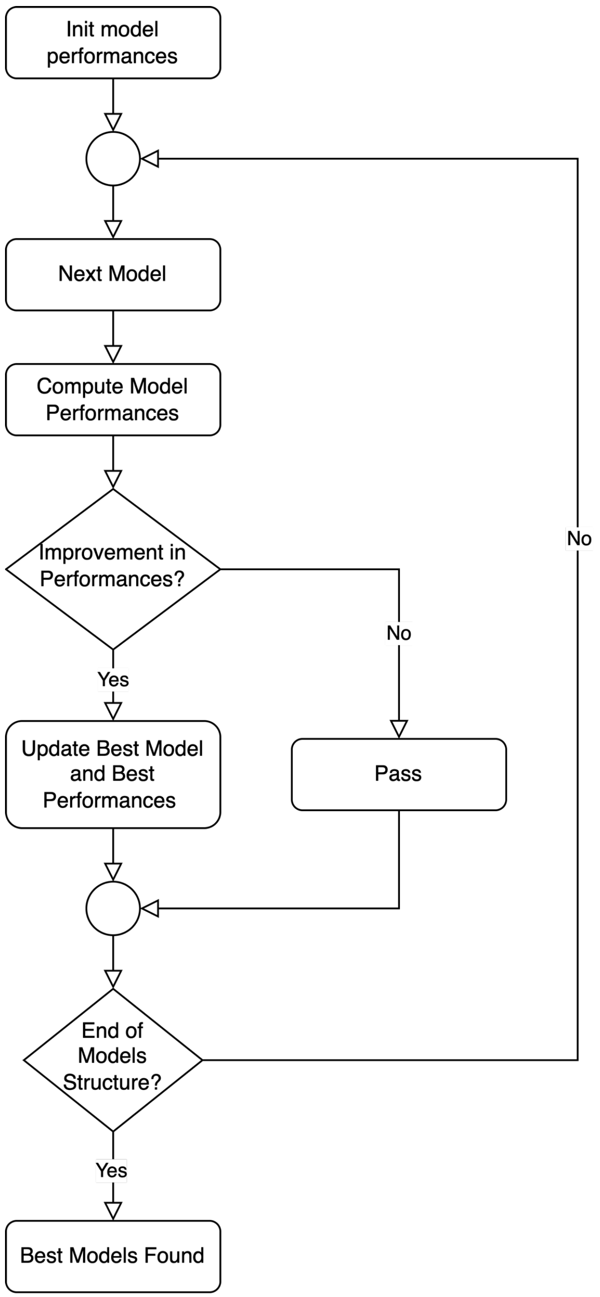
To find the best model, I used an **iterative approach** to train models with different order combinations, both for AR and MA parts. Then I find the model that fitted the best training data.

```
%% Estimate models
na = 1:40;
nc = 1:40;
ct = 1;
models = cell(1,1600);
for i = 1:40
    na_ = na(i);
    for j = 1:40
        nc_ = nc(j);
        models{ct} = armax(my_data_train,[na_ nc_]);
        ct = ct+1;
    end
end
```

02 – Model selection

In the flowchart on the right, there is the description of the algorithm I used to find the best model for each criterion to assess model performances:

- Mean Squared Error
- Root Mean Squared Error
- Akaike Information Criterion
- Minimum Description Lenght
- Mean Absolute Error
- Final Prediction Error
- Mean Absolute Percentage Error
- Percentage of Fitting



02 – Model selection

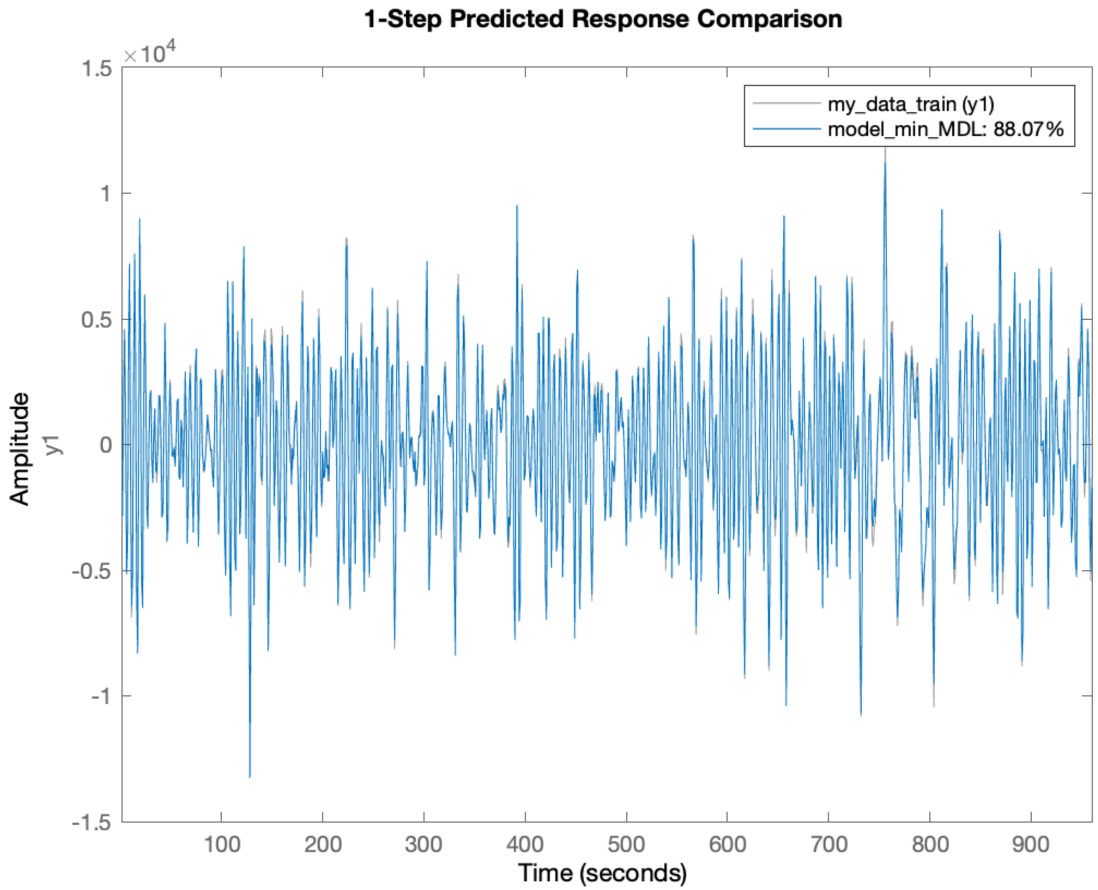
I found best models according to three different criteria:

- By minimizing AIC
- By minimizing MSE
- By minimizing MDL

As the final criterion to choose the optimal fitted model I chose **MDL**, since it is a good trade-off between **model complexity** and **model performances**. In doing so, I tried to find a simpler model and to prevent overfitting. In particular, MDL is defined as

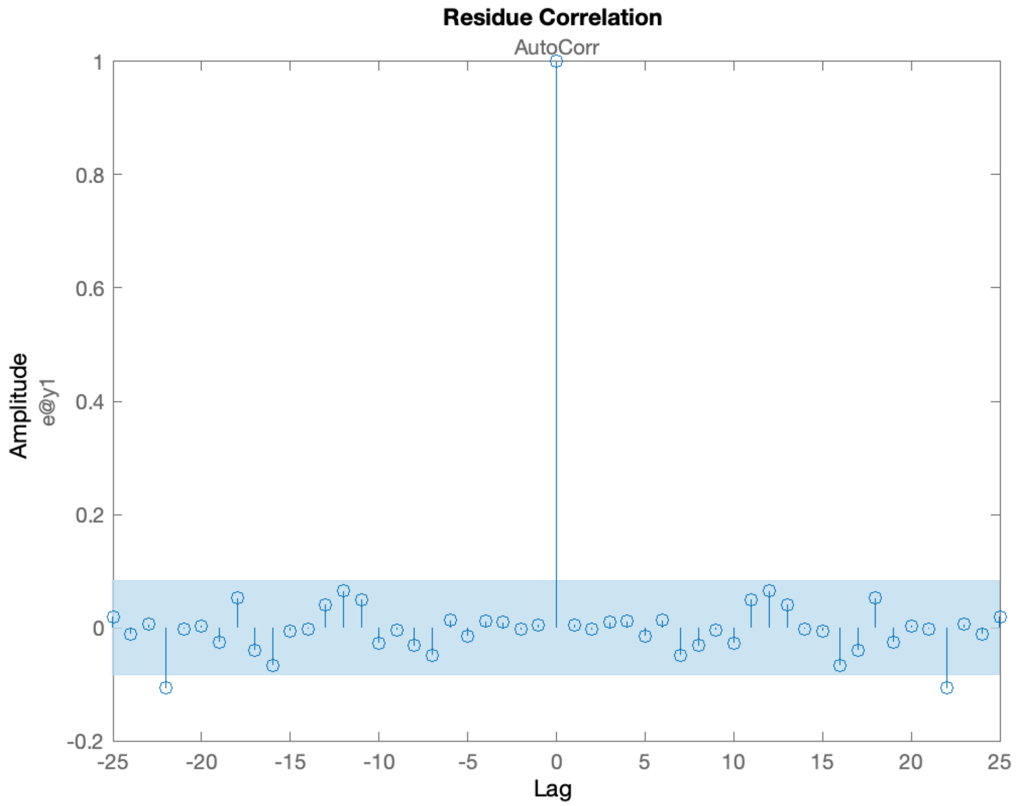
$$MDL = \ln N \cdot \frac{n}{N} + \ln\left(\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2\right)$$

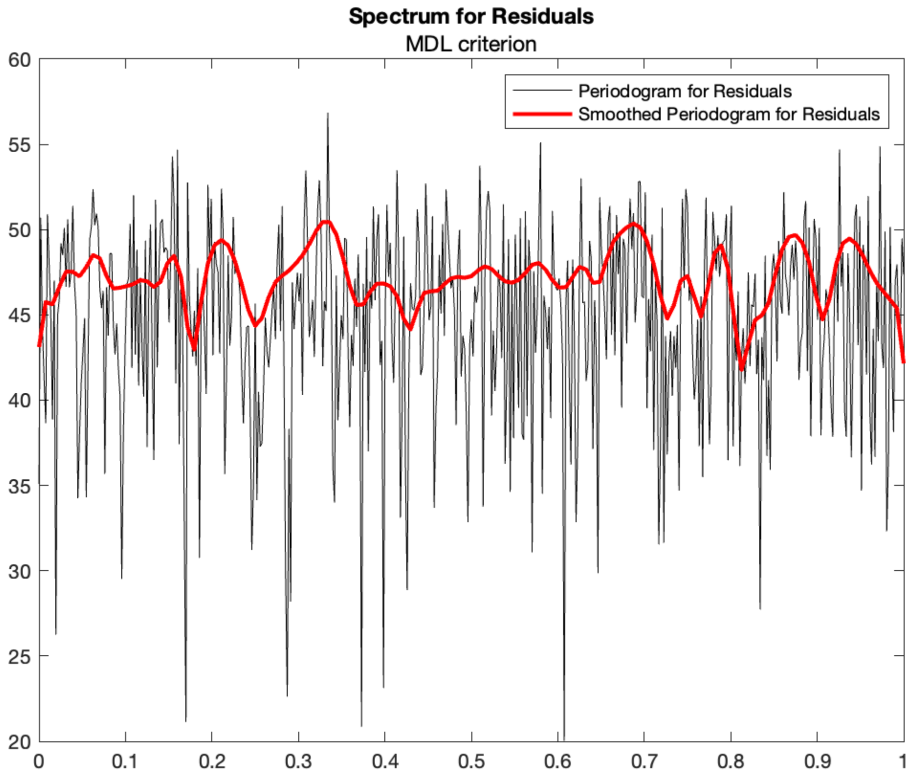
By choosing MDL instead of AIC, I was able to reduce model complexity from an ARMA(39, 37) to an ARMA(9,4) while maintaining a comparable percentage of fitting.



Model	AR order	MA order	n	Fit percentage
model_min_AIC	39	37	76	89.34%
model_min_MSE	39	37	76	89.34%
model_min_MDL	9	4	13	88.07%

02 – Model residuals





02 – Model residuals

Residuals sample mean is 4*.5944,* but let's perform a t-test to assess if there is statistical evidence to reject the null hypothesis of residuals are from a normal distribution from a normal distribution with $\mu = 0$ and with unknow variance.

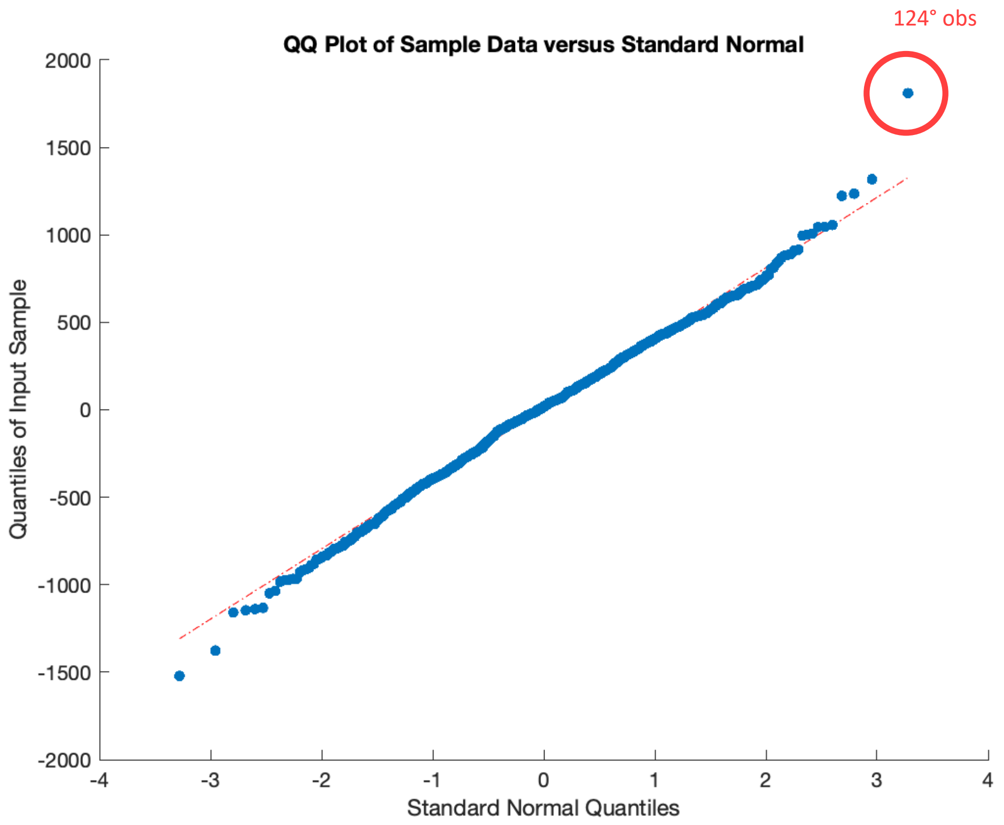
Test **fails to reject** $\boldsymbol{H_0}$ with a really high p-value, so we cannot exclude that residuals are distributed as a white noise

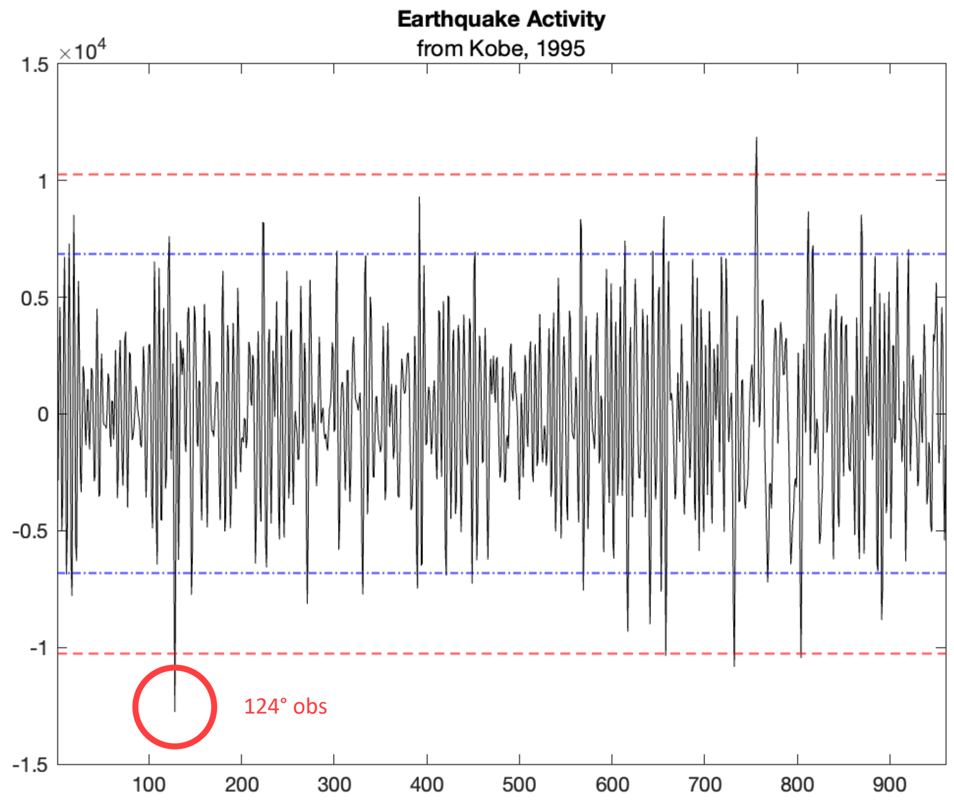
$$resid \sim WN(0, \lambda^2)$$

With an estimated $\widehat{\lambda^2} = 1.6639\text{e}{+}05$.

```
Result of t-test:
h =

     0


p =

    0.7272


ci =

  -21.2417

   30.4304
```

02 – Model residuals





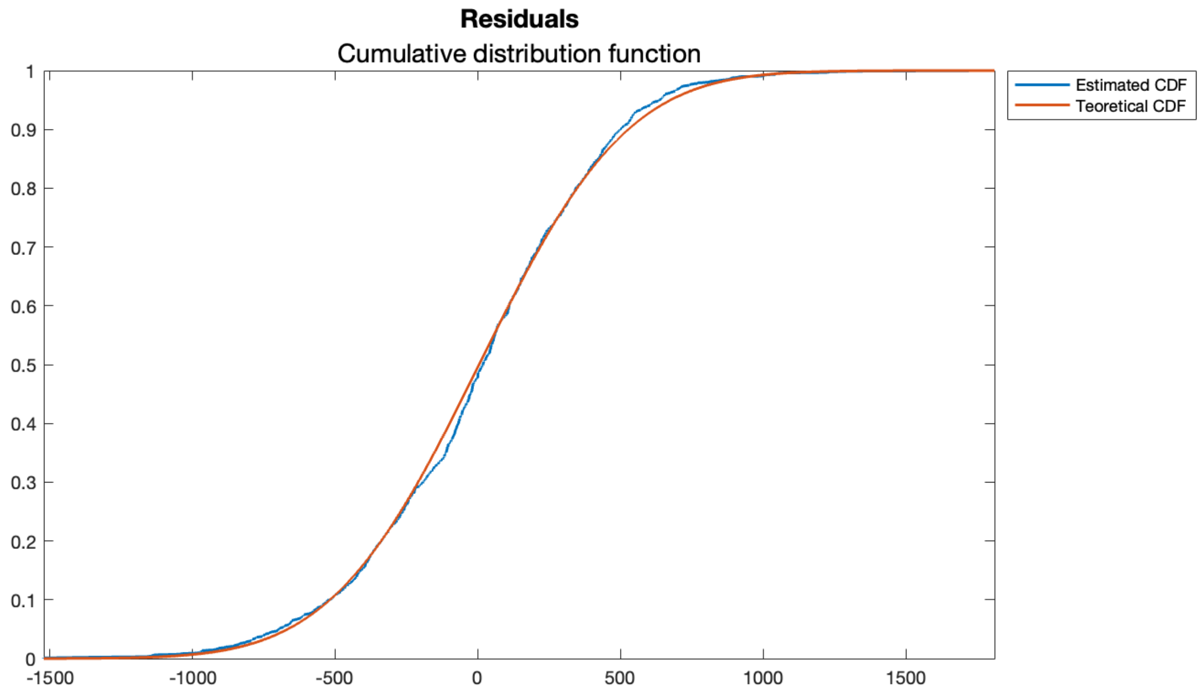
02 – Model residuals

Remember, in our case, `N = 960` which is a **pretty low sample size.**

Critical observations: `[124 121 152 948 130 145 802 145 176 125 745 384 226]`

```
Result of ad-test:

h =

     1


p =

    0.0131
```

```
Result of ad-test without
critical points:

h =

     1


p =

    0.0494
```



03 – Model Validation

- Model validation
- Model performances
- Final model

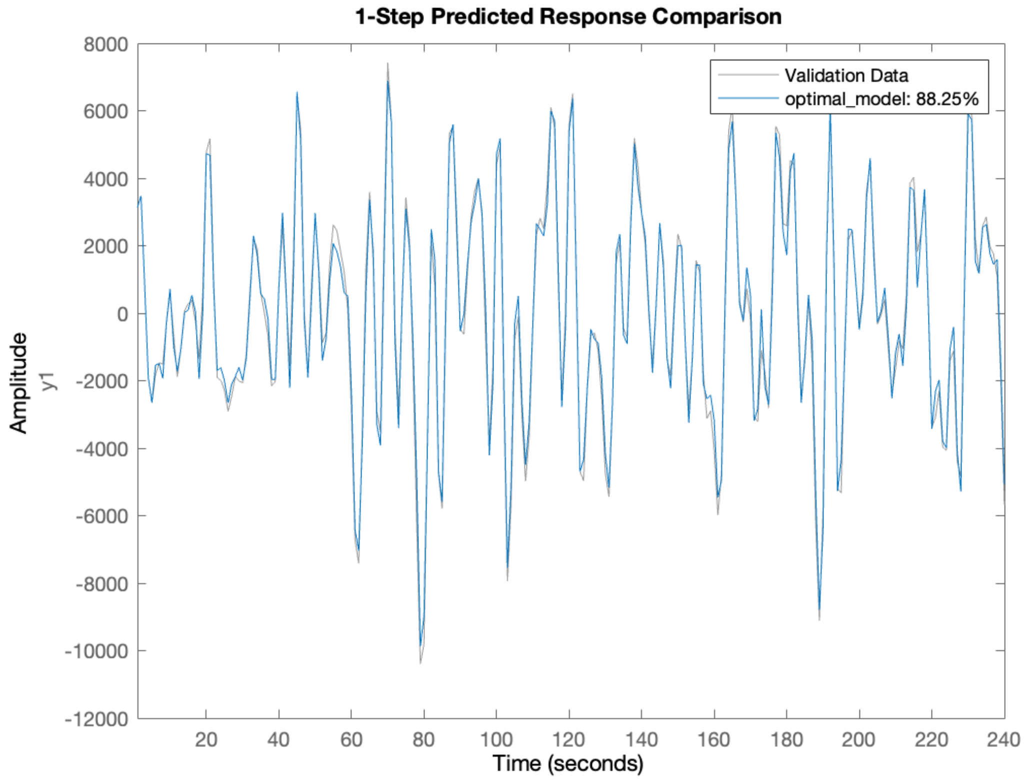
03 – Bias-variance trade-off

We are interested in **minimizing model error** in predicting new observations, so we have to asses the model against the **validation set**. We need to check if the model we find is overfitting our data.

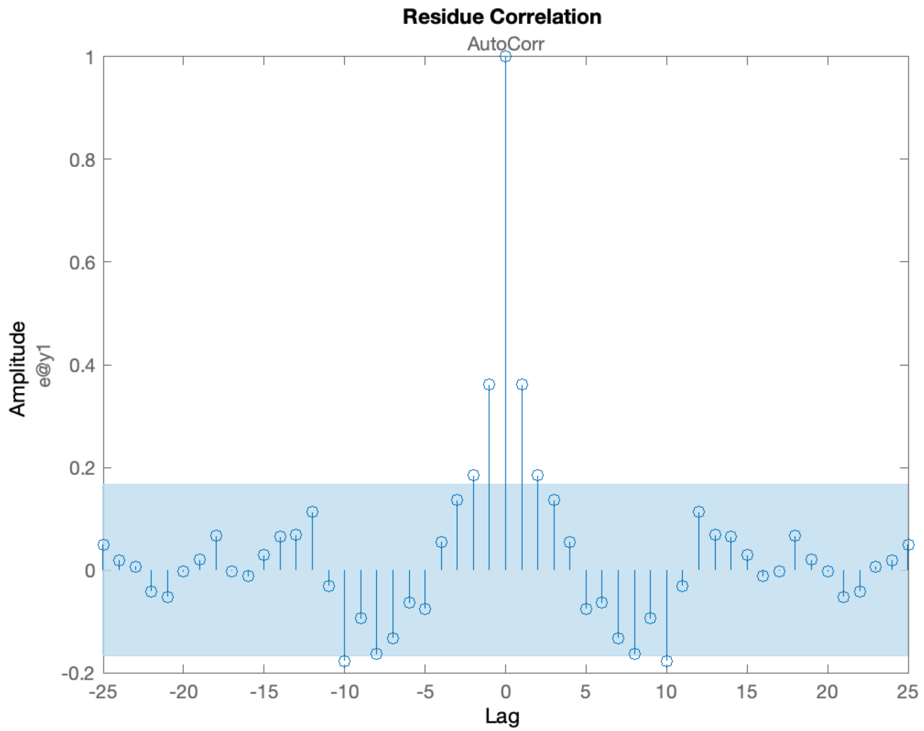
03 – Model performances

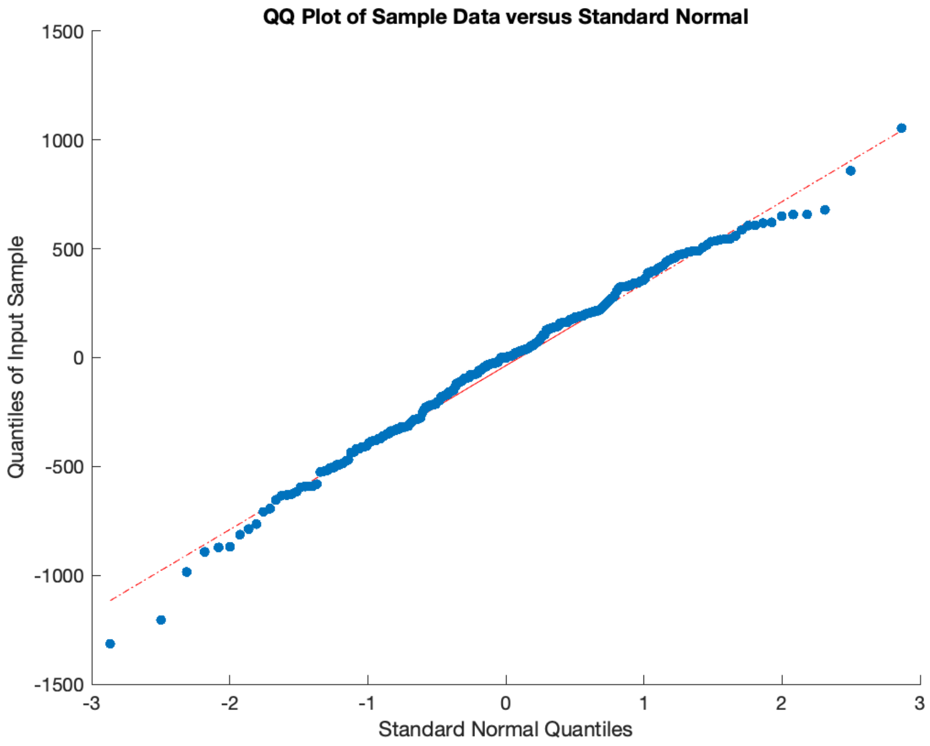
To check If the model I found previously, I decided to assess model performances on the validation set. and model performances, I decided to use MDL criterion. As we can see from the table, performances on both training and validation set of the model are comparable. The model is not overfitting.

Performances	Training Data	Validation Data
MSE	1.6624e+05	1.4725e+05
RMSE	407.7244	383.7377
MAE	318.7848	302.7615
MAPE	1.7889	15.2884
AIC	24.0694	23.9082
MDL	12.1142	12.1968
Fit percentage	88.07%	88.25%



03 – Validation set residuals






```
Result of AD-test:

h =

     0


p =

    0.2665
```

03 – Finally… the model

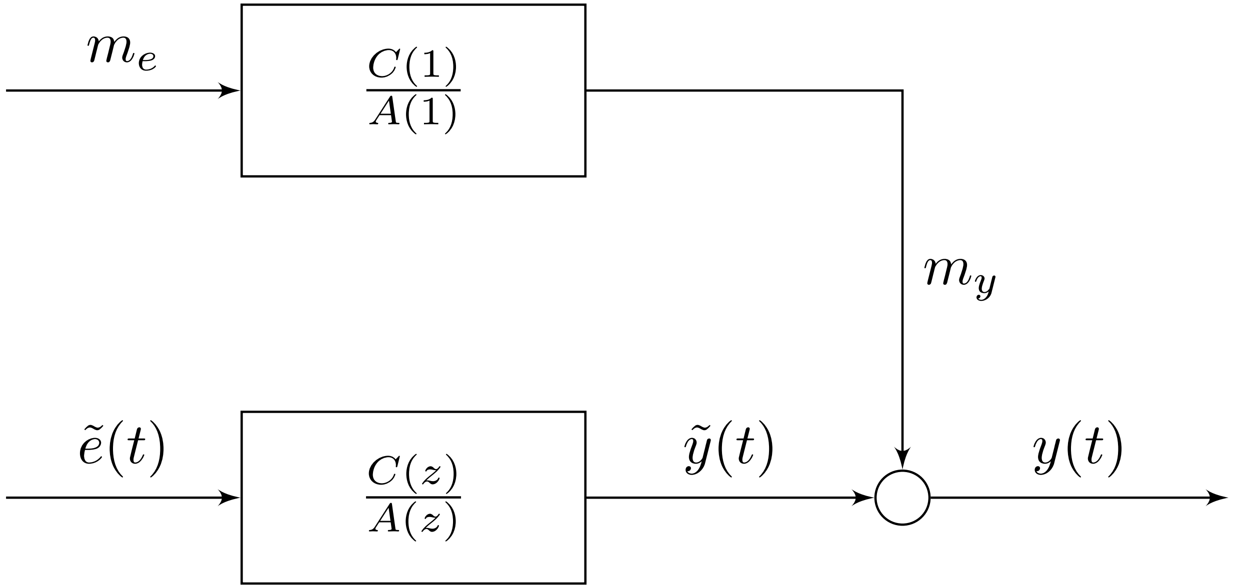
The optimal model is an **ARMA(9, 4)**.

$$W(z) = \frac{1 + 1.791\, z^{-1} - 0.0845\, z^{-2} - 1.744\, z^{-3} - 0.8635\, z^{-4}}{1 - 1.718\, z^{-1} + 1.805\, z^{-2} - 1.215\, z^{-3} + 0.03484\, z^{-4} + 0.6637\, z^{-5} - 0.9679\, z^{-6} + 0.8373\, z^{-7} - 0.3713\, z^{-8} + 0.1345 z^{-9}}$$

03 – Finally… the model

Where $e(t) \sim WN(5.6133{\times}10^3, 1.1486{\times}10^7)$

and $\tilde{e}(t)$ is the unbiased version of $e(t)$.



04 – Model performances

- Model stationarity
- Error to signal ratio vs prediction horizon
- FPE vs prediction horizon

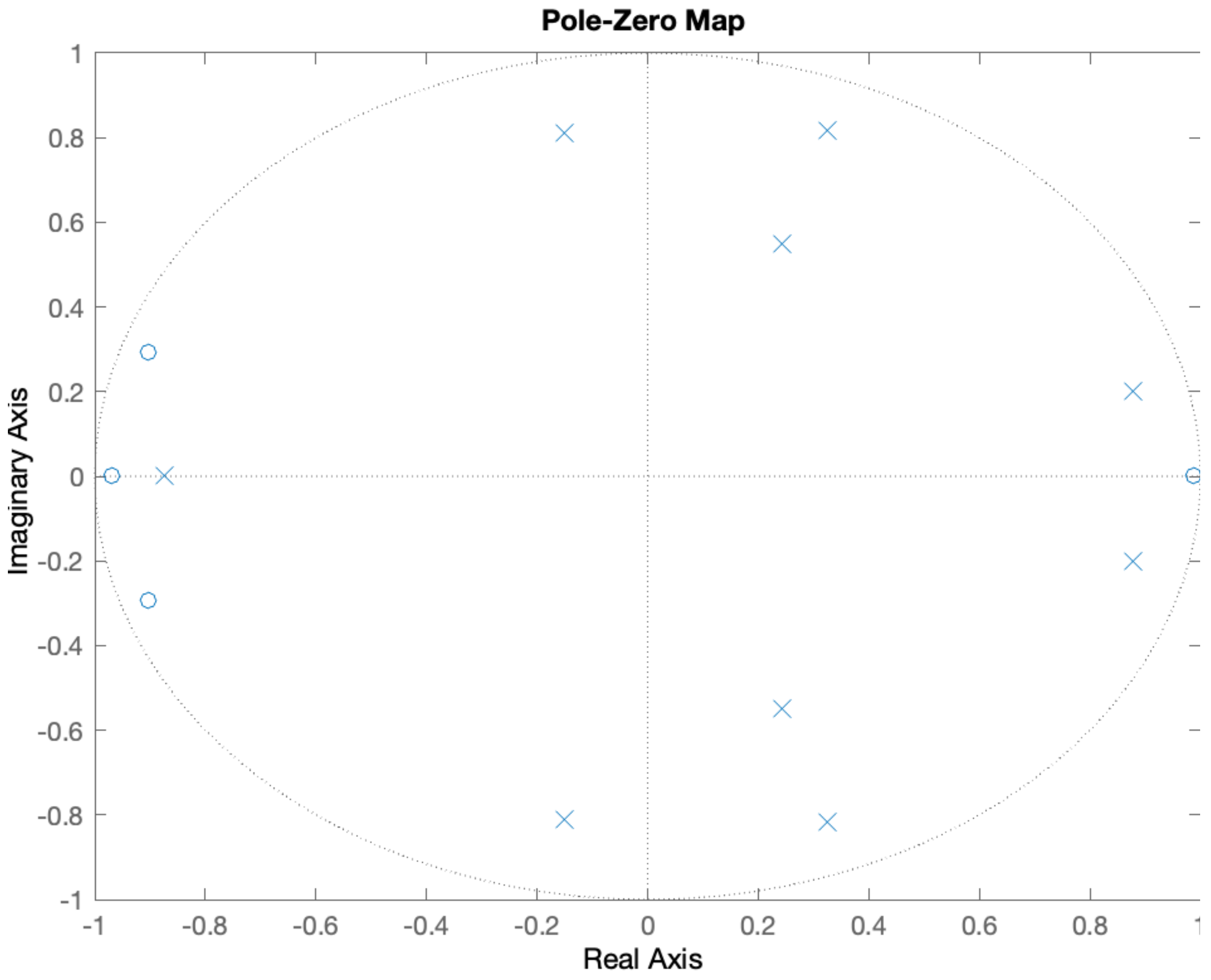
04 – Model stationarity

An ARMA model is stationary if:

- The input is a SSP
- The transfer function is asymptotically stable

The input is a SSP by definition and the transfer function is asymptotically stable since all poles are inside the unitary radius circle.

The model is **stationary**.



04 – Error to signal ratio

1-step-ahead predictor

`Training: 1.42%`

`Validation: 1.38%`

2-step-ahead predictor

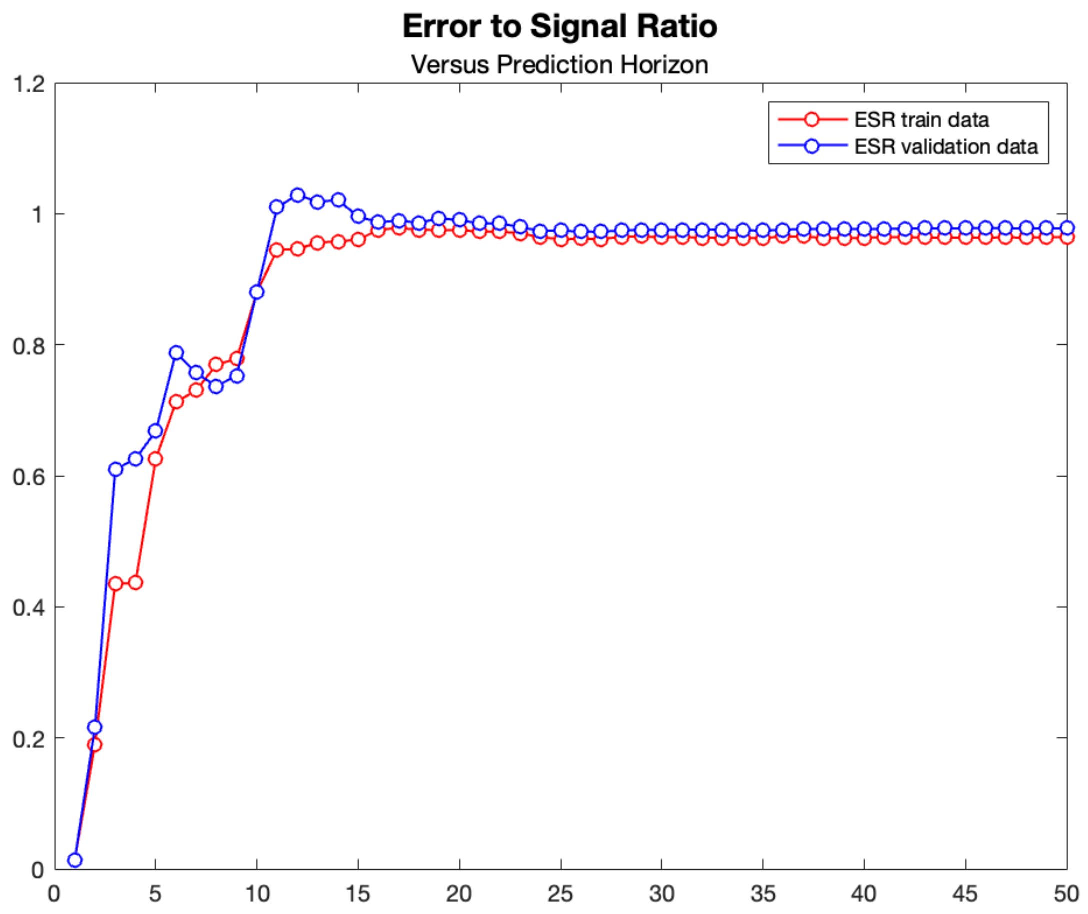
`Training: 18.98%`

`Validation: 21.62%`

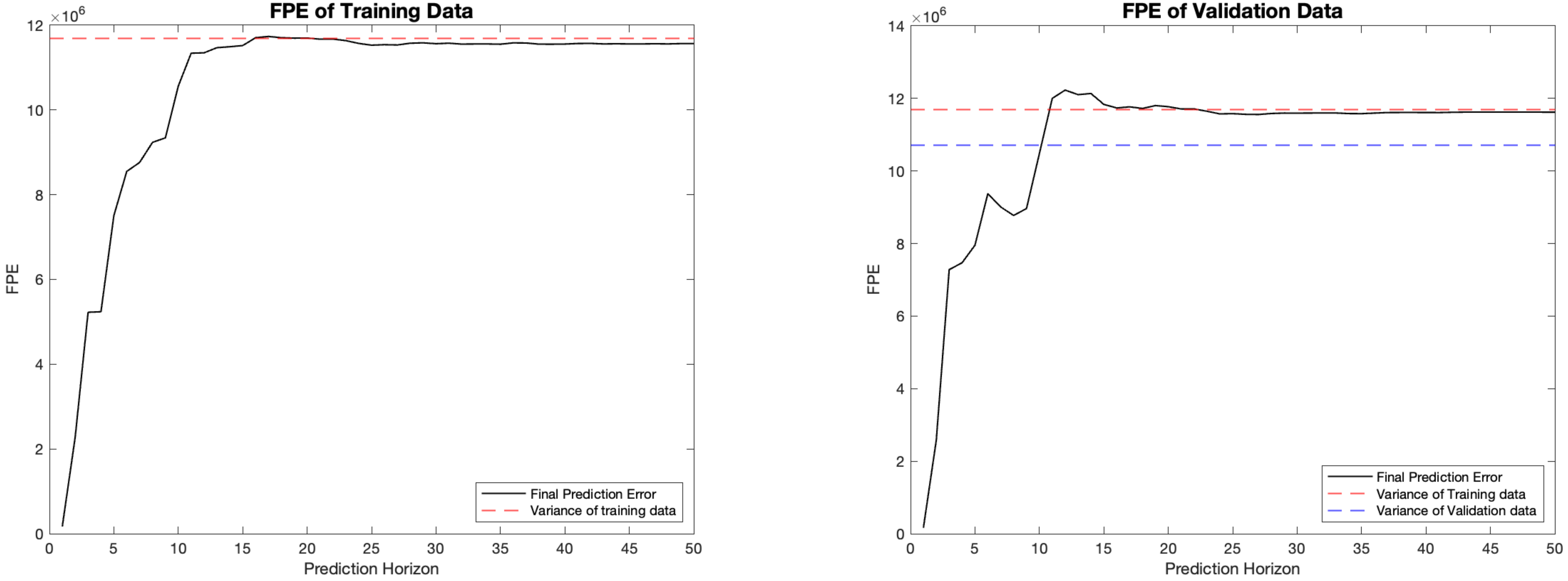
3-step-ahead predictor

`Training: 43.53%`

`Validation: 60.98%`



04 – FPE versus prediction horizon to signal ratio



05 – Can we predict an earthquake?

05 – The idea

I found a model which describes the dynamics of the normal activity. If the **error in prediction increases**, the dynamics of the seismic activity is changed. I chose the MSE as a metric to asses model performances.

05 – Why the mean squared error?

1. Since the nature of the process is oscillatory, I want to prevent error decreasing due to changing in errors sign. The MSE (by nature) avoids this to happening.

2. I want to «amplify» larger error. By using MSE the error in prediction will follow an exponential behaviour, making it more easily identifiable.

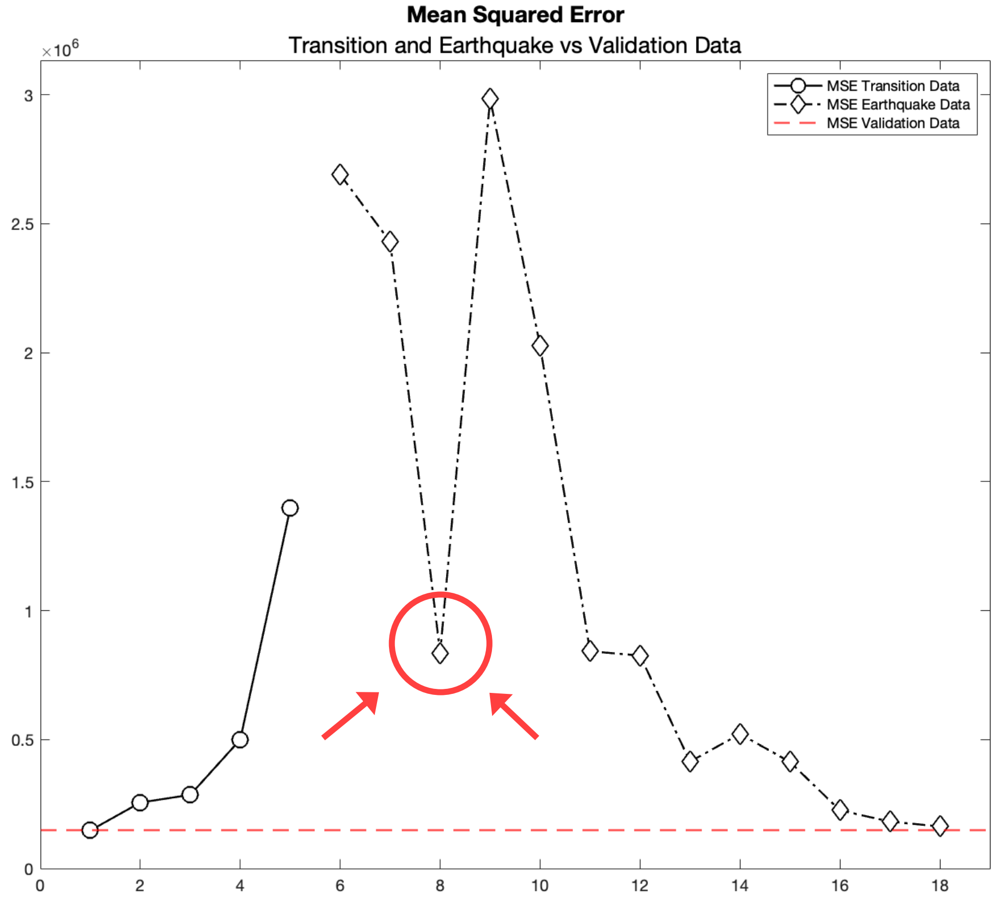
05 – MSE on transition and earthquake activities

-400 sec
to earthquake

The MSE starts increasing

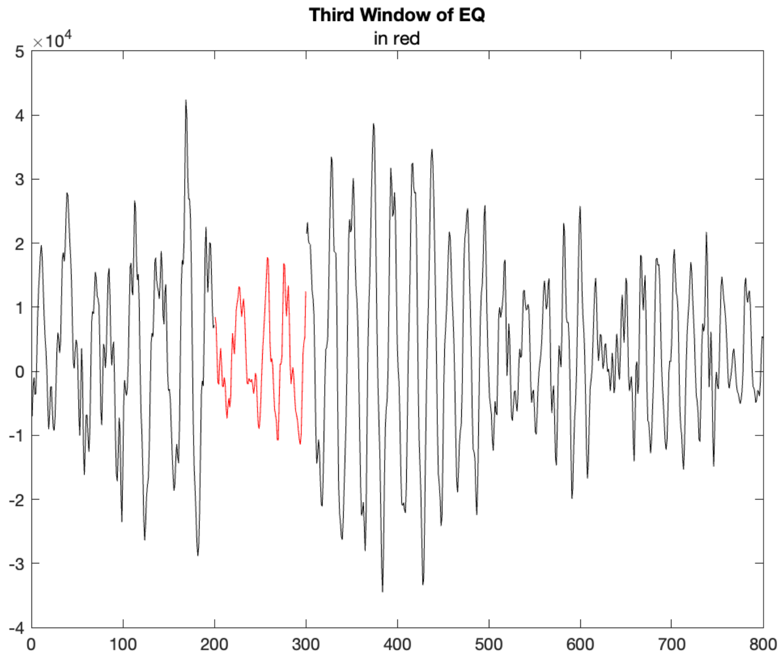
-200 sec
to earthquake

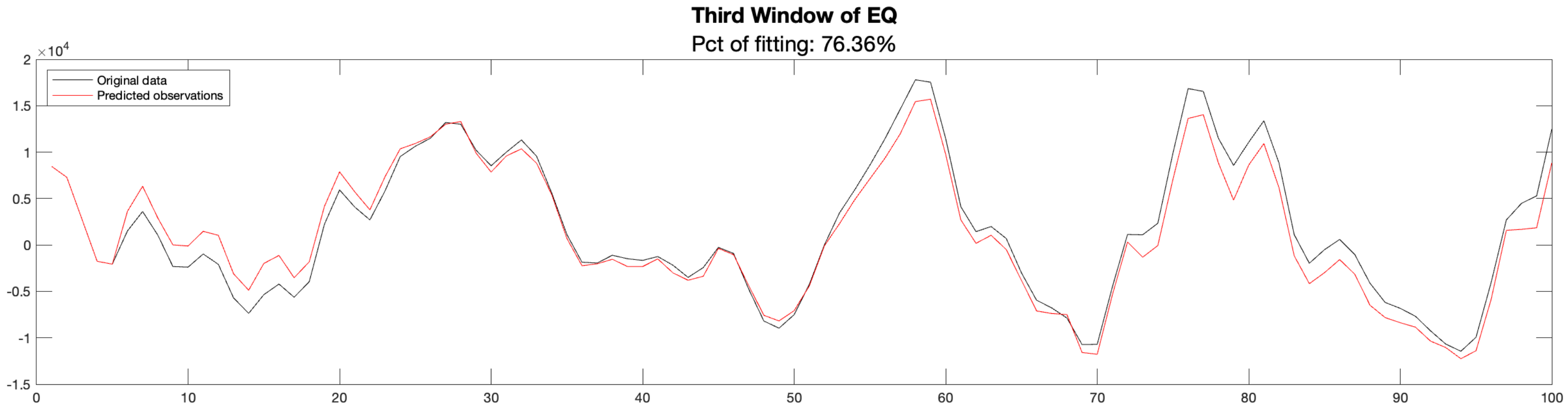
The MSE is sharply increased

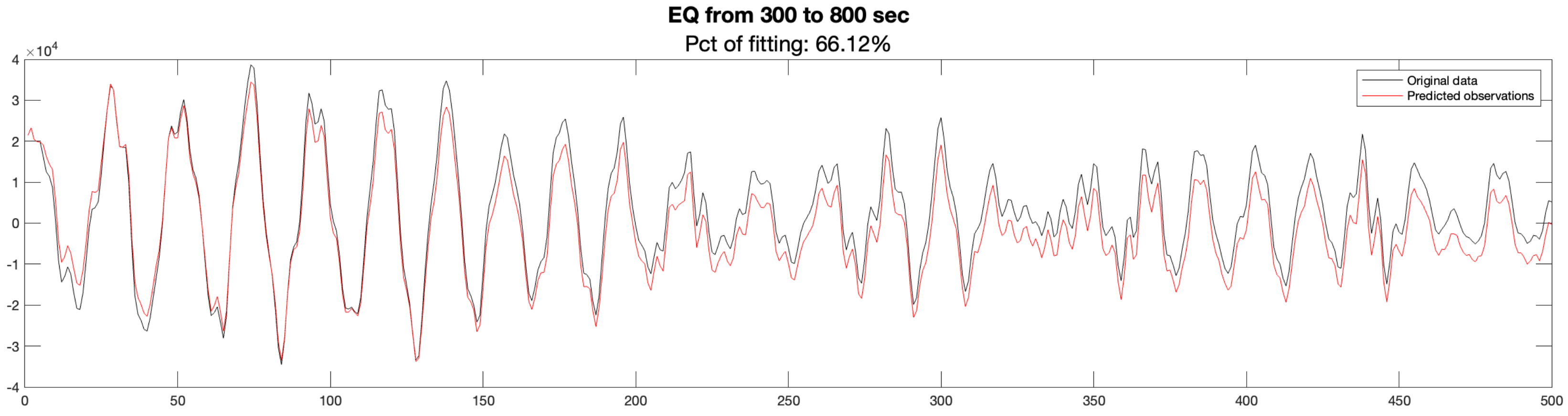


Each time-window is 100 seconds

05 – Third time window of EQ







06 – A control chart approach

- Data normality and false alarm rate
- I-MR control chart
- MA control chart
- EWMA control chart

05 – A control chart approach

Control charts are a statistical tools to check if a **process is under control**.

The idea is to fit a control chart for individual observation (aka I-MR chart) on residuals of the optimal model found in previous steps.

The stages of control chart design

1. Check data normality
2. Define false alarm rate α
3. Find control line
4. Test the chart
5. Update the chart (if needed)

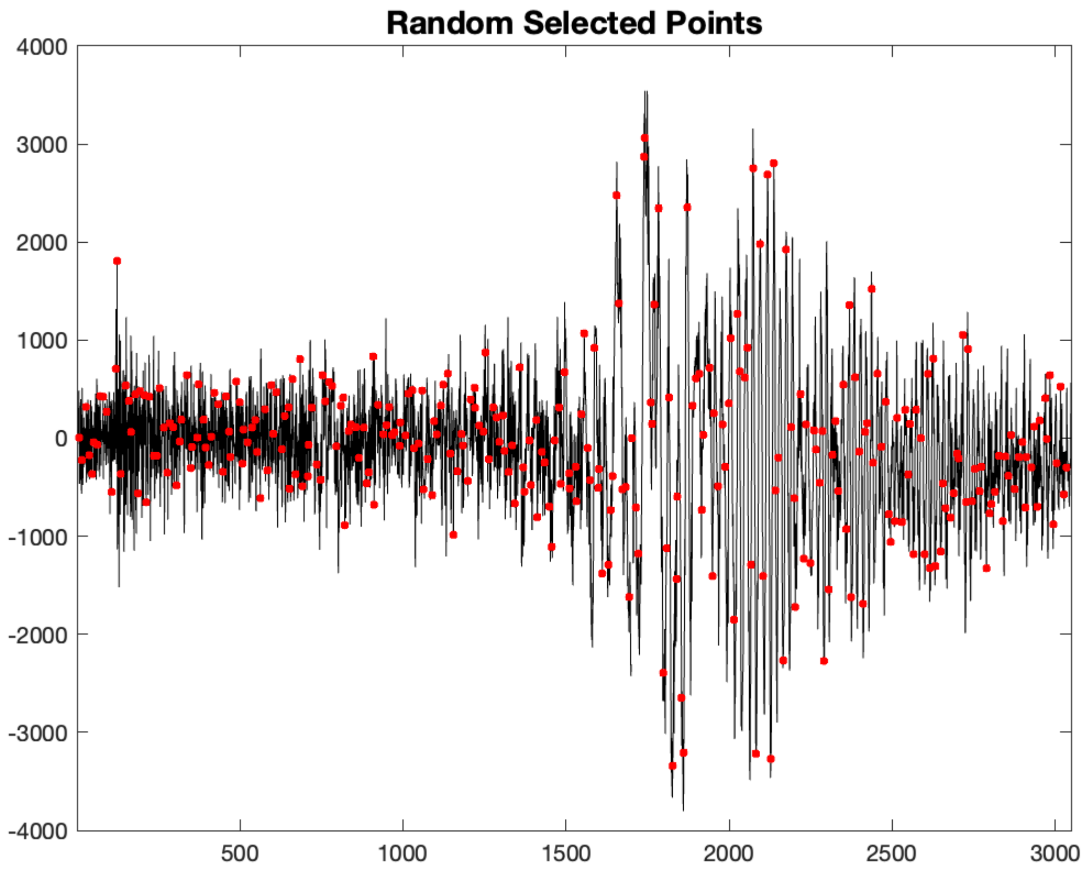
→ **Real time monitoring**

05 – Data normality and false alarm rate

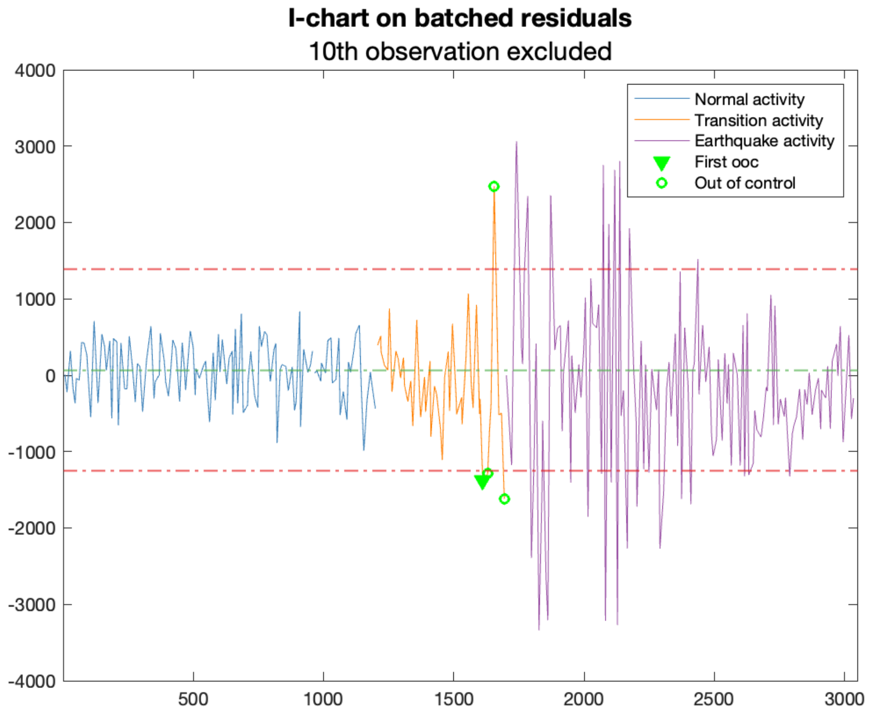
Since there is one measurement per second, I decided to random sample one measurement every ten seconds. By doing so i obtained an AD test $p-value = 0.3626$.

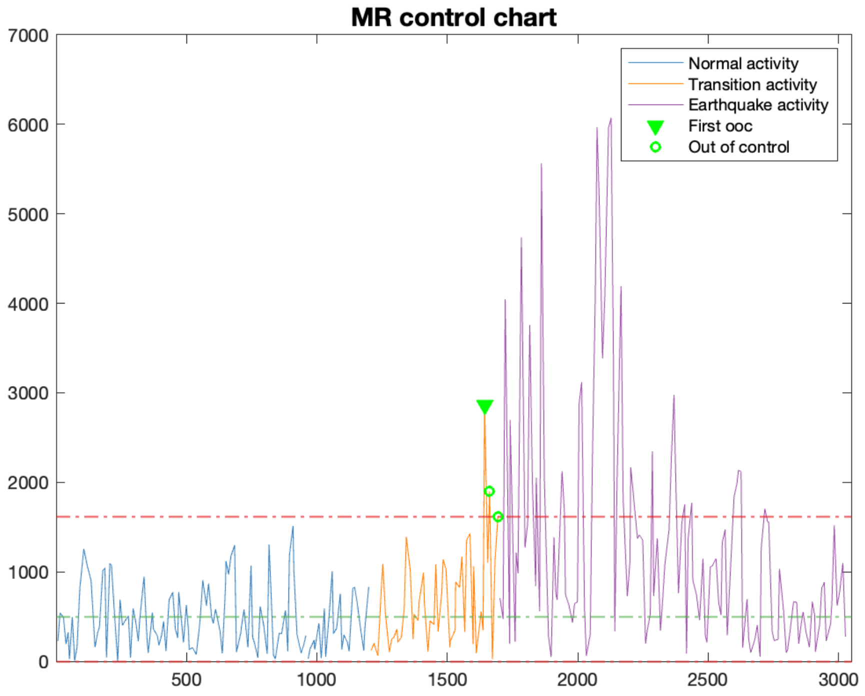
As false alarm rate, I chose the standard $\alpha = 0.0027$. Recall that $\alpha = P(x \notin [LCL; UCL] | H_0)$ where H_0 is "Process is under control".

```
%% Random sampling
WindowLength = 10;
t_train = zeros(floor(length(t_train_or)/WindowLength),1);
e_train = zeros(floor(length(e_train_or)/WindowLength),1);
for idx = 1:floor(length(t_train_or)/WindowLength)
    Block_e = e_train_or((idx-1)*WindowLength+1:idx*WindowLength);
    Block_t= t_train_or((idx-1)*WindowLength+1:idx*WindowLength);
    i = randi([1 WindowLength], 1, 1);
    t_train(idx) = Block_t(i);
    e_train(idx) = Block_e(i);
end
```



05 – I-MR chart





$$UCL = \bar{x} + z_{\alpha/2}\frac{\overline{MR}}{d_2}$$

$$CL = \bar{x}$$

$$LCL = \bar{x} - z_{\alpha/2}\frac{\overline{MR}}{d_2}$$

Anomaly detected at $t = 1611$

t to earthquake $t = 89$

$$UCL = D_4(2)\,\overline{MR}$$

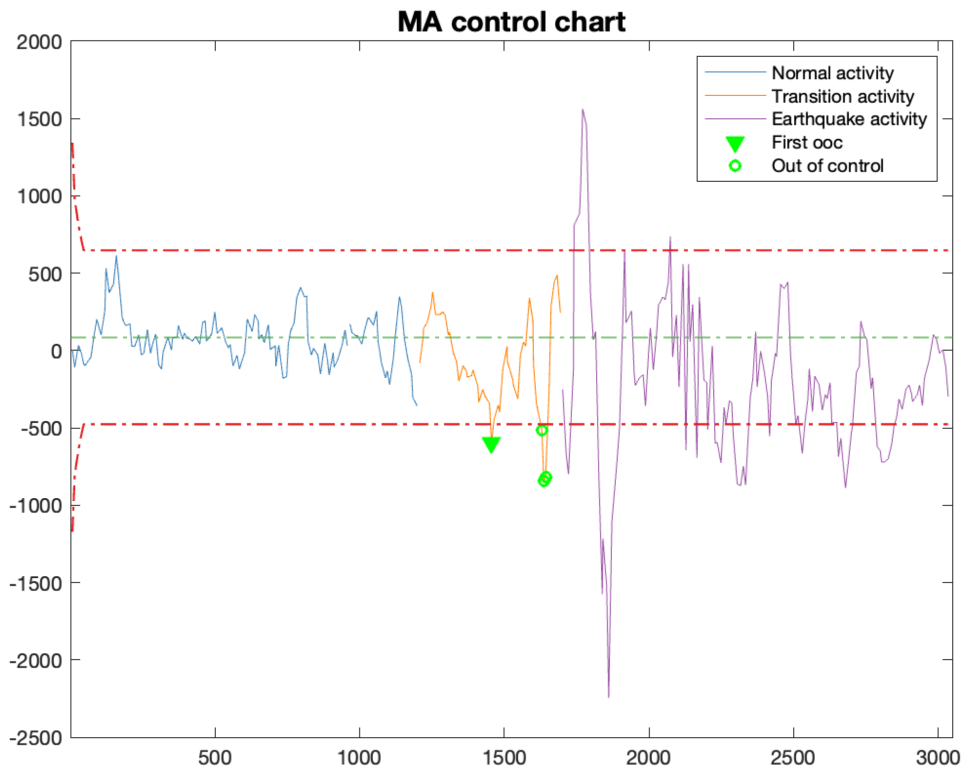
$$CL = \overline{MR}$$

$$LCL = 0$$

Anomaly detected at $t = 1643$

t to earthquake $t = 57$

05 – MA chart



$$M_i = \begin{cases} \dfrac{x_i + x_{i-1} + \cdots}{i}, & \forall i = 1, \dots w - 1 \\ \dfrac{x_i + x_{i-1} + \cdots + x_{i-w+1}}{w}, & i \geq w \end{cases}$$

$$UCL = \begin{cases} \bar{x} + 3\dfrac{\hat{\sigma}}{\sqrt{i}}, & i < w \\ \bar{x} + 3\dfrac{\hat{\sigma}}{\sqrt{w}}, & i \geq w \end{cases}$$

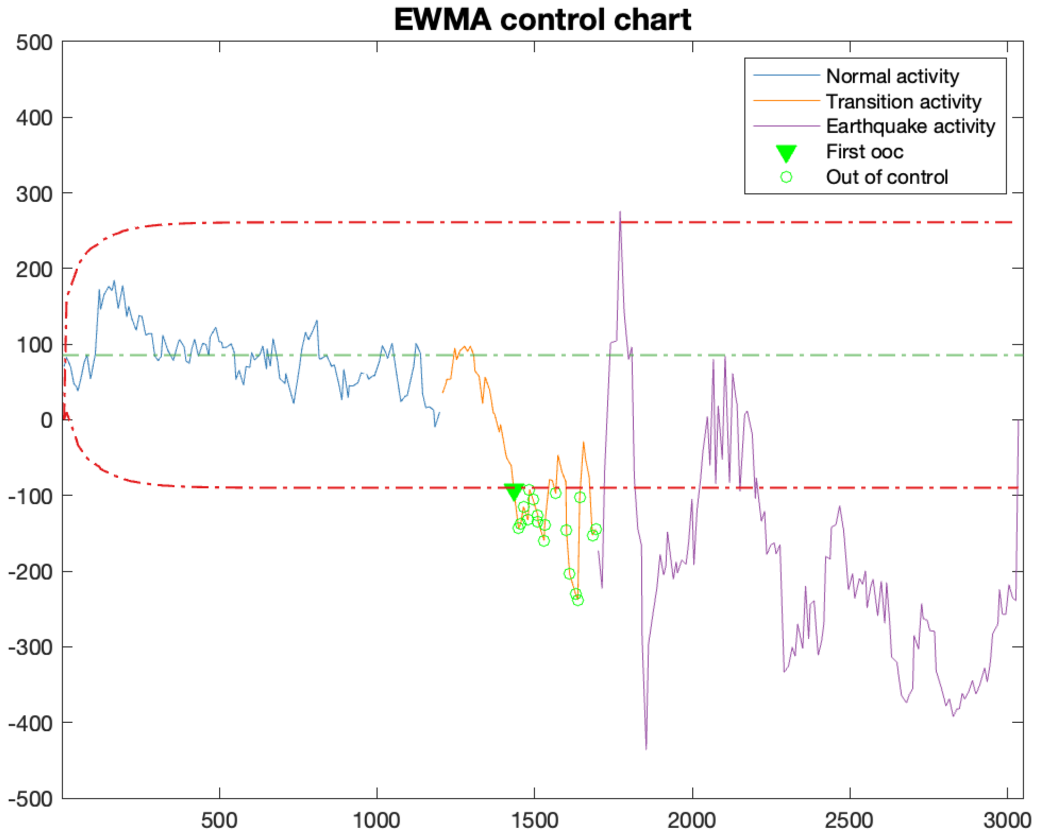
$$CL = \bar{x}$$

$$LCL = \begin{cases} \bar{x} - 3\dfrac{\hat{\sigma}}{\sqrt{i}}, & i < w \\ \bar{x} - 3\dfrac{\hat{\sigma}}{\sqrt{w}}, & i \geq w \end{cases}$$

Anomaly detected at $t = 1456$

t to earthquake $t = 244$

05 – EWMA chart



$$z_i = \begin{cases} \bar{x}, & i = 0 \\ \lambda x_i + (1-\lambda) z_{i-1}, & i \geq 0 \end{cases}$$

$$UCL = \bar{x} + L\hat{\sigma}\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]}$$

$$CL = \bar{x}$$

$$LCL = \bar{x} - L\hat{\sigma}\sqrt{\frac{\lambda}{2-\lambda}[1-(1-\lambda)^{2i}]}$$

Anomaly detected at $t = 1456$

t to earthquake $t = 244$

$\lambda = 0.05$

$L = 2.615$

As suggested by Lucas and Saccucci in 1990 in their paper

07 – Conclusion

05 – Conclusion

I suggested to use a system composed by the joint **monitoring of MSE** and the use of the **EWMA control chart**. The latter allow us to search for a pattern of anomalies, the larger the number of anomalies the greater the probability an earthquake is about to occur.

Actually, **we can save lives**.

Some advices for geographic areas at high seismic risk:

1. Construct the small buildings in such a way that the escape routes are preferably passable in no more than 400 seconds and in any case no more than 200 seconds.
2. Construct big buildings (like offices and hospitals) so that there are safe shelters accessible at less than 400 seconds for visitors and shelters accessible at less than 200 seconds for regular visitors.
3. The main target should be large buildings since escape is more difficult in the presence of many people and in the presence of non-regular users. In small buildings, in fact, it is much more difficult for nonregular users to be present.

Source https://publications.iafss.org/publications/fss/1/523/view/fss_1-523.pdf

A Case Study of Fire and Evacuation in a Multi-Purpose Office Building, Osaka, Japan
S. HORIUCHI

Thanks for the attention.

Any question?