# Model fitting for $CO_2$ measurements

## Mini Project Public Defense

Joël Mateus Fonseca, Francis Damachi

May 21, 2019

COM-503 Performance Evaluation

# Table of contents

# Mini project description

## About the dataset [1]

- sensor network composed of 46 sites from the city of Zurich
- sensor measurements:
  - $CO_2$, temperature and humidity
  - every 30 min
  - entire month of October 2017
- sensor metadata: altitude and zone
- additional information: average daily wind pattern

[1] Taken from the *EE-490 Lab in data science* course.

# Goal and issues

## Prior Knowledge

The exact measurement in each site depends on parameters such as the temperature, the humidity, the wind, the altitude and the traffic around the site.

## Curate the $CO_2$ measurements

The goal is to fit a robust regression model to the $CO_2$ measurements as the pool of sensors are cheap but inaccurate and can therefore be subject to drifts.

## Anomaly

A domain expert indicates that there is a problem with a particular sensor at a given time $t_d$. The fitted model will allow to retrieve the lost measurements and acknowledge the drift.

# Faulty Sensor Modeling

## No mathematical modeling available

The teaching staff was mainly guided by the domain experts.

## Single example at our disposal

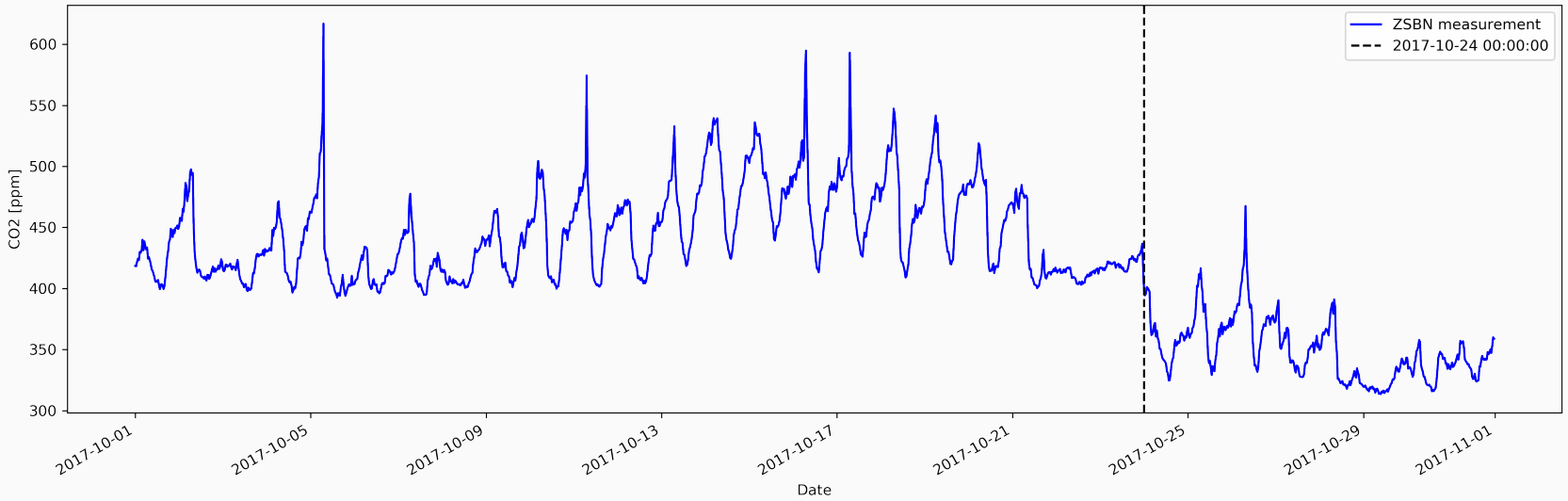



**Figure 1:** ZSBN (faulty sensor) measurements with failure set to the 24th of October.

# Faulty Sensor Modeling

## Proposed definition

A faulty sensor undergoes two random transformations:

- **drift**: up or down, range=[30, 60],
- **rescale** (shrink) in amplitude: based on the std, range=[0.2, 0.3].

**Low influence setting**

Modified measurement will not be directly used.

## Date of failure unchanged

Set to the 24th of October. Main reasons:

- **train/test split**: 75/25 % ideal,
- can be **fatal**: avoid poor training sample.

# Proposed solution method

## Try different Approaches

- Naive
- Zone & Altitude (ZA)
- Wind & Time (WT)
- Brute Force (BF)

## Try different Linear Regressions

- Linear Regression (standard)
- Lasso
- Ridge
- ElasticNet

# Proposed Approaches

## Naive

Rely solely on the faulty sensor. Train on healthy measurements.
Only features are **temperature** and **humidity**.

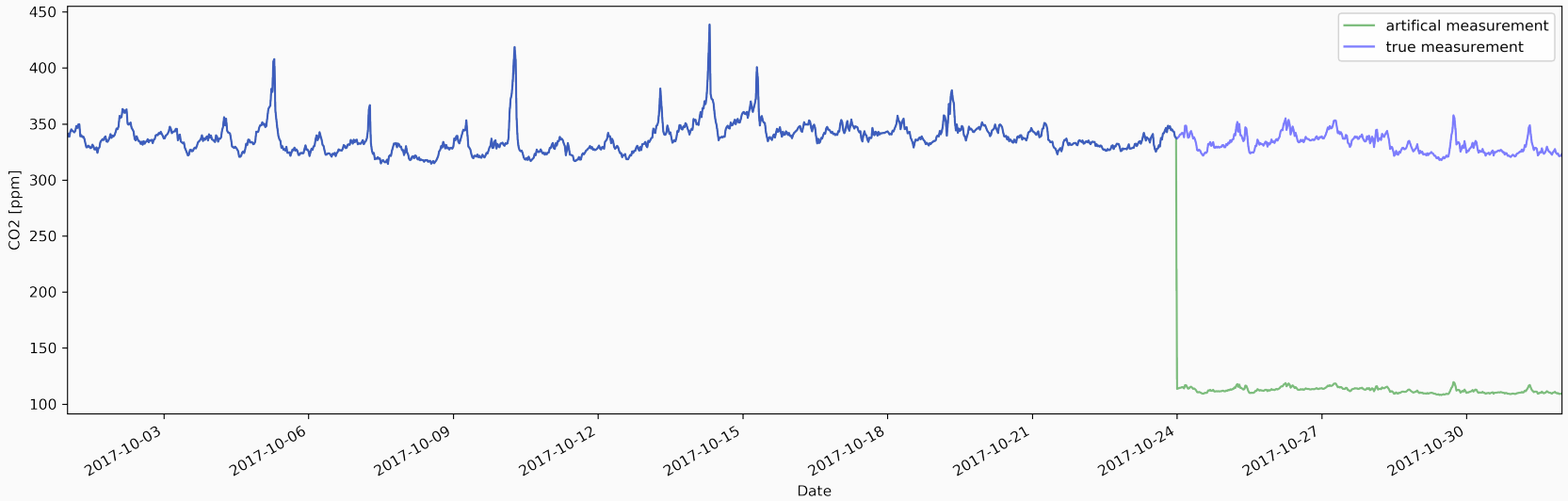



**Figure 2:** AJGR sensor measurements.

# Proposed Approaches

## Zone & Altitude (ZA)

*[includes Naive].* Aggregate sensors with same **zone** and **altitude cluster**. Clustering (K=10) performed on altitude and median CO2 measurement.

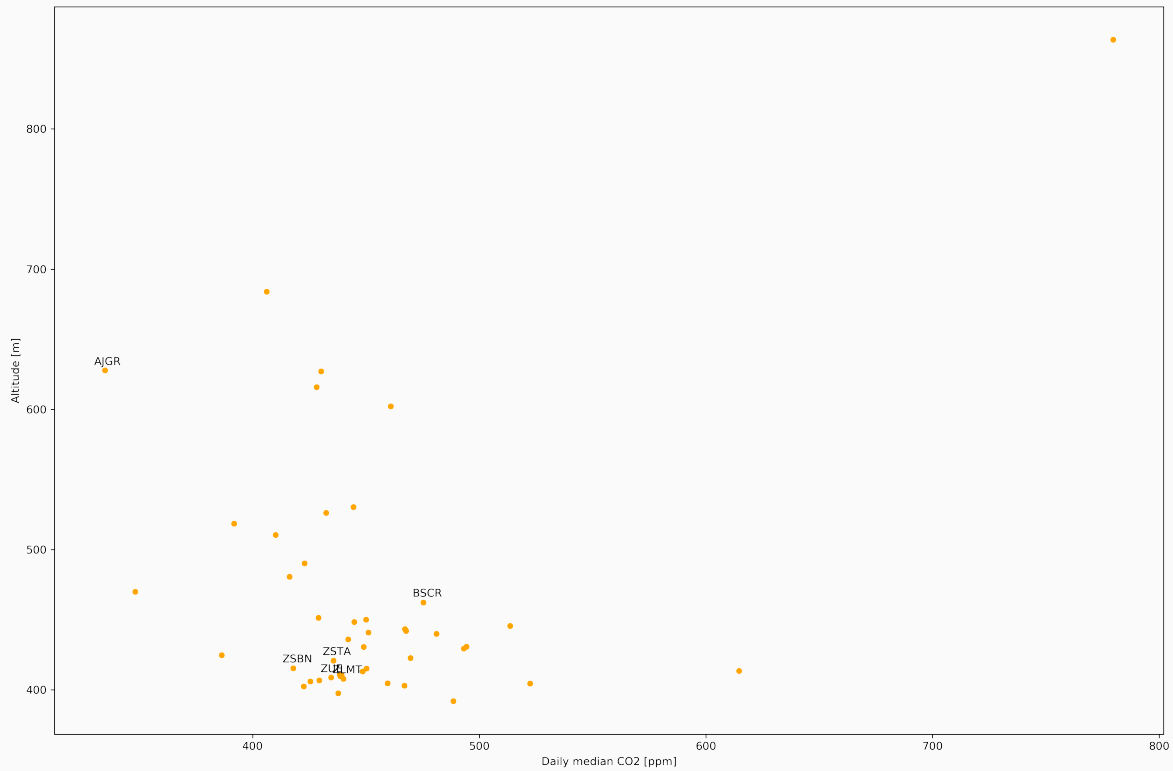



**Figure 3:** Clustering features visualisation.

# Proposed Approaches

## Wind & Time (WT)

*[includes Naive and ZA]*. Train a linear regression based on **wind conditions** and **time** feature [0-47]. Clustering (K=8) performed on 2 PCA components.

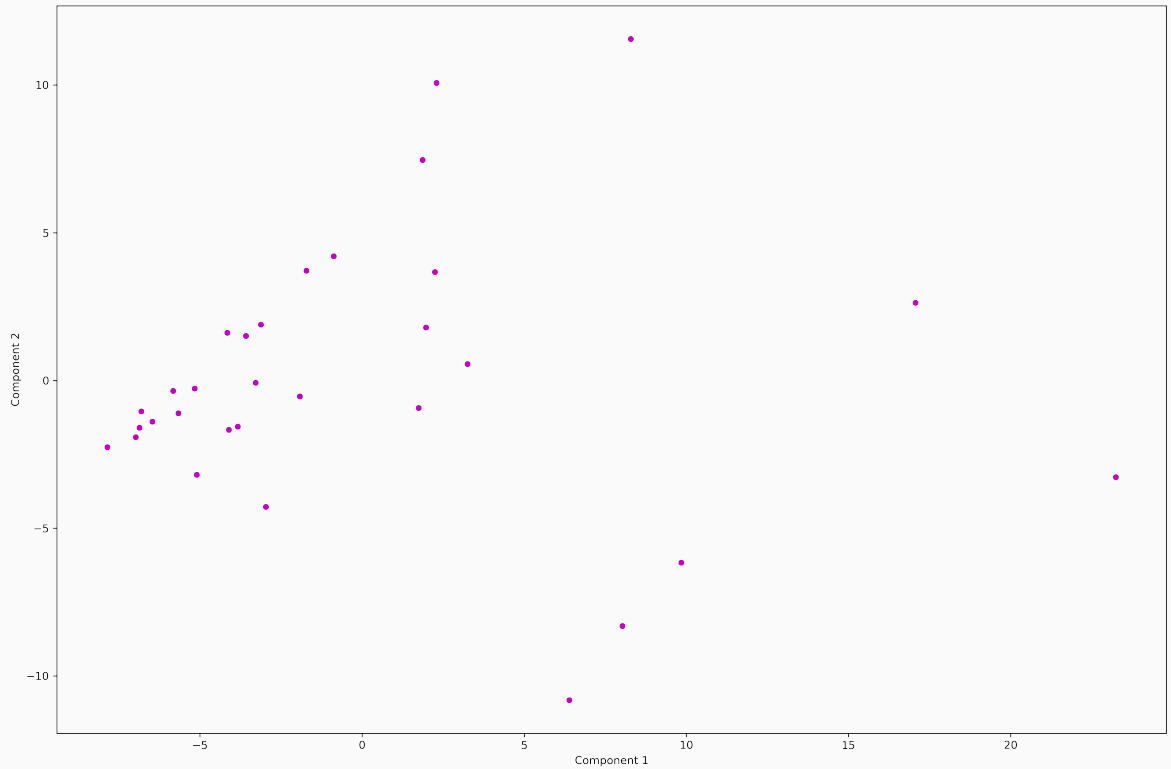



**Figure 4:** Clustering features visualisation .

# Proposed Approaches

## Brute Force (BF)

Only using raw measurements. The features are **temperature**, **humidity**, **altitude** and **wind speed**. Use data of all sensors.

# Proposed Linear Regressions

## Linear Regression (standard)

$$f(X, y|w) = \frac{1}{2N}||y - Xw||_2^2 \tag{1}$$

## Lasso

$$f(X, y|w, \alpha) = \frac{1}{2N}||y - Xw||_2^2 + \alpha||w||_1 \tag{2}$$

## Ridge

$$f(X, y|w, \alpha) = \frac{1}{2N}||y - Xw||_2^2 + \alpha||w||_2^2 \tag{3}$$

## ElasticNet

$$f(X, y|w, \alpha, r) = \frac{1}{2N}||y - Xw||_2^2 + \alpha r||w||_1 + \alpha(1 - r)||w||_2^2 \tag{4}$$

# Results

**Disclaimer ⚠**

Following the homework's instructions one would believe that the WT approach is the most appropriate and robust. But, the whole process suffers some drawbacks:

- focus on a unique sensor only,
- no convincing technique to compare predictions.

# Experiments

## Select interesting sensors

Select sensors that are different in nature to test robustness of the regression models: AJGR, BSCR, ZSTA, ZUE, ZLMT.

## Select appropriate metric

Apply the MSE on the testing set (measurements after the 24th of October).

# Results

**Table 1:** MSE for Linear Regression model.

| Sensor | Naive | ZA | WT | BF |
|---|---|---|---|---|
| $AJGR_1$ | **44.44** | **44.44** | 21782.03 | 31638.90 |
| $BSCR_6$ | 4995.48 | 4677.71 | **4511.09** | 4529.10 |
| $ZSTA_9$ | 1501.71 | 1216.74 | **948.19** | 1031.60 |
| $ZUE_3$ | 1109.11 | 959.13 | **620.86** | 777.01 |
| $ZLMT_1$ | 1065.23 | 1065.23 | 82766.18 | **708.39** |

# Results

| Sensor | Naive | ZA | WT | BF |
|---|---|---|---|---|
| $AJGR_1$ | **44.44** | **44.44** | 21782.03 | 31638.90 |

## Comments

- Temperature and humidity explain very well the CO2 concentration.
- No need to consider external factors.
- One WT reg trained on faulty day.
- AJGR sensor doesn't match centroid sensor.

# Results

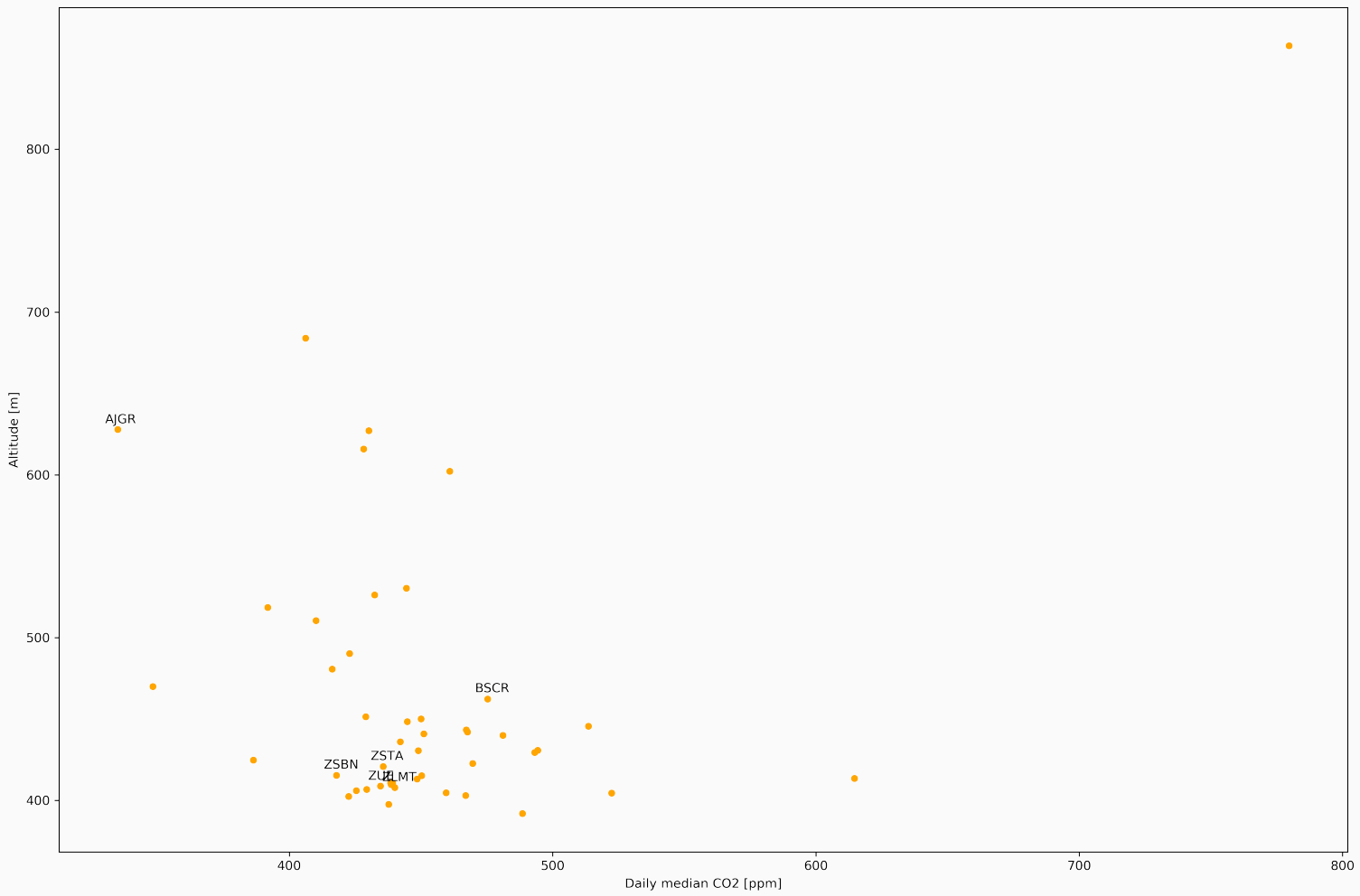



**Figure 5:** Visualisation of sensors based on altitude and median CO2 measurements.

# Results

| Sensor | Naive | ZA | WT | BF |
|---|---|---|---|---|
| $BSCR_6$ | 4995.48 | 4677.71 | **4511.09** | 4529.10 |
| $ZSTA_9$ | 1501.71 | 1216.74 | **948.19** | 1031.60 |
| $ZUE_3$ | 1109.11 | 959.13 | **620.86** | 777.01 |

## Comments

- Be able to rely on similar sensors helps.
- Adding wind feature reduces even more the error.
- Being close to the centroid helps.

# Results

**Table 2:** MSE for Linear Regression model.

| Sensor | Naive | ZA | WT | BF |
|---|---|---|---|---|
| $ZLMT_1$ | 1065.23 | 1065.23 | 82766.18 | **708.39** |

## Comments

- Verifies our hypothesis (being close to centroid helps).

# Results

**Table 3:** MSE for Linear Regression model.

| Sensor | Naive | ZA | WT | BF |
|---|---|---|---|---|
| $AJGR_1$ | **44.44** | **44.44** | 21782.03 | 31638.90 |
| $BSCR_6$ | 4995.48 | 4677.71 | **4511.09** | 4529.10 |
| $ZSTA_9$ | 1501.71 | 1216.74 | **948.19** | 1031.60 |
| $ZUE_3$ | 1109.11 | 959.13 | **620.86** | 777.01 |
| $ZLMT_1$ | 1065.23 | 1065.23 | 82766.18 | **708.39** |

## Deployment solution example (current stage)

Mix Naive and WT based on the support size of the sensor clustering.

# Results

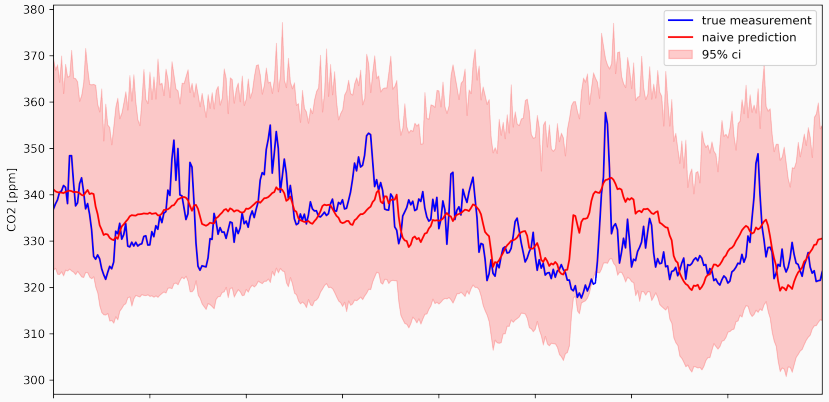
**Table 4:** MSE for WT approach.

| Sensor | Linear Regression | Lasso(1) | Ridge(1) | ElasticNet(1, 0.5) |
|---|---|---|---|---|
| $ZSTA_9$ | 948.19 | 947.33 | 948.20 | **945.02** |
| $ZUE_3$ | 620.86 | 628.25 | **620.85** | 628.05 |

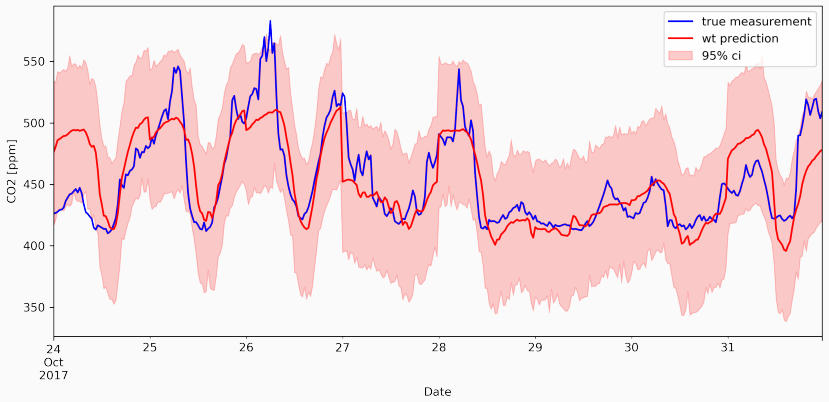
## About regularization

Models not enough complex to benefit from regularization.

# Results




(a) AJGR-Naive




(b) ZUE-WT

**Figure 6:** Prediction measurements with 95% confidence interval using residual bootstrap.

# Results

## About AIC/BIC criterions

Not enough comparable models to use benefits of criterions.

# Conclusion

## “Take home message”

- Always check used scientific method.
- To test robustness, variability of input should be significant.
- Indications of domain expert about result range is crucial.
- “Everything must be made as simple as possible. But not simpler.”

# Further Work

- Ask domain expert about additional external factors not considered and the range of results expected.
- Test robustness with all sensors.

Questions?

# References


- **COM-503 Performance Evaluation**, J.-Y. Le Boudec, EPFL, 2019.
- **EE-490(h) Lab in Data Science**, O. Verscheure, EPFL, 2018.