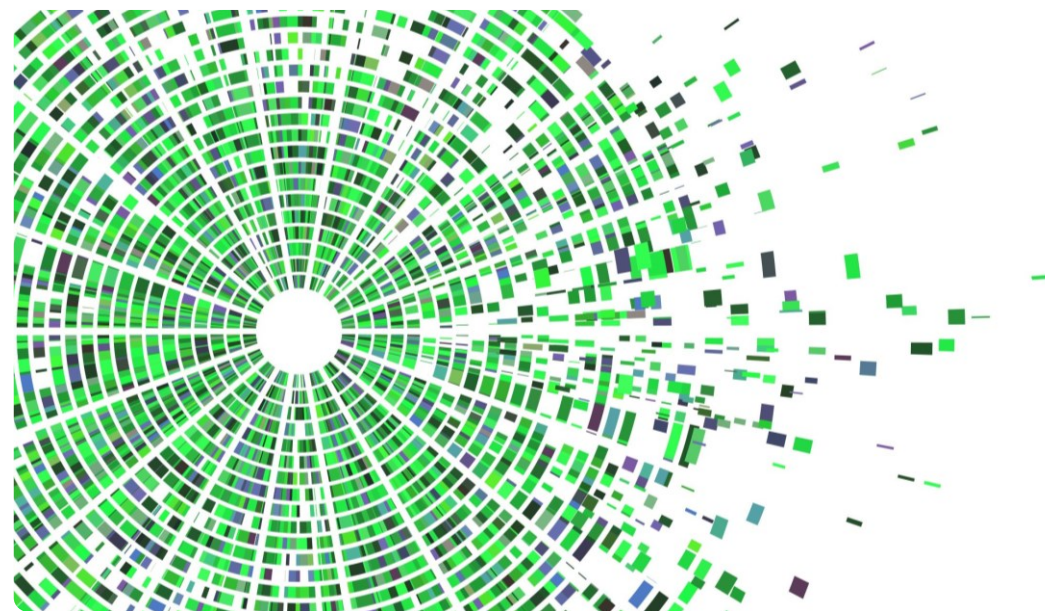


In Field Calibration and Beyond

saverio.devito@enea.it

ISOCS Winter School Bormio 2023





Sections 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?

Forewords

- The content of this lecture is focused on Air Pollutants **QUANTIFICATION** problems
- As such, many concepts are strictly related with regression with respect to the more common artificial olfaction classification problem
- More generally, due to different fields jargon, You may find a different «definition» for the calibration problem
- Here, we focus on **Low cost Air Quality Multisensor Systems**

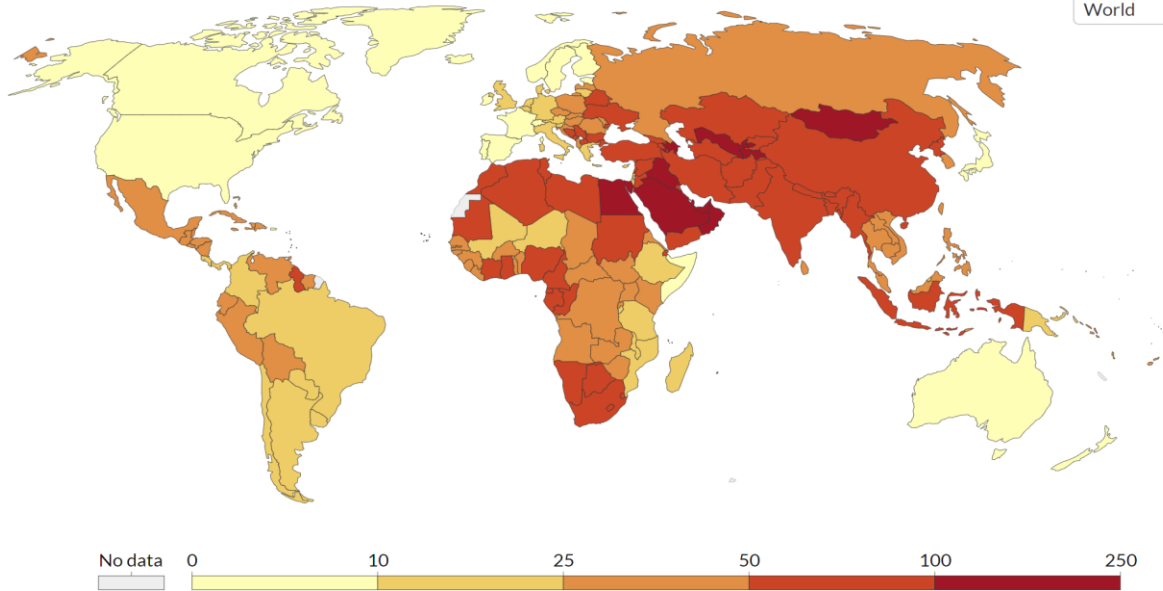
Air Pollution & Health

Death rate from ambient particulate air pollution, 2019

Death rates attributed to ambient particulate matter air pollution, measured as the number of deaths per 100,000 individuals. Death rates are age-standardized and therefore correct for changes in age structure across time and between countries.

Our World
in Data

World



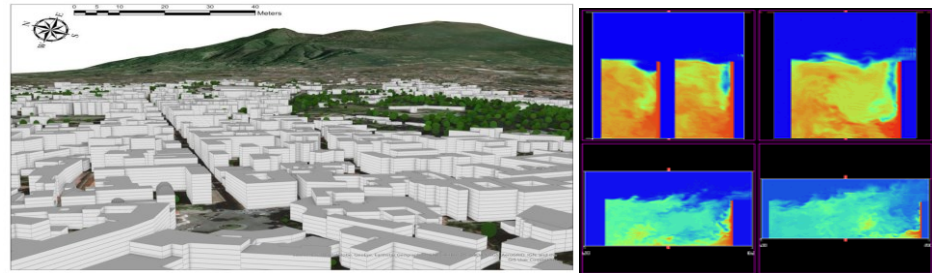
Source: Institute for Health Metrics and Evaluation, Global Burden of Disease (2019)

OurWorldInData.org/air-pollution/ • CC BY

- AQ is the most important environmental factor in determining human health
- While improving in some countries it is still on the rise for many of them
- Knowing the composition (which pollutant?, how much?) along with the spatial (where?) and temporal variance (when?) empowers citizens and administrations to devise the right remediation policies.

However, there is a widely shared fact based opinion about the lack of AQ information!

Regulatory Monitoring Networks: The Naples case.



- Regulatory AQM network in Naples metropolitan area.
- 8 Stations are currently used for AQ monitoring in the Naples urban area (117Km², 955k inhabitants).
- **Roughly, that accounts for one station for each 15Km² and/or one station each 120k+ inhabitants!**
- It is worth to note that this is one of the most dense network in Europe and it is perfectly in line with the regulating EC directive.
- **As a results small towns have limited knowledge of what happens at their urban scale.**

A relatively dense regulatory grade monitoring network which leaves many densely inhabited area with limited knowledge on AQ
 AQ at hyperlocal scale (civic number) is largely unknown

Moving from sparse to dense, hierarchical AQ Monitoring

Advances in IoT and chemical sensors calibration technologies have led to the proposal of **Hierarchical air quality monitoring networks**.

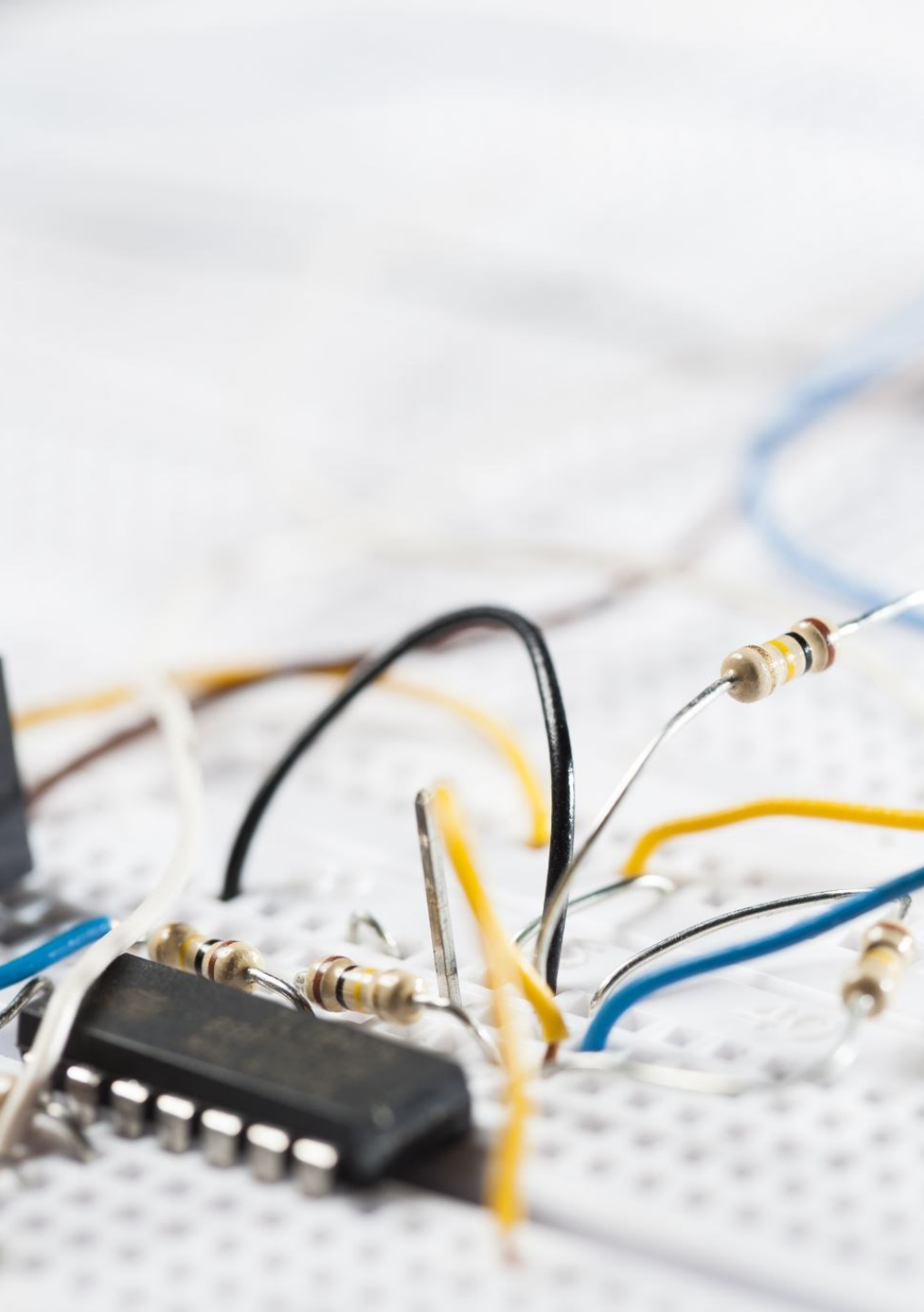


These relies on sensing nodes which differs from size, cost, accuracy, technology, maintenance needs while having the potential to empower communities with increased knowledge on the highly spatiotemporal variance Air Quality phenomenon.



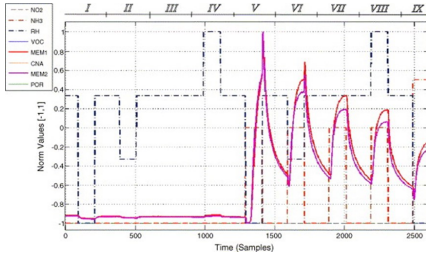
The basic motivation (1)

- In AQ applications, 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/ indirectly suggest a **«one size fit all»** calibration to be used for all sensors of that specific class.



The basic motivation (2)

- 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.



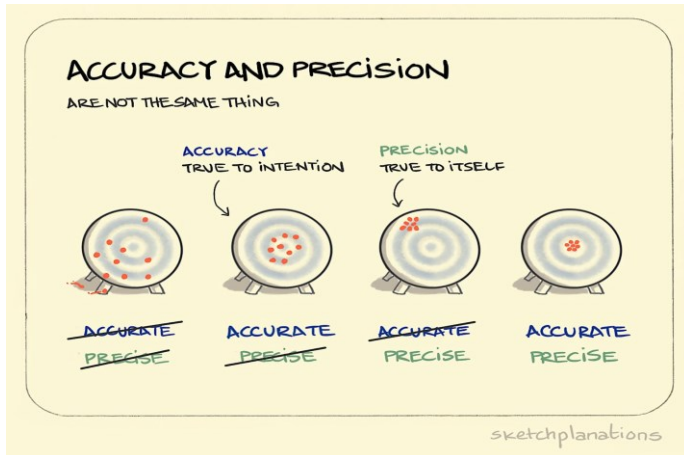
$$Y(t) = f(x(t))$$



The basic motivation (3)



- Out of Box, Chemical and particulate matter sensors are subject to:
 - Fabrication Variance (specs change from unit 2 unit)
 - Interferences from Non-Target gases
 - **Interferences from environmental parameters**
 - (Individual) Drift (Ageing, Poisoning)

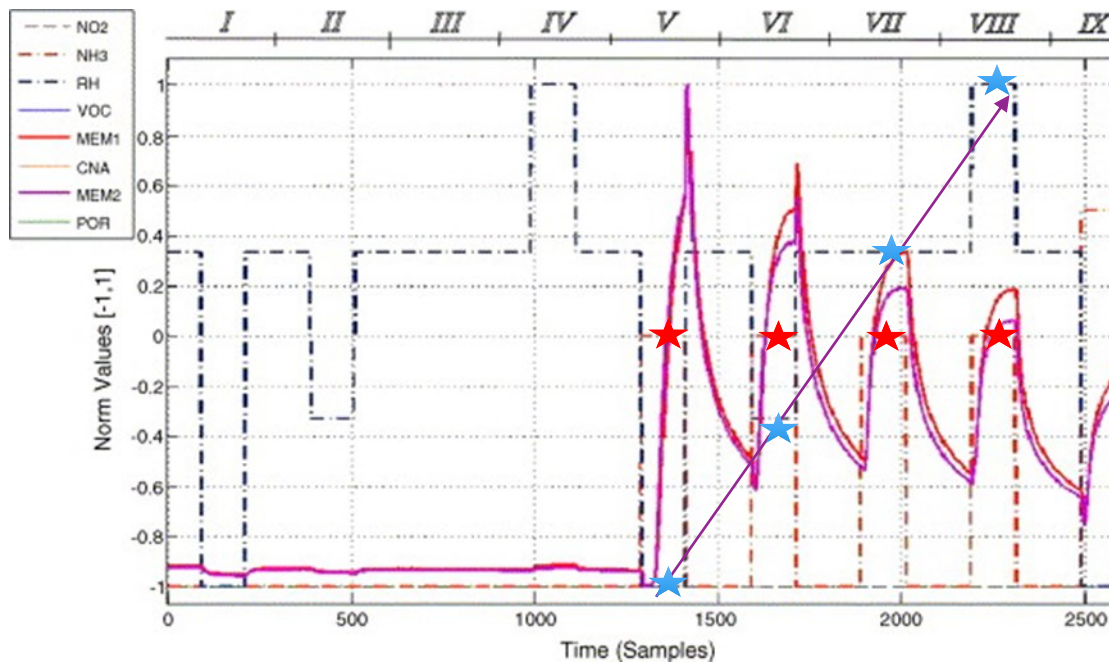


All these factors contribute to hinder original «one size fit all» (if any) vendor calibration strongly limiting the sensors accuracy & precision!

..... But...is it always true? Sometimes not (will get back to this)

Example of Environmental Interference (1)

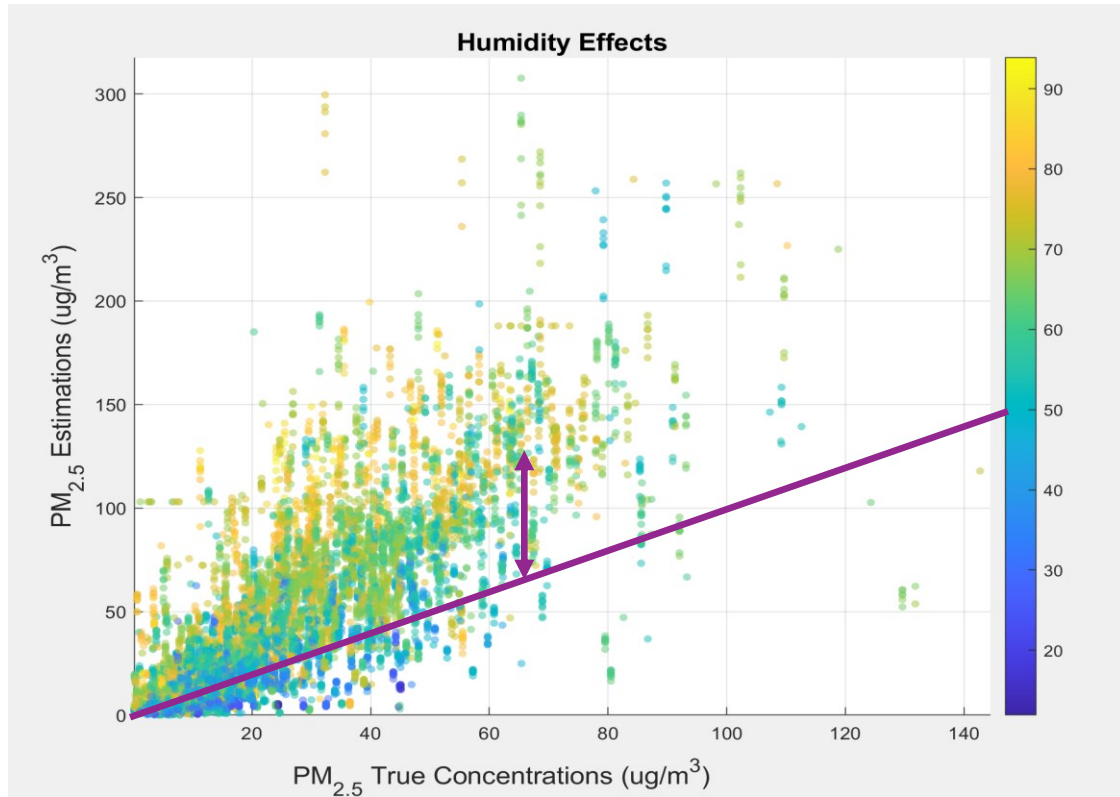
- Here, a MOX sensor is exposed to NH_3 in a humid carrier.
- We note that depending on RH levels, the sensor response to the same concentration of NH_3 is quite different!



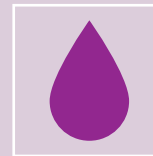
★ NH_3 Concentration

★ RH Level (%)

Example of Environmental Interference (2)

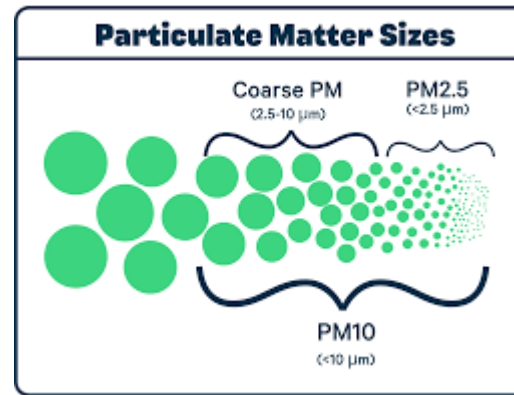
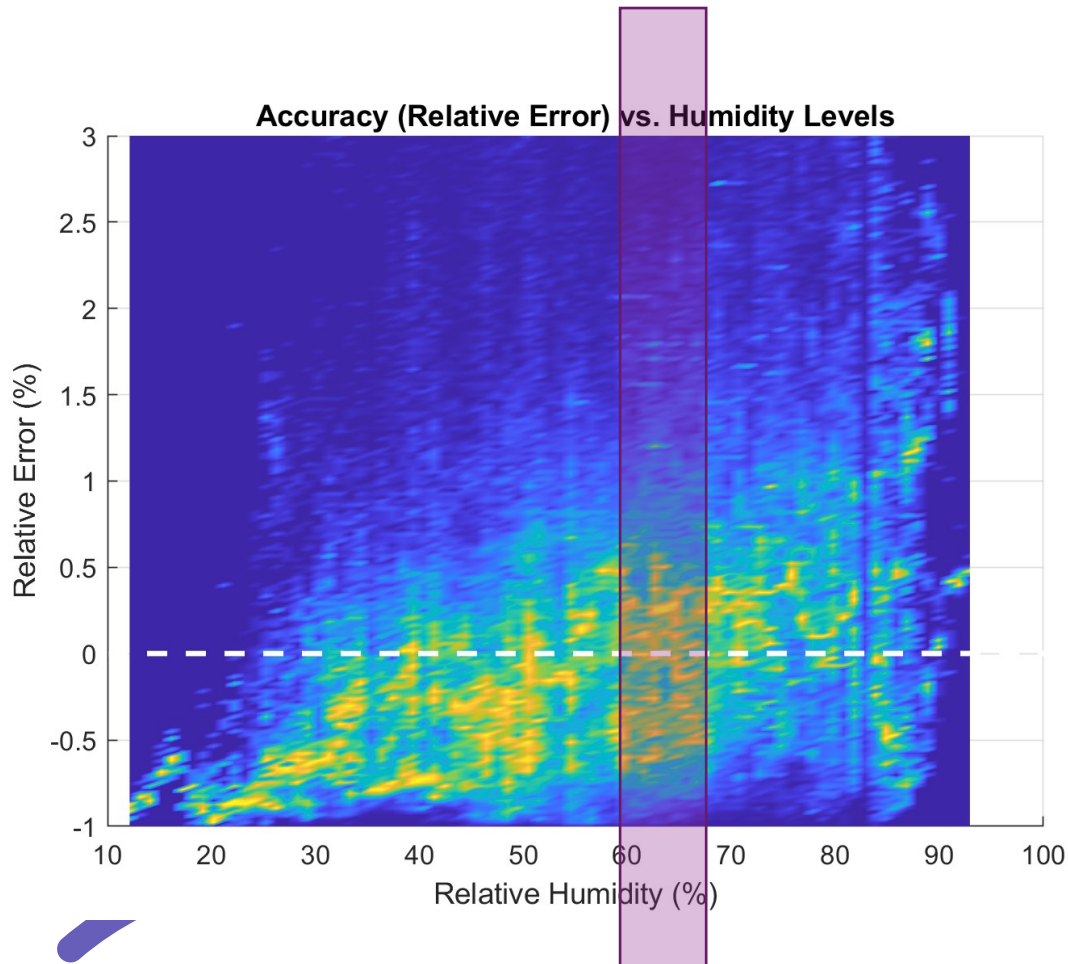


Here, a Plantower PMS7003 OPC is exposed to different concentrations of Particulate in the field.



We note that, the sensor response to the same concentration of PM_{2.5} is statistically and positively correlated with RH concentrations.

Example of Environmental Interference (2)



PM Sensors are optical detector which detect and measure the size of particles. As a result we obtain a size partitioned mass concentration

At high relative humidity, a water layer recover the particle surface modifying its density and disrupting the density hypothesis on which thier mass is estimated. As a result, vendor calibration accuracy is severely hampered.

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 $x(t) = [k_1(t), \dots, k_i(t), \dots, k_n(t)]$ a vector of all relevant **(known and observable)** interferences or information about them*

... 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 suit 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**



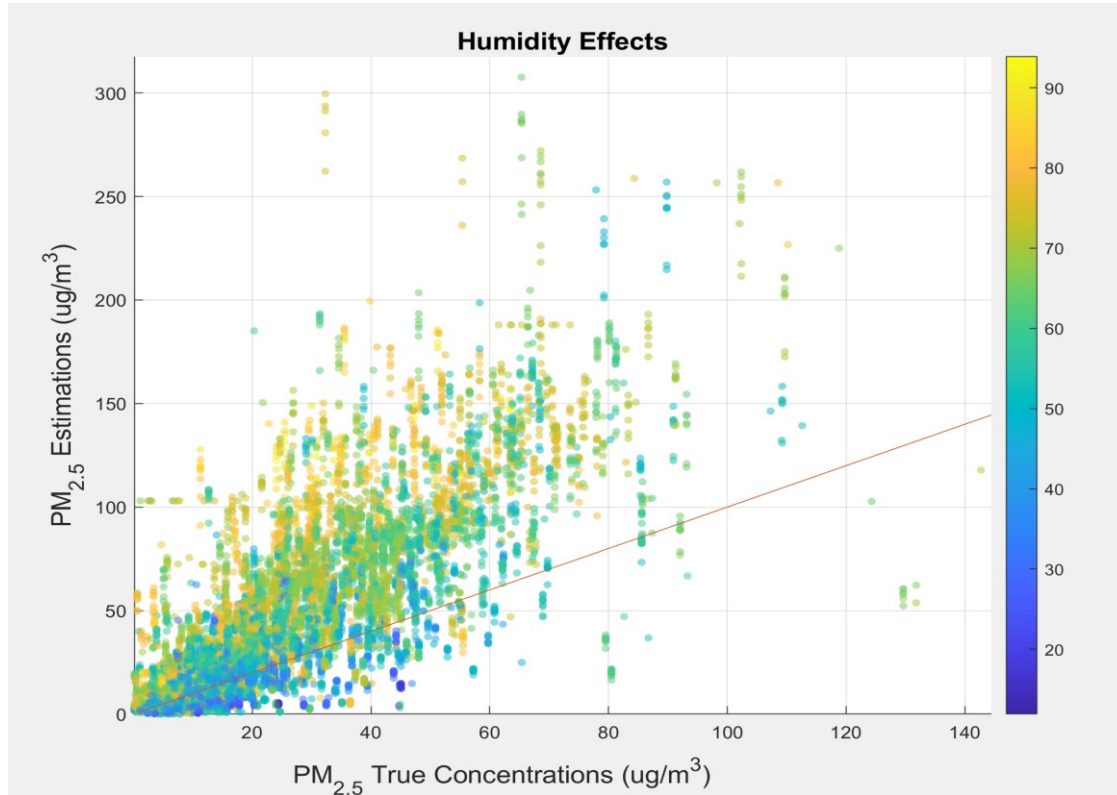
Laboratory calibration vs Field calibration

- 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 (indirectly by cross sensitivity)



Example :
PMS7003
Calibration

PMs7003 calibration towards PM_{2.5}



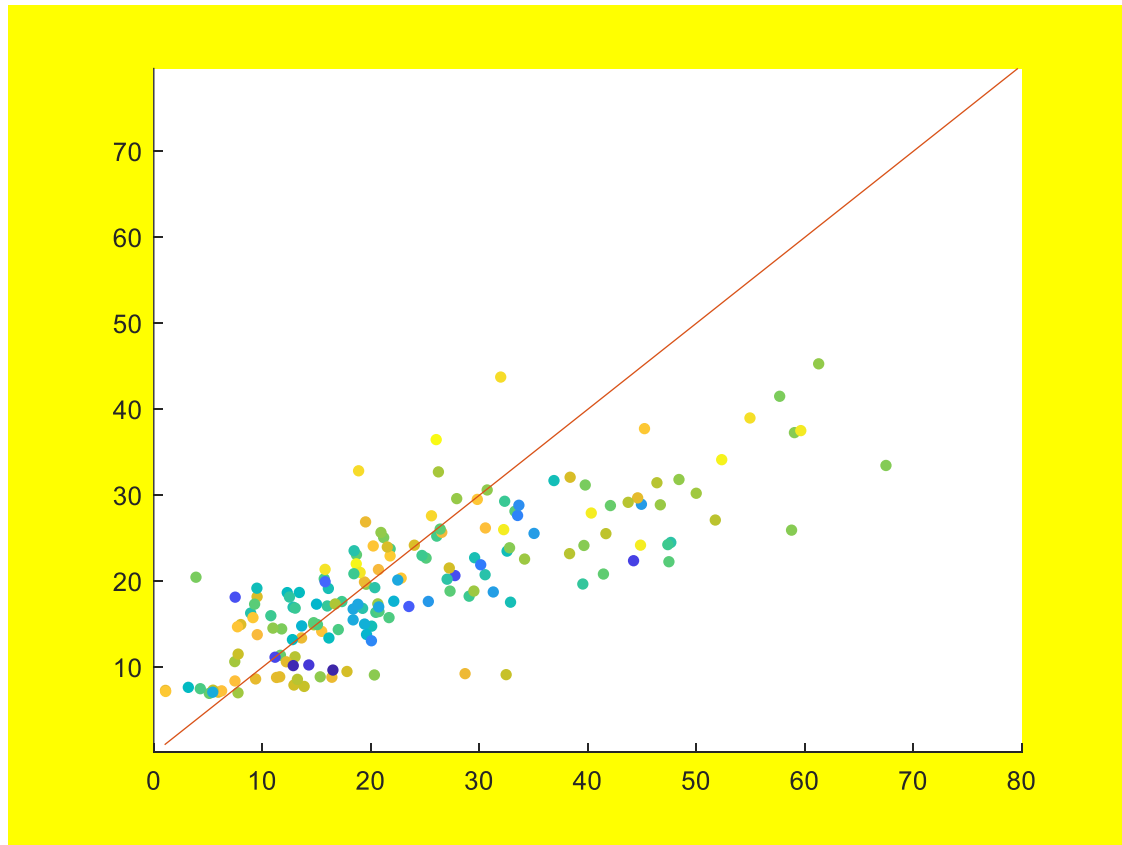
- Let's try to correct:
 - Overestimation tendency
 - Humidity interference
- We may try to derive a black box model based on linear response hypothesis:

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

... with a conventional OLS method

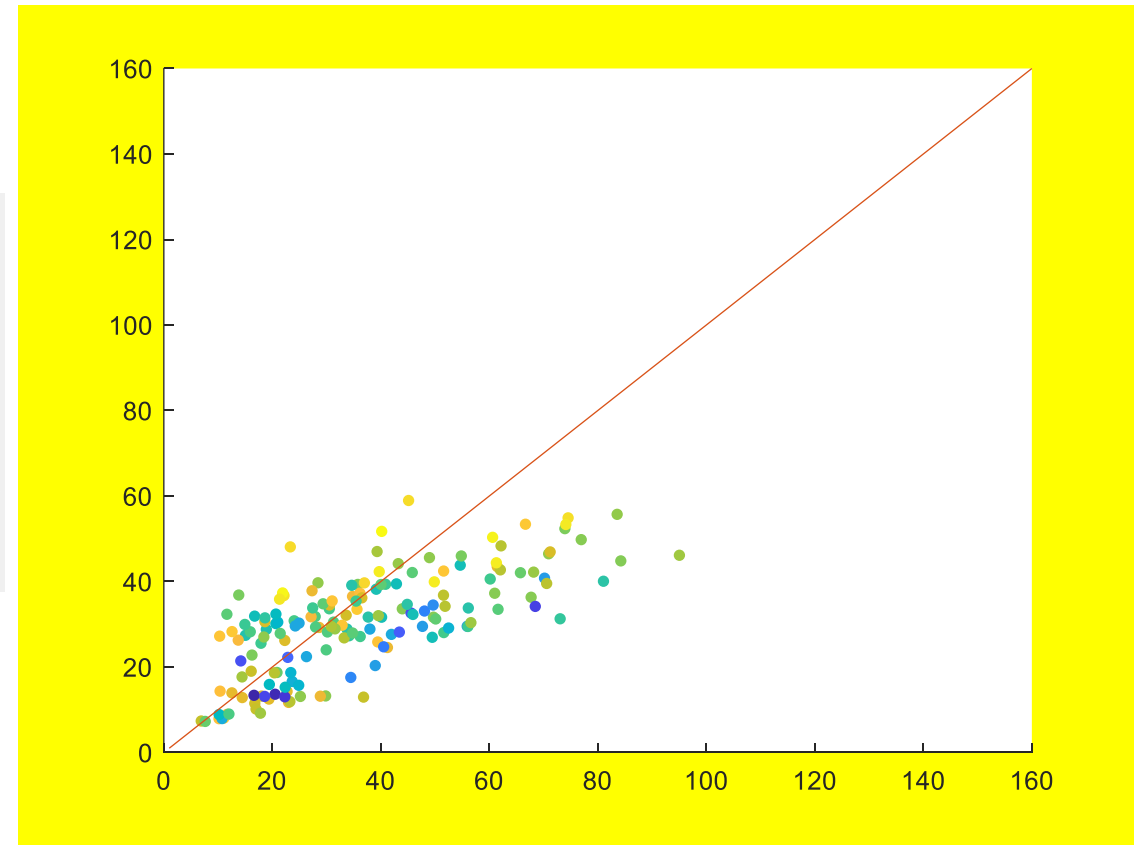
PMs7003 calibration towards PM_{2.5}

PM2.5



PM_{2.5} True Concentrations (ug/m³)

PM10



PM_{2.5} True Concentrations (ug/m³)



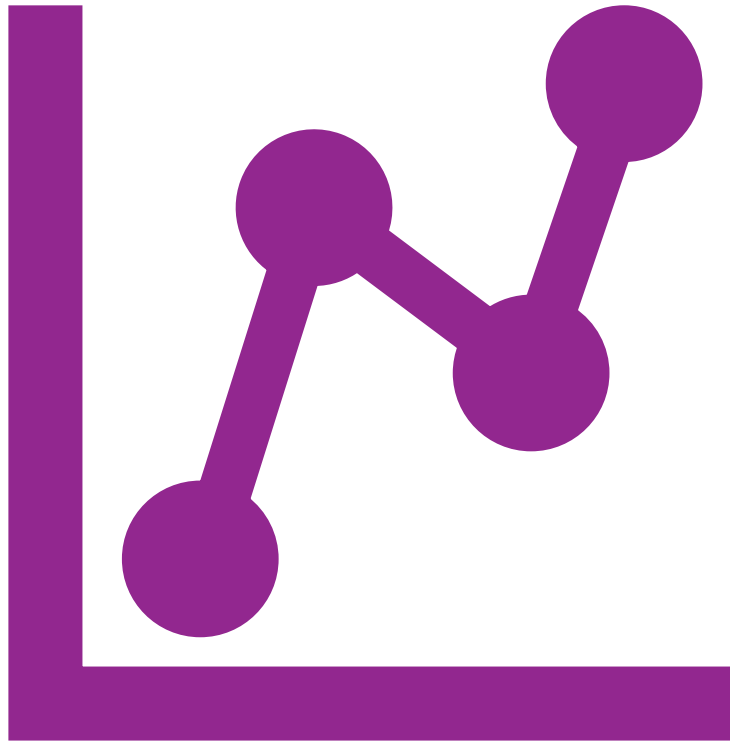
Quantitative Analysis

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 scientifically 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 – Pearson's' 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

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

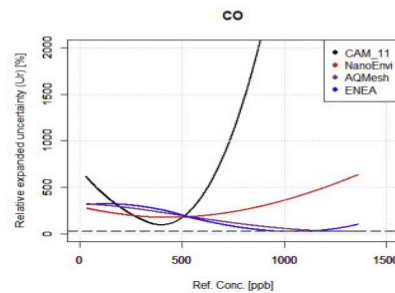
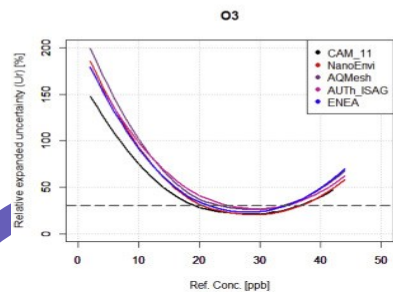
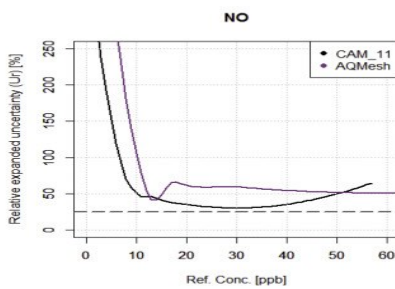
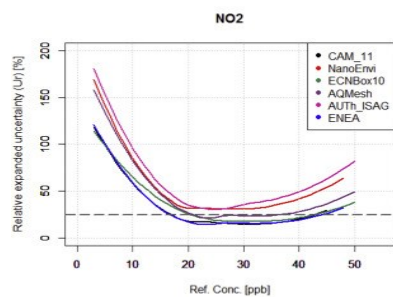
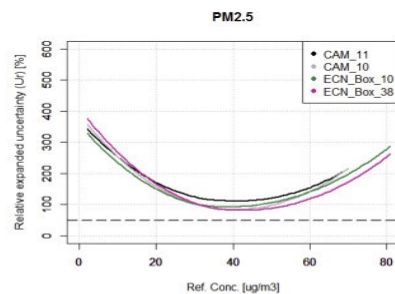
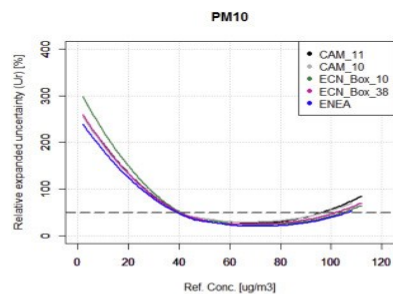
Warning: None of these indicators is perfect, too. RMSE is sensitive to outliers and the same concerns for MAE also applies. FOEX has the same value for perfect fit and total underestimation. R² should be careful tested for the actual definition that has been used. Pearson's r just measure a linear relationship 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 (ug/m3)	R ²	RMSE (ug/m3)	CRMSE (ug/m3)	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

EU Regulation Reference



Expanded Relative Uncertainty

$$U_r(y_i) = \frac{2 \left(\frac{RSS}{(n-2)} - u^2(x_i) + [b_0 + (b_1 - 1)x_i]^2 \right)^{1/2}}{y_i}$$

$$RSS = \sum (y_i - b_0 - b_1 x_i)^2$$

Defined in EU AQ Directive 2008

Gives an outlook on relative error at different concentrations

«Legal» basis for acceptability for selected applications (Indicative measurements)



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.

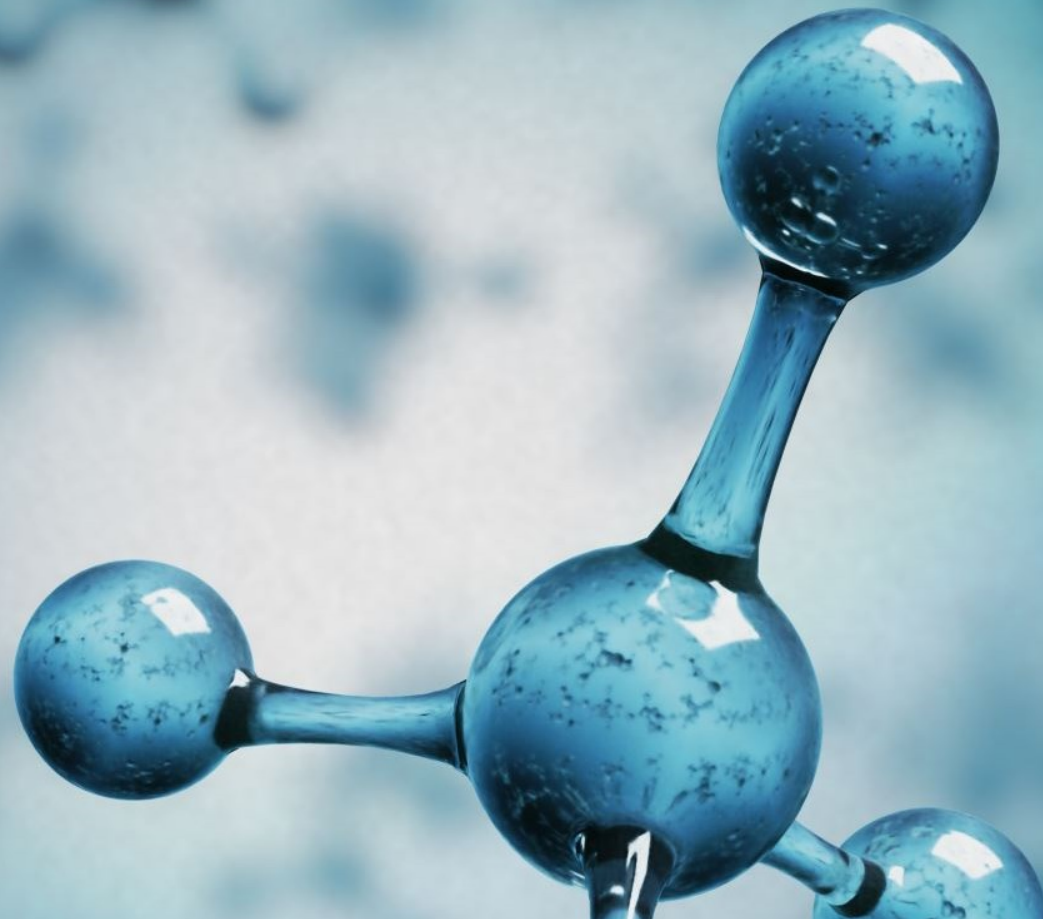
While field calibration remains the most accurate approach we should go beyond. More on this later....



(a)



(b)

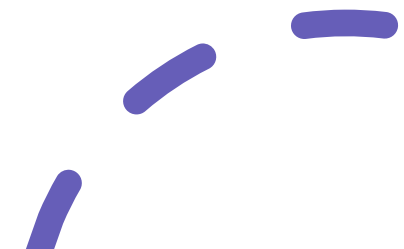


Gas Sensors calibration Example

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 fo 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 interferents
- It does not take into account sensor fabrication variability

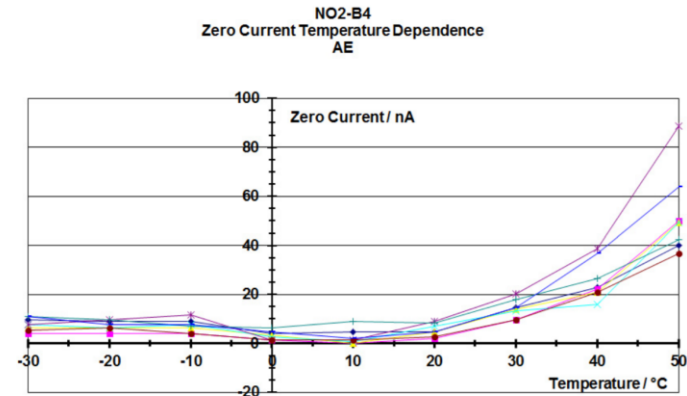


Fig. 4 Plots of zero currents for the AEs in NO2-B4 sensors as a function of temperature

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

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.

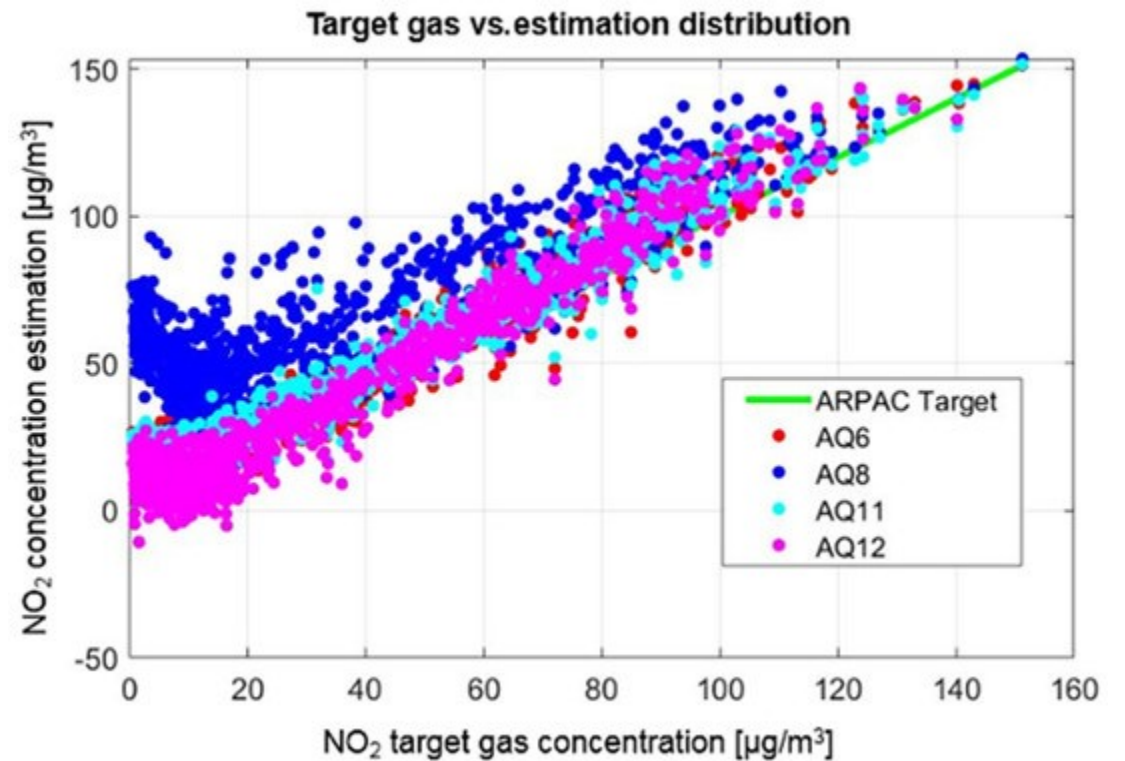
Results for MLR:

Model:

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

Target: NO₂

Primary Interferent: Temperature





Models Evaluation & Selection

Example: Using performances indices for comparing different models: MLR and SNN

Short Term performance, 3 Months Colocation

Winter Time -> High Pollutant concentrations

4 MONICA Devices based on Alphasense A4s

- Determining optimal training length: 3-4 Weeks are optimal
- Best MAE ranges from about 5 $\mu\text{g}/\text{m}^3$ to 12 $\mu\text{g}/\text{m}^3$
- Best R^2 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 (R^2) for NO_2 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.

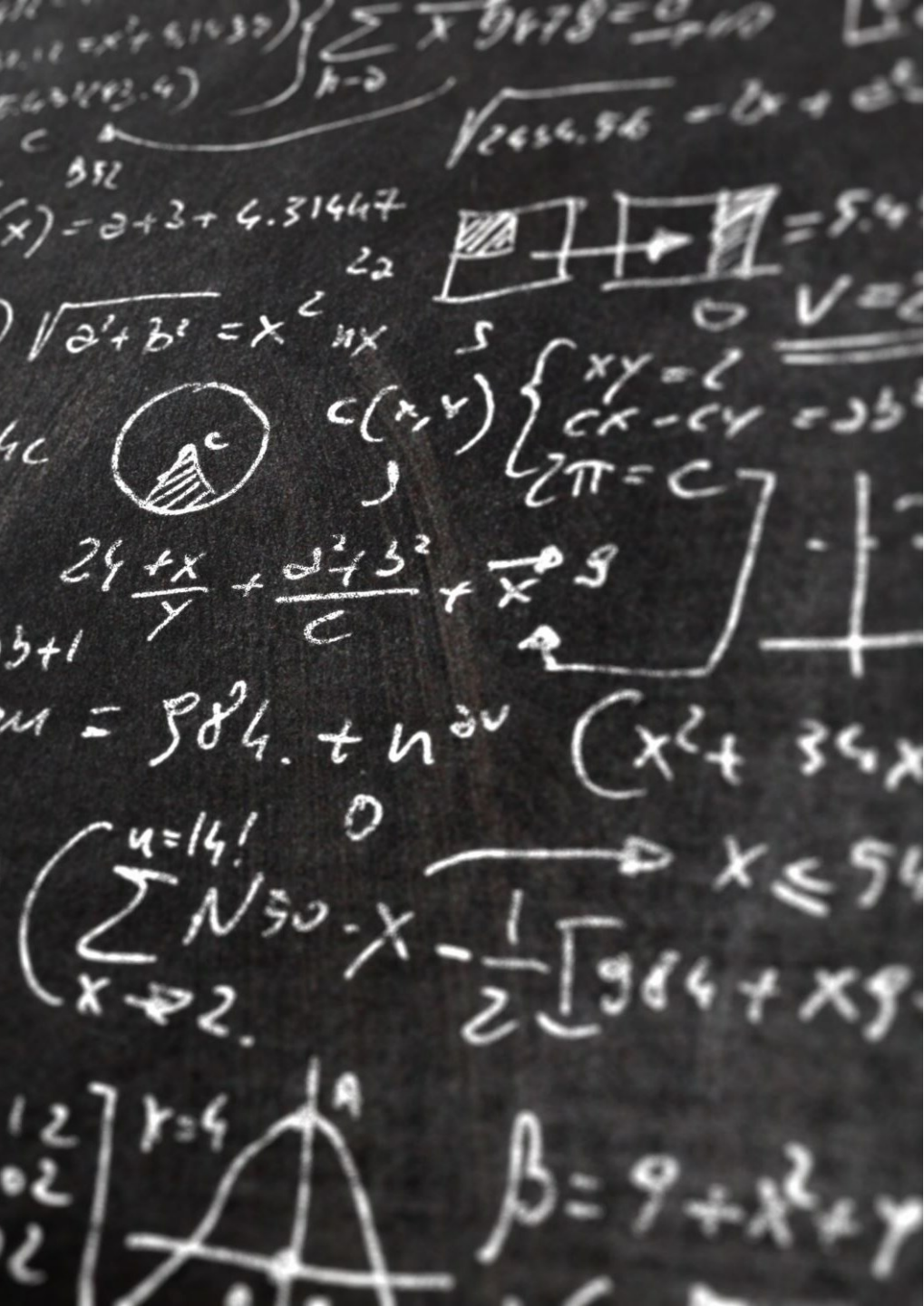
L	Mean Absolute Error (MAE) [$\mu\text{g}/\text{m}^3$]							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	11.7	7.94	21.94	23.36	8.20	7.78	12.23	6.55
2	7.53	7.70	25.64	16.78	10.07	9.51	8.82	6.92
3	8.89	7.73	19.48	13.30	10.09	8.86	8.33	6.49
4	8.74	7.56	11.71	12.63	10.24	9.88	7.08	6.31
5	7.98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

Evaluating and Selecting Models

- Several black box model have been reported for use. Comparisons highlighted that many hold similar results. **You still have to compare them and select their parameters.**
- How? So Far, the presented results and methodology omits to describe the evaluation process.
- 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....?



How about the calibration model?

- Typical examples are **Multilinear regression** (see before), **ANN** (mostly shallow architectures), **Random Forests** (shown to provide great but with bad generalization properties), **SVMs**
- So far comparison literature showed no clear «winner», if adequately optimized with fair choice of hyperparameters values they provide similar results
- First order considerations may help to rule out some models (when sensors are not linear than purely linear models are to be ruled out)
- Depending on applications, recurrent architecture may provide a performance boost in **fast transients**. Beware: Hardware for operation and reference can lead the choice.

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?

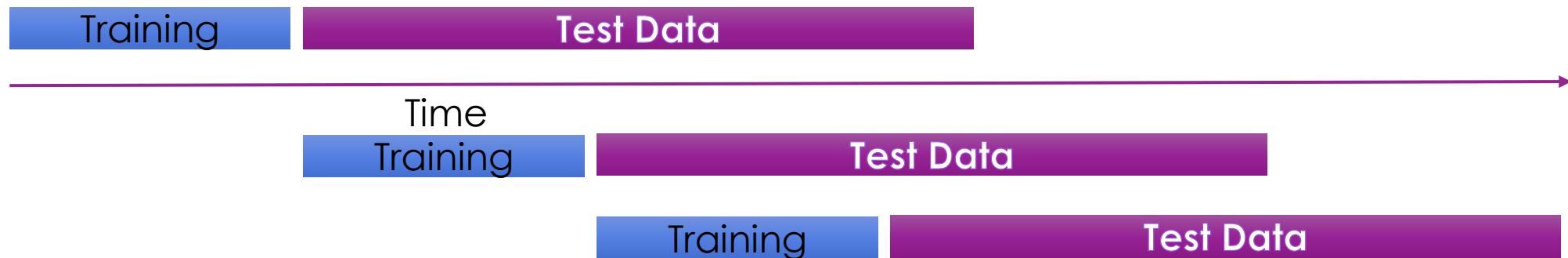
- First rule!
 - Avoid correlation between training and test data (more on this later)
- If I have enough data:



- 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

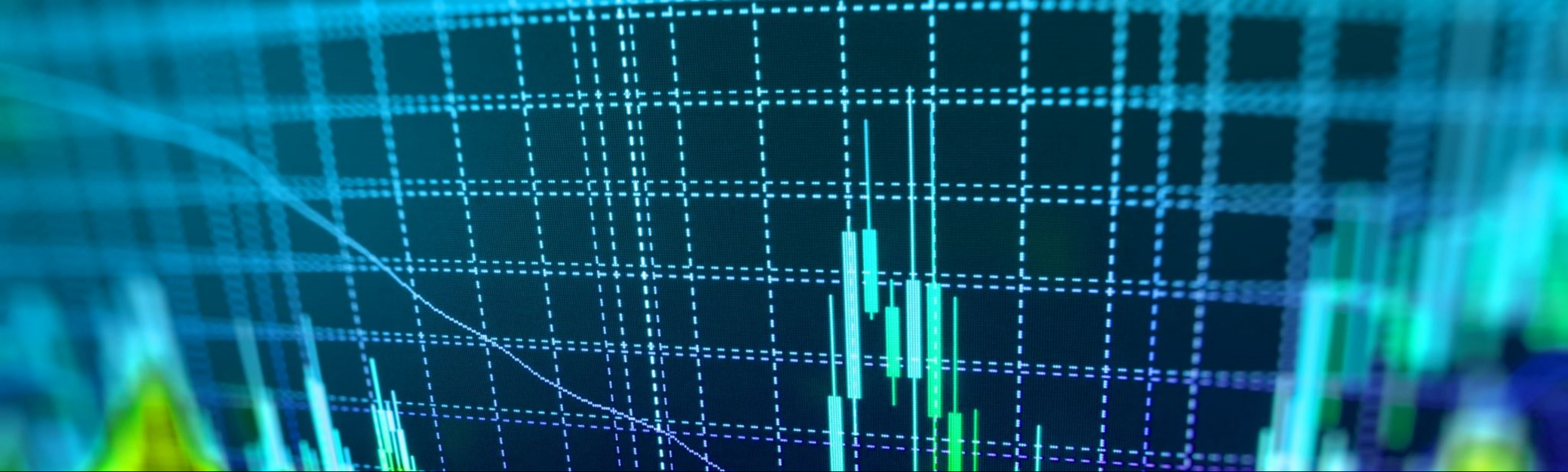
Table 5. Calibration performance indicators for NO₂ estimations obtained using two calibration models with different choices of training length.

(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?

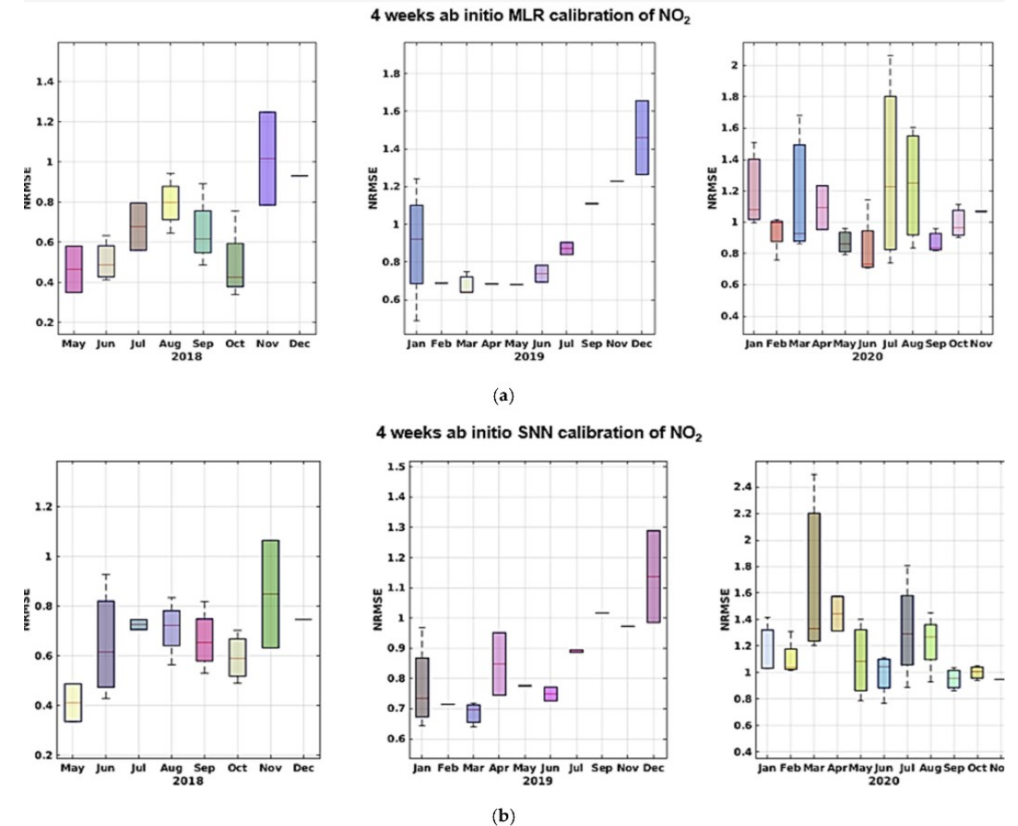


...why performance of FC systems drops in the long term anyway?



...why these losses seem to happen even when no sensors drift is detected?

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.



The reasons behind...

The reasons lie behind change:

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

In a few words: Change in the *response eliciting forcers joint distribution*

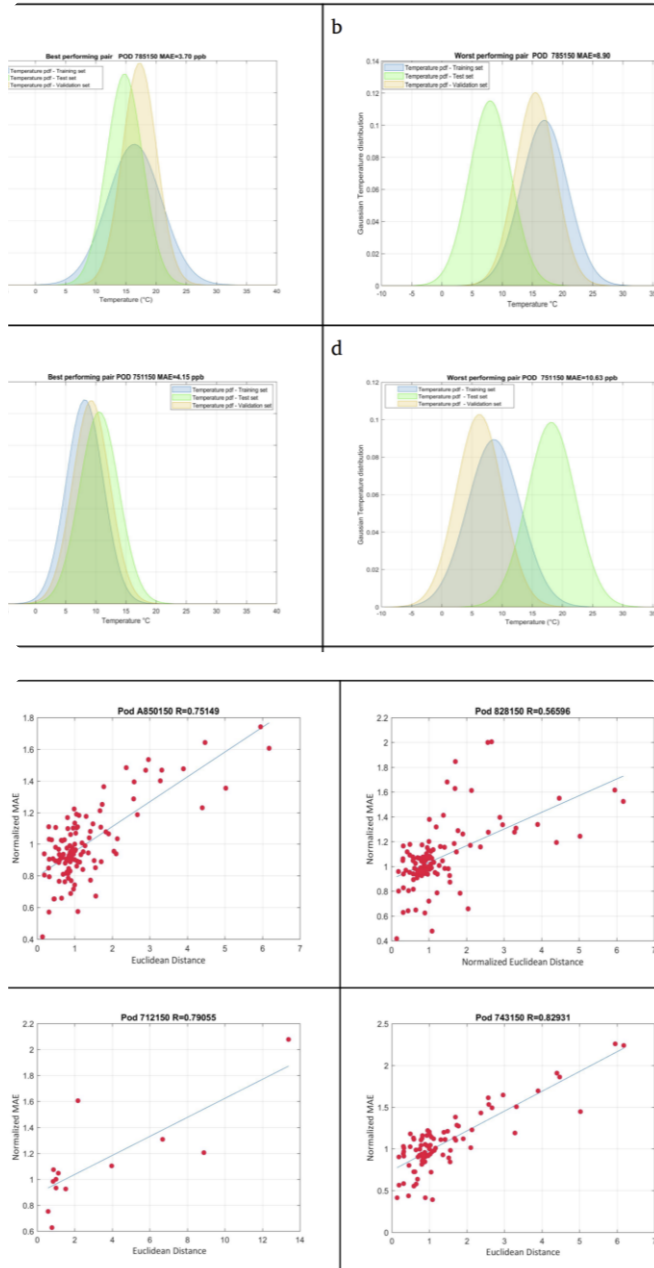


Figure 6c: Correlation plots showing the actual relationship between normalized MAE estimation and Euclidean distance applied to joint empirical distribution p(T, NO2) for the 4 pods relocated in Akebergveien rd.

Robustness boosting strategies

Most strategies depends on:

-Boosting the calibration set completeness so to be able to face the conditions variance (**long term colocation**, **multiple site colocation**, **recalibration** either with **co-location** or **remote** calibration).

or

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



Robustness boosting strategies

- **Repeat Short Term Calibration & Use** – (Hagan et al., *Atmos Meas Tech*, 2018, **11**, 315–328)
[Application Dependent, Highly Costly]
- **Long Term Colocation/Calibration** – (Bigi et al. *Atmospheric Meas. 568 Tech.*, 2018, **11**, 3717–3735)
[Effective if lasting for multiple seasons when operating with similar sources, Highly Costly, Sensor Drift? Sensor Lifetime?]
- **Multi-Site Long Term** - (Vikram et al., *Atmospheric Meas. Tech.*, 2019, **12**, 578 4211–4239)
[Effective for relocation in different environment/sources and for multiple seasons, Huge cost, Sensor Drift? Sensor Lifetime?]
- **Calibration Transfer** - (Mailings et al., *Atmospheric Meas. Tech.*, 2019, **12**, 903–920)
[Potentially very effective, Limit the costs, Sensor Drift?]
- **Laboratory ‘Temperature Binned’ Calibration** - (Wei et al., *Atmos. Environ.*, 2020, **230**, 117509)
[Improved generalization, High cost, Sensor drift?]
- **Remote data exploitation/OSINT** - (Miskell et al., *Atmos. Environ.*, 2019, **214**, 116870)
[Globally effective, Low cost, Sensor drift robust, Less accurate in the short time]



Going beyond Field Calibration

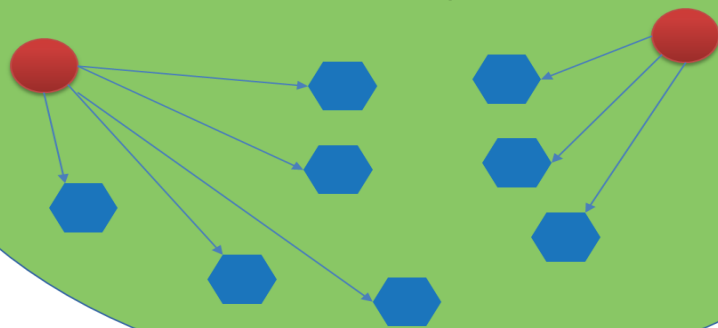
Summarizing, recently (last 3-5yrs) several strategies have been proposed to overcome the scalability issues:

- 1. Remote Calibration schemes**
- 2. Global, General Calibration schemes**
- 3. Calibration Transfer schemes**

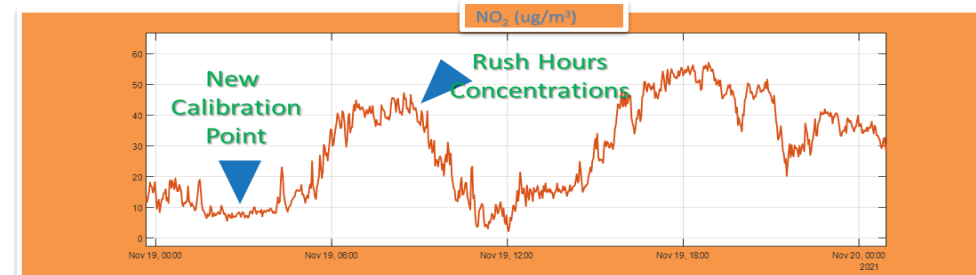
Remote Calibration

Continuous (re)-calibration scheme relying on reference data from remote stations exploiting particular conditions / hypothesis.

Since 2005, Researchers have suggested to exploit data coming from remote regulatory grade stations, in low spatial variance conditions, to correct for drifting network of low cost systems [Tsuji et al., Sens. Act. B Chem. 110 (2005) 304-311]



Resorting to special conditions may limit the dataset variance and knowledge improvement (Miskell et al., *Atm. Env.*, 214, 2019, 116870). As an example, using nighttime concentrations allows to correct for drifting bias but no info is obtained about sensitivity drift induced by changing environmental conditions at higher concentrations during daytime.



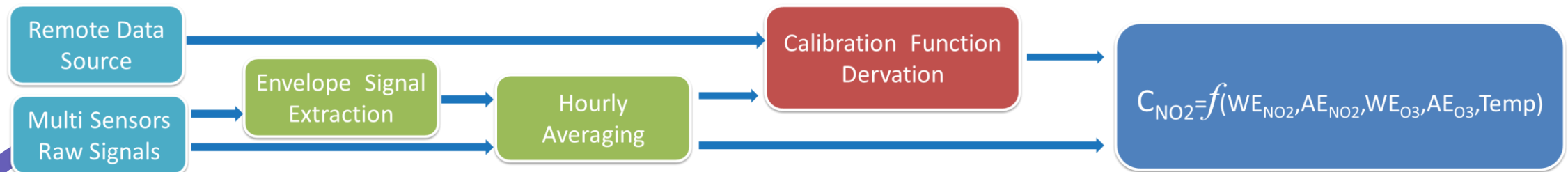
Selecting the remote station

Very recently, We have started to exploit the entire stream of data.
This approach is divided in several steps:

- a) Selecting a similarity measure (e.g. correlation)**
- b) Finding the best candidate remote reference station (using the similarity measure)**
- c) Calibrating using remote data**
 - 1. Matching some moments (e.g. mean and variance)**
 - 2. Matching each sample by devising a transfer function**

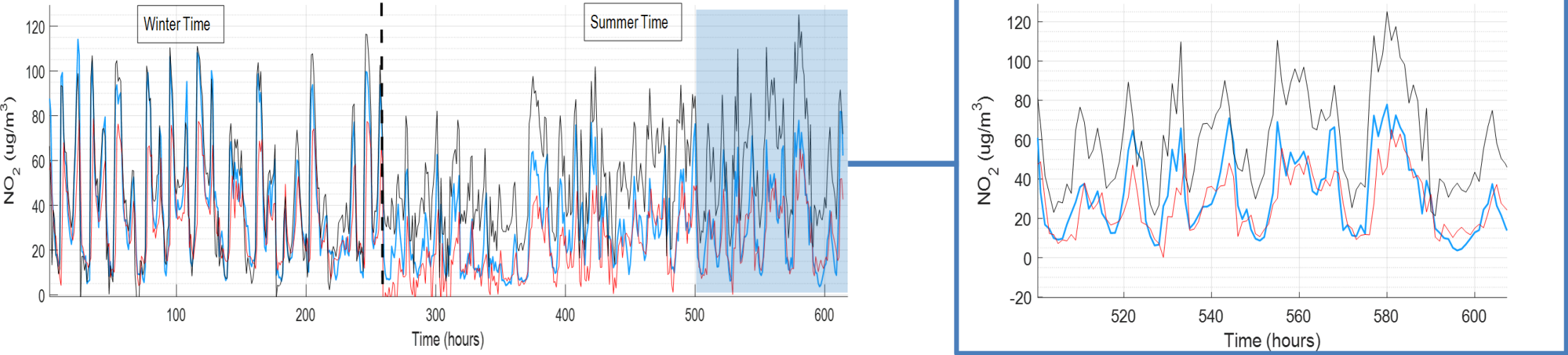
Two examples for remote calibration strategy:

- Moment Matching: Rebasing and Normalizing the raw sensor response to match location and spread moments of the remote station short term distribution.
 - On which time scale? Must take into account of anthropogenic and natural induced cyclostationarity
- Extracting relevant segment of portion of data based on the closer or most relevant remote source of data.



Typical Results obtained with remote calibration strategies

Fig. Comparison of conventional 2 weeks ab-initio filed calibration estimates with hourly updated continuous calibration by remote (regional background) data (Results form a single Multisensor device).



Acceptable, but You will loose performance in the short term, do not match a full recalibration by colocation

Global/General Calibration Schemes

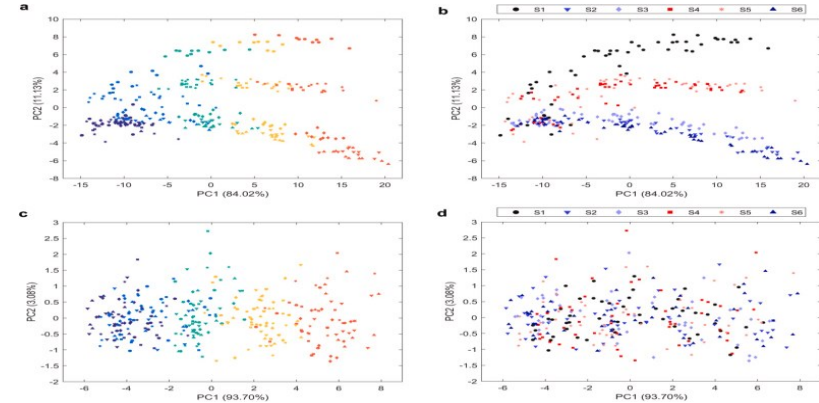
- Calibration Transfer vs. Multiunit calibration
- First is suitable for controlled environments in which You can set similar conditions and record transfer samples. Need a «master» tool to receive a complete calibration to derive the general calibration.
- Second more suitable for field conditions or uncontrolled setup, need some response normalization anyway or the development of a virtual sensor model.

MultiUnit Calibration Schemes

One approach proposed by **Mailings et al., 2019**, dealing with EC Sensors, aims to calibrate a virtual sensor model using **the median of the appropriate input signal across a subset of sensor**

You will end with a single calibration function which may seamlessly be applied to all sensors. Of course you need limited fabrication variance!

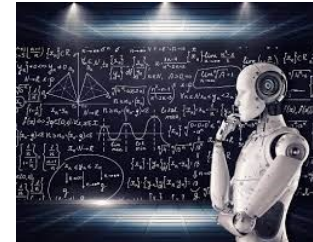
Miquel-Ibarz et al. dealing with TM MOX sensors finds **fabrication variance robust features**, derive a multiunit calibration that can be used with other sensors. However **their signals should be normalized** before applying the derived calibration. This underlies the need to operate in a somehow controlled setup in which operative conditions do not differs from training ones.



A. Miquel-Ibarz, et al., Global calibration models for temperature-modulated metal oxide gas sensors: A strategy to reduce calibration costs, Sensors and Actuators B: Chemical, Volume 350, 2022.

Mailings et al., Development of a general calibration model and long term performance evaluation of low cost sensors for air pollutant gas monitoring, Atmos.Meas. Tech., 12, 903-920,2019

Global Calibration Strategy



COLOCATION FOR
 A SUBSET OF
 SENSORS



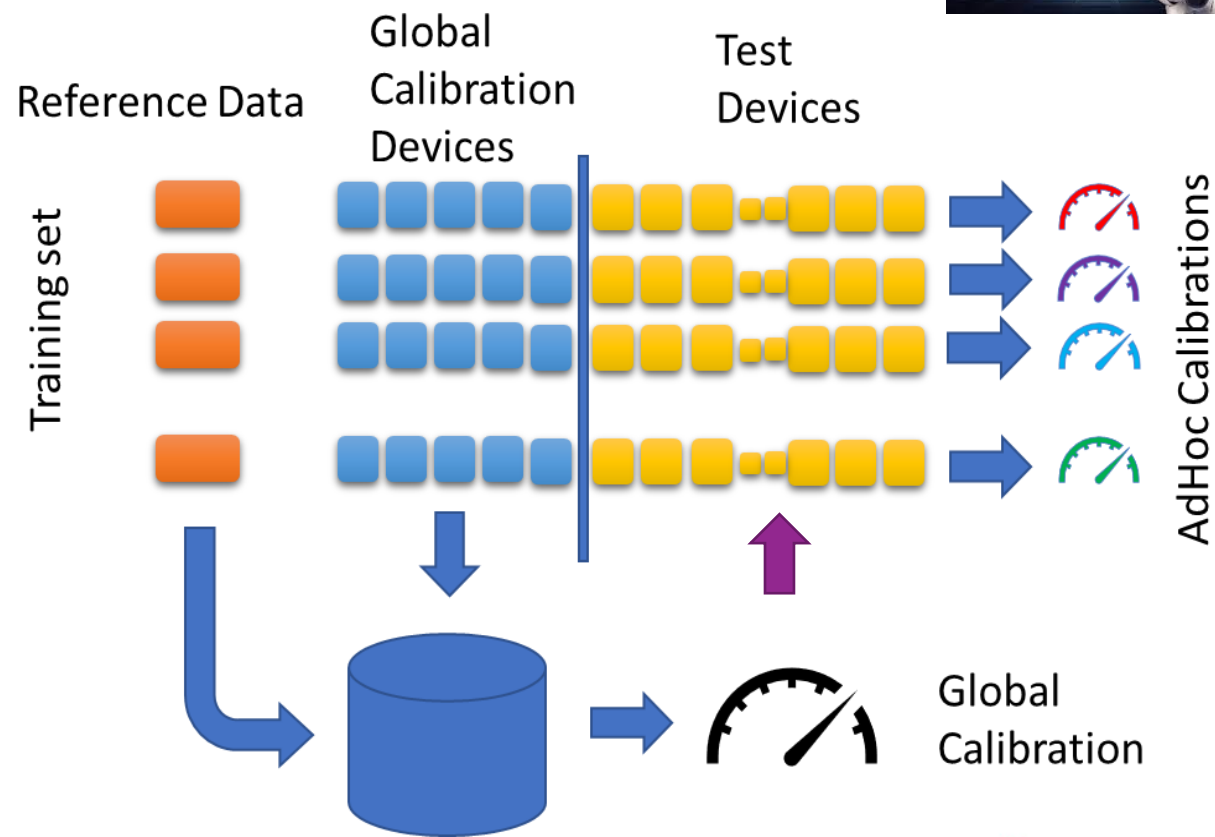
DERIVING A SINGLE
 CALIBRATION
 USING ALL SENSORS
 DATA....



...WITH DATA
 DRIVEN
 ALGORITHMS



COMPARING
 PERFORMANCE WITH
 AD-HOC CALIBRATION
 (ONE FOR EACH
 SENSOR)



An PM focused example

- OPC PM Sensors usually works very well with PM2.5 estimations. They have limited fabrication variance
- It is possible to extract a global calibration just by joining (U) the training dataset from several sensors.
- This work surprisingly well reducing the need for calibrating each device but a subset (5-10 seems to work well woth PMS7003)
- May result in more robust models in the long term

Drawback: In the end, we will need to tackle drift (sensors+concept)!



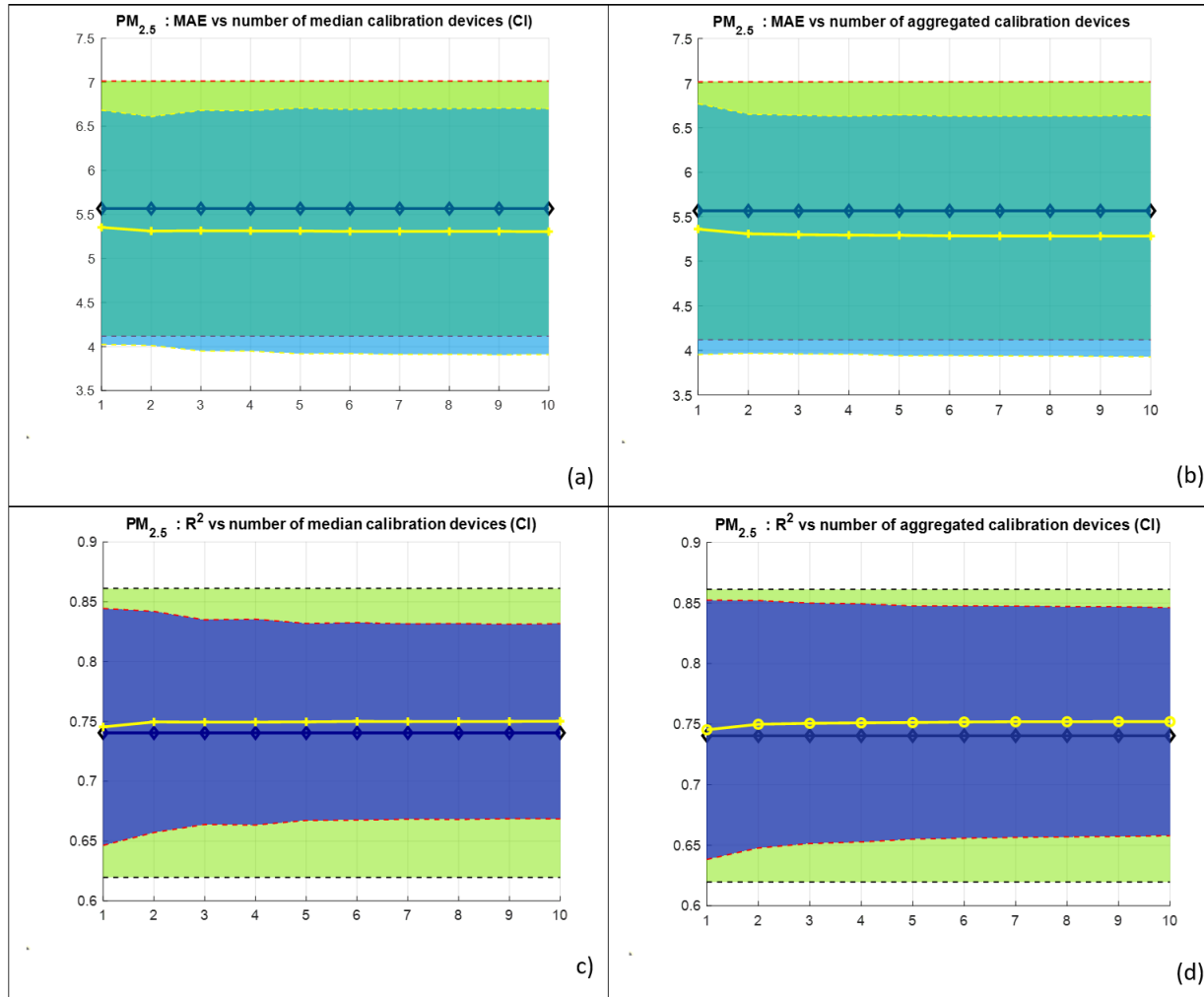


Figure 8: Global Calibration approaches (yellow) versus Ad-Hoc Calibration (black) average MAE (a,b) and R^2 (c,d) figures at different no. (n) of involved devices along with uncertainty bars (CI).

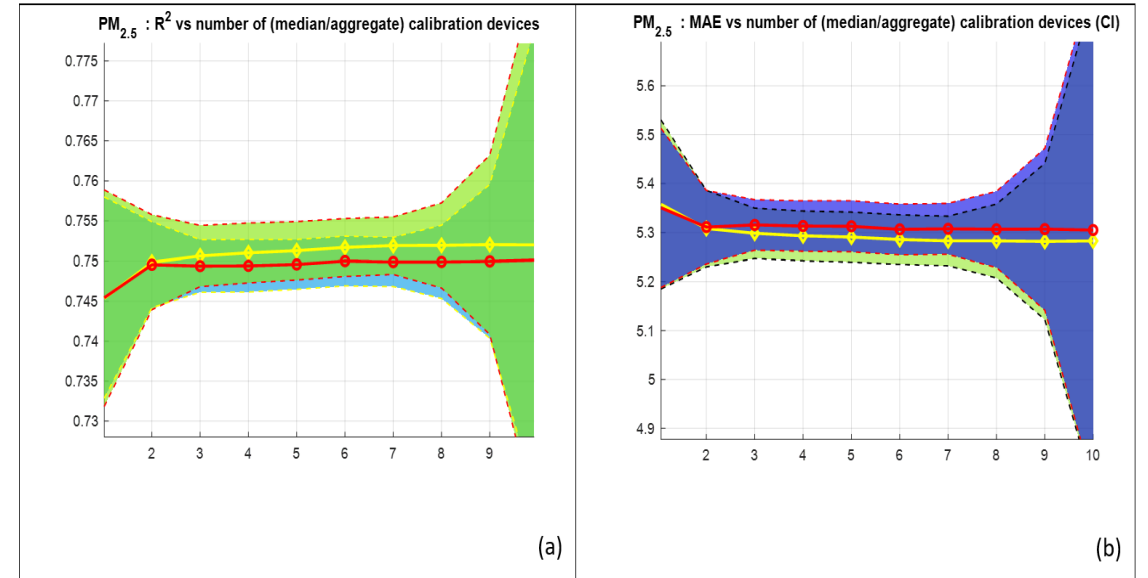


Figure 9: Comparison of the two data fusion models (red: median; yellow: aggregate) for the proposed global calibration methodology using uncertainty bars (0.95 CI). While aggregated models consistently but very slightly outperform median based models, uncertainty bands of both performance index population means are actually largely superimposed.


Beyond Field calibration: A critical recall

- Problem to solve: **Derive a calibration function for generically non linear time variant dynamical unique transducers units.**
- FC have locally and temporarily *nice but inherently limited* performances
- A Global (multiunit) Calibration or Calibration transfer may address fabrication variance and can deal with seasonal changes (you can derive it once for all) allowing scalability but..... *Sensors may drift themselves sooner or later*
- Remote calibration inherently deals with sensors and concept drifts but relies on «wrong teachers» (*inexact calibration data*)



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 but needs improvements
- 



Dr. Saverio De Vito
ISOCS President

Thank You for Your Attention
saverio.devito@enea.it