Bayesian calibration of sensors for air and water pollution monitoring

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 LOTUS

1. Introduction
2. Sensor Calibration
3. Regression and inversion step
4. Application
5. Conclusion

Context

- Environmental pollution causes more than 8 million deaths per year worldwide [a].
- Need for deployment of accurate low-cost sensors in air and water pollution monitoring.
- Innovative materials based sensors : a possible solution.

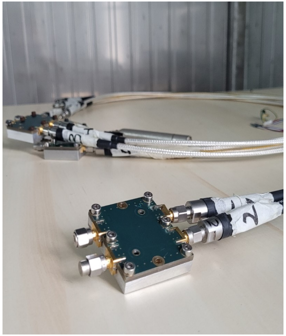
[a] *Pollution and health: a progress update*, The Lancet, 2022

Pollution on a ring road[a]

[a] afp.com/Francois Guillot

Classical sensor issues

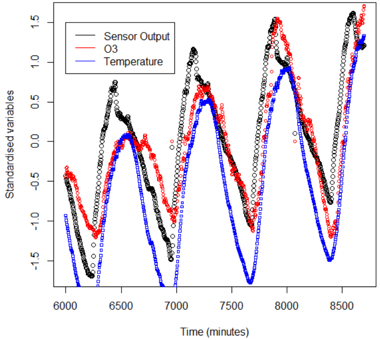
- **Innovative materials based sensors: higly sensitive to desired and undesired pollutants (interferents).**
- **Different kinds of uncertainties (unmeasured quantities and noise).**
- Potentially non negligeable response time.
- Temporal drift.



Polymer based sensor

The difficulty of moving from the laboratory to an uncontrolled environment

- Highly sensitive to pollutants **but few specificity.**
- Identification of sensitivity on laboratory.
- Strong correlation between environmental variables.



Sensor Modelisation

- *Known Outputs :*
  - y : sensor outputs
- *Known Inputs :*
  - z : environmental variables
- *Unknown Inputs :*
  - x : observed and sensitive variables
- Unmeasured Inputs
  - u : error, interfering potential («What we know that we don't know...»)

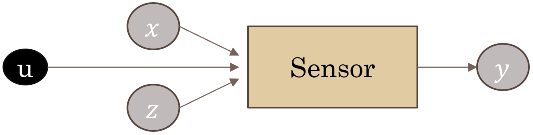


Figure: Representation of a sensor

Hypothetis : a model $\mathcal{M}$ of parameters θ exists such that : $\mathbf{y} = \mathcal{M}(\mathbf{x}, \mathbf{z}, u, \theta)$

Principle of sensor calibration

We are interested in the deployment of $\boldsymbol{d_y}$ sensors for monitoring $\boldsymbol{d_x}$ pollutants to estimate with $\boldsymbol{d_z}$ influents known quantities.

- Available information : $\mathcal{D}_n := (\mathbf{x}_i^{\text{mes}}, \mathbf{z}_i^{\text{mes}}, \mathbf{y}_i^{\text{mes}})_{i=1}^n$ to construct the model $\mathcal{M}$,
- Hypothesis : $\forall 1 \leq i \leq n$ and a new value $\star$:

$$\mathbf{y}_i^{\text{mes}} = \mathcal{M}(\mathbf{x}_i^{\text{mes}} + \varepsilon_i^x, \mathbf{z}_i^{\text{mes}} + \varepsilon_i^z, u_i, \boldsymbol{\theta}) + \varepsilon_i^y \quad (1)$$

$$\mathbf{y}_\star^{\text{mes}} = \mathcal{M}(\mathbf{x}_\star, \mathbf{z}_\star^{\text{mes}} + \varepsilon_\star^z, u_\star, \boldsymbol{\theta}) + \varepsilon_\star^y \quad (2)$$

- Objective : estimate pollutants concentrations $\mathbf{x}_\star$ knowing $\mathcal{D}_n, \mathbf{z}_\star^{\text{mes}}, \mathbf{y}_\star^{\text{mes}}$.

Bayesian formalism

In this framework, all unknown quantities are modeled by random variables and estimating $\mathbf{x}_\star$ amounts to estimate :

$$\pi[\mathbf{x}_\star|\mathbf{y}_\star^{\text{mes}}, \mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n] \propto \pi[\mathbf{y}_\star^{\text{mes}}|\mathbf{x}_\star, \mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n]\pi[\mathbf{x}_\star|\mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n] \quad \text{(Bayes)} \tag{3}$$

$$\propto \int \pi[\mathbf{y}_\star^{\text{mes}}|\mathbf{x}_\star, \mathbf{z}_\star^{\text{mes}}, \boldsymbol{\theta}]\pi[\boldsymbol{\theta}|\mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n]\mathrm{d}\boldsymbol{\theta}\; \pi[\mathbf{x}_\star|\mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n] \tag{4}$$

Resolution in two step :

- **Regression step** : to estimate the law of hyperparameters $\pi[\boldsymbol{\theta}|\mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n]$ and the *a priori* law $\pi[\mathbf{x}_\star|\mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n]$,
- **Inversion step** : to deduce the law $\pi[\mathbf{x}_\star|\mathbf{y}_\star^{\text{mes}}, \mathbf{z}_\star^{\text{mes}}, \mathcal{D}_n]$.

A first model : Linear Regression with model error (SLR or GLR + ME)

- Hypothesis :

$$\mathbf{y}_i^{\text{mes}} = \mathcal{M}(\mathbf{x}_i^{\text{mes}} + \varepsilon_i^x, \mathbf{z}_i^{\text{mes}} + \varepsilon_i^z, u_i, \boldsymbol{\theta}) + \varepsilon_i^y, \tag{5}$$

For each sensor j at time i :

$$(\mathbf{y}_i^{\text{mes}})_j = \mathbf{h}_j(\mathbf{x}_i^{\text{mes}}; \mathbf{z}_i^{\text{mes}})^T \boldsymbol{\beta}_j + (\varepsilon_i^y)_j + (\varepsilon_i)_j \tag{6}$$

- $\mathbf{h}_j$ is a vector-valued function,
- $\boldsymbol{\beta}_j$ is a vector of parameters modeled by a gaussian vector known parameters (*a priori*),
- $(\varepsilon_i)_j$ a model error modeled by a gaussian vector of unknown variance.

marine.dumon@univ-eiffel.fr

A second model : Gaussian process regression with model error (GPR + ME)

- Previous model :

$$(\mathbf{y}_i^{\text{mes}})_j = \mathbf{h}_j(\mathbf{x}_i^{\text{mes}}; \mathbf{z}_i^{\text{mes}})^T \beta_j + (\varepsilon_i^y)_j + (\varepsilon_i)_j \tag{7}$$

- Separate the model error into two :

  - $\varepsilon_j^{\text{mod}}$: to quantify the approximate character of the proposed function $\mathbf{h}_j$ modeled by a Gaussian Process with mean and covariance to estimate,

  - $(\delta_i)_j$: to quantify the impact of not taking unobserved quantities $\mathbf{u}_i$ into account modeled by a gaussian random variable with variance to estimate,

$$(\mathbf{y}_i^{\text{mes}})_j = \mathbf{h}_j(\mathbf{x}_i^{\text{mes}}; \mathbf{z}_i^{\text{mes}})^T \beta_j + (\varepsilon_i^y)_j + \varepsilon_j^{\text{mod}}(\mathbf{x}_i^{\text{mes}}; \mathbf{z}_i^{\text{mes}}) + (\delta_i)_j. \tag{8}$$

A third model : input uncertainties (GPR + IU)

- Previous model

$$(\mathbf{y}_i^{\text{mes}})_j = \mathbf{h}_j(\mathbf{x}_i^{\text{mes}}; \mathbf{z}_i^{\text{mes}})^T \boldsymbol{\beta}_j + (\varepsilon_i^y)_j + \varepsilon_j^{\text{mod}}(\mathbf{x}_i^{\text{mes}}; \mathbf{z}_i^{\text{mes}}) + (\boldsymbol{\delta}_i)_j. \tag{9}$$

- Handling input incertainties

$$(\mathbf{y}_i^{\text{mes}})_j = \mathbf{h}_j(\mathbf{x}_i^{\text{mes}} + \varepsilon_i^x; \mathbf{z}_i^{\text{mes}} + \varepsilon_i^z)^T \boldsymbol{\beta}_j + (\varepsilon_i^y)_j + \varepsilon_j^{\text{mod}}(\mathbf{x}_i^{\text{mes}} + \varepsilon_i^x; \mathbf{z}_i^{\text{mes}} + \varepsilon_i^z) + (\hat{\boldsymbol{\delta}}_i)_j. \tag{10}$$

Linearisation and Gaussian approximation

- Even if all parameters are gaussian, the law of $(\mathbf{y}_i^{mes})_j$ is not explicit (composition of gaussian variables).
- Assumptions :
  - The measurement errors are small enough to make a linearisation with an approximation by Taylor expansion,
  - The law is still not Gaussian (product of gaussian variables) : we approximate the measured sensor outputs by the Gaussian distribution of the same mean and covariance matrix.
- Estimation of hyperparameters by log-likelihood maximisation.

Inversion step

- Estimate the law using $\mathcal{M}$:

$$\pi[\mathbf{x}_\star|\mathbf{y}_\star^{\text{mes}},\mathbf{z}_\star^{\text{mes}},\mathcal{D}_n] \propto \pi[\mathbf{y}_\star^{\text{mes}}|\mathbf{x}_\star,\mathbf{z}_\star^{\text{mes}},\mathcal{D}_n]\pi[\mathbf{x}_\star|\mathbf{z}_\star^{\text{mes}},\mathcal{D}_n] \quad \text{(Bayes)}$$
$$\propto \int \pi[\mathbf{y}_\star^{\text{mes}}|\mathbf{x}_\star,\mathbf{z}_\star^{\text{mes}},\boldsymbol{\theta}]\pi[\boldsymbol{\theta}|\mathbf{z}_\star^{\text{mes}},\mathcal{D}_n]\mathrm{d}\boldsymbol{\theta}\ \pi[\mathbf{x}_\star|\mathbf{z}_\star^{\text{mes}},\mathcal{D}_n]$$

- Not explicit : linearisation and gaussian approximation of $(\mathbf{y}_\star^{\text{mes}})_j$,
- θ is known based on the regression step,
- MCMC methods to approximate the PDF of $\mathbf{x}_\star|\mathbf{z}_\star^{\text{mes}},\mathbf{y}_\star^{\text{mes}},\mathcal{D}_n$.

3 examples of applications

- Simulated data
  → $d_x = 2$pollutants, $d_z = 2$ known environmental variables, $d_y = 5$ sensors simulated,
  → Noisy data.
- Experimental dataset
  → For water quality [1]: Calibration of pH and chlorine ($d_x = 2$), knowing temperature ($d_z = 1$) with $d_y = 20$ sensor outputs.
  → For air quality [2] : Calibration of RH ($d_x = 1$) sensors based on radio frequency in lab conditions with $d_y = 2$ sensors outputs and different z possible.
  → For air quality : a dataset for ozone prediction, in progress.

[1] G. Perrin B. Lebental, IEEE sensor journal, 2023
[2] B.B. Ngoune, Article will be submitted to IEEE sensor journal

Results on the simaluted dataset

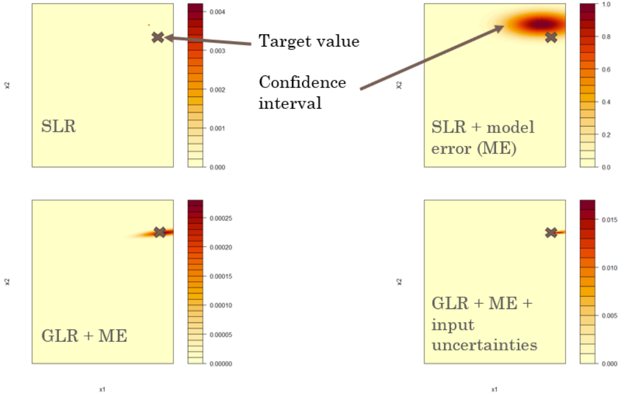


Figure: Graphical results for the prediction of two simulated pollutants at one time

Some results on water quality (G. Perrin's work)

- A dataset made from lab experiments in drinking water loop for a pH and chlorine sensor based on carbon nanotube.
- Uncertainties on chlorine was as large as the response of PH variation.
- Dataset of 25 points (small data) analyzed through a leave-one-out (LOO) approach.

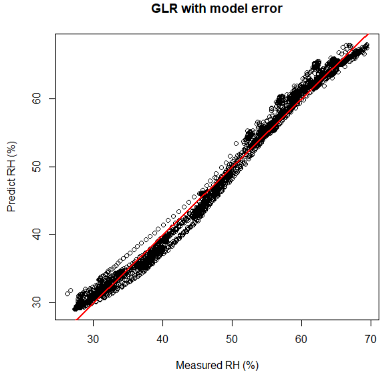
Method	MAE_1(HClO)	MAE_2(pH)
SLR	0.056	1.36
SLR+ME	0.054	1.75
GLR+ME	0.064	0.872
GLR+ME+IU	0.068	1.057
GPR+ME	0.054	1.75
GPR+ME+IU	**0.039**	**0.254**

Results with a LOO approach for Lotus project data

Some results on air quality

ML models	Evaluation metrics			
	MAE (%RH)	MAPE (%)	R^2 (%)	Prediction time (ms)
LR	2.2	5.5	93.8	2
SVM	1.1	2.4	98.6	864
RF	1.0	2.3	98.6	64
KNN	1.0	2.2	98.7	16
GLR + ME	1.1	2.5	98.8	1
GPR + ME	0.9	2.2	99.1	12000

Comparison of different calibration methods



Conclusion and perspective

- Searching for the right model $\mathcal{M}$ with uncertainties.
- Choice of x, z (causal discovery).
- Improve the calibration model by adding a temporal part in the model.

Thank you for your attention