European Commission | Horizon 2020 European Union funding for Research & Innovation

CAPTOR

Collective Awareness Platform for Tropospheric Ozone Pollution

Work package	WP2
Deliverable number	D2.3
Deliverable title	Software tool for Ozone Concentration Estimation Development
Deliverable type	R
Dissemination level	PU (Public)
Estimated delivery date	M9
Actual delivery date	30/10/2016
Actual delivery date after EC review	15/09/2017
Version	2.0
Comments	

Authors
Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Manel Guerrero-Zapata (UPC)
Anna Ripoll, Mar Viana (CSIC)

Project funded by the European Union's Horizon 2020 research and innovation programme under Grant Agreement N° 688110

Document History

Version	Date	Contributor	Comments
V0.1	20/07/2016	Jose M. Barcelo-Ordinas (UPC)	Outline
V0.2	10/09/2016	Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Manel Guerrero-Zapata (UPC)	First draft
V0.3	28/09/2016	Anna Ripoll, Mar Viana (CSIC)	Peer review
V1.0	30/10/2016	Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Manel Guerrero-Zapata (UPC)	Final version
V2.0	15/09/2017	Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal (UPC)	Final version after EC review

Table of Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
MLR	Multivariate Linear Regression
PR	Palau Reial reference Station
RDF	Resource Description Framework
RH	Relative Humidity
RSS	Residual Sum of Squares
RSE	Residual Standard Error
RRSE	Relative Residual Standard Error
T	Temperature
XML	eXtensible Markup Language

Executive Summary

Description of the work

This deliverable describes how the data taken from the CAPTOR ozone sensor nodes is processed in order to produce ozone data with certain degree of quality. The processed data is the raw data taken by the resistor of the ozone sensor and the output data of the process is the believed true ozone concentration with certain degree of quality. Ozone concentrations are given with a relative error with respect ground truth data measured by accurate reference stations.

Objectives

The main objectives of the deliverable are:

- Describe a method to calibrate low-cost sensors
- Assess the method defined by means of a quality parameter, the RRSE (Relative Residual Standard Error)

1. Research Context

The purpose of this deliverable is to describe the process of transforming raw data taken from the CAPTOR sensor nodes to real ozone concentrations with the best possible quality in terms of relative error with respect ground truth data measured by accurate reference stations. The **ground truth data** is defined as the data taken by direct observation, i.e., by a reference station, in contrast to that one that is provided by inference.

CAPTOR nodes are built with low-cost sensors, e.g., metal-oxide ozone sensors that are not calibrated in specialized laboratories like the reference stations. When the low-cost sensor interacts with the pollutant its resistor measures a value that represents the ozone concentration in terms of electric resistance. A multivariate linear regression is then used in order to calculate the ozone concentrations. In this deliverable, it is described how to obtain ozone concentrations by regressing over ground truth surface ozone concentrations measured by reference station instrumentation.

2. Technological Context

Ambient air quality is routinely monitored by networks of air monitoring stations or reference stations, usually deployed and operated by public administrations. Reference stations measure pollution data with high accuracy. The equipment costs of these reference stations are from the tens of thousands of Euros to the hundred of thousands of Euros. Thus, the number of reference stations in an area is low.

Due to its high cost of purchase and operation, normally only few monitoring stations are deployed in a given area, meaning that the data they produce has a limited spatial resolution. On the other hand, in some geographical areas of the world, the information on air quality is either incomplete or non-existent.

The calibration process in the examples of this deliverable has been done using data from the Palau Reial reference station, in Barcelona, Spain (41°23’14’’N, 2°6’56’’E), of the “Xarxa de Vigilància i Previsió de la Contaminació Atmosfèrica (XVPCA)”[1], operated by the Catalan local government, Figure 1.(a).

Each CAPTOR node[2], Figure 1.(b), consists of:

i. Sensing subsystem in charge of sampling sensing data. This subsystem contains a sensor board with M ozone (MiCS 2614) sensors (typically M=3,…,5), 1 Temperature (MCP9700A) sensor and 1 Humidity (808H5V5) sensor. Sensors sample the area every period T_{sample} and produce values, called from now ***raw data***, which represent the resistor value of the sensor, Figure 2.

ii. Communication subsystem in charge of communicating the data to the CommSensum platform[3]. The communicating subsystem is based on WiFi technology or on cellular technology. The raw data is sent in XML/RDF format and stored in the CommSensum MySQL database.

[1] http://mediambient.gencat.cat/ca/05_ambits_dactuacio/atmosfera/qualitat_de_laire/avaluacio/xarxa_de_vigilancia_i_previsio_de_la_contaminacio_atmosferica_xvpca/index.html

[2] Deliverable D2.2, “Release of electro-chemical board design (a)”

[3] Deliverable D2.4, “Open Link Data repository development”

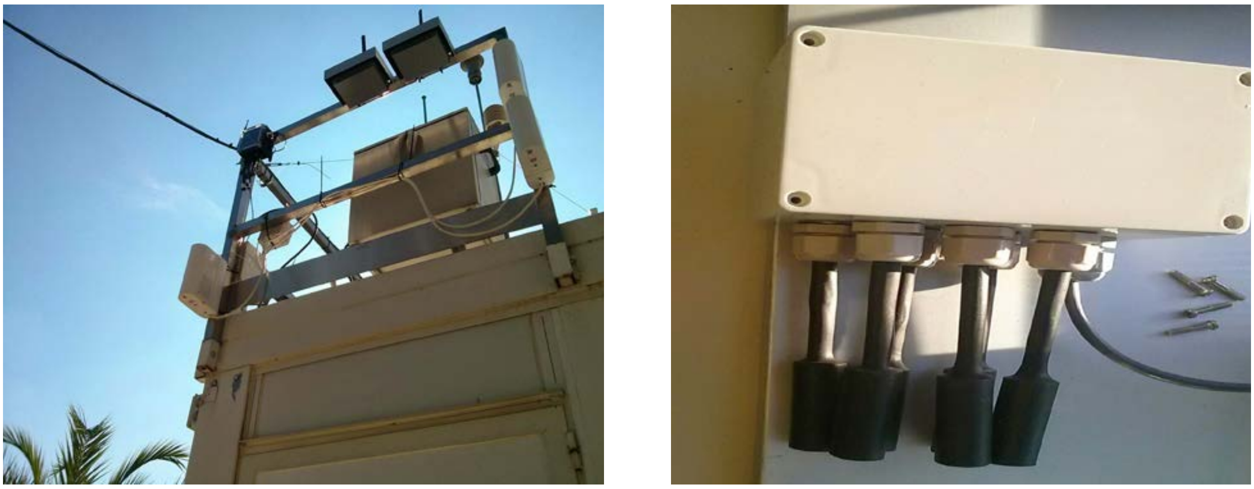


Figure 1. (a) Palau Reial reference Station, (b) CAPTOR node with 5 ozone sensors, and 1 Temperature/Humidity sensor

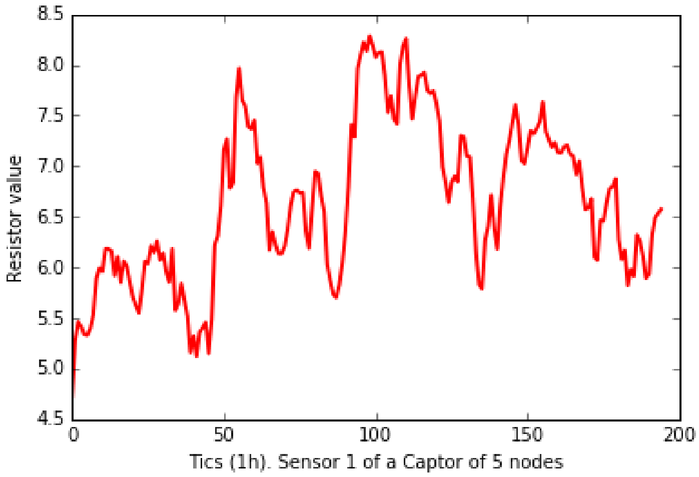


Figure 2. Resistor values for sensor 1 (ozone) taken from CAPTOR-ID34 node.

On the other hand, the data obtained by reference stations and downloaded from official repositories are given in $\mu g/m^3$, Figure 3.

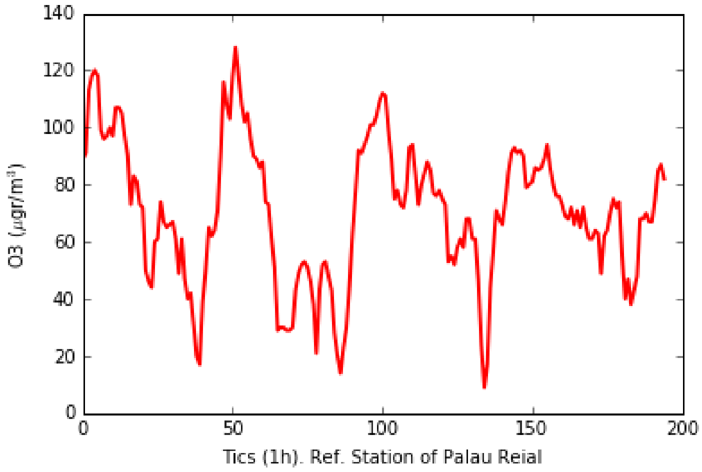


Figure 3. Ozone true data from Palau Reial reference station.

A way of visually checking that the data follows the same pattern is normalizing both parameters or attributes. The reason of normalizing is that both measures are on different units. Normalization means centering and scaling the data so that a unit of attribute number 1 means the same thing as a unit of attribute number 2. Let $\boldsymbol{x} \in R^n$ be a vector of dimension n, where n is the size of the sample set, representing the ozone in case of the reference station or resistor in case of a CAPTOR sensor. Let $\boldsymbol{\mu}_x$ be the average, σ^2_x be the variance and σ_x be the standard deviation of **x** respectively. The normalized variable y of x is defined as:

$$y = \frac{x - \mu_x}{\sigma_x} \qquad (1)$$

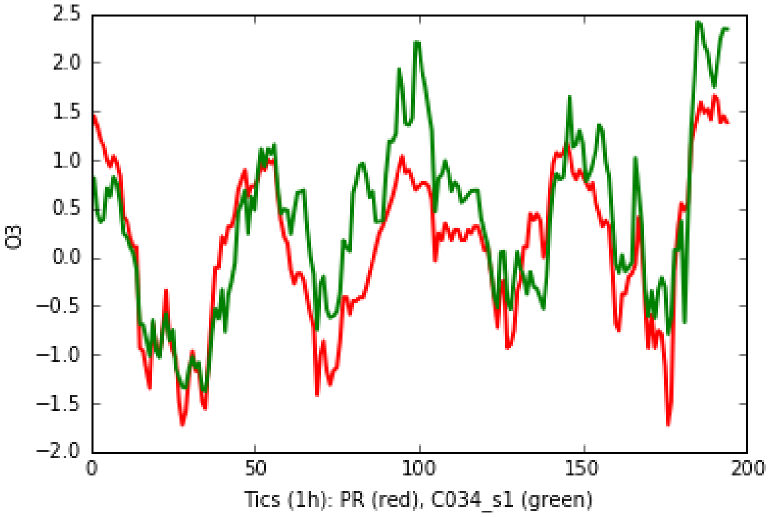


Figure 4. Normalized ozone data from Palau Reial reference station and normalized resistor values for sensor 1 (ozone) taken from CAPTOR-ID34 node.

The normalization adjusts the variables so that they all have zero mean and a standard deviation of one. Figure 4 shows the normalized ozone true data from Palau Reial reference station and normalized resistor values for sensor 1 (ozone) taken from CAPTOR-ID34 node, where CAPTOR-ID34 was deployed on the roof of the reference station. As it can be observed, both measures follow a similar pattern since they are measuring the same physical phenomena. Although it is observed that the pattern is not exactly the same, it is known that measuring the ozone from the low-cost sensor data is not enough, since other parameters such as Temperature and Humidity also impacts the reconstruction of the ozone concentrations. The conclusion is that the uncalibrated low-cost sensor may be calibrated from the reference station data.

3. Calibration of CAPTOR ozone nodes

3.1 Mathematical background

As it has been mention in the previous section, each CAPTOR node is deployed on the roof of the reference station during a period time of at least 3 weeks. In general, the calibration of a sensor means to approximate the true value Y by a function f(X):

$$Y = f(X) + \varepsilon \qquad (2)$$

Where *f* is a fixed but unknown function, X is a vector of *p* predictors or input variables and ε is a random error term distributed as a zero mean Gaussian random variable with variance σ^2, i.e, $\varepsilon \sim N(0, \sigma^2)$ and independent of X. In this approximation, eq (2) is modelled by saying that we are **regressing Y on X** (or Y onto X).

In order to find a regression of the data, we may consider linear combinations of fixed non-linear functions of the input variables, of the form:

$$f(X) = \beta_0 + \sum_{i=1}^{p} \beta_i \phi_i(X)$$

where $\phi_i(X)$ are known as **basis functions** and β_0 is the **slope or intercept** and the β's (i=1,…,p) are the **regression coefficients**. The most basic model is using a Multivariate Linear Regression (MLR) in which each basis function ϕ_i is linear with respect X_i i.e., $\phi_i(X)=X_i$:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon = \sum_{i=0}^{p} \beta_i X_i + \varepsilon = \beta X + \varepsilon \qquad (3)$$

It is to say, Y is approximated by a linear combination of the predictors. Note that the dimension of Y and X_i (i=1,…,p) is N, the size of the sample set ($X_i, Y \in R^N$ or $X \in R^{N \cdot (p+1)}$), where X has been extended by a vector X_0 of 1's and the slope β_0 has been integrated in the β. More complicated basis functions may be used, such as powers of x, $\phi_i(X)=X^j_i$ or polynomial functions of several features.

In this first version of the calibration process for the CAPTOR project, the Multivariate Linear Regression (MLR) will be used. In order to regress Y on X, the coefficients β have to be approximated by β'. Our goal is to obtain coefficient estimates β' such that the linear model of eq (3) fits the available data well, that is, so that $y \approx \beta' X$. In other words, we want to find those coefficients β' where by the resulting line is as close as possible to the N data points. We refer to James et al.[4] for finding the β', e.g., minimizing the least squares criterion.

Let $y_i' = \beta' X_i$ be the prediction of y_i based on the value of x_i. The difference between the estimated value y_i' and the original value y_i is called the **residual**, $e_i = y_i - y_i'$. We define the Residual Sum of Squares (RSS) as:

$$RSS = e_1^2 + \ldots + e_n^2 = \sum_{i=1}^{n} (y_i - y_i^{'})^2 = \sum_{i=1}^{n} (y_i - \beta_i^{'} x_i)^2 \qquad (4)$$

We wonder how close are the β' from the real true β. In computing the standard errors in β', they depend on the variance σ^2 of the error ε. However, this variance is unknown. A way of estimating this variance is to define the Residual Standard Error (RSE), defined as:

$$RSE = \sqrt{\frac{RSS}{n - p + 1}} \qquad (5)$$

The quality of a linear regression fit is typically assessed using the **Residual Standard Error (RSE)** or the **Relative Residual Standard Error (RRSE)** obtained by normalizing the RSE with respect the mean of y.

3.2 CAPTOR Calibration Procedure for a node with a single ozone sensor

Let us assume that the data set for calibration has size N. We assume that each of the M ozone

[4] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, "An introduction to statistical learning, with applications in R", Springer, 2013.

sensors is independent of each other. The data consist of:

- The Pal reference station data Y∈R^N,
- The ozone (O3) data captured by each sensor X_1=X∈R^N, with M ozone sensors,
- The Relative Humidity data captured by the sensor X_2=HR∈R^N,
- The Temperature data captured by the sensor X_3=T∈R^N,

The MLR model used is, then:

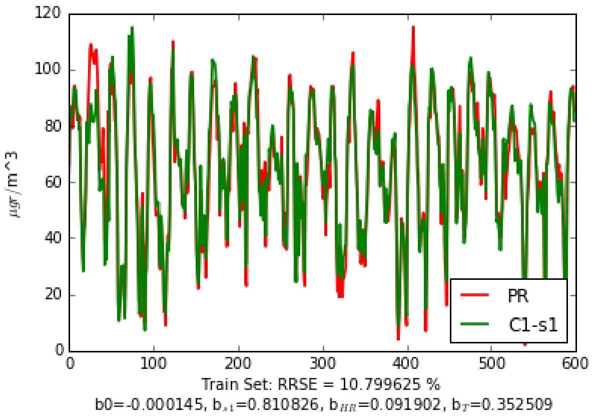
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon = \beta_0 + \beta_1 X + \beta_2 HR + \beta_3 T + \varepsilon \qquad (6)$$

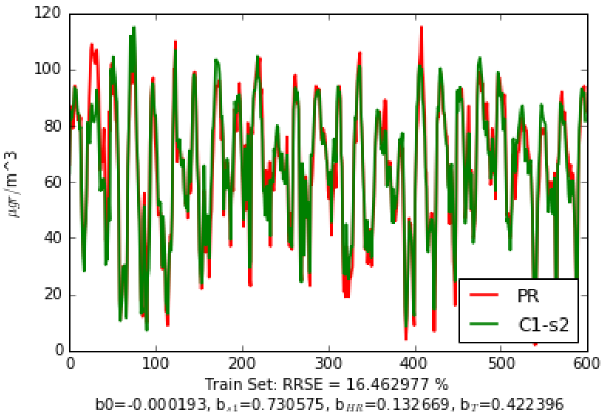
Where we have recalled X_1=X (ozone), X_2=HR (Relative Humidity) and X_3=T (Temperature) for commodity. In order to calibrate a CAPTOR node with a single ozone sensor, we proceed as follows:

- The data set N is divided in two sets: the **training set** of size N_1 and the **test or validation set** of size N_2.
- Obtain the β' by minimizing the least squares criterion over the training set and obtain the RRSE as quality parameter of the training set by using the RSS of the training set.
- Predict the y'= β_0' + β_1' X + β_2' HR + β_1' T where X,HR,T∈R^{N2} are data of the validation set. Obtain the RRSE of the validation set by using the RSS of the test set.

At the end of the process, each M individual sensor is calibrated per each CAPTOR node. Now the question is which one represents best the CAPTOR node. The sensor that has less validation RRSE is taken as reference sensor for that node.

For showing how the method works, let us take a set of N=1200 samples. Figure 5 shows the calibration of individual sensors (j=1,…,5) for the Training set, N_1=600 samples, while Figure 6 shows the prediction over another N_2=600 new samples. The figures show the coefficients obtained for each individual sensor calibration and the training and validation RRSE.





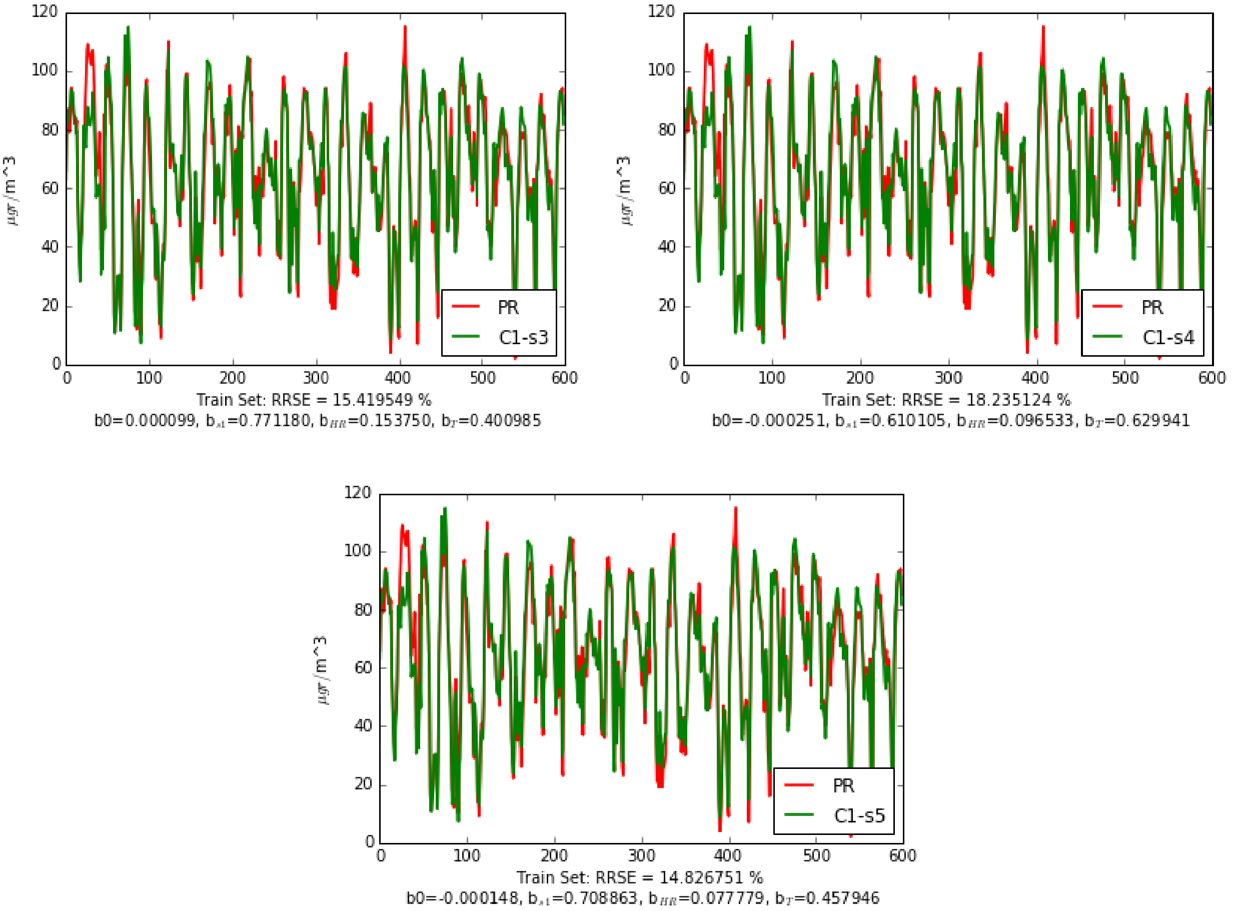
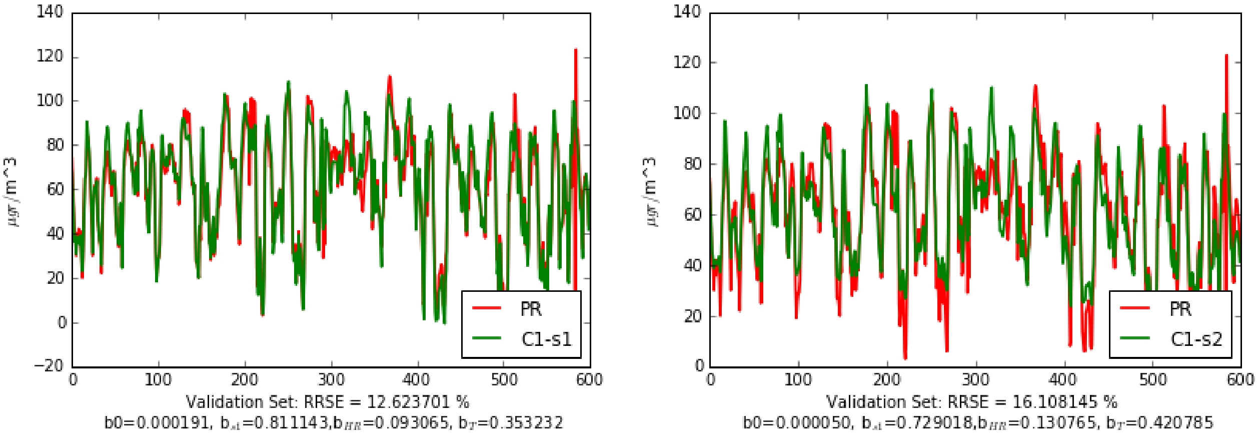


Figure 5. Train calibrated data for CAPTOR C1, 5 ozone + 1 Humidity + 1 Temperature sensors for individual sensors s1, s2, s3, s4 and s5.



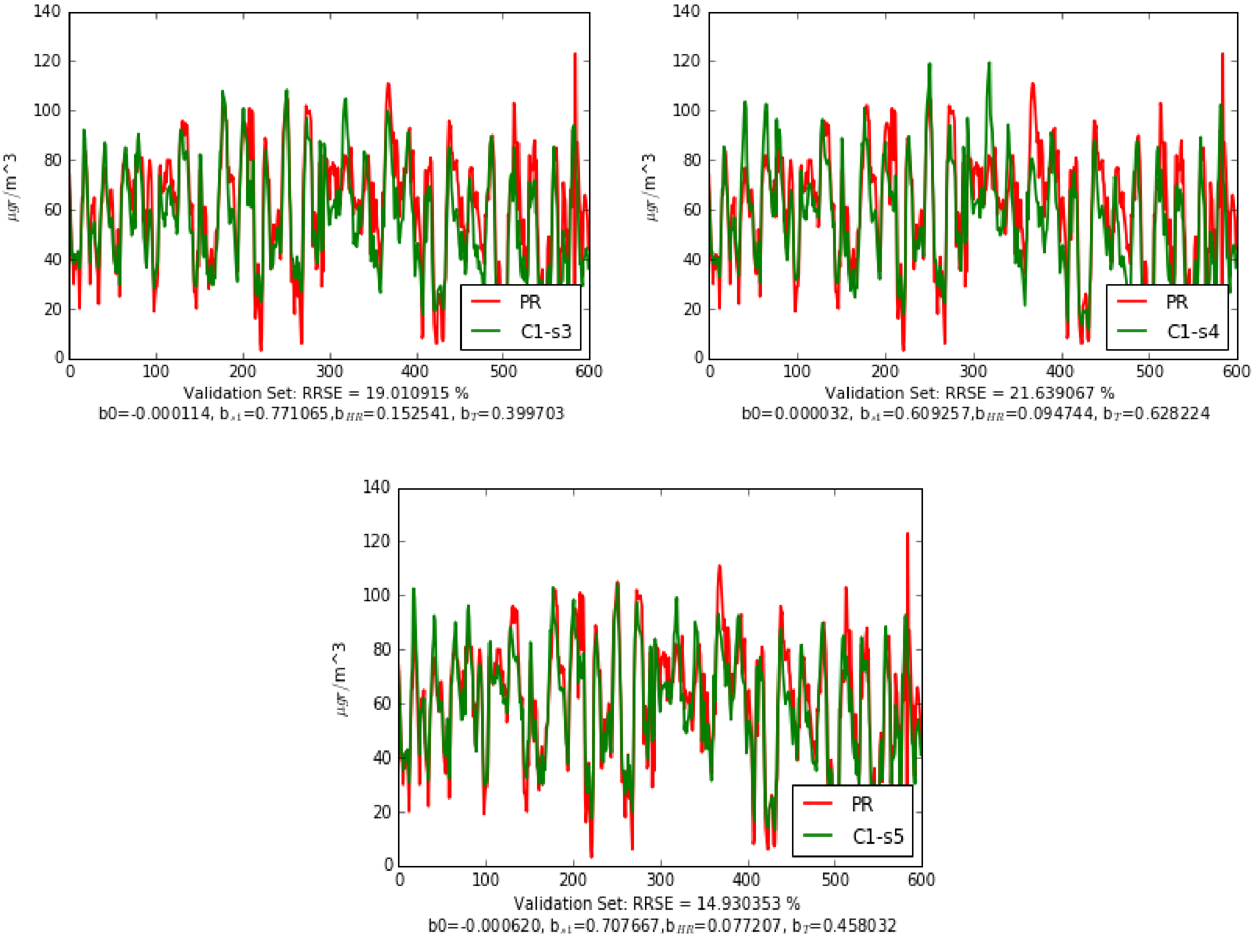


Figure 6. Validation calibrated data for CAPTOR C1, 5 ozone + 1 Humidity + 1 Temperature sensors for individual sensors s1, s2, s3, s4 and s5.

	Sensor S1	**Sensor S2**	**Sensor S3**	**Sensor S4**	**Sensor S5**	**Fusion**
Training RRSE	10.8%	16.46%	15.42%	18.23%	14.82%	9.20%
Validation RRSE	12.62%	16.10%	19.01%	21.64%	14.93%	11.74%

Tabla 1. Training and Validation RSSE for calibration using individual sensors and fusion of sensors.

As it can be observed in Table 1, the minimum validation RRSE is produced by sensor 1, thus, this one would be selected for calibrating the CAPTOR node.

3.3 CAPTOR Calibration Procedure by averaging the different sensors

Let us note that a similar way of calibrating the data is to average over the M ozone sensor data. Now, the MLR model would be:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon = \beta_0 + \beta_1 X_{AVG} + \beta_2 HR + \beta_3 T + \varepsilon \qquad (7)$$

Where we have recalled X_1=X_{avg} (Average ozone for all the sensors), X_2=HR (Relative Humidity) and X_3=T (Temperature). We have tested this model and conclude that it does not give better results than taking the best sensor among the M sensors.

3.4 CAPTOR Calibration Procedure for a node by fusion of ozone sensor data

Let us assume that the data set for calibration has size N. The data consist of:

- The Palau Reial reference station data $Y \in R^N$,
- The ozone (O3) data captured by each sensor X_1=$X \in R^N$, with M ozone sensors,
- The Relative Humidity data captured by the sensor X_2=$HR \in R^N$,
- The Temperature data captured by the sensor X_3=$T \in R^N$,

The MLR model used is, then:

$$Y = \beta_0 + \sum_{j=1}^{M} \beta_j X_j + \beta_{M+1} X_{M+1} + \beta_{M+2} X_{M+2} + \varepsilon = \beta_0 + \sum_{j=1}^{M} \beta_j X_j + \beta_{M+1} HR + \beta_{M+2} T + \varepsilon \tag{8}$$

Where we have recalled X_j=X (ozone), with j=1,…M, X_{M+1}=HR (Relative Humidity) and X_{M+2}=T (Temperature) for commodity. In order to calibrate a CAPTOR node by fusion of ozone sensor data, we proceed similarly as before as follows:

- The data set N is divided in two sets: the **training set** of size N_1 and the **test or validation set** of size N_2.
- Obtain the β' by minimizing the least squares criterion over the training set and obtain the RRSE as quality parameter of the training set by using the RSS of the training set.
- Predict the **$y' = \beta_0' + \Sigma\, \beta_j' X_j + \beta'_{M+1}$ HR + β'_{M+2} T** where X_j,HR,$T \in R^{N2}$ are data of the validation set. Obtain the RRSE of the validation set by using the RSS of the test set.

At the end of the process, there is a virtual sensor calibrated for the CAPTOR node. For showing how the method works, let us take a set of N=1200 samples. Figure 7 shows the calibration of fusion of sensors (j=1,…,5) for the Training set, N_1=600 samples, while Figure 8 shows the prediction over another N_2=600 new samples. The figures show the coefficients obtained for each fusion calibration and the training and validation RRSE.

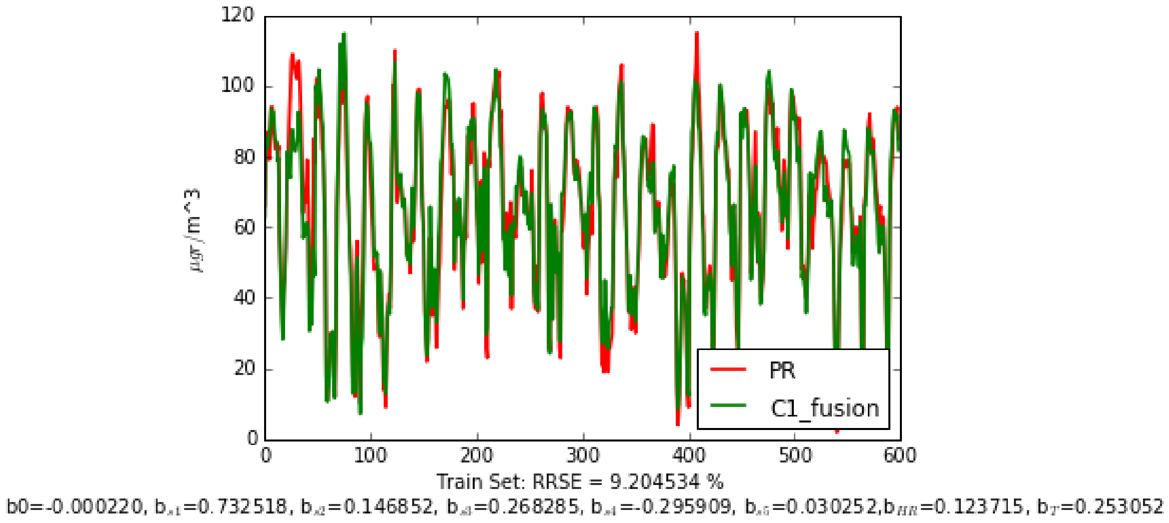


Figure 7. Train calibrated data for CAPTOR C1, 5 ozone + 1 Humidity + 1 Temperature sensors for fusion of 5 sensors.

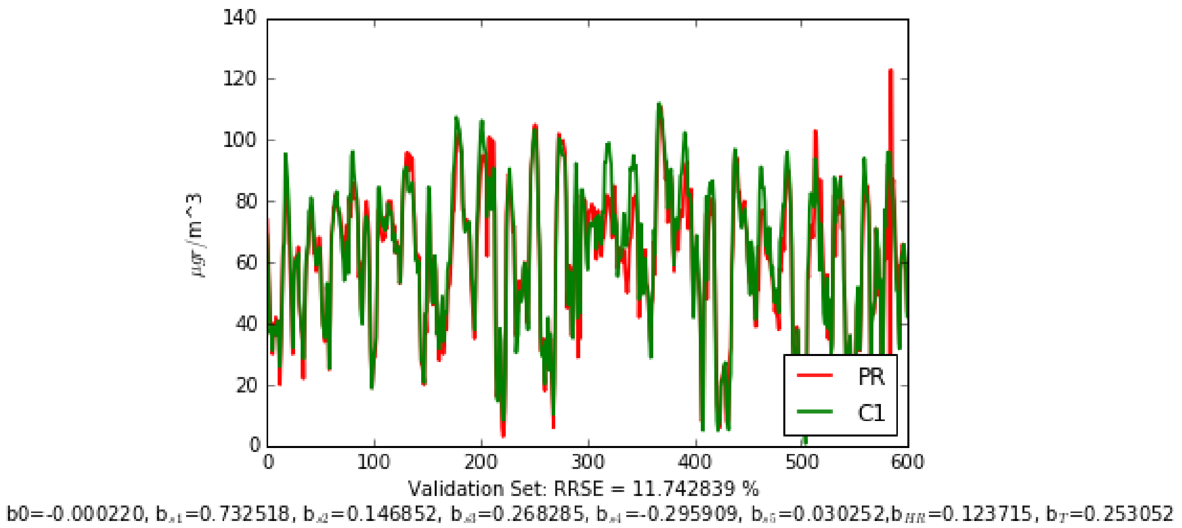


Figure 8. Validation calibrated data for CAPTOR C1, 5 ozone + 1 Humidity + 1 Temperature sensors for fusion of 5 sensors.

As it may be observed in Table 1, the fusion of data obtains better results than using individual sensors. The intuition behind this fact resides in that the multivariate regression chooses automatically, i.e., “regress”, over the whole number of sensors selecting the coefficients for each sensor that minimizes the least squares criteria.

4. Integration with the CommSensum platform

The CommSensum platform[5] provides an open link data repository for data collected by reference stations and CAPTOR nodes. As explained in section 3, each captor node is calibrated according to the fusion of sensors mode since it gives the best RRSE. The β’ coefficients obtained from the N samples obtained during the calibration phase are stored in the CommSensum Database in per node basis.

Now, deployed CAPTOR nodes send the raw data they sample to the CommSensum platform. This raw data is a vector of (x’$_1$, …,X’$_M$, HR’, T’) per sample. In order to obtain the true ozone of the captor node, the CommSensum platform software calculates the ozone concentration according to the following formula:

$$y = \beta'_0 + \sum_{j=1}^{M} \beta'_j \, x'_j + \beta'_{M+1} HR' + \beta'_{M+2} T' \qquad (9)$$

The new value y is stored in the CommSensum platform. Now it can be plotted and visualized in the CommSensum platform or shown in the AirCat application[6].

[5] Deliverable D2.4, “Open Link Data repository development”

[6] Deliverable D2.5, “Mobile app development”