# Data Fusion for the Improvement of Low-Cost Air Quality Sensors

Theodosios Kassandros, Evangelos Bagkis, and Kostas Karatzas

**Abstract** Aim of this study is to develop a calibration procedure through Machine Learning to upgrade the low-cost air quality sensor performance and investigate the generalization of this function over a specific area towards air quality data fusion.

**Keywords** Air quality · Low-cost sensors · Machine learning · Data fusion

## 1 Introduction

Bad air quality (AQ) has a negative impact on peoples' quality of life. The small number of monitoring stations used for the official AQ monitoring and the operationally available air pollution modelling tools still leave open space for improving local AQ knowledge. The KASTOM project (www.air4me.eu) is developing a versatile and flexible air quality monitoring and forecasting system by deploying an IoT-oriented network of low-cost AQ sensor nodes (LCAQSN), while in parallel developing a state-of-the-art emission modeling module combined with state-of-the-art three-dimensional AQ models. LCAQSN can cover larger areas due to their low cost but are lacking the necessary accuracy.

## 2 Materials

The Greater Thessaloniki Area (GTA) is the second largest urban agglomeration in Greece hosting more than 1 million inhabitants. The KASTOM project has installed 33 LCAQSN in the GTA including: (a) Particle sensors (PM10–PM2.5: PMS5003, Beijing Plantower Co., Ltd.), (b) sensors for gaseous pollutants ($NO_2$, $O_3$ and CO: Alphasense Ltd., U.K.) and meteorological sensors (Air Temperature, Relative Humidity and Air Pressure, BME280 Bosch Sensortec, Germany).

T. Kassandros (✉) · E. Bagkis · K. Karatzas
Environmental Informatics Research Group, School of Mechanical Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: tkassand@physics.auth.gr

C. Mensink and O. Jorba (eds.), *Air Pollution Modeling and its Application XXVIII*,
Springer Proceedings in Complexity, https://doi.org/10.1007/978-3-031-12786-1_24

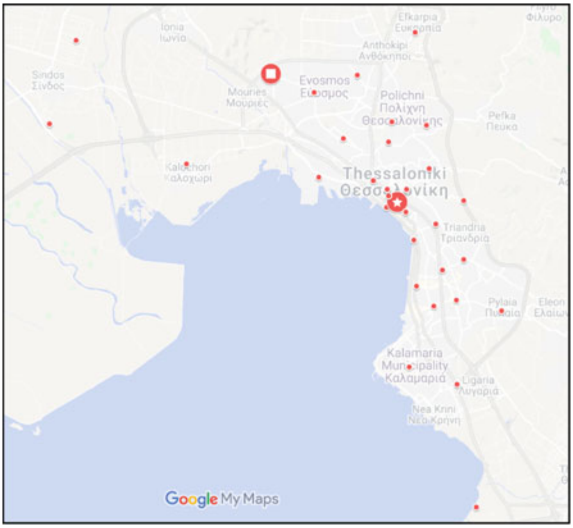

**Fig. 1** LCAQSN network in the GTA (Dots). □: KORD ✪: AGSOF

**Table 1** Dataset description

| NodeSet (Nset) | FusionSet (Fset) | Target variables |
|---|---|---|
| PM10, PM2.5, PM1, CO, $O_3$, $NO_2$, relative humidity, temperature, pressure | NodeSet + wind speed, wind direction, precipitation, friction velocity, saturation, saturation ratio, traffic | PM10, $O_3$, $NO_2$ |

In this study, we have collocated six nodes with two reference stations (Fig. 1) in Agias Sofias (AGSOF) and Kordelio (KORD) areas, classified by the European Environment Agency as an urban traffic and urban industrial station respectively.

The initial dataset (NodeSet) consists of six nodes measurements (Node1–3 located in AGSOF and Node4–6 located in KORD) for the period of 21/12/2019–10/03/2020 and the reference stations measurements for PM10, $O_3$ and $NO_2$ ($NO_2$ measurements in KORD omitted due to missing value problems). The additional dataset (FSet) included meteorological modeling (WRF) and free traffic flow data (Salanova et al., 2018). All variables are presented in Table 1.

## 3 Methods

The first step of the computational procedure aimed at generating a set of features, capturing the maximum amount of information. We therefore applied time lags (from 1 to 12 h) and rolling—aggregation statistics (6 and 12 h) to all the variables, leading to

161 features for the Nset and 401 features for the Fset. To reduce noise introduced by features, a feature reduction procedure was followed employing the Random Forest Feature Importance (RFFI) method. We then employed a Machine Learning (ML)-oriented modeling approach, making use of the reference station measurements as target parameters (PM10, $O_3$ and $NO_2$) to calibrate and upgrade the KASTOM nodes performance. Models were trained in the two subsets, for each sensor and location. A Gradient Boosting algorithm was used (Friedman, 2001), combining the outputs sequentially from individual regression trees, where each new tree helps to correct errors made by a previously trained tree.

To evaluate the initial performance of the LCAQSN, the Pearson Correlation Coefficient (r) and Coefficient of Divergence (CoD) were calculated. The ML models were evaluated using a fivefold time forward cross validation on a rolling basis, using the Coefficient of Determination ($R^2$) and the Relative Expanded Uncertainty (REU), following the methodology described in the Guide to the Demonstration of Equivalence of Ambient Air Monitoring Methods (EUD, 2008). According to the European Air Quality Directive, uncertainties for “class 1 sensor” or indicative measurements are 50, 25, 30% and for “class 2 sensor” or objective measurements are 100, 75, 75% for PM10, $NO_2$ and $O_3$ respectively.

## 4 Results

Field calibration of an LCAQSN network requires the individual nodes to perform identical to each other, this being the first condition to apply the same calibration function. This was checked with the aid of the CoD versus Pearson (Fig. 2). All PM10 sensors scored very high Pearson and very low CoD thus behaving identical, but the gas sensors, and especially $O_3$ sensors in the AGSOF, displaying a more diverse behavior therefore suggesting that in this case, the generalization of the calibration functions could be more challenging.

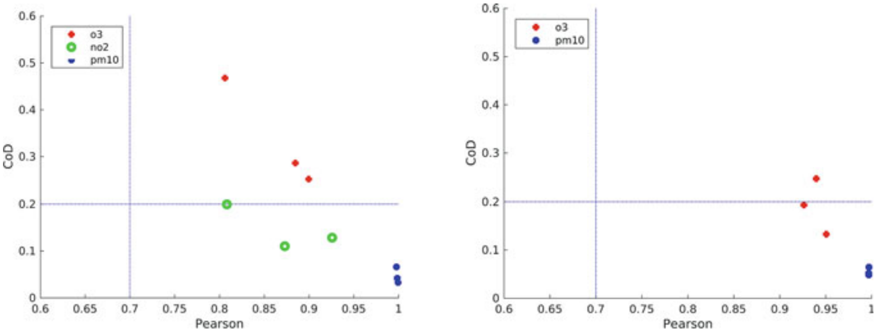

**Fig. 2** Pearson against CoD. Left: AGSOF, right: KORD

The RFFI selected the most relevant features, mostly the ones deriving from the KASTOM nodes' measurements, but also meteorological factors deriving from modeling (Fig. 3). On the other hand, traffic related features are only chosen in the AGSOF location (an urban traffic station). Also, traffic features seem to influence more NO$_2$ and PM10 than O$_3$.

.

While raw measurements display extremely poor scores against reference measurements, the computational procedure and the XGBoost shows promising results (Table 2). In most of the cases the use of the Fset leads to better output than the use of the Nset, though by a small margin.

In terms of REU, the calibrated PM10 can be considered as "class 1 sensor" in both locations, while the calibrated O3 are above the desired threshold but have still improved their performance and be considered as "class 2 sensor" (Fig. 4).

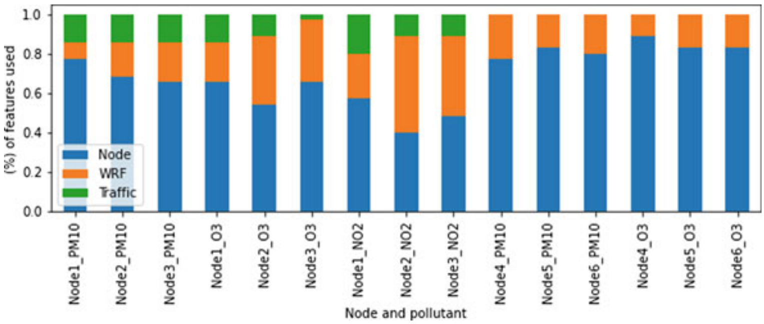

**Fig. 3** Feature selected from the RFFI by category for each node

**Table 2** R$^2$ score for XGBoost and raw measurements. Bold: best performance per sensor

| Set | Node | PM10 | NO$_2$ | O$_3$ | Node | PM10 | O$_3$ |
|---|---|---|---|---|---|---|---|
| Fset | 1 | **0.82** | **0.69** | **0.76** | 4 | **0.8** | **0.72** |
| Nset | 1 | 0.78 | 0.61 | 0.75 | 4 | 0.78 | **0.72** |
| Raw | 1 | −27.96 | −4.28 | −1.95 | 4 | −25.93 | −12.18 |
| Fset | 2 | **0.81** | **0.65** | **0.69** | 5 | **0.8** | **0.74** |
| Nset | 2 | **0.81** | 0.6 | **0.69** | 5 | 0.79 | 0.69 |
| Raw | 2 | −24.43 | −10.46 | −13.95 | 5 | −23.95 | −1.36 |
| Fset | 3 | **0.82** | **0.67** | **0.74** | 6 | **0.77** | **0.74** |
| Nset | 3 | 0.78 | 0.6 | 0.73 | 6 | **0.77** | **0.74** |
| Raw | 3 | −33.35 | −1.35 | −0.3 | 6 | −20.54 | 0.67 |

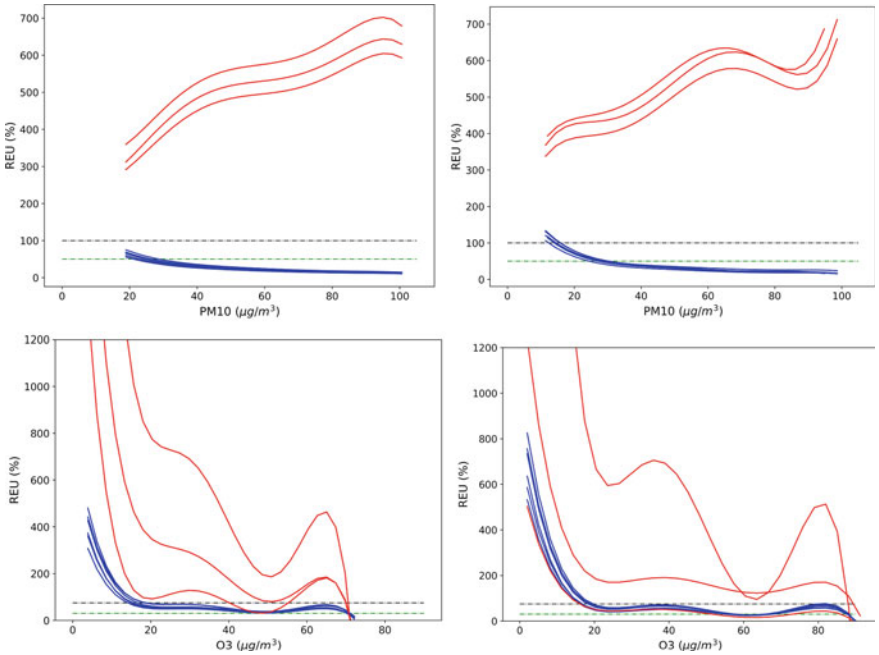

**Fig. 4** REU for PM10 (up) and $O_3$ (down) in AGSOF (left) and KORD (right). Red lines: raw measurements, blue lines: calibrated measurements, black line: "class 2 sensor" threshold, green line: "class 1 sensor" threshold

# 5 Conclusions

The intercomparison of LCAQSN for a small time period, proves that PM10 sensors are behaving similar in the same locations and the proposed computational calibration procedure can upgrade their performance as indicative measurements for regulatory purposes, while it may be possible to apply the same approach to the rest of the network. For $NO_2$ and $O_3$, while the calibration functions can improve the sensors' response, the desired REU levels couldn't be reached. In every case data fusion is improving results and therefore more data sources and additional effort towards better fusion should be considered.

**Acknowledgements** This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH—CREATE—INNOVATE. Project code T1EDK-01697; project name Innovative system for air quality monitoring and forecasting (KASTOM, www.air4me.eu).

## References

EUD. (2008). Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. *Official Journal of the European Union L152*

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.https://doi.org/10.1214/aos/1013203451

Salanova et al., 2018Salanova Grau J. M., Mitsakis E., Tzenos P., Stamos I., & Aifadopoulou, G. (2018). Multisource data framework for road traffic state estimation. *Journal of Advanced Transportation*, 1–9. https://doi.org/10.1155/2018/9078547

## Questions and Answers

QUESTIONER Zhaoyue Chen

QUESTION Thanks, how could you determine lagged hour length when enriching feature space?

ANSWER The lagged length was determined after trial-and-error experiments, while it has been observed from previous computational exercises by our group that no more than 24 hours lagged is important for low-cost sensor nodes calibration.

QUESTIONER Bas Mijling

QUESTION Low-cost sensors are calibrated at two different sites. What would happen if the sensor location snapped? Does the calibration obtained at site 1 is applicable at site 2?

ANSWER This is a very interesting question and can be answered thoroughly only if further research is applied. From our knowledge of the field and ongoing experiments, applying a calibrated function from Agias Sofias to Kordelio and vice versa is yielding good results in terms of uncertainty and $R^2$ for PM2.5 and PM10, and acceptable metrics for $O_3$. Although the question about the spatial generalizability of the calibration function cannot be answered with only two reference stations collocated with the low-cost sensors. Data from a third collocated reference station, not included in this study, show more ambiguous behavior and thus applying functions by proximity or by type of station (urban, suburban, traffic, background, etc.) or applying one generalized calibration function trained in all available locations, would be considered for calibrating the whole network of 33 devices.