Assessment of bias removal techniques used in case of sensor parametric modelling

Pl Nikovski
University of food technologies – Plovdiv, Technical faculty, bul. Mariza 26, Plovdiv, Bulgaria
E-mail: plmnn@abv.bg

Abstract. The study assesses how two techniques for bias removing influence the parameterization of a humidity sensor model with known structure. Experimental data consist of SHT 31-DIS sensor responses to a step change of humidity from 12 %RH to 65 %RH. The test performed after parameter estimation shows that the time-series subtraction unbiased method produces a model with a better coefficient of determination. This and other results obtained are useful when a data pre-processing for black-box modelling of humidity sensors is used.

1. Introduction
Environmental sensors have been applied to a variety of fields such as industrial and home automation (figure 1), environmental monitoring, energy management etc. [1-6]. The increasing advances in CMOS technology have resulted in availability of high accuracy and small size digital sensors used to track temperature, humidity, particulate matter, VOC and CO₂ [7-10]. However, the precise sensor construction is not always a decisive factor when smart devices are designed. Particularly applicable to data mining and machine learning, sophisticated algorithms for analysis or decision making need an adequate mathematical model describing the real-time data processing in the sensor.

Available from the Swiss company Sensirion SHT 31-DIS is a humidity sensor that integrates in the same chip a polymer-based capacitive sensing element and an electronic circuit [11]. Under different conditions polymer material adsorbs or desorbs some amount of water vapor molecules

Figure 1. Sensirion SHT31 Smart Gadget [5] for humidity and temperature measurements.
depending on the relative humidity of the surrounding environment. As a result, the relative dielectric permittivity of the polymer and the capacitance of the sensing element are changed. An electronic circuit measures this change in capacitance and after correction and linearization, sampled values of relative humidity can be read through the I²C digital interface. SHT 31-DIS does not require any user calibration and is applicable for measuring humidity from 0 %RH to 100 %RH. In most cases, its error does not exceed ±2 %RH [11].

Due to the complexity of the processes in SHT 31-DIS, the lack of sufficient information about the polymer properties and the electronic circuit, it is impossible to obtain a physics-based sensor model. In this case, it is more appropriate to implement black box modelling [12]. Black box modelling uses experimental input-output data and does not require any knowledge about the internal structure and processes in the sensor. However, its effectiveness strongly depends on the data and their pre-processing. The main purpose of this study is to assess how two techniques for bias removing influence the parameterization of a humidity sensor model.

2. Materials and methods

By experiments, several step responses of the sensor SHT 31-DIS were obtained (output was sampled every 0,3 s): four of them when the input humidity rises rom $\varphi_A$ to $\varphi_B$ and four when it drops from $\varphi_B$ to $\varphi_A$. The relative humidities $\varphi_A$ and $\varphi_B$, approximately equal to 12 %RH and 59 %RH, were reproduced by saturated salt solutions. The collected data were used to estimate a sensor model of the following type:

\[ y[k] = -a_1 y[k - 1] - a_2 y[k - 2] - a_3 y[k - 3] + b_1 u[k - d - 1] + b_2 u[k - d - 2] + b_3 u[k - d - 3] + e[k] + a_1 e[k - 1] + a_2 e[k - 2] + a_3 e[k - 3] \] (1)

where backward shift operator is denoted by $q^{-1}$, $u[k]$ and $y[k]$, $k=0,1,2,\ldots$ are samples of the sensor input and output, $\theta = [b_1, b_2, b_3, a_1, a_2, a_3]^T$ is the parameter vector, $e[k]$ is an additive white noise and $d = 1$ is the input-output delay in samples. The order of the difference equation (1) and time delay are based on results of the preliminary data analysis.

Usually immediate use of the acquired data to build the model is not recommended. In addition to the noise, many factors such as outliers, missing values, offsets, drifts etc. degrade the quality of the data and may have a significant negative effect on the estimation process [13, 14]. Pre-processing of the raw data is an important step that allows many difficulties to be avoided later. The visual inspection of the data records shows that the number of gross errors like outliers and missing values is negligible (less than $10^{-5}$ %) and therefore a simple strategy for their correction is applied: a destroyed value is replaced by the previous neighboring value. The situation with the other impacts is considerably more complicated. It can be seen in figure 2a that under the repeatability experimental conditions the initial values of separate realization are different (this is also true for the final values in figure 3a). These values involve not only the true steady state sensor response but also the low-frequency disturbances, drifts and offsets from both the sensor and the experimental setup. By removing the bias from the experimental data, most of these unwanted components can be eliminated. Data, treated in this way, will no longer contain information about the actual values of the sensor steady states. But in most cases this does not affect seriously the estimation procedures.

3. Bias removing and sensor model estimation

Many approaches are used to remove the bias from the experimental data but two of them are fundamental [13, 14]. When the input bias $u_B$ and the output bias $y_B$ are known, they can just be subtracted from the experimental raw data $u[k]$ and $y[k]$. Now model parameters are determined by using deviations:

\[ \begin{align*}
  u[k] - u_B \\
  y[k] - y_B
\end{align*} \] (2)
Figure 2. Raw (a), unbiased (b) and scale unbiased (c) output experimental data before input rise.

Figure 3. Raw (a), unbiased (b) and scale unbiased (c) output experimental data after input rise.
For a linear time-invariant object the use of absolute value \( u[k] \) and \( y[k] \) or deviations (2) will lead to the same parameter vector \( \mathbf{\theta} \). Removing the bias as described above zeroing the initial data values. The final values also change. As can be seen from figure 3b they continue to be different from each other and the unbiased step responses will have varying heights. When the model parameters are found by minimizing the mean squared error (i.e. using the prediction error method) this is not desirable because the individual responses will have a different weight in the estimation. To solve this problem, unbiased responses are scaled:

\[
\begin{align*}
    u^*[k] &= \frac{u[k] - u_b}{u[\infty] - u_b} \\
y^*[k] &= \frac{y[k] - y_b}{y[\infty] - y_b}
\end{align*}
\]

(3)

where \( u[\infty] \) and \( y[\infty] \) are the final values after a sufficiently long time. In the case being considered here, the biases and the final values are calculated as sample means from groups of 60 observations at the beginning and the end of each raw step response. Taking into account equations (2) and (3), the model (1) can be further re-written as

\[
y^*[k] = -a_1 y^*[k - 1] - a_2 y^*[k - 2] - a_3 y^*[k - 3] \\
+ b_{1s} u^*[k - d - 1] + b_{2s} u^*[k - d - 2] + b_{3s} u^*[k - d - 3] \\
+ c (e[k] + a_1 e[k - 1] + a_2 e[k - 2] + a_3 e[k - 3])
\]

(4)

where

\[
c = \frac{u[\infty] - u_b}{y[\infty] - y_b}
\]

(5)

\[
b_{is} = C b_i, \quad i = 1, 2, 3.
\]

(6)

The unbiased and scaled sensor data are shown in figure 2c and figure 3c.

The implementation of the second approach does not require prior knowledge of the bias. Static components are removed by a special case of high-pass filtering realized by using the differences

\[
\begin{align*}
    u[k] - u[k - 1] &= (1 - q^{-1}) u[k] \\
y[k] - y[k - 1] &= (1 - q^{-1}) y[k]
\end{align*}
\]

(7)

instead of the variables \( u[k] \) and \( y[k] \). Scaling the new variables for the reasons previously discussed,

\[
\begin{align*}
    u^{**}[k] &= \frac{u[k] - u[k - 1]}{u[\infty] - u_b} \\
y^{**}[k] &= \frac{y[k] - y[k - 1]}{y[\infty] - y_b}
\end{align*}
\]

(8)

and having in mind the linearity of the difference equation (1), we obtain a model with the same structure as (4):

\[
y^{**}[k] = -a_1 y^{**}[k - 1] - a_2 y^{**}[k - 2] - a_3 y^{**}[k - 3] + b_{1s} u^{**}[k - d - 1] \\
+ b_{2s} u^{**}[k - d - 2] + b_{3s} u^{**}[k - d - 3] \\
+ c (e[k] + a_1 e[k - 1] + a_2 e[k - 2] + a_3 e[k - 3])
\]

(9)

Parameterization of the model can be done in several different ways [13, 14]. According to the Prediction error method (PEM), the “best” parameters will be those, by means of which the output can be accurately predicted at any moment of time. Most often, an estimate of \( \mathbf{\theta} \) (denoted by \( \hat{\mathbf{\theta}} \)) is determined
by minimizing the sum of the error squared between the pre-processed $y^*[k]$ and one-step ahead predicted value

$$
\hat{y}^*[k|k-1] = -a_1y^*[k-1] - a_2y^*[k-2] - a_3y^*[k-3] + b_1u^*[k-d-1]
+ b_2u^*[k-d-2] + b_3u^*[k-d-3]
$$

(10)

over a set of $N$ samples:

$$
\theta^* = \text{argmin}_{\theta} \sum_{k=0}^{N-1} (y^*[k] - \hat{y}^*[k|k-1])^2
$$

(11)

The difference

$$
\varepsilon^* = y^*[k] - \hat{y}^*[k|k-1]
$$

(12)

is known as a prediction error. An advantage of PEM is its ability to find good approximation when there is a mismatch between the user defined model and the real dynamics of the studied system [13, 14]. The solving equation (11) with $N - 1 = 10000$ yields the model

$$
y^*[k] = 2.917y^*[k-1] - 2.834y^*[k-2] + 0.918y^*[k-3] + 0.066u^*[k-2]
- 0.131u^*[k-3] + 0.065u^*[k-4] + e[k] - 2.917e[k-1]
+ 2.834e[k-2] - 0.918e[k-3]
$$

(13)

Using the same approach and taking into account equations (7) - (9), the second model was also evaluated:

$$
y^{**}[k] = 2.930y^{**}[k-1] - 2.859y^{**}[k-2] + 0.930y^{**}[k-3] + 0.060u^{**}[k-2]
- 0.120u^{**}[k-3] + 0.060u^{**}[k-4]
+ \frac{1}{1 - q^{-1}}(e[k] - 2.930e[k-1] + 2.859e[k-2] - 0.930e[k-3])
$$

(14)

or after some transformation

$$
y^{**}[k] = 3.930y^{**}[k-1] - 5.789y^{**}[k-2] + 3.789y^{**}[k-3] - 0.930y^{**}[k-4]
+ 0.060u^{**}[k-2] - 0.180u^{**}[k-3] + 0.180u^{**}[k-4]
- 0.060u^{**}[k-5] + e[k] - 3.930e[k-1] + 5.789e[k-2]
- 3.789e[k-3] + 0.930e[k-4]
$$

(15)

The impact of the compared methods for bias removing on the quality of the sensor model can be assessed in different ways [13, 15, 16]. The figure 4 and figure 5 show the prediction error in both cases, calculated on fresh data not used in the identification process.

When data are preprocessed by the time-series subtraction method, the maximum prediction error does not exceed 0.028. The model, obtained by the time-series differencing method has approximately the same magnitude of error: -0.031. Another widely used metric in model estimation is the coefficient of determination $R^2$:

$$
R^2 = 1 - \frac{\sum_{k=0}^{N-1} \varepsilon^2[k]}{\sum_{k=0}^{N-1} (y[k] - \bar{y})^2}
$$

(16)

where $\bar{y}$ is the mean value of $y[k]$. $R^2$ is a number, usually between 0 and 1. In an ideal case, i.e. when $\bar{y}[k] = y[k]$, $R^2 = 1$ and it decreases as the discrepancy between the original system and the model output increases. In the first case the coefficient of determination is $R^2 = 0.99$ and in the second case $R^2 = 0.81$: the first model fits the data better than the second one. In addition, its structure is simpler (see equations (13) and (15)).
4. Conclusion
This paper discusses a case where the parameters of a humidity sensor SHT 31-DIS discrete model with a known structure are estimated applying two different unbiasing techniques. Experimental data consist of sensor responses to a step change of humidity from 12 %RH to 65 %RH. Separate realizations have some bias and two techniques have been applied to remove it. In the first case the bias stated in the raw data before input changing is subtracted and after that the obtained variation is scaled. In the second case, the raw data are first differentiated and after that they are scaled. Replacing the original variables with pre-processed variables yields the new model with unbiased input and output. Its parameterization is done by the prediction error method. After estimation, a test was conducted with fresh data not used in the identification process. The coefficients of determination for the time-series subtraction and time-series differencing unbiased methods were $R^2 = 0.99$ and $R^2 = 0.81$ respectively. Therefore, in cases when step response data are used to parametrize a model with the considered structure it is more appropriate to apply the first approach for bias removing.

References
[1] Hancke G, Silva B and Hancke Jr G 2013 The role of advanced sensing in smart cities. Sensors 13, 393-425
[2] Kim G, Son Y, Lee J et al. 2016 Air Pollution Monitoring and Control System for Subway Stations Using Environmental Sensors. Journal of Sensors 2016
[3] Nedev N, Nenova Z and Ivanov S 2014 Virtual instruments for humidity and temperature measurements. Information Technology Based Higher Education and Training, York pp 1-6
[4] Kumar A and Hancke G 2014 An Energy-Efficient Smart Comfort Sensing System Based on the IEEE 1451 Standard for Green Buildings. IEEE Sensors Journal 14 4245-52
[5] Sensirion 2015 SHT31 Smart Gadget User Guide. (Revision 1) p11
[6] Nenova Z and Dimchev G 2013 Compensation of the Impact of Disturbing Factors on Gas Sensor Characteristics. Acta. Polytech. Hung. 10 97–111
[7] Ueno K, Hirose K, Asai T and Amemiya Y 2006 Ultralow-Power Smart Temperature Sensor with Subthreshold CMOS Circuits. International Symposium on Intelligent Signal Processing and Communications (Tottori) pp 546-549
[8] Passeri D et al. 2003 High-resolution CMOS particle detectors: design and test issues. IEEE Nuclear Science Symposium (Portland) pp 331-334
[9] Maeng S, Guha P, Udrea F et al. 2008 SOI CMOS-based smart gas sensor system for ubiquitous sensor networks. *ETRI Journal* **30** 516–525
[10] Nizhnik O, Higuchi K and Maenaka K 2012 Self-calibrated humidity sensor in CMOS without post-processing. *Sensors* **12** 226–232
[11] Sensirion 2019 Datasheet SHT3x-DIS Humidity and Temperature Sensor. (Version 6) p 22
[12] Ljung L 2001 Black-box models from input-output measurements. *IMTC 2001* (Budapest) pp 138-146
[13] Ljung L 2014 *System identification: Theory for the user*. (Prentice Hall) p 640
[14] Tangirala A 2014 *Principles of system identification: Theory and practice*. (NW: CRC Press) p 908
[15] Botchkarev A 2018 Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology. *(Available online: https://arxiv.org/abs/1809.03006 accessed on 9 September 2018)*
[16] Tian Y, Nearing G, Peters-Lidard C et al. 2016 Performance Metrics, Error Modeling, and Uncertainty Quantification. *Monthly Weather Review* **144** 607-613