Uncertainty-based performance evaluation of a carbon nanotube-based sensor array monitoring pH and active chlorine in drink water

Berengere Lebental, Guillaume Perrin

Uncertainty-based performance evaluation of a carbon nanotube-based sensor array monitoring pH and active chlorine in drink water

1st Bérengère Lebental
COSYS-LISIS, Université Gustave Eiffel
77420 Champs-sur-Marne, France
berengere.lebental@univ-eiffel.fr

2nd Guillaume Perrin
COSYS-LISIS, Université Gustave Eiffel
77420 Champs-sur-Marne, France
guillaume.perrin@univ-eiffel.fr

***Abstract*—Current challenges in the field of air and water pollution monitoring require the capability to detect simultaneously a large variety of chemical compounds at very low concentration using low-cost, compact sensor nodes. While carbon nanotube-based (CNT) sensor arrays have long been proposed as a solution to this challenge, their sensing performances usually suffer from the large number of interferents in real-life conditions. Here we discuss an uncertainty-based calibration and prediction framework which allows to recover multi-parameter sensing even in a highly perturbed environment. We study a 10×2 CNT-sensor array for pH and active chlorine monitoring in drink water. While in deionized water pH and active chlorine are easily monitored, in tap water only the active chlorine level can be recovered by standard calibration. By contrast, using our Bayesian approach, both active chlorine and pH are recovered with mean absolute error comparable with reference sensors.**

***Index Terms*—Sensor calibration, uncertainty quantification, carbon nanotubes, drink water monitoring**

I. Introduction

The development of nanosensors for the monitoring of air and water chemicals is the focus of much attention to achieve dense and real-time monitoring of pollution in urban or environmental context. Among these, carbon-nanotubes-based electronic noses and tongues are particularly promising [1] for their high sensitivity, their adaptability to a wide range of chemicals and their integrability into highly compact sensor arrays. However, while intense efforts have been focused on achieving selectivity of these sensors against interfering parameters (environmental factors such as temperature, or other chemicals in the air or water matrix), nanosensors performances are almost always degraded in real context compared to lab calibration due to these interferents. Moreover, they may display significant device-to-device variability in sensitivities from the manufacturing process (5% to 30%). Often requiring low-signal electronics, they may often be significantly affected by measurement noise.

Calibration allows to quantify the response of such sensors. To achieve manageable duration of calibration work, calibration datasets usually contain a number of input-output pairs which is relatively small compared to the number of influential factors. The inputs include measurements of environmental variables, as well as concentrations of one or more chemicals of interest measured by reference sensors. The outputs are the responses of the sensors, noting that to monitor the concentration of several chemicals, several types of sensors are required. The consequence to the limited number of calibration points is that calibration models may be quite inaccurate or may not integrate fully the perturbing factors, especially in a context with significant measurement uncertainties.

In the present paper, we propose a method based on a Bayesian framework for calibration and performance assessment of multi-parameter sensor arrays. It allows to account reliably for perturbing factors and high measurement uncertainties. We apply the method to the use case of drink water quality monitoring. The goal is to detect both active chlorine (HClO) and pH in tap water using a sensor array containing 20 carbon-nanotube (CNT) based chemistors of two different types in a single 1cm² chip mounted into a 15cm long,

The authors thank Thomas Vezin for dataset preparation and Adel Bendib for helpful discussions. The authors acknowledge support from H2020 Project LOTUS number 820881 and from Agence National de la Recherche project CARDIF reference ANR-19-CE04-0010-05.

2.5cm wide sensor node computer-connected through USB [5]. In this specific use case, during lab calibration experiments, the two types of devices showed strong, differentiated sensitivities to HClO and pH in deionized water. However, in tap water, the differentiation between types was strongly reduced and the sensitivity to pH appeared about 10 times lower than HClO. Henceforth, while HClO could still be easily predicted from the sensor array, the pH response could not be easily recovered by standard calibration approaches such as two-parameter regression. This paper shows that the Bayesian formalism summarized hereunder also allows to recover reliable pH predictions not accessible by usual methods.

II. Theoretical background

A. Notations and objectives

The following notations are introduced.

- $\boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^2$ is the vector of interest gathering the concentration of HClO and pH,
- $\boldsymbol{y} \in \mathbb{Y} \subset \mathbb{R}^{20}$ is the vector gathering the 20 sensor outputs,
- $\boldsymbol{z} \in \mathbb{Z} \subset \mathbb{R}^{d_z}$ is the vector gathering $d_z \geq 1$ environmental quantities (such as temperature) expected to influence the measurements.
- $\left\{\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{y}^{(i)}\right\}_{i=1}^{n}$ gathers n observed input-output triplets, such that:

$$\boldsymbol{y}^{(i)} = \mathcal{M}(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{\varepsilon}_x^{(i)}, \boldsymbol{\varepsilon}_z^{(i)}, \boldsymbol{\varepsilon}_y^{(i)}), \qquad (1)$$

with $\mathcal{M}$ an unknown function, and $\boldsymbol{\varepsilon}_x^{(i)}$, $\boldsymbol{\varepsilon}_z^{(i)}$ and $\boldsymbol{\varepsilon}_y^{(i)}$ three measurement uncertainties, respectively affecting the values of $\boldsymbol{x}^{(i)}$, $\boldsymbol{z}^{(i)}$ and $\boldsymbol{y}^{(i)}$.

The objective is therefore to estimate function $\mathcal{M}$ given the available data and uncertainties (training phase), in order to be able to predict the value of $\boldsymbol{x}^\star$ in another input-output triplet $\{\boldsymbol{x}^\star, \boldsymbol{z}^\star, \boldsymbol{y}^\star\}$, for which only $\boldsymbol{z}^\star$ and $\boldsymbol{y}^\star$ are observed (testing phase).

B. Bayesian formalism

In order to integrate the different sources of uncertainty affecting the system, a Bayesian formalism is proposed for the calibration and operation of sensors. It is based on the assumptions that there is some unknown probability distribution over the product space $\mathbb{X}\times\mathbb{Z}\times\mathbb{Y}$, and that the training set is made up of samples from this probability distribution. The relationship between the inputs and the outputs of each sensor is written, for $1 \leq j \leq 20$:

$$y_j^{(i)} = \boldsymbol{h}(\boldsymbol{x}^{(i)} + \boldsymbol{\varepsilon}_x^{(i)}, \boldsymbol{z}^{(i)} + \boldsymbol{\varepsilon}_z^{(i)})^T \boldsymbol{b}^{(j)} + \varepsilon_{j,y}^{(i)} + \xi_j, \quad (2)$$

where $\boldsymbol{h}$ is a chosen vector-valued function (e.g. the calibration model), $\boldsymbol{\beta} := (\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{20})$ gathers the unknown parameters of the calibration model, and $\boldsymbol{\xi} := (\xi_1, \ldots, \xi_{20})$ is the model error, modeled by a random quantity. Whereas the statistical properties of $\boldsymbol{\varepsilon}_x^{(i)}$, $\boldsymbol{\varepsilon}_z^{(i)}$, $\boldsymbol{\varepsilon}_y^{(i)}$ are supposed to be known, it is important to notice that the distribution of $\boldsymbol{\xi}$ is unknown.

Under these formalism and assumptions, standard Bayesian techniques (see [2], [3] for more details) can be used to first estimate the distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\xi}$ given labeled data $(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}, \boldsymbol{y}^{(i)})$, and secondly to estimate the distribution of $\boldsymbol{x}^\star$ given $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{y}^\star, \boldsymbol{z}^\star$.

C. Performance analysis

For the performance assessment, we propose in the following to extract two kinds of quantities from the a posteriori distribution of $\boldsymbol{x}^\star$, as illustrated in Figure 1:

- the most likely values of the target chemicals, noted $\boldsymbol{x}^\star_{\text{MAP}}$, i.e. the value of $\boldsymbol{x}$ maximizing the distribution of $\boldsymbol{x}^\star$,
- the 95% confidence intervals for each component of $\boldsymbol{x}^\star$.

The performance of the ensemble (sensors+calibration/prediction method) will thus be evaluated against three sets of metrics.

- The error-based metrics, such as mean absolute error (MAE) or slope, offset and regression coefficient of the linear regression between predicted and measured quantities, allow to quantify the "global" error between the predicted chemical concentrations $\boldsymbol{x}^\star_{\text{MAP}}$ and their known value $\boldsymbol{x}^\star_{\text{ref}}$ measured using reference devices (see [4] for more metrics).
- These global error-based metrics may hide strong disparities in the dataset (e.g. outliers). To detect these, probability distribution of errors may be used. They are particularly useful from an application perspective (e.g. percentage of errors below a target threshold, percentage of outliers).
- We will also consider uncertainty-based metrics, that is global statistics on the confidence intervals (mean, standard deviation), as well as their probability distribution, to be compared with the measurement uncertainties provided by the reference devices.

III. Application

A. Description of the dataset

We consider a dataset made of 23 points out of 3 calibration experiments in tap water at ambient temperature. The free chlorine concentration is increased from 0 up to about 2mg/L over 7 to 8 points, at different pH ranging from 6 to 8. The uncertainty on pH is estimated at

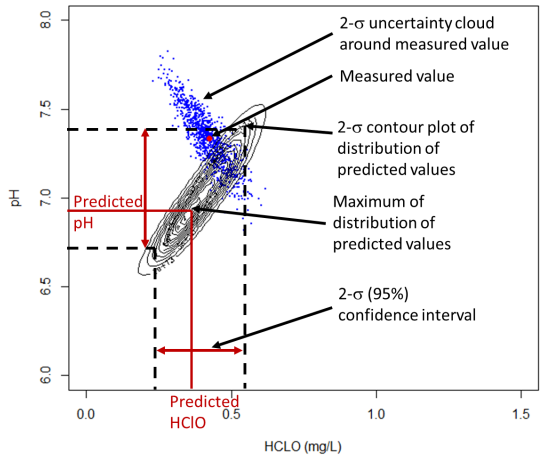

Figure 1. Graphical evaluation of the predictive capacity of the model.

0.15, on chlorine at 0.05mg/L. The target chemical is active chlorine, HClO, whose mass concentration c_{HClO} (in mg/L) is derived from chlorine mass concentration c_{Chl} (in mg/L), pH and temperature (T, in °C) according to the following formula:

$$c_{\mathrm{HClO}} = c_{\mathrm{Chl}} \left(1 + 0.98 \times 10^{\mathrm{pH}-\mathrm{pKa}(T)}\right)^{-1}, \qquad (3)$$

with $\mathrm{pKa}(T) = 7.5 - 0.01 \times (T - 30)$. The uncertainty on HClO concentration is derived as the standard deviation of a set of HClO values calculated from 1000 sets of (chlorine, pH, temperature) values, themselves randomly generated using their known uncertainties.

B. Identification of the calibration model $\boldsymbol{h}$

The form of the calibration model is not known beforehand. A polynomial form is selected - for generality sake - for each of the variables (HClO, pH, temperature). Then we look empirically, using training and testing equal to the full dataset, for the polynomial form with the lowest degree of freedom enabling both HClO and pH prediction with non-null slope and regression coefficients between measured and predicted values. Second order polynomials for each of the variables is found to provide the most consistent results.

C. Inverse problem - Performance analysis

For prediction, the dataset is split randomly at 70%/30% between training and testing and prediction performances are calculated after applying different conditions to the dataset. To ensure representative results, the metrics are calculated over 50 or 20 random splits. They are then averaged after (if needed) removal of large outliers. The number of large outliers is indicative of the stability of the tested parametrization of the method. An excerpt of the metrics of 4 interesting test configurations is provided in Table I. They show, as observed also when using a standard calibration process on HClO, that HClO is predictable for all configurations; the performance is significantly better when testing over a small concentrations of HClO. Unlike with traditional calibration, pH is here clearly predictable with good overall performance, but only when an upper threshold is placed on the absolute error on HClO. This is attributed to the much lower sensitivity in pH than HClO: when high uncertainties are accepted on HClO values, the sensitivity to pH is masked by the measurement uncertainty on HClO concentrations.

Table I
PREDICTION RESULTS FOR HCLO AND PH.

#repetitions	50	50	20	20
#points	23	19	21	17
Range HClO	No Thr.	No Thr.	<1mg/L	<1mg/L
Accur. HClO	No Thr.	<0.08mg/L	No Thr.	<0.08mg/L
#outlier	10%	6%	0%	5%
Slope HClO	1.4±0.5	1.4±0.5	1.05±0.32	1.24±0.36
R^2	0.73±0.15	0.81±0.12	0.68±0.26	0.81±0.17
MAE (mg/L)	0.16±0.08	0.13±0.09	0.09±0.03	0.07±0.03
≤0.1mg/L	52%	66%	63%	77%
≤3-σ error	44%	26%	50%	51%
Slope pH	0.65±0.37	1.0±0.2	0.36±0.43	0.94±0.22
R^2	0.2±0.56	0.85±0.08	-0.8±1.9	0.65±0.25
MAE	0.42±0.17	0.18±0.05	0.50±0.21	0.23±0.07
≤ 0.25	57%	75%	48%	71%
≤ 3-σ error	11%	39%	8%	42%

IV. CONCLUSIONS

To summarize, this paper shows that using a Bayesian framework to carry out sensor array calibration enables to recover reliably low sensitivity parameters which are hidden by measurement uncertainties when using other methods. Such an approach could thus open new and very interesting possibilities for a low-cost and more systematic monitoring of water quality.

REFERENCES

[1] Qi, P. et al. (2003). Nano letters, 3(3), 347-351.
[2] T. Hastie et al. "The Elements of Statistical Learning : Data Mining, Inference, and Prediction", Springer, New York, 2001.
[3] R. Y. Rubinstein, "Simulation and the Monte Carlo Method", Wiley, 1981.
[4] Delaine, F. et al. Sensors, 20(16), 4577
[5] Zucchi, Gaël, et al. U.S. Patent Application No. 16/604,423.