ROYAL OBSERVATORY OF BELGIUM

Using Machine Learning Algorithms for Automated Data Cleaning of GNSS Position Time Series Based on Data Quality Indicators

Fikri Bamahry, Juliette Legrand, Carine Bruyninx, Eric Pottiaux and Andras Fabian

Royal Observatory of Belgium

EUREF Symposium 2023

23 May 2023

Background

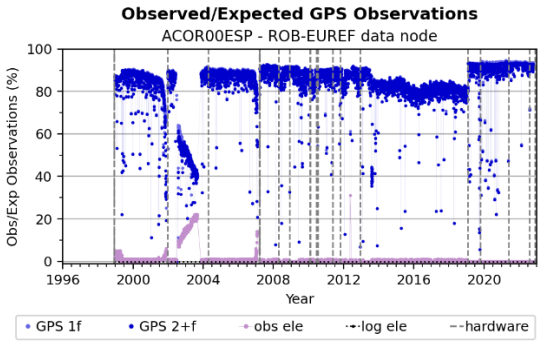
- Data cleaning in GNSS position time series analysis is a critical step that can affect the accuracy and reliability of GNSS position and velocity estimates.
- We know that data quality degradation is a crucial factor affecting the quality of GNSS position estimates.
- The data quality indicators plots are used to help understanding outliers in the position time series.
- Therefore, we need to develop a new algorithm which can automatically identify degraded GNSS quality that cause outliers in position time series.

How to achieve the goal?

We investigated the correlations between daily GNSS data quality indicators and daily position estimates

Feature

GNSS data quality indicator

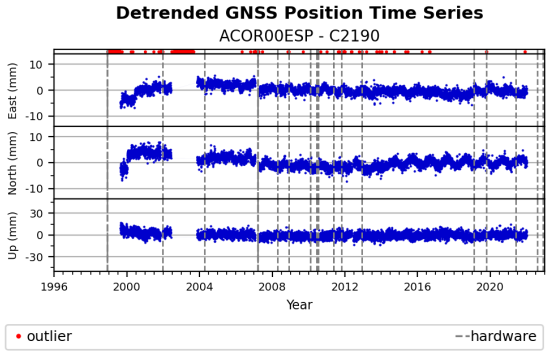


Algorithm

Machine Learning

Label

GNSS position time series

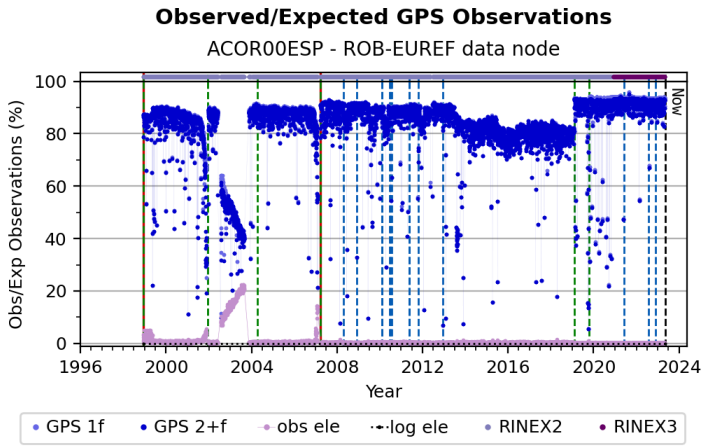


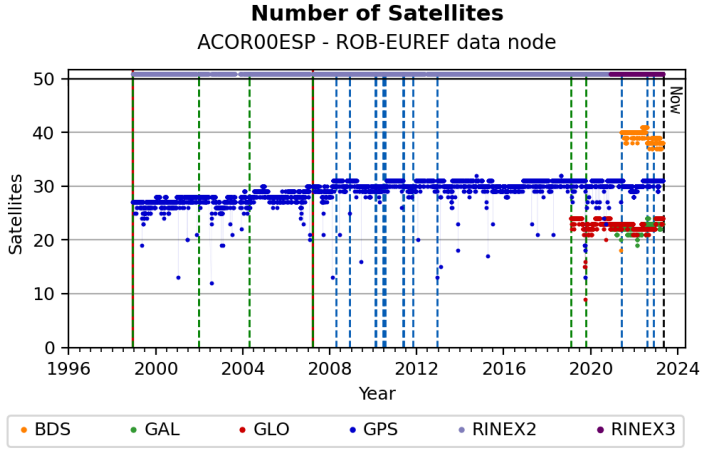
Multi-year reference frame solution based on daily combined EPN position solutions

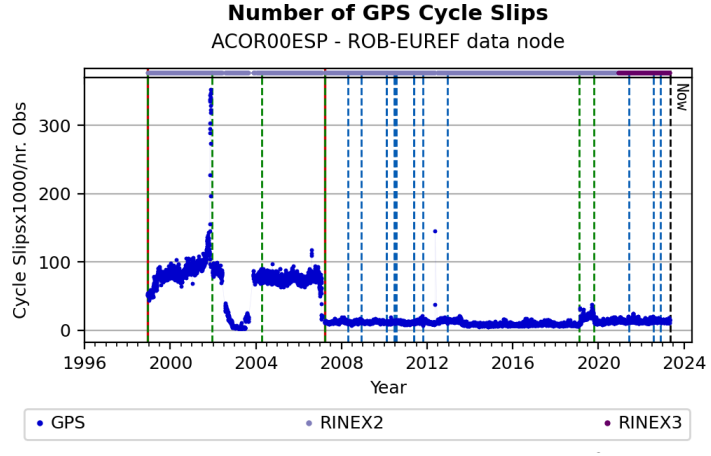
a suitable data-driven model based on the correlation between data quality indicators and position time series

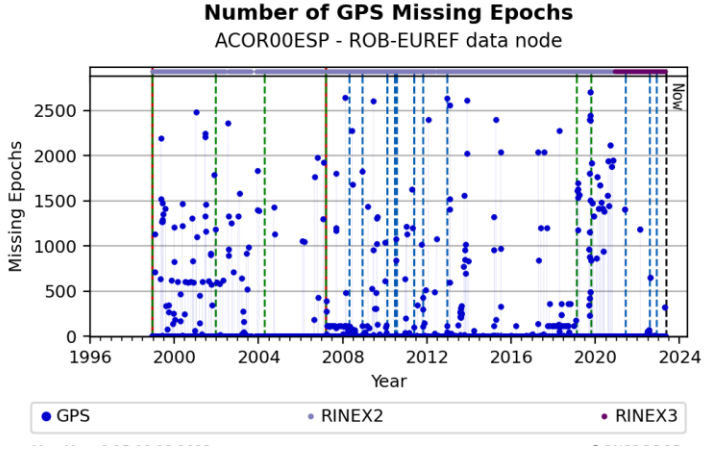
GNSS Data Quality Indicators

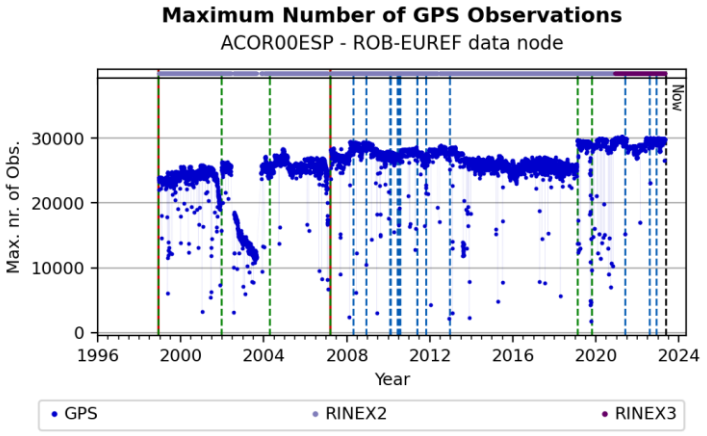
www.epncb.oma.be

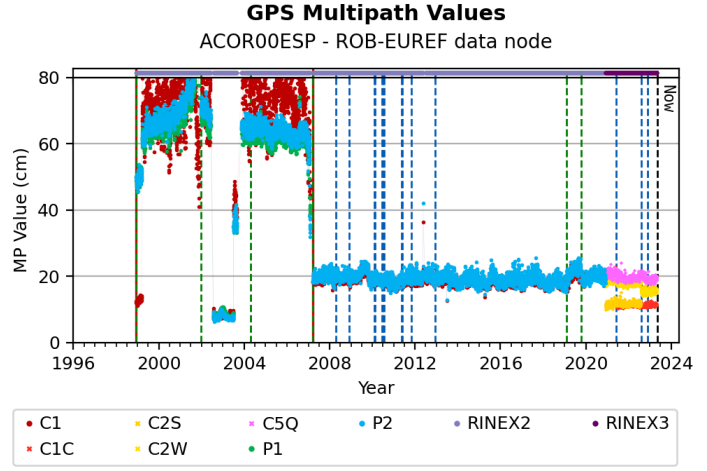






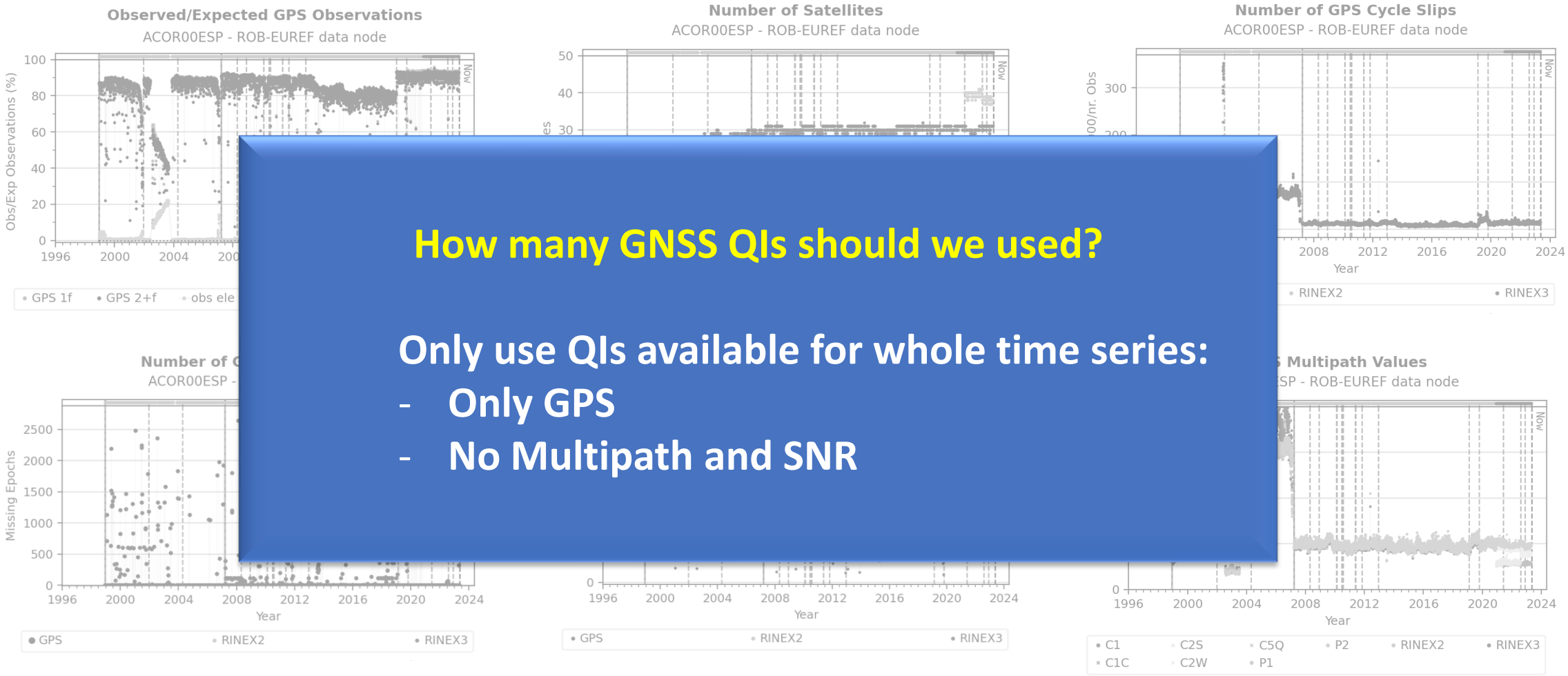






GNSS Data Quality Indicators

www.epncb.oma.be

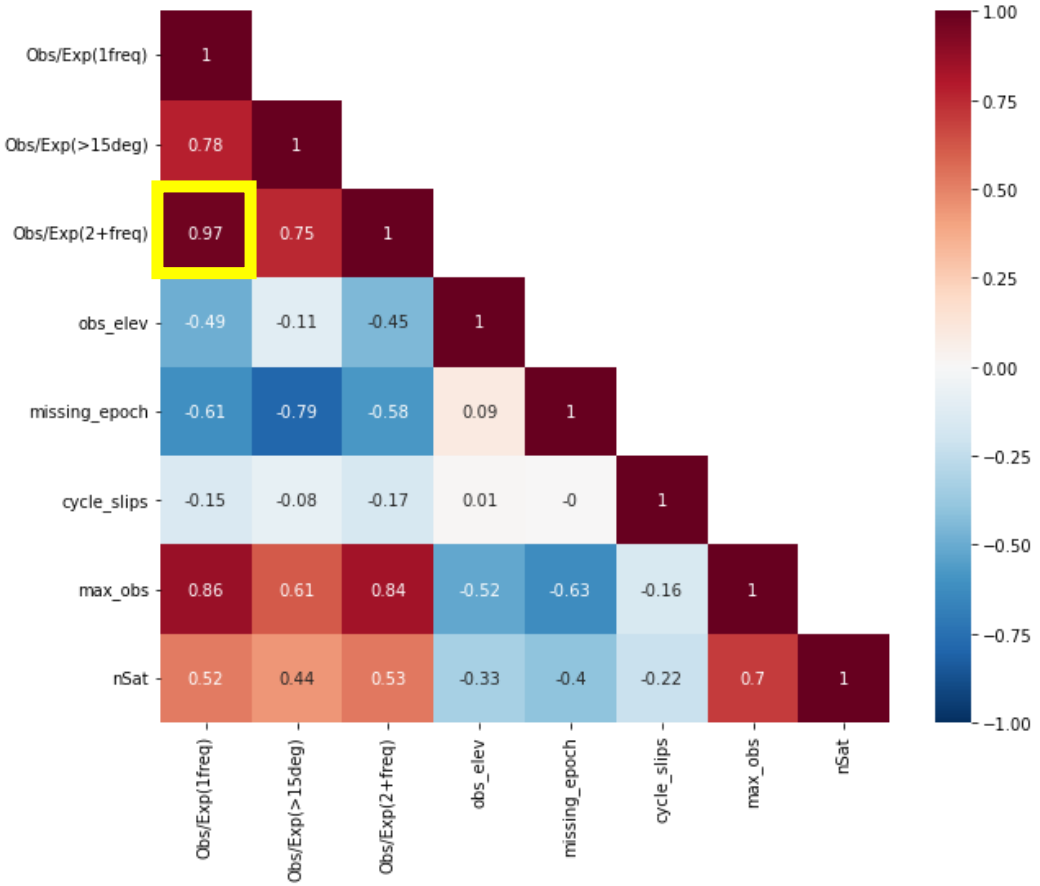


How many GNSS QIs should we used?

Only use QIs available for whole time series:

- **Only GPS**
- **No Multipath and SNR**

Correlation between GNSS QI

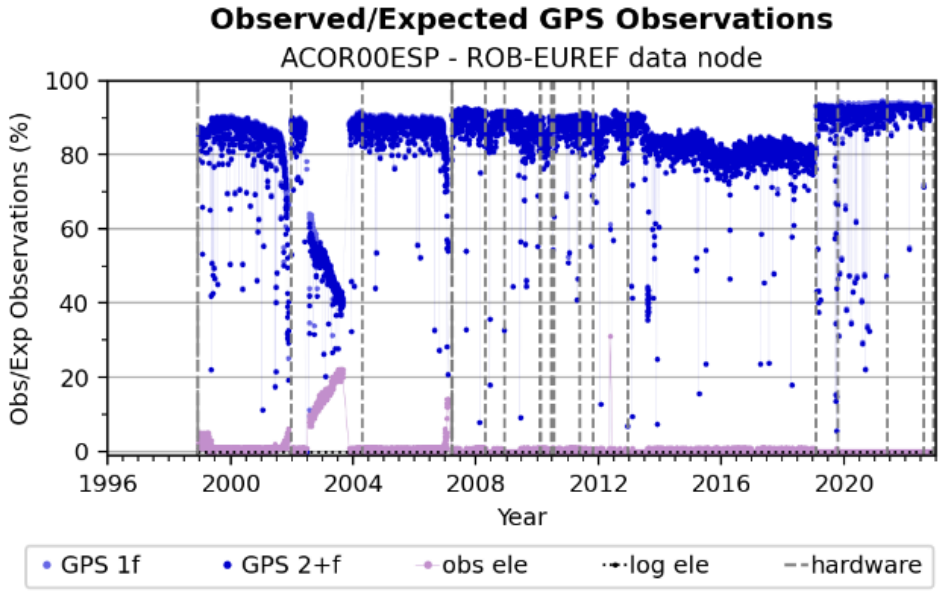


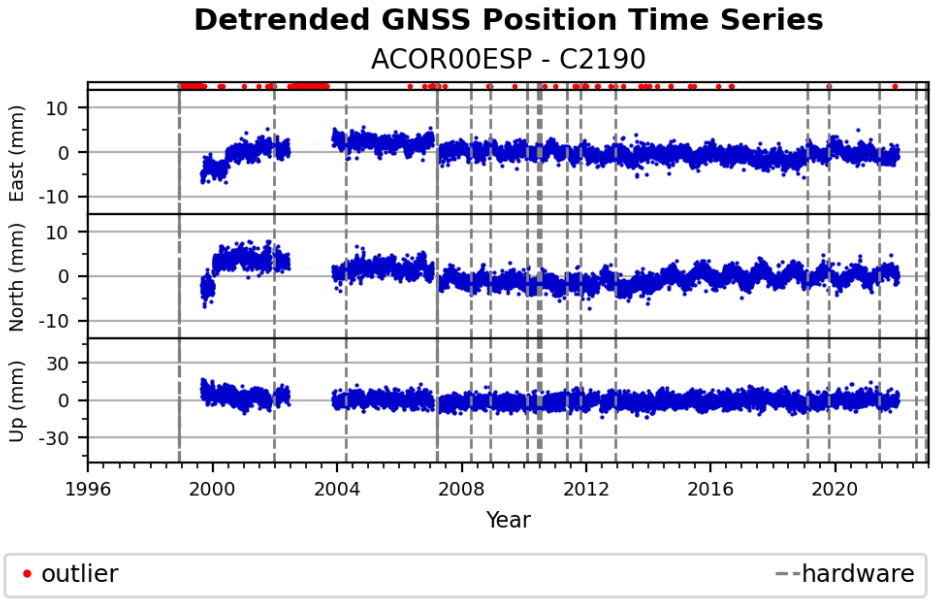
High correlation between Observed vs Expected Observations (1freq) and (2+freq):

→ Discard Observed vs. Expected Observations (1freq)

Only 7 GPS QIs will be used in this study

Supervised Machine Learning





Features

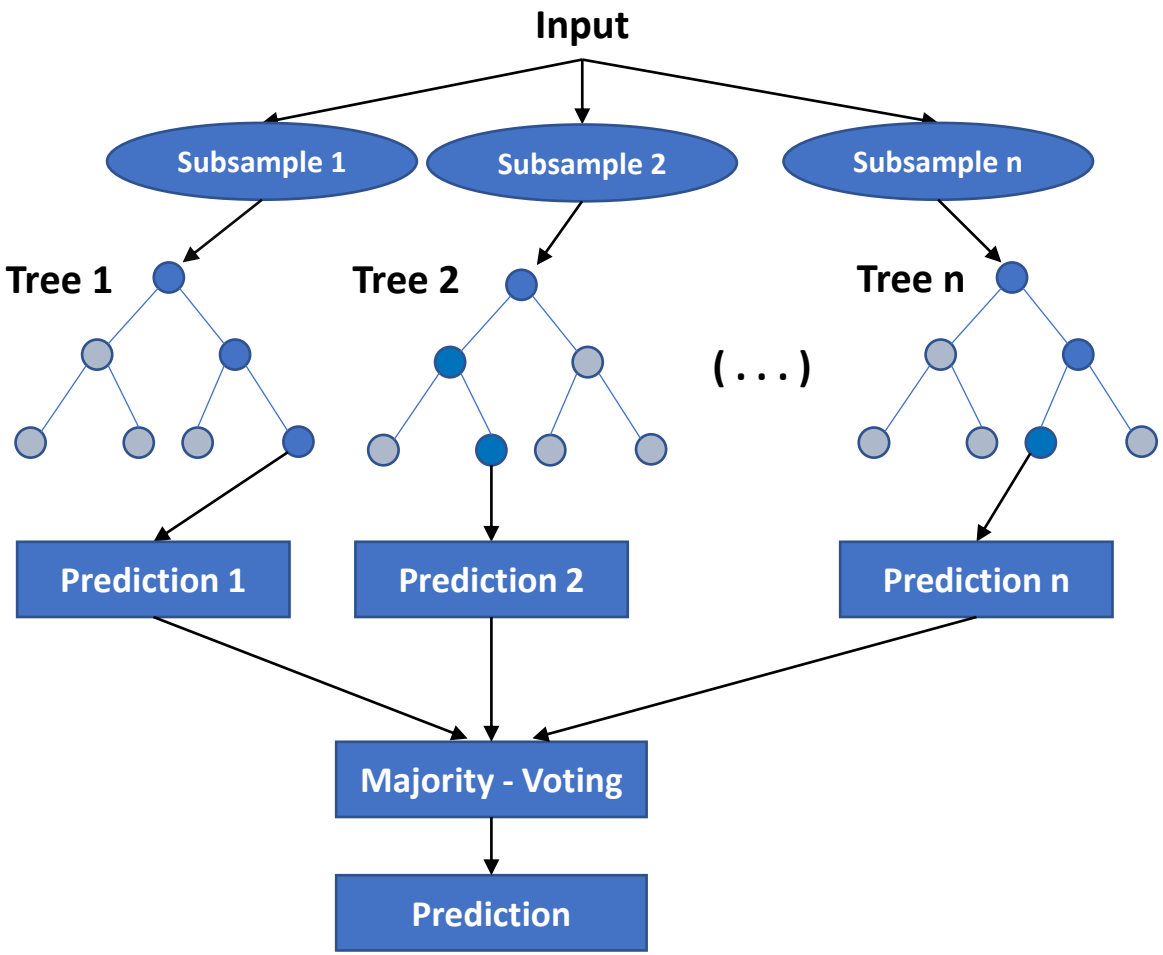
- Obs/Exp observations (2+freq)
- Obs/Exp observations (1freq) above 15 deg
- Lowest elevation cut-off observed
- Missing epochs
- Number of satellites
- Maximum observations
- Number of cycle slips

Label

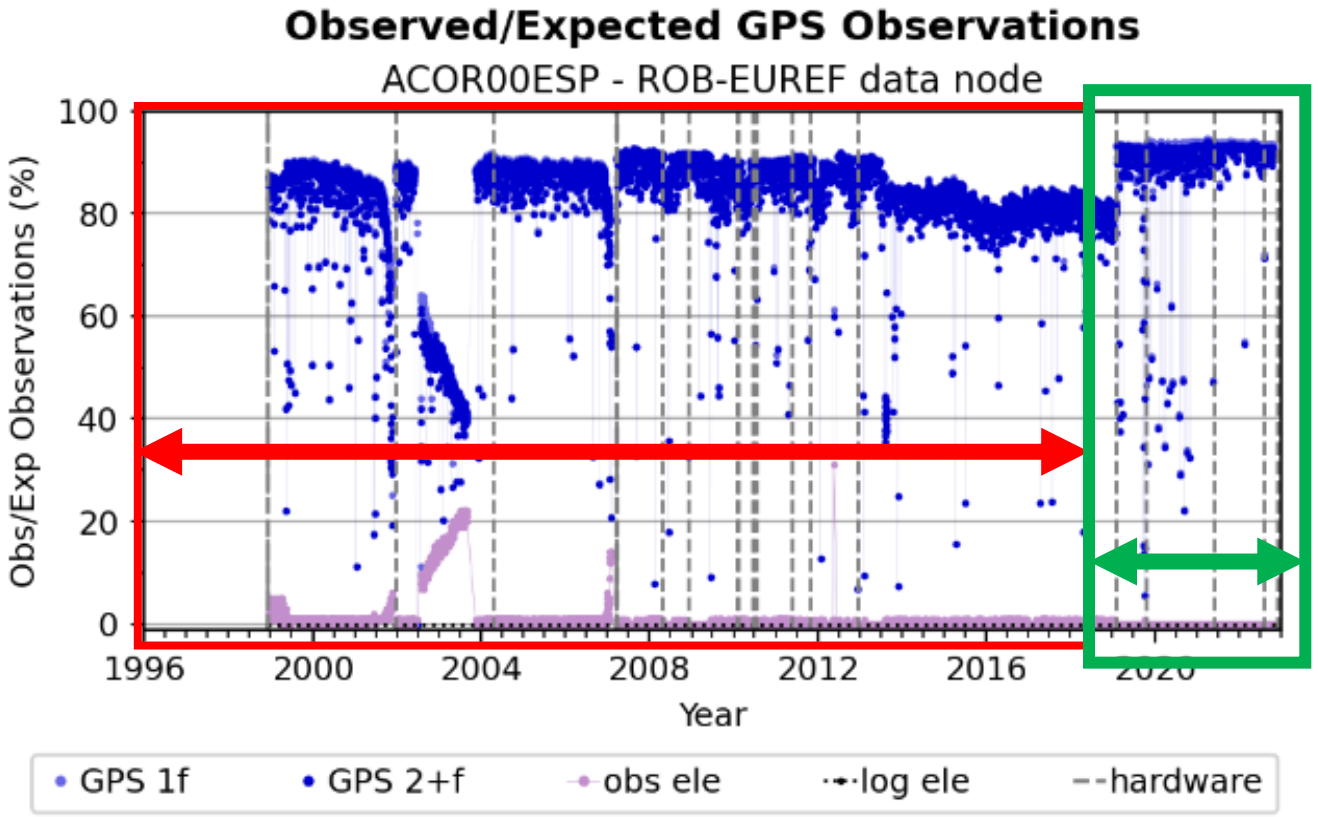
If daily GNSS position is in the cleaned time series → **not an outlier**
If daily GNSS position is not in the cleaned time series → **an outlier**

Random Forest

Random Forest



- A supervised learning algorithm consisting of many decision trees
- Choosing random subsample from the input
- Building several trees based on each subsample
- Combining the results of all decision trees to get prediction



(80%) Training dataset

(20%) Test dataset

RANDOM

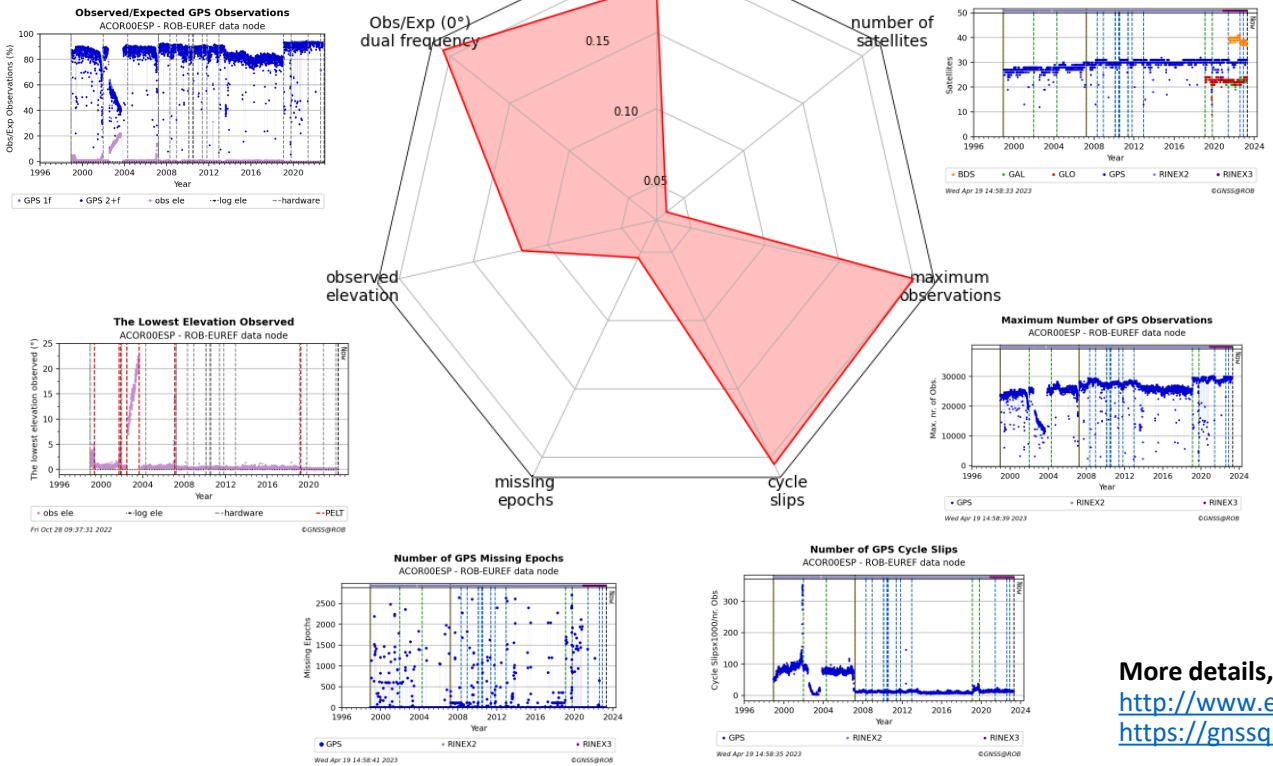
Training dataset is used to allow algorithm to make classification into 2 categories:

- Good GPS data → reliable station position estimate
- Bad GPS data → outlier

Test dataset is used to assess quality of model after training process.

Validation dataset is used as independent dataset to see how well model works.

Feature Selection



The most important features driving this model:

- **Observed vs expected observations (2+freq)**
- **maximum observations**
- **cycle slips**

More details, visit:
http://www.epncb.oma.be
https://gnssquality-epos.oma.be

Assessment

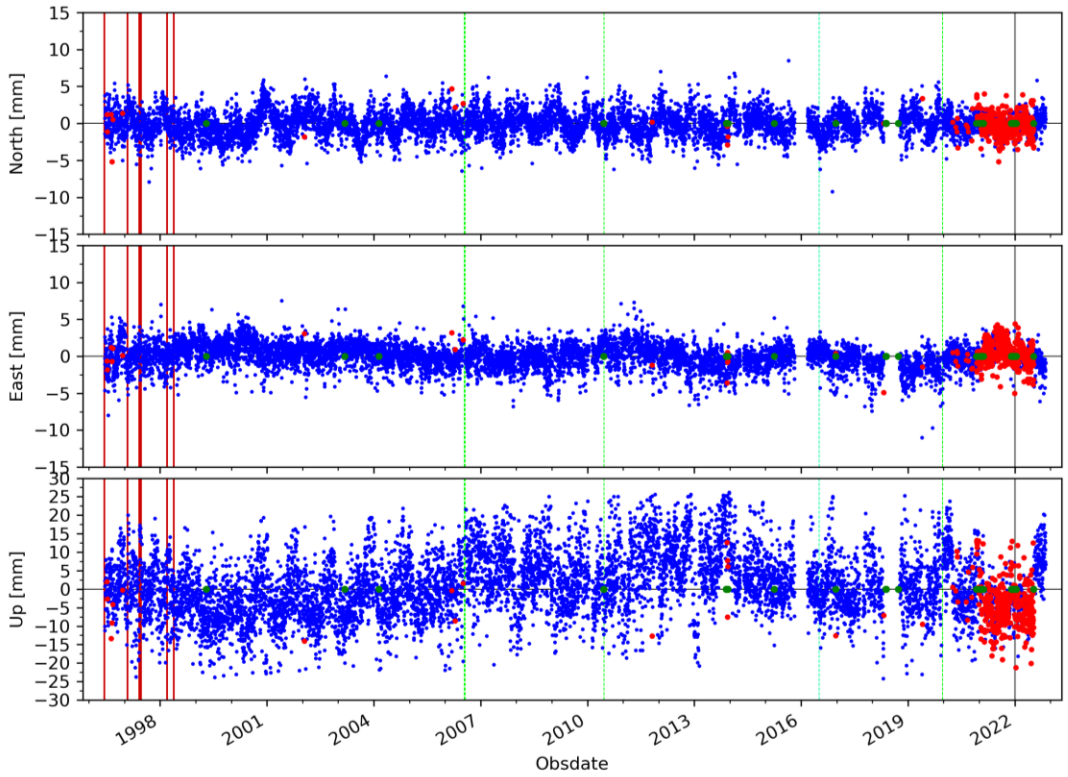
No significant impact:

- Eliminating features that are less important (Number of Satellites)
- Changing the size of the training vs. test datasets (using random selection)
- Training different group of stations

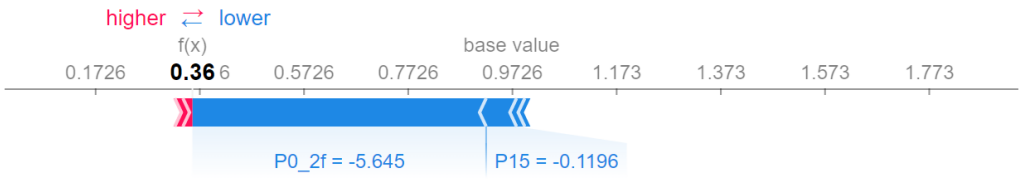
The model is consistent

Evaluate the prediction (Work in Progress)

Residual Position Time Series
MOPI00SVK

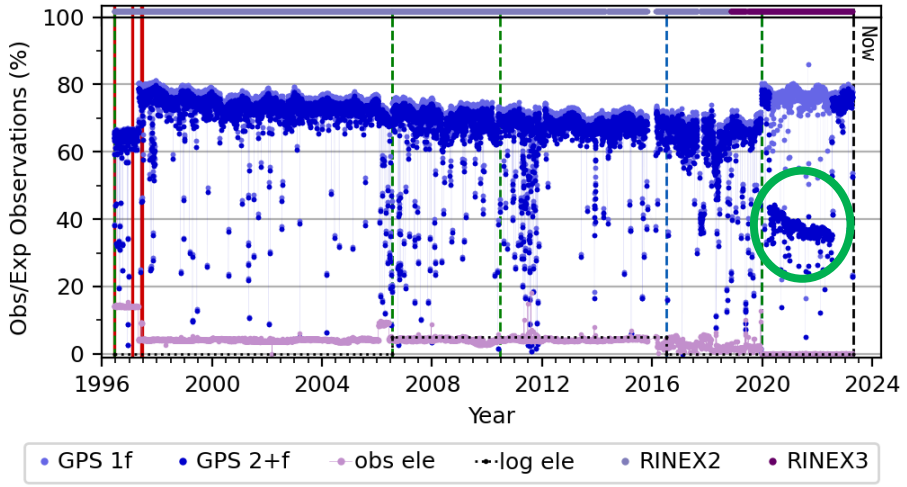


SHAP force plot (MOPI: 2021-05-03)



Observed/Expected GPS Observations

MOPI00SVK - ROB-EUREF data node



Summary

✓ Most important parameters driving our model:

- '**Observed vs expected observations (2+freq)**'
- '**maximum observations**'
- **'cycle slips**'

✓ Based on current tests:

- Training on different sizes of training and test datasets had no significant impact
- Training on different groups of stations had no significant impact

✓ Based on current evaluation:

The model can detect the degraded GNSS quality, however it is not good enough to identify possible outliers in position time series → Under investigation

Future work

- To label the position time series into multi-class category
- To investigate using different datasets (NGL, IGS, or our solution)
- To use GNSS QIs that are not available for whole time series? (Multi-GNSS, Multipath values, and SNR values)

Thank You

Contact: fikri.bamahry@oma.be

GNSS Team ROB

Changing the size of the training vs. test datasets