Calibration methods of innovative sensors for monitoring pollutants for air and water quality

Marine Dumon[1], Bérengère Lebental[1], Guillaume Perrin[1]

[1] *Université Gustave Eiffel, COSYS, F-77454 Marne-la-Vallée, France*

Financed by ANR project CARDIF and H2020 project HS4U

Table of contents

Context

3.4 million deaths directly related to water quality

9 million premature deaths related to air quality

Context

3.4 million deaths directly related to water quality

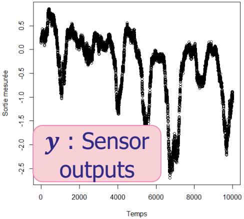
9 million premature deaths related to air quality

→ Need for accurate, reliable, but low-cost and small, sensors for local monitoring of air and water quality

Sensor Calibration in the lab

x : Pollutants to predict

Sensor



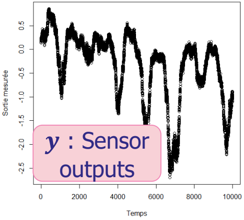
y : Sensor outputs

The difficulty of moving to an open environment

x : Pollutants to predict

Sensor

Interferents



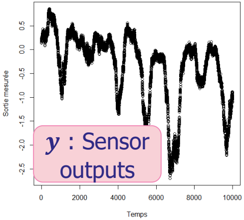
y : Sensor outputs

Other quantities affecting y

The difficulty of moving to an open environment

x : Pollutants to predict

Sensor



y : Sensor outputs

Interferents

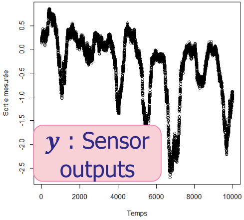
z : Known environmental variables

Other quantities affecting y

The difficulty of moving to an open environment

x : Pollutants to predict

Sensor



y : Sensor outputs

Interferents

z : Known environmental variables

Other quantities affecting y

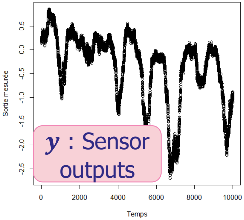
u : Unmeasured and/or unknown variables

?

The difficulty of moving to an open environment

x : Pollutants to predict

Sensor



y : Sensor outputs

Interferents

z : Known environmental variables

Other quantities affecting y

u : Unmeasured and/or unknown variables

?

$\epsilon^x, \epsilon^z, \epsilon^y$:

- Measurement uncertainties unknown

The difficulties of the sensor calibration

- Highly sensitive to pollutants **but partial selectivity**
- Unmeasured or unknown variables
- Real data inaccessible: only access to the **measurements**
- Real relation between environmental variables and sensors **unknown** and increasingly **non-linear**
- Strong **correlation** between environmental variables

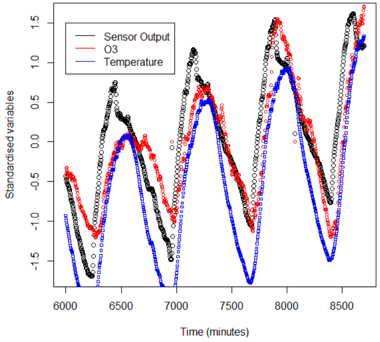


Figure: Sensor variables

Summary of the method

- **Small data** context: **physical information** is used for mitigation purposes
  - Two-step process: **calibration** + inversion
  - **Bayesian formalism** to handle all uncertainties

Summary of the method

- Two-step process: **calibration** + inversion
  - **Derivation of the calibration model** on the training data

$$\mathit{Outputs} = \mathcal{H}(\mathit{Targets}, \mathit{Interferents})$$

  - **Inversion** of the calibration model on the testing data

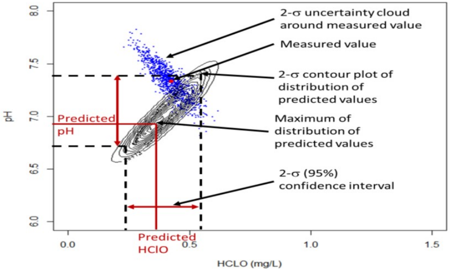
$$\mathit{Targets} = \widetilde{\mathcal{H}}^{-1}(\mathit{Output}, \mathit{Interferents})$$

  - By contrast, in 'regular' ML/IA, $\widetilde{\mathcal{H}}^{-1}$ is directly learnt

$\rightarrow$ Provides **physico-chemical information** on the sensors

Summary of the method

- **Bayesian formalism** to handle all uncertainties :
  - Targets, ie. the reference measurements (up to 100% noise in chemical sensing!)
  - Interferents, eg. temperature, humidity (measured or not measured, possibly even unknown)
  - Outputs, ie. the innovative sensors (often EM noise)
  - The calibration model itself (linearity is the exception, not the rule!)



Summary of the method

- **Small data** context: **physical information** is used for mitigation purposes
  - Two steps process: **calibration** + inversion
  - **Bayesian formalism** to handle all uncertainties

→ **Grey/White Box** method

A grey-box method

- For each sensor j at time i, the *a priori* model is:

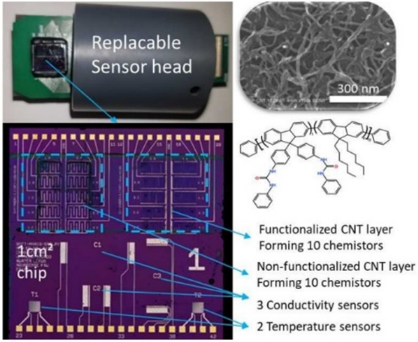
$$(\mathbf{y}_i^{\text{mes}})_j = \mathbf{h}_j(\mathbf{x}_i^{\text{mes}} + \varepsilon_i^x ; \mathbf{z}_i^{\text{mes}} + \varepsilon_i^z)^T \boldsymbol{\beta}_j + (\varepsilon_i^y)_j + \varepsilon_j^{\text{mod}}(\mathbf{x}_i^{\text{mes}} + \varepsilon_i^x ; \mathbf{z}_i^{\text{mes}} + \varepsilon_i^z) + (\delta_i^{\text{mod}})_j,$$

- The calibration model $\mathbf{h}_j$ is explicite (at least through Taylor expansion), enabling physical interpretation
- The model errors $\varepsilon_j^{\text{mod}}, (\delta_i^{\text{mod}})_j$ are explicite (chosen covariance functions)
- The uncertainties can be derived from the training set

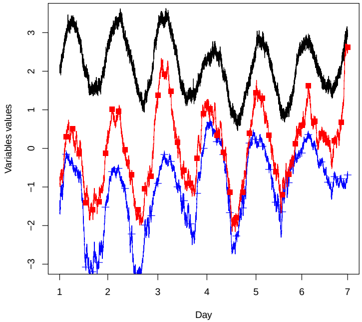
Presentation of the two cases

1. Experimental data
   - Sensor based on **carbon nanotubes deployed in an open environment** during 57 days
   - Exhaustive search to obtain sensor influence parameters (O_3, CO, RH, T).
2. Simulated data
   - Created to mimic as closely as possible the experimental data.

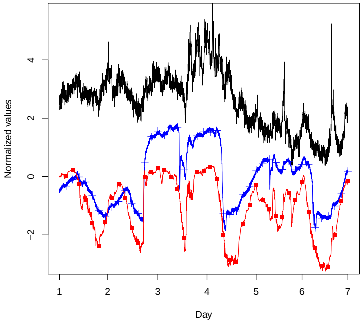
$$\mathbf{y}_j = (4+\beta_j^1)\log(\beta_j^2 x_1 + 1 - \min(\mathbf{x}_1)) + \beta_j^3 \arctan(\beta_j^4 \mathbf{z}_1) + \beta_j^4 \arctan(\beta_j^5 \mathbf{x}_2 + \beta_j^6 \mathbf{z}_2) + \alpha_u \mathbf{u}.$$



Similarity of the two datasets



Simulated data



Experimental data

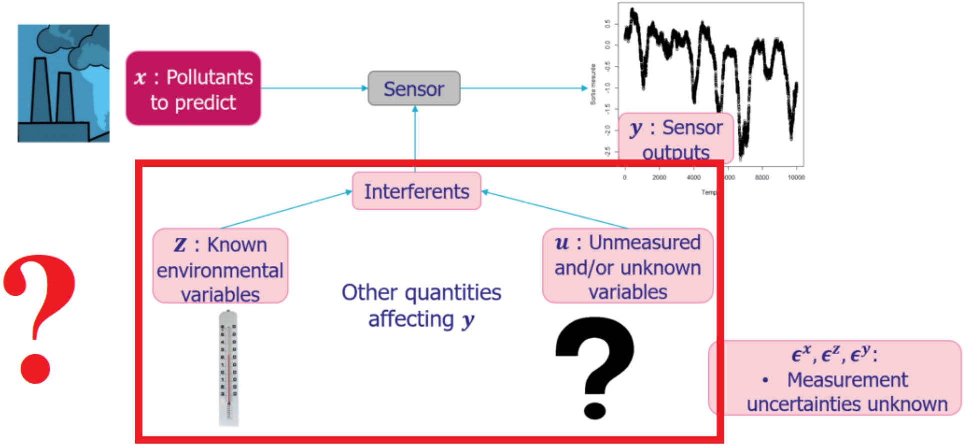
- Representation of the time evolution over one week of one sensor output, one environmental variable and **one target pollutant (black)**. For the experimental data, the target pollutant is the *CO* and the values are normalized.

Comparison with classical methods

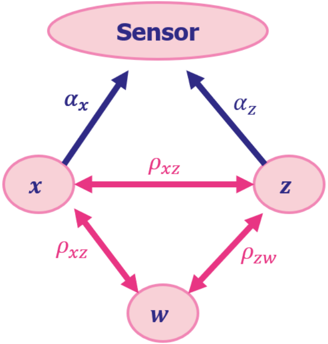
Method	Indicators on simulated data					
	R_1^2	R_2^2	MAE_1	MAE_2	$\mathcal{L}_1^{95\%}$	$\mathcal{L}_2^{95\%}$
GLR	0.87	0.83	0.39	0.49	1.8	2.7
GPR	0.98	0.88	0.14	0.38	0.95	1.9
GPR+IU	0.98	0.88	0.15	0.38	0.86	1.8
Method	**Indicators on experimental data**					
	$R_{O_3}^2$	R_{CO}^2	MAE_{O_3}	MAE_{CO}	$\mathcal{L}_{O_3}^{90\%}$	$\mathcal{L}_{CO}^{90\%}$
GLR	0.55	0.74	5.1	0.030	18	0.083
GPR	0.65	0.79	4.4	0.023	18	0.079
GPR+IU	0.73	0.79	4.2	0.022	17	0.073

Table: Performances indicators of the methods on simulated and experimental data. For the experimental data, the pollutants vary from 15 to 83 ppb for O_3 and from 7 to 8 ppm for CO. The results of the indicators are presented in ppb for O_3 and in ppm for CO. GLR: Generalized Linear Regression. GPR: Gaussian Process Regression. IU: Input Uncertainties.

Correlation versus causality in the field of sensors

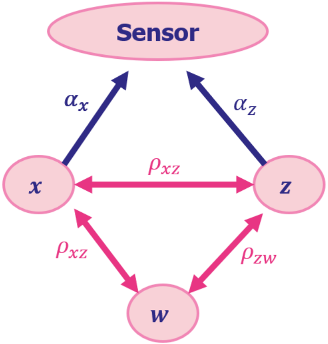


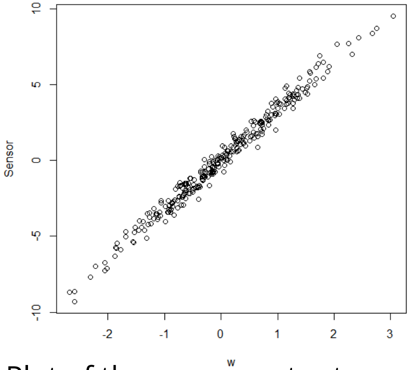
Correlation versus causality in the field of sensors



x and ***z*** has a causal impact on the sensor,
w does not have a causal impact but is
correlated with ***x*** and ***z***

Correlation versus causality in the field of sensors





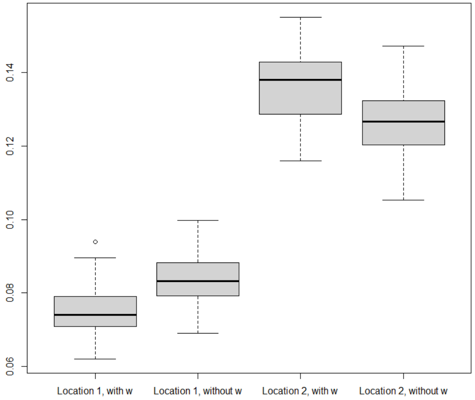
- ***x*** and ***z*** have a causal impact on the sensor, ***w*** does not have a causal impact but is correlated with ***x*** and ***z***.
- Plot of the sensor outputs according to ***w***, **the relationship is clear but does not imply causality**

Why do we need to distinguish correlation and causality?

- Using a non-causal, correlated-only variable may improve calibration model performance! $\rightarrow$ **why bother?**
  - Environmental variables are **highly correlated** (chemical reactions, dayly cycles)
  - The correlation between variables depends on **deployment specificities** (location, time, circumstances)
  - Using non-causal variables **reduces model transferability**

Why do we need to distinguish correlation and causality?

- Illustation on simulated data (Boxplots of the mean absolute error)
  - With different values of **the correlation between variables** (location 1 or 2)
  - With and without using the non-causal but correlated variable **w**



Classical sensitivities analysis: an example with Sobol

- **Problem: Classical sensitivity analysis techniques (eg. Sobol) do not differentiate causality and correlation**
- Illustration: $\begin{pmatrix} x \\ w \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} S^2\right)$ and a linear relation $y = \alpha x + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- We get: $y|w \sim \mathcal{N}\left(\alpha\rho w, \alpha^2 S^2(1-\rho^2) + \sigma^2\right)$
- Sobol first-order indice (>0 if the variable has influence)

$$S_w := \frac{\text{Var}\left(\mathbb{E}[y|w]\right)}{\text{Var}\left(\mathbb{E}[y]\right)} = \frac{\alpha^2\rho^2 S^2}{\alpha^2 S^2 + \sigma^2} > 0$$

Causality defined through the calibration model

- Our approach: treat causality in the context of the calibration $\rightarrow$ **no causality if the calibration model does not depend on the variable**
- In the Generalized Linear Regression approximation of the calibration model, there is no causality if the corresponding **coefficient is equal to 0**
- Existing tests: **Student and Fisher indices**
- BUT formally not developped for situations with noise on inputs

Our new contributions

- Our improvment: Handling uncertainties in the Student test for one variable and generalisation to the Fisher test

- Achieved by **handling model error and measurement noise in the law of coefficients** $\widehat{\beta}$ (in the GLR)

  - Asymptotically follows the law $\mathcal{N}\left(\beta, \boldsymbol{C}^{\mathrm{IC}}\right)$

  - Where $\boldsymbol{C}^{\mathrm{IC}} = \boldsymbol{C}_\beta + \mathcal{C}_\beta^{\mathrm{IC}}$ with $\boldsymbol{C}_\beta$ the covariance on the noise-free GLR method and $\mathcal{C}_\beta^{\mathrm{IC}}$ the contribution associated with the propagation of uncertainties

  - Ignoring $\mathcal{C}_\beta^{\mathrm{IC}}$ leads to underestimation of the error

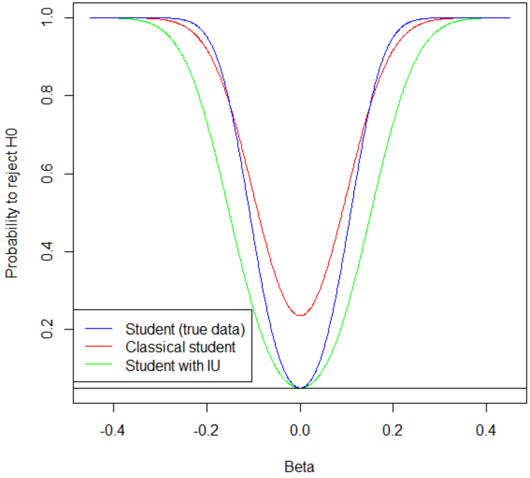
Statistical test formalism

- **Null Hypothesis**: $H_0 : \{\beta_w = 0\}$,
- **Construction of an estimator** $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} := (\beta_x, \beta_z, \beta_w)$ and the model error $\widehat{\sigma}^2$ by least squares approaches and find the **Statistical properties**: $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{C}^{\mathrm{IC}})$,
- **Statistical test**: Under $H_0, \zeta(\mathcal{D}_n) := (\boldsymbol{C}_w^{\mathrm{IC}})^{-\frac{1}{2}} \widehat{\beta}_w$,
- **Definition of the region of rejection**: the set of W_n of realisations such that $W_n = \{|\zeta(\mathcal{D}_n)| > a\}$ and **select** a to maximise the power,
- **Without noise**: classical Student framework; **With noise**: adaptation of the formalism with linearisation to account for input noise.

Results

- Let's see the **statistical power** (the probability to reject H_0 knowing the value of β_w) of the student test on β_w.

- Comparison of the power of the student test by using classical student test on the non noisy values, on noisy values and by using the Student test with IU on noisy values.



1. Introduction
2. A calibration process in a bayesian framework: Application and study of performances
3. Causality problems in the field of sensors - From a grey box to a white box method
4. Conclusion

Conclusions and prospects

- **A calibration process in a Bayesian framework**
  - Environmental sensors calibration is a multi-variate problem with very small data volumes
  - A Bayesian framework expliciting the calibration model, the uncertainties and the model error is proposed
  - It significantly improves calibration performances compared to standard approaches
  - The proposed model is a Grey Box: influence factors can be identified
- **Causality problems in the field of sensors**
  - Differentiating causality and correlation is challenging (but would improve model performance)
  - Several methods are proposed (and under test) to separate causality from correlation in that context
- **Further improvements lie in accounting for time effect (response time and drift)**

Communication

- **Conference**
  → IEEE sensors 2023 (Vienna, Austria) and 2024 (Kobe, Japan) conference (oral)
  → SIAM conference 2024 (poster)
  → MascotNum conference 2023 and 2024 (poster)
  → Workshop MascotNum 2023 (oral)
- **Publication**
  → Proceeding IEEE sensors 2023 and 2024
  → Co-author on a paper published in IEEE sensors journal
  → An article in review about the method
- **Prevision of publication**
  → Publication of a package about the method
  → Co-author of a patent
  → Upcoming: A future application paper on CNT sensors
  → Upcoming: A future paper on causality

Thank you for your attention