

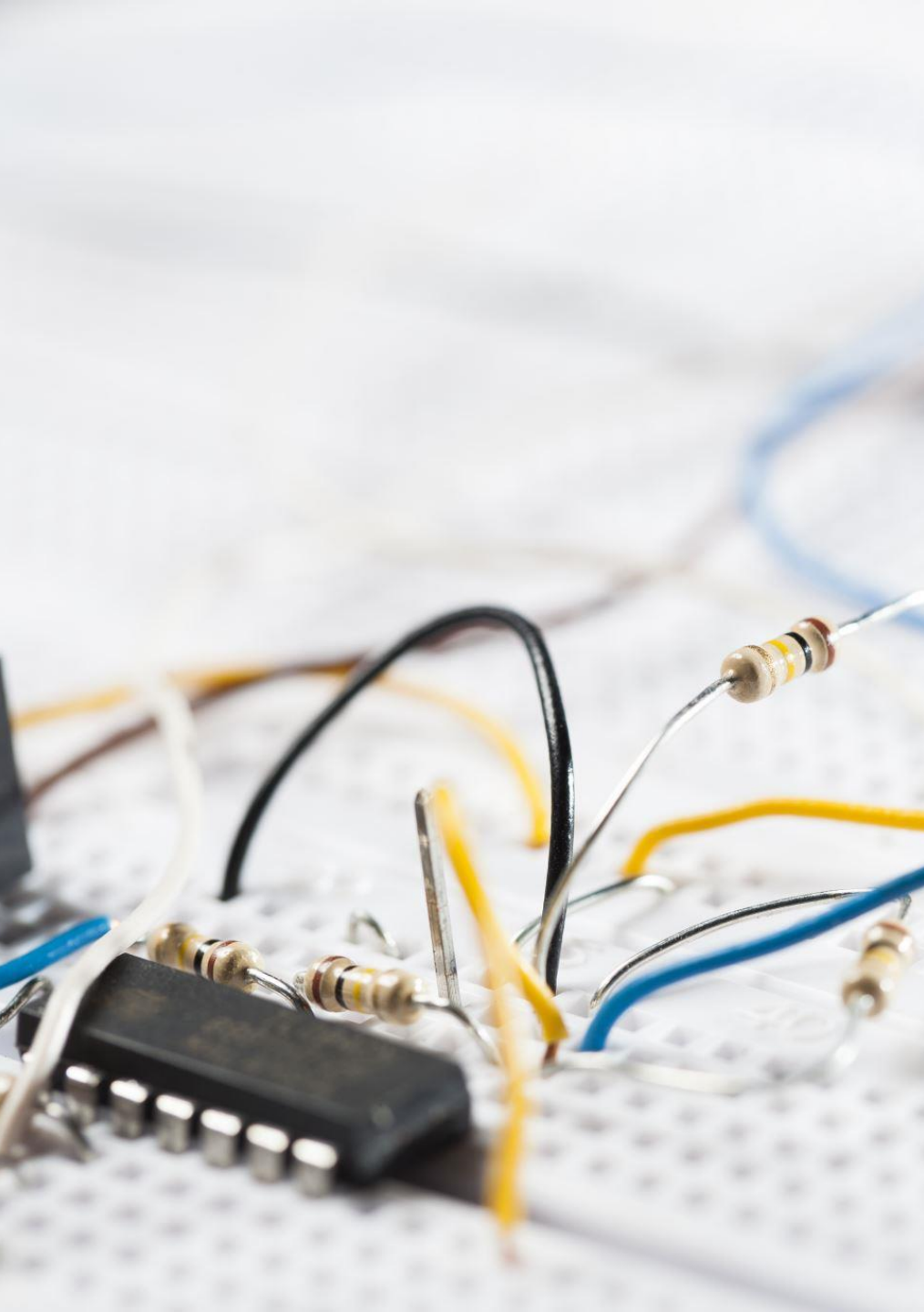


In Field Calibration and Beyond (Part 1)

saverio.devito@enea.it

Topics

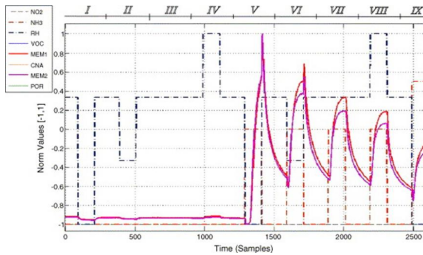
- Motivation: Why we need to calibrate sensors?
- Calibration basics and PM targeted example
- A Gas targeted example
- Calibration models selection
- Limits of Field Calibration
- What lies beyond?



The basic motivation

- Chemical and particulate sensors translate target concentrations into variable(s) which should be translated back in concentration estimations by inverting a sensor «model» i.e. by using a calibration function.
- Some vendors suggests how to derive this calibration function or directly or indirectly suggest a «one size fit all» calibration to be used for all sensors of that specific class.

The basic motivation (2)



$$Y(t) = f(x(t))$$



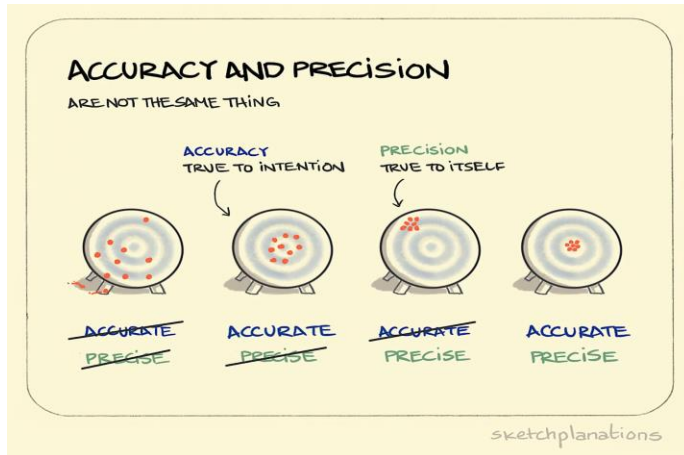
- Chemical and particulate sensors translate target concentrations into variable(s) which should be translated back in concentration estimations by inverting a sensor «model» i.e. by using a calibration function.
- Some vendors suggests how to derive this calibration function either directly or indirectly suggest a «one size fit all» calibration to be used for all sensors of that specific class.

The basic motivation (3)



- Out of Box, Chemical and particulate matter sensors are subject to:

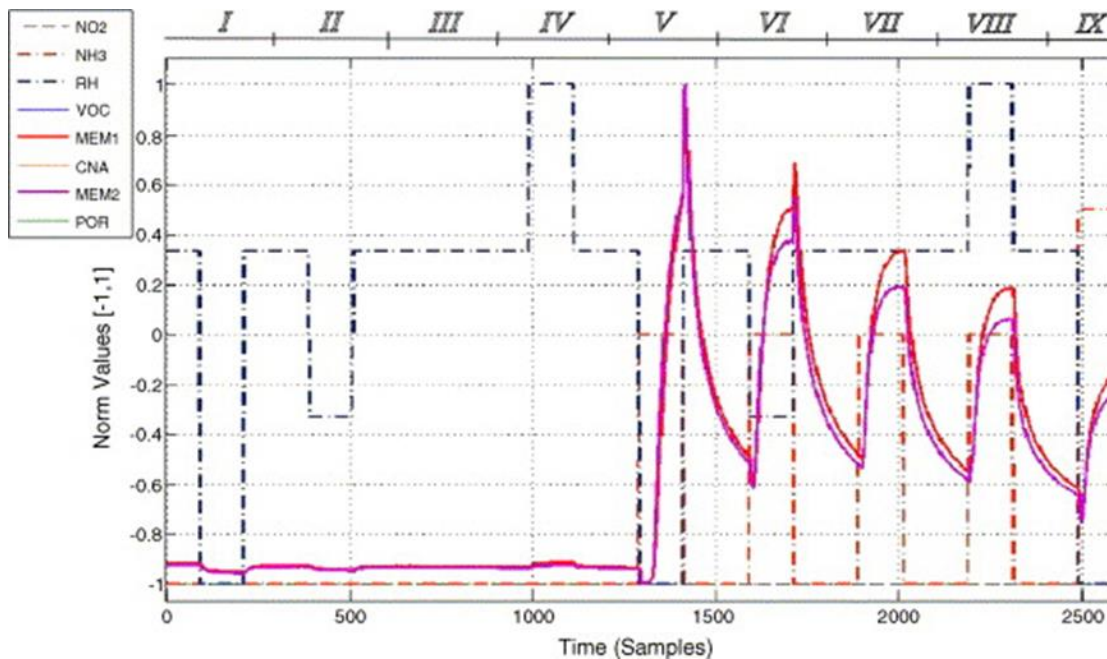
- Fabrication Variance
- Interferences from Non-Target gases
- **Interferences from environmental parameters**
- Drift (Ageing, Poisoning)



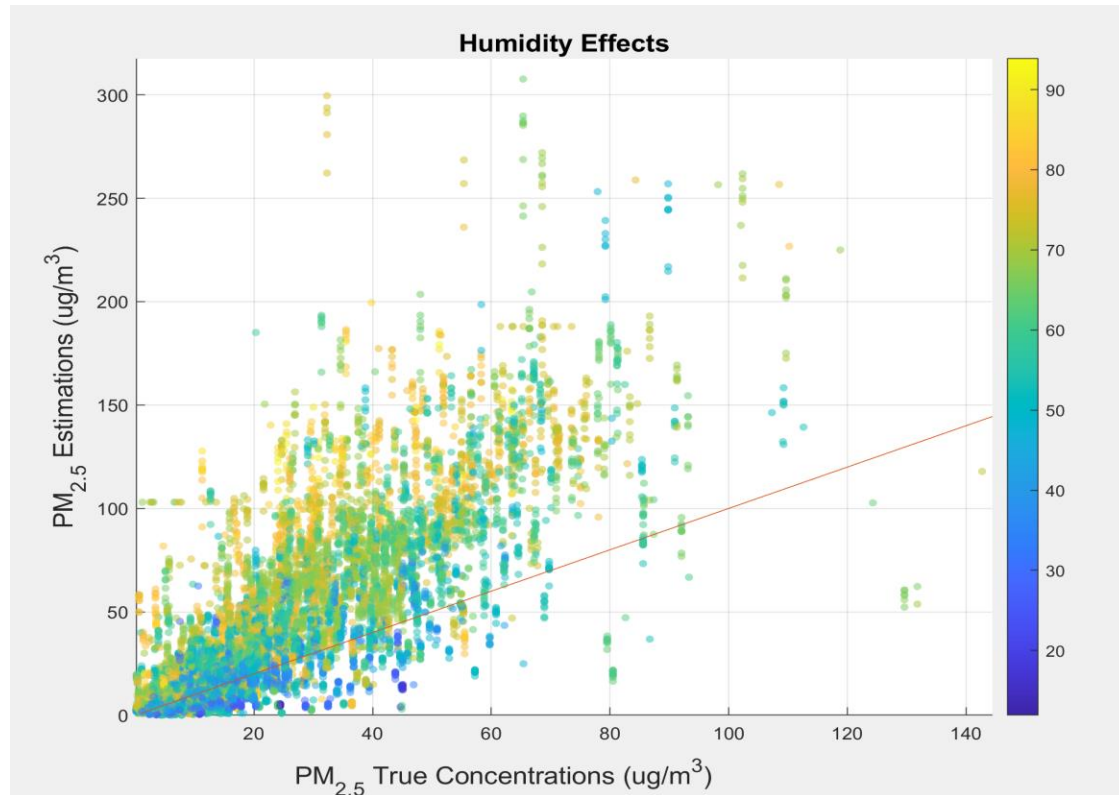
All these factors contribute to hinder original «one size fit all» vendor calibration strongly limiting the sensors accuracy!

Example of Environmental Interference (1)

- Here, a MOX sensor is exposed to NH₃ in a humid carrier.
- We note that depending on RH levels, the sensor response to the same concentration of NH₃ is quite different!



Example of Environmental Interference (2)



- Here, a Plantower PMS7003 is exposed to different concentrations of Particulate in the field.
- We note that, the sensor response to the same concentration of PM_{2.5} is statistically and positively correlated with RH concentrations.

A good news

Most of these issues can be tackled with a one-shot ad-hoc calibration process which derives a specific calibration function for each sensor:

$$Y(t) = f(x(t), k(t))$$

With $x(t) = [x_1(t), \dots, x_i(t), \dots, x_n(t)]$ a vector of all relevant sensors raw outputs

And $k(t) = [k_1(t), \dots, k_i(t), \dots, k_n(t)]$ a vector of all relevant **(known and observable)** interferences

... and the bad news

- F is not easy to derive at all.....
- You need a suitable physical/chemical inspired model of Your sensor or a «black box» model which suite Your sensor response function
- You need a sufficiently complete dataset to practically and accurately derive f relevant parameters.
- You need access to reference data to compare your sensor response to and correctly derive your calibration function
- All these components, in principle, should be obtained and put together in a low cost and scalable process



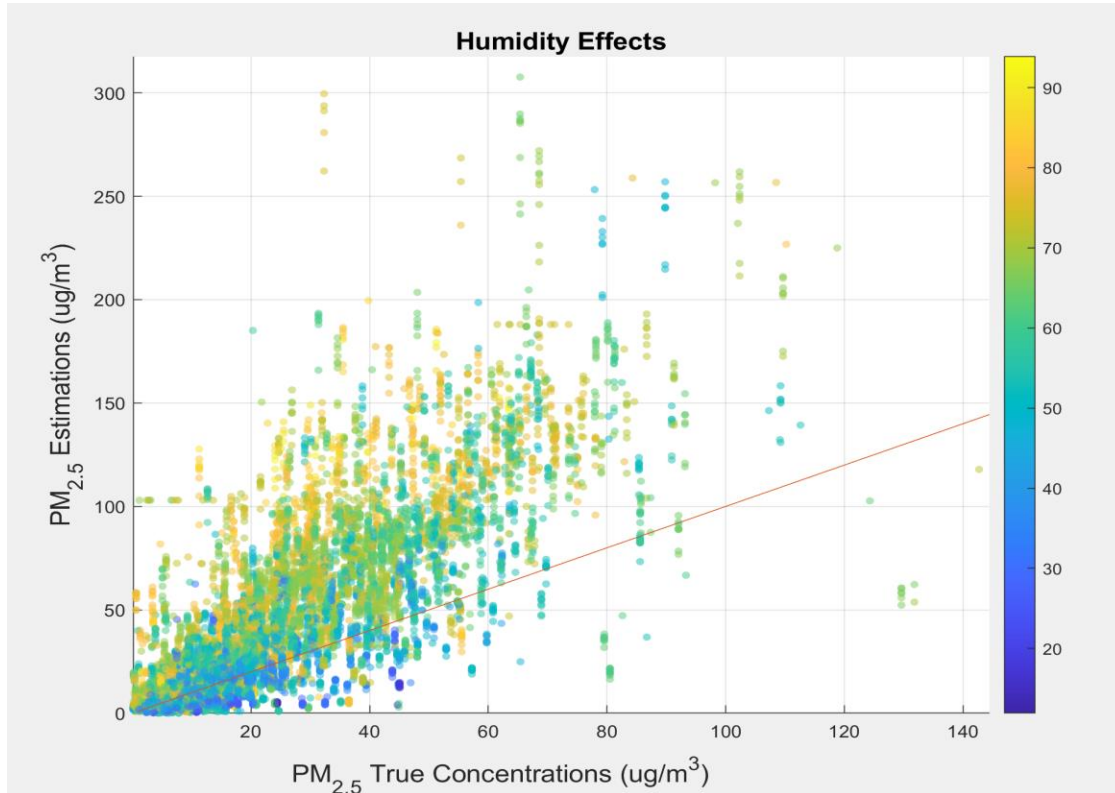
Laboratory calibration vs Field calibration

- Relies on controlled atmosphere chambers
 - Controlled conditions (concentrations and interferent span)
 - **Unobservable and Unknown interferences not taken into account**
- Relies on reference stations
 - **Uncontrolled conditions: Span and mix depends on local conditions**
 - Unknown and Unobservable interferences are partially taken into account



Example :
PMS7003
Calibration

PMs7003 calibration towards PM_{2.5}



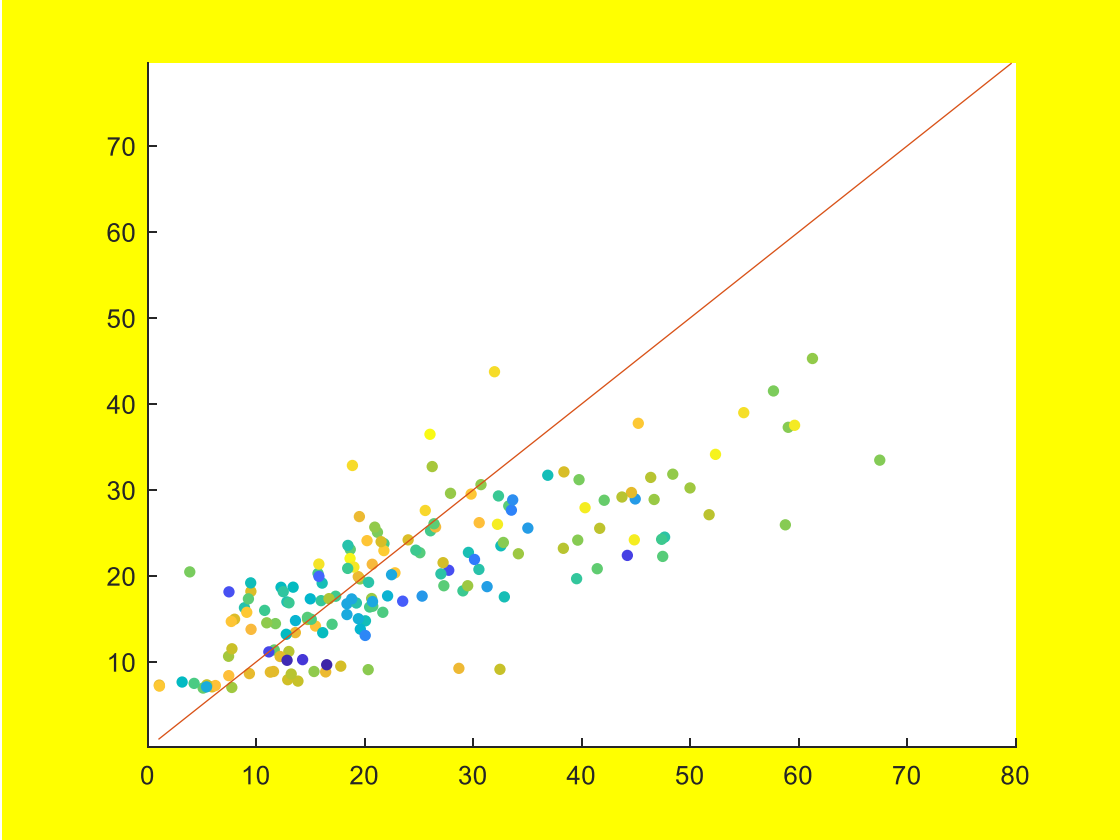
- Let's try to correct:
 - Overestimation tendency
 - Humidity interference
- We may try to derive a black box model based on linear response hypothesis:

$$Y(t) = a PM(t) + b RH(t)$$

... with a conventional OLS method

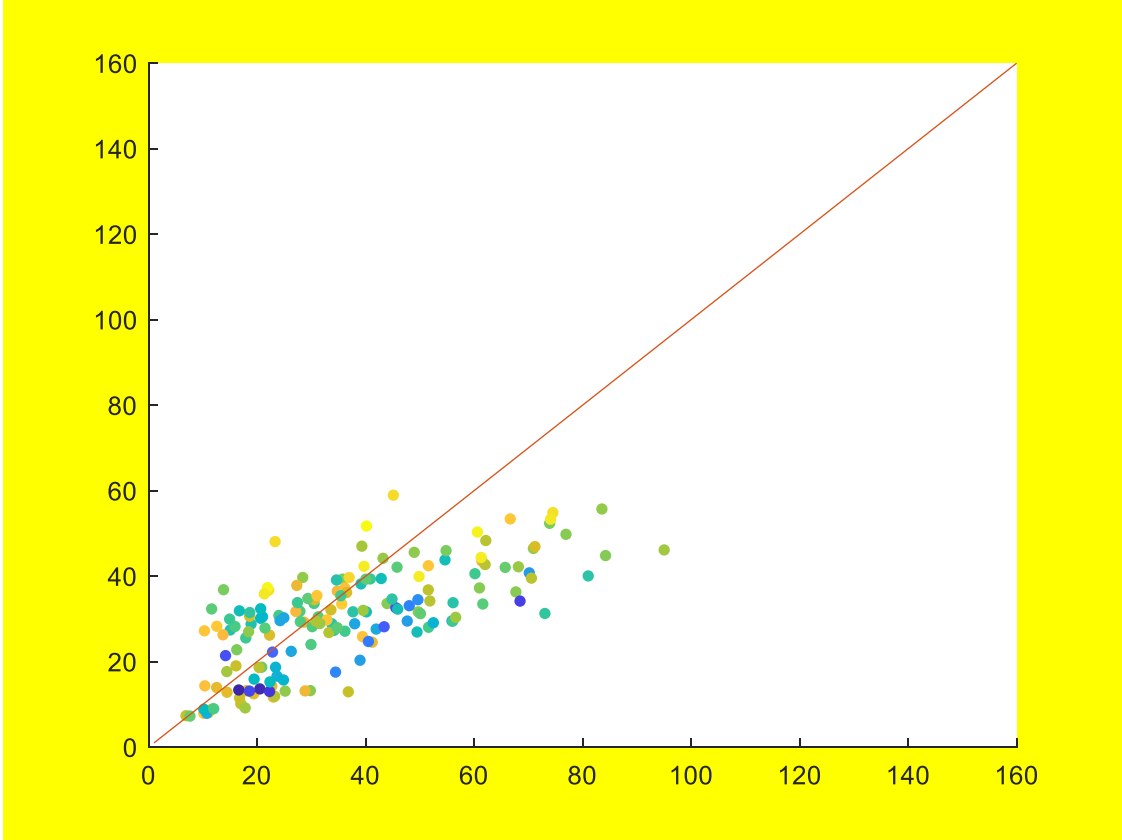
PMs7003 calibration towards PM_{2.5}

PM2.5



PM_{2.5} True Concentrations (ug/m³)

PM10

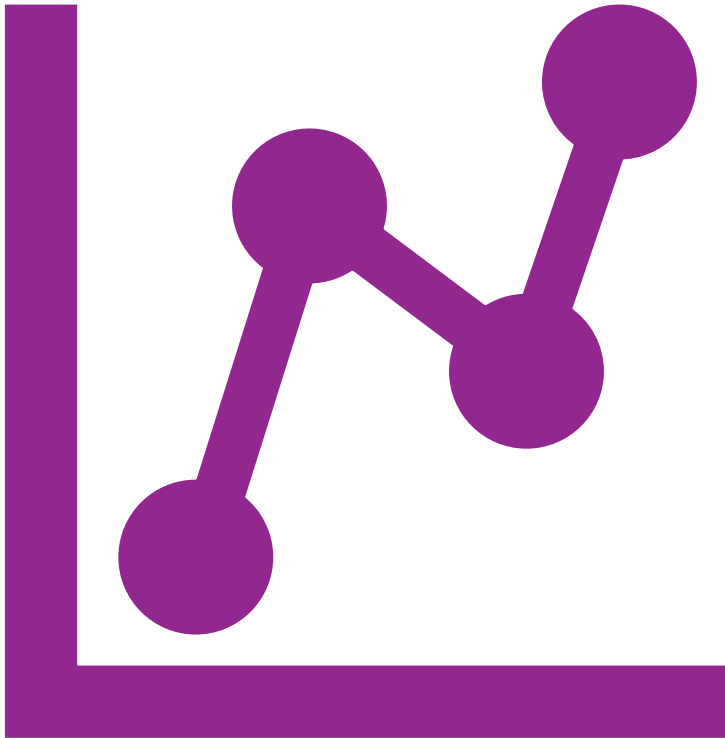


PM_{2.5} True Concentrations (ug/m³)



Quantitative Analysis

How to quantitatively capture the performances?



- We need a set of indicators which can capture both precision and accuracy.
- We need them to be «universally» recognized and grasped by stakeholders

Resorting to scientific literature and regulatory standards we usually find:

MAE, RMSE (and CRMSE, NRMSE), MRE, MBE, MAPE, R, R², REU, etc.

Some relevant performance indicators

MAE – Mean Absolute Error

Capture the accuracy (and precision) by evaluating the average absolute estimation error

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

MBE – Mean Bias Error

Highlights the existence of a bias, a systematic under/over estimation issue

$$MBE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n}$$

MPE – Mean Relative/Percentage Error

Normalize the absolute error for each estimation to the true value

$$MPE = \frac{100\%}{n} \sum_{t=1}^n \frac{a_t - f_t}{a_t}$$

MAPE – Mean Absolute Percentage Error

Normalize the mean absolute error to the range of the relevant target

$$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

Warning: No indicators is perfect, low MAE can reflect an unnoticed high relative error when dealing with low end values of the target distribution; MAPE may become extremely high when dealing with values close to 0. Scaling MAPE with the range of possible target values may help.

Some relevant performance indicators

MAE – Mean Absolute Error

Capture the accuracy (and precision) by evaluating the average absolute estimation error

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

MBE – Mean Bias Error

Highlights the existence of a bias, a systematic under/over estimation issue

$$MBE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{n}$$

MPE – Mean Relative/Percentage Error

Normalize the absolute error for each estimation to the true value

$$MPE = \frac{100\%}{n} \sum_{t=1}^n \frac{a_t - f_t}{a_t}$$

MAPE – Mean Absolute Percentage Error

Normalize the mean absolute error to the range of the relevant target

$$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right|$$

Warning: No indicator is perfect neither complete, low MAE can reflect an unnoticed high relative error when dealing with low end values of the target distribution; MAPE may become extremely high when dealing with values close to 0. Scaling MAE (NMAE) with the range of possible target values may help.

You may also find that some indicators definitions differ according to different authors!

Some relevant performance indicators

r – Pearsons' correlation factor

Capture the strength of a linear relationship between the estimation and reference time serie.

$$\frac{\frac{1}{n} \sum_{i=1}^n (M_i - \bar{M})(M_i - \overline{RM})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (M_i - \bar{M})^2 \frac{1}{n} \sum_{i=1}^n (RM_i - \overline{RM})^2}}$$

R²- Coefficient of determination

Assess the fraction of target variance explained by the model

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

FOEX- Factor of Exceedance

Measures the over or under estimation of studied measurements against reference data.

$$100 \times \left[\frac{N(M_i > RM_i)}{N_{\text{total}}} - \frac{1}{2} \right]$$

RMSE – Root Mean Squared Error

Magnitude of Error, sensitive to outliers

$$\sqrt{\frac{1}{n} \sum_{i=1}^n (M_i - RM_i)^2}$$

Warning: None of these indicators is perfect, too. RMSE is sensitive to outliers and the same concerns for MAE also applies. FOEX has the same value for perfect fit and total underestimation. R² should be careful tested for the actual definition that has been used. Pearson's r just measure a linear relationship strength but accuracy may be low due to bias and so on.... Only a set of indicators may contribute to a regression analysis.

PM2.5 MultiLinear Correction Results

Results obtained with averaging time stratified set crossvalidation performances (2weeks training, 1 week test) over 30 MONICA devices

CalFunctional	MAE	R ²	RMSE	CRMSE	NMAE
Original	11.0823	0.0639	16.8503	0.9619	0.1170
MLR	7.8503	0.5454	11.0363	0.6667	0.0969
GMLR	9.1176	0.0885	14.2759	0.8386	0.1102
NN	8.5958	0.4638	11.9292	0.7151	0.1056



Is that truly simple? And scalable?

We should not fool ourselves.

This simple calibration approach required

1. a multiweeks colocation experiment with reference analyzers
2. deriving the calibration for multiple devices

This strongly limits the scalability especially when dealing with hundreds of analyzers.

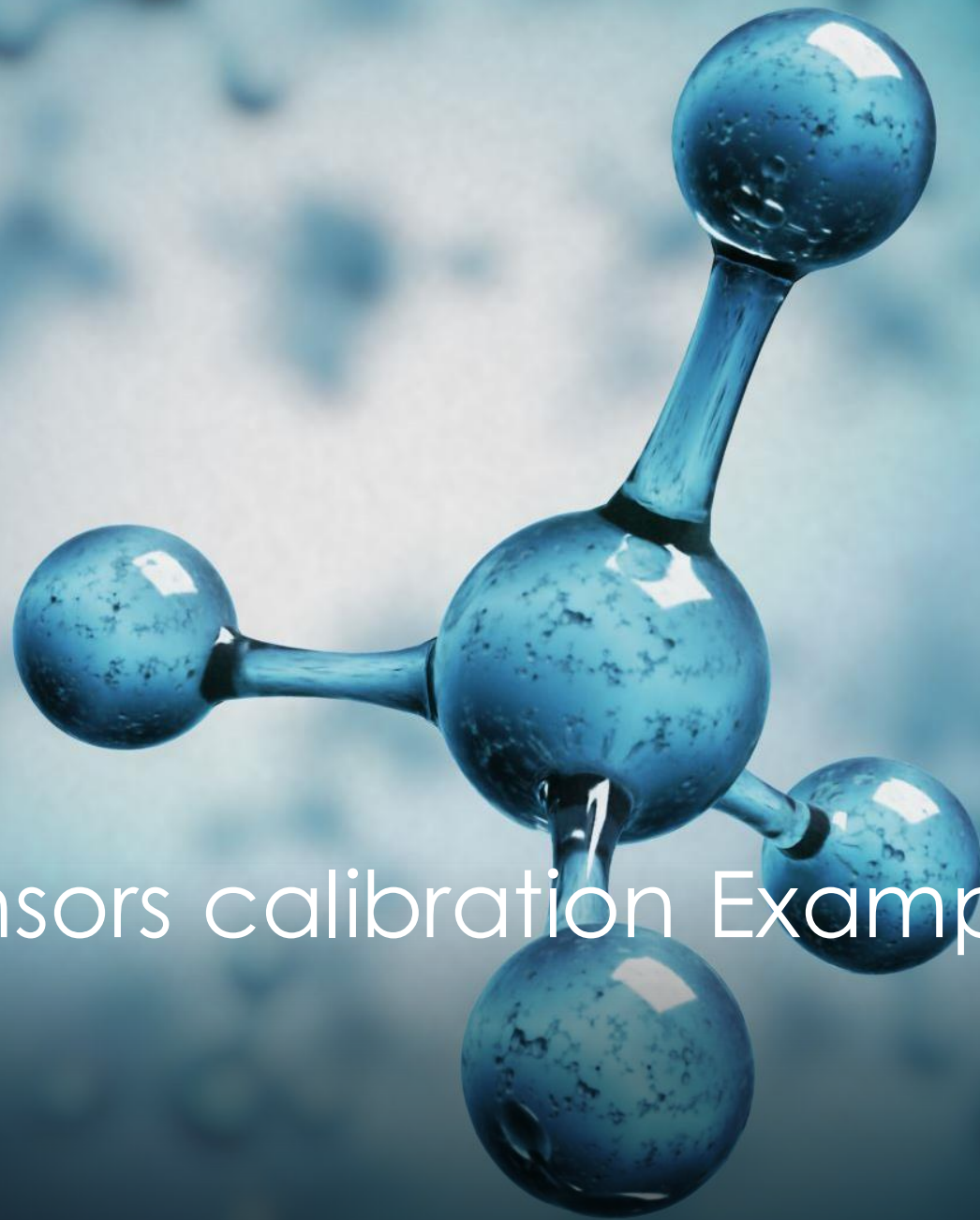
While field calibration remains the most accurate approach we should go beyond. More on this later....



(a)



(b)



Gas Sensors calibration Example

Calibrating Gas Sensor : The EC case

- EC are the most reliable and accurate solution for outdoor AQ monitoring up to now.
- They are at the core of several proven commercial solutions
- Unfortunately they are prone to cross interferences and environmental sensitivity

Let' s have a look to possible solutions

EC Sensors characterization: The Alphasense case

- Alphasense A4/B4 classes are one of the most tested sensors class in the literature
- Their estimations are based on Working Electrode potential wrt Reference electrode.
- One of the most common interferent is temperature (but known interferents are also RH and T transients, Pressure and non target gases e.g. NO₂ with O₃ sensor)
- An Auxiliary electrode provides for an estimation of temperature influence on WE potential. But correction is not exact!



How to derive a field calibration for EC Sensors

Vendor distribute calibrated sensors which reports sensitivity S and Zero air response on both WE and AE. As such we can derive the following simple calibration scheme:

$$\text{Concentration (ppb)} = \frac{1}{S} [(Vwe_{measured} - Vae_{measured}) - (Vwe_{zero} - Vae_{zero})]$$

Some basic correction is also provided by using a Look Up table which allow to correct the V_{AE} for temperature interference.

So far, this approach just won't work in the field:

- It does not take into account non target interferences
- It does not take into account sensor fabrication variability

You may also try to build Your own LUT by measuring Your sensor in the lab!

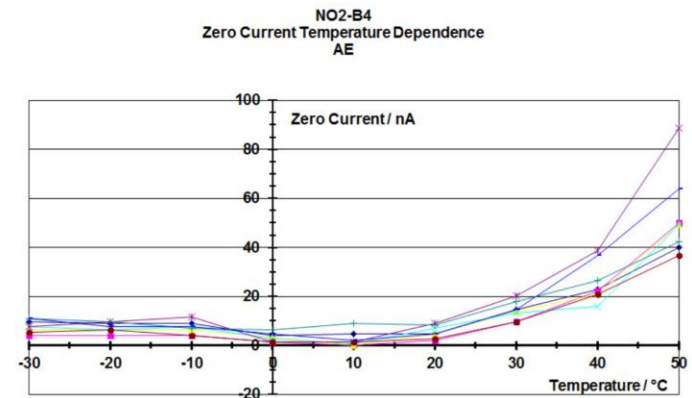


Fig. 4 Plots of zero currents for the AEs in NO2-B4 sensors as a function of temperature

A Data driven approach

- Tune a black box model using field/lab calibrated data:

$$C = f(X), \quad X = \begin{bmatrix} WE_{NO_2} \\ AE_{NO_2} \\ T \end{bmatrix}.$$

And try different models e.g. MLR, Shallow Neural Networks, RF, etc.

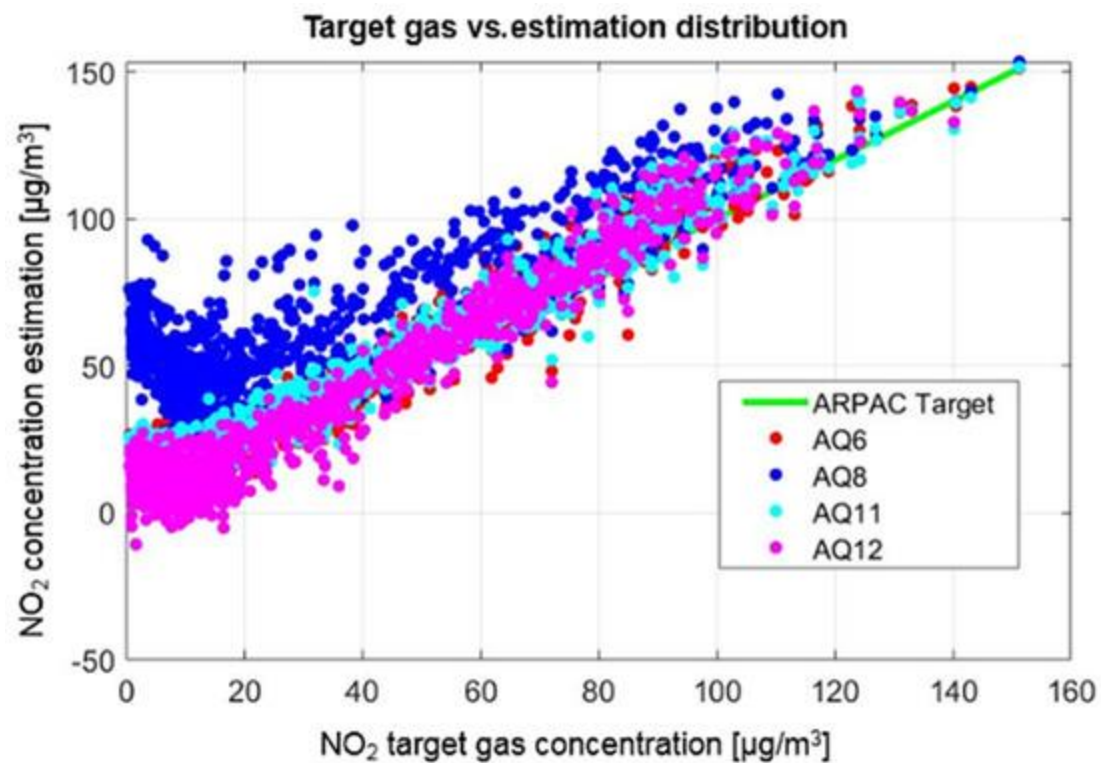
Results for MLR:

Model:

$$y = X\beta + c,$$

Target: NO₂

Primary Interferent: Temperature



Compare MLR and SNN:

Short Term performance, 3 Months Colocation

Winter Time -> High Pollutant concentrations

4 MONICA Devices based on Alphasense A4s

- 3-4 Weeks are optimal
- Best MAE ranges from about 5ug/m³ to 12ug/m³
- Best R² ranges from 0.7 to more than 0.9
- Similar results obtained by MLR and ANN

Table 2. Mean absolute errors: (a) Mean Absolute Error, (b) Pearson correlation coefficient and (c) coefficient of Determination (R²) for NO₂ estimations obtained using two calibration models with different choices for the training length (L, in weeks) for each node. Bold indicates the performance level that was best achieved.

L	Mean Absolute Error (MAE) [$\mu\text{g}/\text{m}^3$]							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	11.7	7.94	21.94	23.36	8.20	7.78	12.23	6.55
2	7.53	7.70	25.64	16.78	10.07	9.51	8.82	6.92
3	8.89	7.73	19.48	13.30	10.09	8.86	8.33	6.49
4	8.74	7.56	11.71	12.63	10.24	9.88	7.08	6.31
5	7.98	7.63	13.15	11.37	9.6	9.65	5.79	5.15

L	Pearson Correlation Coefficient r							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	0.93	0.97	0.94	0.93	0.97	0.97	0.93	0.98
2	0.97	0.97	0.92	0.94	0.97	0.97	0.98	0.98
3	0.97	0.98	0.93	0.94	0.97	0.97	0.98	0.98
4	0.97	0.98	0.95	0.95	0.98	0.98	0.98	0.98
5	0.98	0.96	0.96	0.96	0.98	0.98	0.98	0.98

L	Coefficient of Determination R ²							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	0.79	0.91	0.47	0.41	0.91	0.92	0.78	0.94
2	0.91	0.9	0.22	0.62	0.85	0.88	0.88	0.92
3	0.88	0.89	0.49	0.74	0.86	0.88	0.89	0.92
4	0.87	0.88	0.77	0.75	0.84	0.84	0.91	0.93
5	0.88	0.88	0.75	0.81	0.87	0.87	0.94	0.95

A close-up, shallow depth-of-field photograph of a clothing rack. The rack is filled with various shirts, including solid colors like red, pink, and blue, as well as patterned shirts with small dots. The shirts are hanging on dark-colored hangers. The text 'Models Evaluation & Selection' is overlaid in white, sans-serif font across the lower portion of the image.

Models Evaluation & Selection

Evaluating and Selecting Models

- So Far, the presented results and methodology omits to describe the evaluation process.
- Several black box model have been reported. Comparisons highlighted that many hold similar results.
- To avoid overoptimistic results, indicators have been computed on estimation performed during a set-apart «test set» as opposed to the so called «training set»
- But how to select the appropriate partition of the dataset....?

Which base model?

- So far comparison literature showed no clear «winner», if adequately optimized with fair choice of hyperparameters values they provide similar results
- Typical examples are Multilinear regression (see before), ANN (shallow architectures), Random Forests (shown to provide bad generalization properties), SVMs
- Hardware for operation can lead the choice.
- Depending on applications, recurrent architecture may provide a performance boost in fast transients.

One Clear lesson: Keep it simple!

Simple models provide better generalization avoiding overtraining.



Dataset partitions, How to?

- The main goal is to provide realistic evaluation of the accuracy so to:
 - Avoid overoptimistic conclusions coming from overtraining
 - Selecting the right amount of needed data (cost/accuracy trade off)
 - Selecting the best model in terms of generalization to real world conditions
- The most important question is How will I use the model?
 - During short term (≤ 3 Months campaigns?)
 - During long term campaigns?
 - Have I (or will I have) multiple seasons data?
- Then.... How many data do I Have?

Dataset partitions, How to?

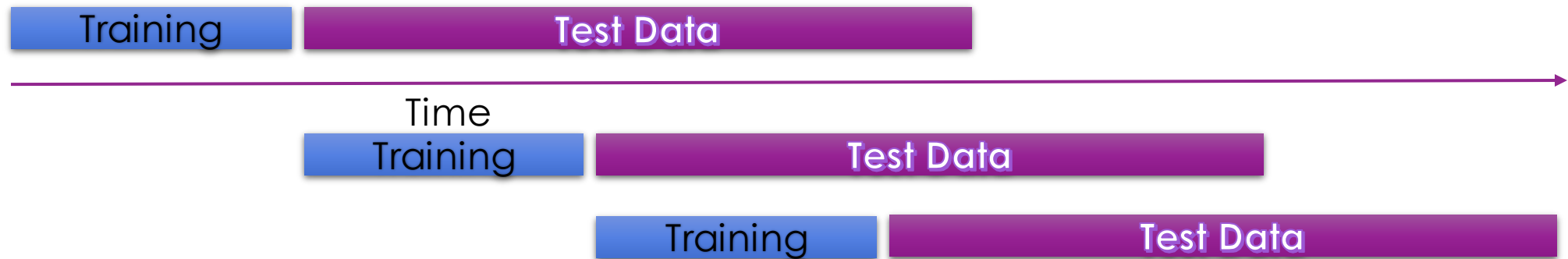
- First rule!
 - Avoid correlation between training and test data (more on this later)
- If I have enough data:



- However performance may depends on the peculiar conditions during training and test time periods.....

Dataset partitions, How to?

- Second rule:
 - Avoid being dependent on specific training/test conditions -> cross validate!
- If You have enough data:



- Average Your performance indicators across different training/test cycles.

Dataset partitions, How to?

- Second rule:
 - Avoid being dependent on specific training/test conditions -> cross validate!
- If You feel, You **don't** have enough data:



- Slightly overoptimistic but one of the best approach in these conditions
- Average Your performance indicators across different training/test cycles.

An Example of case 2

- NO₂ targeted calibration
- Comparing MLR and SNN
- Long term (from one year to >2yrs)
- 1 yr ->SNN and MLR hold similar results
- 4 weeks obtain best figures
- R² falls significantly on 2 yrs exp.
- 2yrs -> MLR offers better generalization

Table 5. Calibration performance indicators for NO₂ estimations obtained using two calibration models with different choices of training length.

(a) NO ₂ calibration with cross-validation (CV) (April 2018–July 2019).														
Training Set Length		MAE (µg/m ³)		STD		RMSE (µg/m ³)		NRMSE		R ²		R		
		MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	
1 week	(Mean)	16.91	15.64	14.59	12.92	22.35	20.33	0.92	0.84	-0.15	0.20	0.72	0.69	
	(Median)	13.85	13.73	12.22	11.97	18.40	18.29	0.76	0.76	0.42	0.43	0.78	0.74	
2 weeks	(Mean)	13.90	14.89	12.08	13.00	18.43	19.79	0.76	0.81	0.40	0.25	0.76	0.69	
	(Median)	13.80	13.61	11.23	11.88	17.87	17.93	0.74	0.74	0.46	0.44	0.79	0.74	
3 weeks	(Mean)	14.42	13.85	12.98	12.30	19.42	18.55	0.80	0.76	0.23	0.39	0.76	0.72	
	(Median)	12.81	12.85	10.92	11.28	16.84	16.98	0.69	0.70	0.52	0.51	0.79	0.75	
4 weeks	(Mean)	13.02	13.34	11.40	11.92	17.33	17.91	0.71	0.73	0.49	0.42	0.78	0.74	
	(Median)	13.33	11.87	10.69	10.50	17.03	15.80	0.70	0.65	0.51	0.58	0.80	0.78	
(b) NO ₂ calibration with cross-validation (CV) (July 2019–November 2020).														
Training Set Length		MAE (µg/m ³)		STD		RMSE (µg/m ³)		NRMSE		R ²		R		
		MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	
1 week	(Mean)	18.04	18.19	14.12	13.94	22.92	22.94	0.99	0.99	-0.04	-0.05	0.60	0.54	
	(Median)	16.30	16.80	12.78	12.96	20.96	21.20	0.90	0.92	0.18	0.15	0.65	0.58	
2 weeks	(Mean)	16.13	17.73	12.63	13.62	20.50	22.39	0.89	0.97	0.19	-0.01	0.63	0.56	
	(Median)	15.59	17.11	12.10	13.12	19.79	21.68	0.86	0.94	0.27	0.11	0.68	0.60	
3 weeks	(Mean)	15.20	17.06	12.05	13.78	19.41	21.95	0.84	0.95	0.27	0.04	0.66	0.56	
	(Median)	14.46	15.71	11.55	12.65	19.01	20.06	0.83	0.87	0.32	0.24	0.68	0.63	
4 weeks	(Mean)	13.76	14.73	10.99	11.83	17.62	18.90	0.76	0.82	0.41	0.31	0.71	0.65	
	(Median)	13.96	14.59	10.99	11.53	17.74	18.53	0.77	0.80	0.41	0.35	0.71	0.66	
(c) NO ₂ calibration with cross-validation (CV) (April 2018–November 2020).														
Training Set Length	Test Set Length	MAE (µg/m ³)		STD		RMSE (µg/m ³)		NRMSE		R ²		R		
		MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	
4 weeks	4 weeks CV	(Mean)	15.09	16.55	12.40	13.96	19.54	21.67	0.82	0.90	0.32	0.12	0.70	0.61
		(Median)	14.91	15.59	12.08	12.91	19.41	20.61	0.81	0.86	0.34	0.25	0.72	0.66
(d). NO ₂ calibration ab initio (April 2018–November 2020).														
Training Set Length	Test Set Length	MAE (µg/m ³)		STD		RMSE (µg/m ³)		NRMSE		R ²		R		
		MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	MLR	SNN	
4 weeks	4 weeks	(Mean)	14.72	15.68	11.22	10.78	18.56	19.07	0.86	0.89	0.18	0.11	0.69	0.60
		(Median)	14.93	15.83	10.78	10.95	17.41	19.20	0.84	0.88	0.28	0.22	0.70	0.62



Field Calibration
Robustness

What happens if?

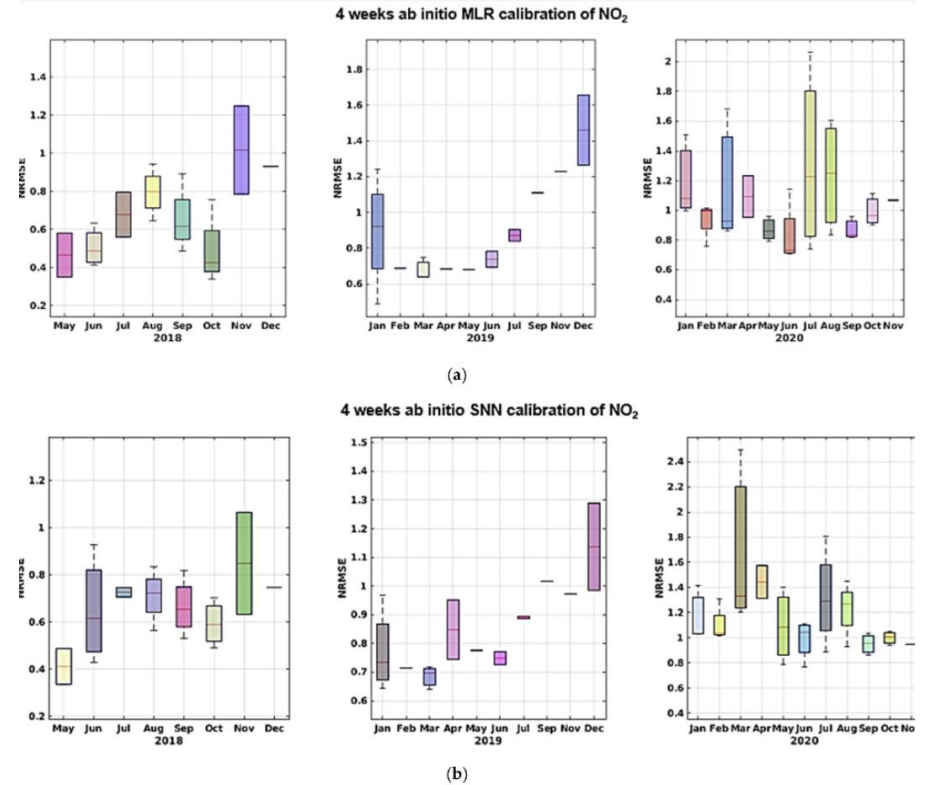


...we relocate the calibrated station? Why FC performance drops?



...why performance of FC systems drops in the long term?

Figure 18. NRMSE trends shown by monthly boxplot for ab initio calibration of NO₂ for MLR (a) and SNN (b); the latter shows slightly better figures during the first and last year.



What happens if?

The reasons lie behind change:

- Change in the pollutant ranges
- Change in the pollutant mix
- Change in the particulate composition

In one word: Change in the response eliciting forcers joint distribution

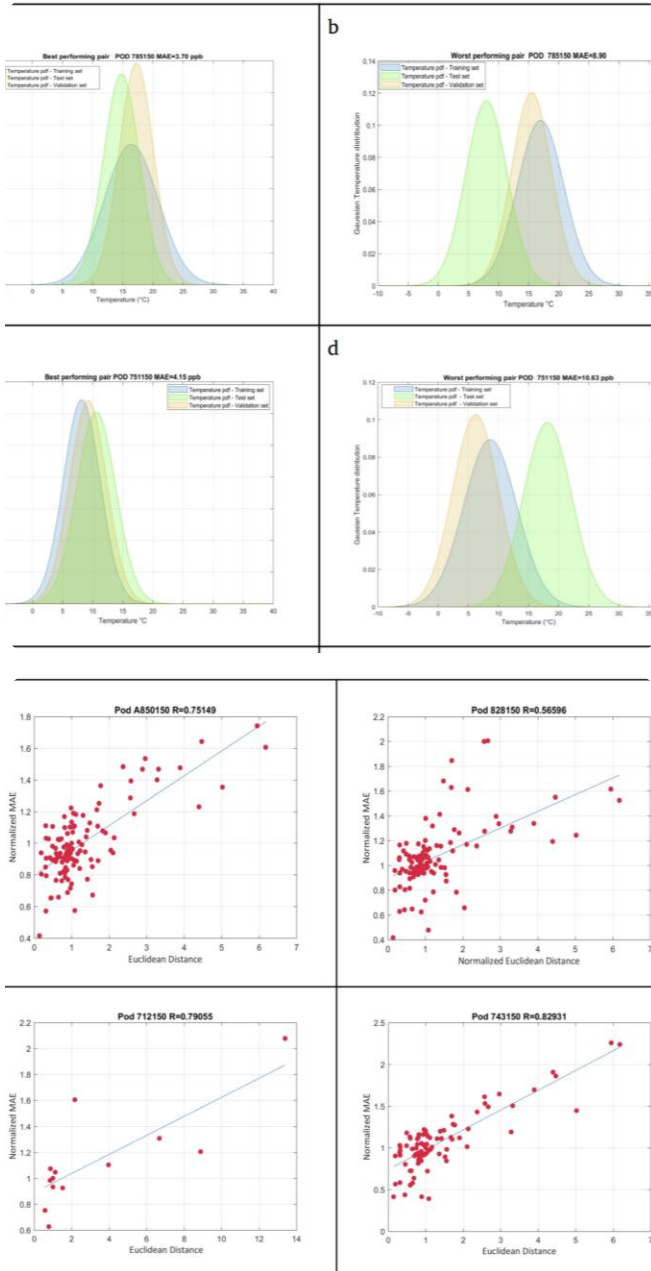


Figure 6c : Correlation plots showing the actual relationship between normalized MAE estimation and Euclidean Distance applied to joint empirical distribution $p(T, NO_2)$ for the 4 pods relocated in Akebergveien rd.

Robustness boosting strategies

Most strategies depends on boosting the calibration set completeness so to be able to face the conditions variance (recalibration either with co-location or remote calibration).

or

Improve generalization capabilities of the model (e.g. temperature dependent multiple calibration models).





Wrap Up

Take home lessons

- Chemical and PM sensors needs calibrations to optimize performances
 - Field calibration obtain the most for operating in the wild
 - Seasonalities + Relocation (and any **changes in the forcers distribution** wrt field calibration conditions envelope) -> Performance losses
 - Fabrication variance + Needs for local and periodic recalibration hinder **scalability**
 - **Remote & Global** calibration models are promising approaches to obtain the sought scalability
- 