

UNIVERSITY OF SALERNO



DEPARTMENT OF INDUSTRIAL ENGINEERING

*Ph.D. Course in Industrial Engineering
Curriculum in Electronic Engineering - XXXV Cycle*

Ph.D. thesis entitled:

ADVANCED PROCEDURES FOR ON-FIELD CALIBRATION OF LOW-COST AIR QUALITY MONITORING SYSTEMS

Supervisor

Prof. Paolo Sommella


Ph.D. student

Gerardo D'Elia



Scientific Referees

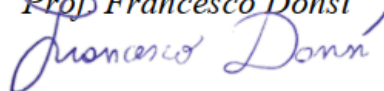
Prof. Consolatina Liguori

Eng. Saverio De Vito, Ph.D. (ENEA)

Eng. Sergio Ferlito (ENEA)

Ph.D. Course Coordinator

Prof. Francesco Donsì



i

**Advanced procedures for on-field
calibration of low-cost air quality
monitoring systems**

Gerardo D'Elia

UNIVERSITY OF SALERNO



DEPARTMENT OF INDUSTRIAL ENGINEERING

*Ph.D. Course in Industrial Engineering
Curriculum in Electronic Engineering - XXXV Cycle*

Ph.D. thesis entitled:

ADVANCED PROCEDURES FOR ON-FIELD CALIBRATION OF LOW-COST AIR QUALITY MONITORING SYSTEMS

Supervisor

Prof. Paolo Sommella

Ph.D. student

Gerardo D'Elia

Scientific Referees

Prof. Consolatina Liguori

Eng. Saverio De Vito, Ph.D. (ENEA)

Eng. Sergio Ferlito (ENEA)

Ph.D. Course Coordinator

Prof. Francesco Donsì

Academic year 2021/2022

Vol. 1 ISBN 88-7897-150-2

List of Publications

1. **G. D’Elia**, S. De Vito, M. Ferro, V. Paciello, P. Sommella, “*Simulation of WSN for Air Particulate Matter Measurements*” AISEM 2020, Springer International Publishing, DOI: 10.1007/978-3-030-69551-4_23
2. E. Esposito, **G. D’Elia**, S. Ferlito, A. Del Giudice, G. Fattoruso, P. D’Auria, S. De Vito, G. Di Francia “*Optimal Field Calibration of Multiple IoT Low Cost Air Quality Monitors: Setup and Results,*” Computational Science and Its Applications – ICCSA 2020. ICCSA 2020. Lecture Notes in Computer Science, vol 12253. Springer, DOI: 10.1007/978-3-030-58814-4_57
3. S. De Vito, E. Esposito, **G. D’Elia**, A. Del Giudice, G. Fattoruso, S. Ferlito, P. D’Auria, F. Intini, G. Di Francia, E. Terzini, “*High Resolution Air Quality Monitoring with IoT Intelligent Multisensor devices during COVID-19 Pandemic Phase 2 in Italy*” 2020 AEIT International Annual Conference (AEIT), 2020, pp. 1-6, DOI: 10.23919/AEIT50178.2020.9241144.
4. S. De Vito, E. Esposito, E. Massera, F. Formisano, G. Fattoruso, S. Ferlito, A. Del Giudice, **G. D’Elia**, M. Salvato, T. Polichetti, P. D’Auria, A. M. Ionescu, G. Di Francia, “*Crowdsensing IoT Architecture for Pervasive Air Quality and Exposome Monitoring: Design, Development, Calibration, and Long-Term Validation*” Sensors, vol. 21, no. 15, Jul. 2021, DOI: 10.3390/s21155219.
5. S. De Vito, **G. D’Elia**, G. Di Francia, “*Global calibration models match ad-hoc calibrations field performances in low-cost particulate matter sensors*” 2022 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN), 2022, pp. 1-4, DOI: 10.1109/ISOEN54820.2022.9789669.
6. S. De Vito, G. Fattoruso, **G. D’Elia**, E. Esposito, S. Ferlito, A. Del Giudice, E. Massera, G. Loffredo, G. Di Francia, “*Hyper resolved Air Quality maps in urban environment with crowdsensed data from intelligent low-cost sensors*” 2022 IEEE International Symposium on Olfaction and Electronic Nose (ISOEN), 2022, pp. 1-4, DOI: 10.1109/ISOEN54820.2022.9789614.

7. **G. D'Elia**, M. Ferro, P. Sommella, S. De Vito, S. Ferlito, P. D'Auria, G. Di Francia, "*Influence of Concept Drift on Metrological Performance of Low-Cost NO₂ Sensors*" in *IEEE Transactions on Instrumentation and Measurement*, 2022, DOI: 10.1109/TIM.2022.3188028.
8. S. De Vito, A. Del Giudice, **G. D'Elia**, E. Esposito, G. Fattoruso, S. Ferlito, F. Formisano, G. Loffredo, E. Massera, G. Di Francia, P. Bellucci, F. Ciarallo, "*Continuous Measurement of Air Pollutant Concentrations in a Roadway Tunnel in Southern Italy*", EAI SmartGov 2022 - 4th EAI International Conference on Smart Governance for Sustainable Smart Cities, November 16-18, 2022, Braga, Portugal

Table of Contents

List of Publications

Table of Contents	i
List of Figures	iii
List of Tables	viii
Abstract	ix
Introduction	xi
Why air quality monitoring is important.....	xi
Air Quality Monitoring networks.....	xiii
The arising of the field low-cost sensors calibration problem	xv
Unsolved questions: Aims of thesis	xvi
Chapter 1	1
Low-Cost Air Quality Monitoring Systems (LCAQMS)	1
1.3 The MONICA node.....	7
1.3.1 MONICA hardware architecture	7
1.3.2 IoT infrastructure.....	8
1.4 Factory calibration and Laboratory characterization.....	10
Chapter 2	13
Field Calibration of LCAQMS	13
2.1 Reasons for field calibration.....	13
2.2 Background to supervised machine learning algorithms.....	14
2.2.1 Multivariate Linear Regression	14
2.2.2 Neural Network	14
2.3 Evaluation Metrics	15
2.4 Experimental framework.....	16
2.4.1 Field calibration setup	16
2.5 Sensor calibration models	18
2.6 Experimental results	18
2.7 Citizen science campaign in Portici during COVID 19 Phase 2	22

2.7.1 Campaign details.....	22
2.7.2 Air Quality Geo-Mapping algorithm	24
2.7.3 Citizen science campaign results	25
2.8 Global Calibration Methodology	28
Chapter 3.....	31
Influence of Concept Drift on Metrological Performance of Low-Cost NO₂ Sensors	31
3.1 A neglected phenomenon in field calibration: The concept drift.....	31
3.2 Calibration metrics and data quality requirements.....	32
3.3 Methodology	33
3.3.1 Concept Drift.....	33
3.3.2 Statistical tools for detecting the Concept Drift Events	34
3.3.3 Estimation of Relative Expanded Uncertainty.....	35
3.3.4 The proposed approach	35
3.4 Results and discussion.....	36
3.4.1 Experimental Data and MLR Calibration	37
3.4.2 Validation of the proposed methodology.....	39
3.4.3 Auto-Detection of the Concept Shift.....	47
3.5 Concluding remarks	48
Chapter 4.....	49
Strategies for Concept Drift Mitigation in Low-Cost Air Quality Networks	49
4.1 Calibration update triggered via concept drift detector.....	49
4.1.2 Remote calibration	49
4.1.3 Reference data selection for calibration update in presence of Concept Drift.....	51
4.2 Calibration update without reference data.....	55
4.2.1 General calibration model.....	55
4.2.2 Importance weighting calibration model.....	56
4.2.3 Extending the calibration validity	57
4.3 Handling concept drift with Stacking Ensemble model.....	58
Conclusion.....	61
Reference.....	63
Appendix A	71
A.1 REU plots of reference data selection for calibration update.....	71
A.2 REU plots of general calibration model	79
A.3 REU plots of importance weighting calibration model.....	82
A.4 REU plots of stacking ensemble calibration model	87

List of Figures

Figure I. 1 Top 10 noncommunicable diseases causing deaths attributable to the environment. (Picture from: EEA – Healthy environment, healthy lives, 2018 based on WHO (2016)).	xii
Figure I. 2 Typical fixed-site air quality monitoring station a); Mobile air quality monitoring station b).	xiii
Figure I. 3 Air quality monitoring stations subdivision by type and zone (Picture from Technical Report of EEA).	xiv
Figure I. 4 Graphical description of the sensor calibration process.	xvi
Figure 1. 1 New approved pollutant level in WHO AQG 2021 vs WHO AQG 2005 (Picture from: www.iqair.com).	1
Figure 1. 2 Inside an electrochemical gas sensor.	4
Figure 1. 3 Exploded view of the MONICA node with Alphasense gas sensors, AFE and other paramount parts as well.	6
Figure 1. 4 Main functional blocks of a single MONICA 2.0 node.	7
Figure 1. 5 Main functional blocks of a single MONICA 3.0 node.	8
Figure 1. 6 Scheme of IoT architecture in stationary setup.	9
Figure 1. 7 Picture of a user session accessed through web service backend. PM plots are visible and also the user path on google maps.	9
Figure 1. 8 MONICA IoT architecture in mobile setup.	10
Figure 1. 9 LVTC with eight MONICA node under laboratory calibration procedure.	12
Figure 1. 10 a) Typical pulse with the three different phases of a calibration procedure (i.e., CO). b) Sensitivity curve and the linear regression for the sensor output in the range 0-5 ppm during adsorption (black dots) and desorption (red dots).	12
Figure 2.1 Functional layers of a neural network.	14
Figure 2.2 Diagram of node activation in a neural network.	15
Figure 2.3 Field calibration scheme with a Google Earth view of the co-location and a photo of 4 MONICA 2.0 mounted on the roof of the mobile laboratory.	16

Figure 2.4 Box-plot representation of NO ₂ gas concentration distribution along the entire co-location period.....	17
Figure 2.5 NO ₂ hourly concentration estimations versus target gas concentration starting from the 4th week of the co-location period.....	21
Figure 2.6 NO ₂ gas concentration estimation computed for each node, along the entire co-location period versus target gas concentration line. The differences among the four sensors performances become apparent when considering low true concentration of the target pollutant.....	21
Figure 2.7 NO ₂ gas concentration as measured by ARPAC mobile laboratory within the city of Portici and specifically in winter 2020 co-location at the Waterfront building site. Linear trend is highlighted in black.	22
Figure 2.8 Monthly boxplot of the daily average concentrations of PM _{2.5} measured by ARPAC mobile Lab at Waterfront site in Portici.	23
Figure 2.9 The four pre-set monitoring paths (red, blue, green, orange) along with the mobile laboratory location (blue dot).	23
Figure 2.10 Measurements density plot shows slightly uneven density patterns and unforeseen measurements paths.	26
Figure 2.11 IDW averaged CO concentration pattern is characterized by localized hotspots near main crossroads or streets characterized by heavy traffic load (arrows). Unforeseen hotspots have also arised prompting for ad-hoc measurements campaigns (ellipse).	26
Figure 2.12 IDW averaged NO ₂ concentration pattern.	27
Figure 2.13 IDW averaged O ₃ concentration pattern shows generally lower spatial variance with average values that reach or overcome the regulatory threshold.	27
Figure 2.14 Procedure for creating the training set of the general calibration model.....	29
Figure 3. 1 Typical life cycle of LCAQMS: on-field calibration, instrument operation, and maintenance.....	32
Figure 3. 2 NO ₂ concentration (reference station) during the co-location period and time slot partition. Two abrupt changes in time series are marked with x.....	37
Figure 3. 3 MAE and MAPE of examined devices in the respective slot time. While a smooth trend is visible from T1 up to T4, in range T5 - T8 the increasing values of both the metrics point out a worst model performance.	38
Figure 3. 4 Boxplot of NO ₂ concentrations measured by reference station and temperature during the co-location period in each time slot.	39
Figure 3. 5 Probability Density Function (Lognormal fitting) of the reference and estimated NO ₂ concentrations during time slots T2-T4 compared with the training data set (time slot T1). In T3 there is no AQ12 data due to loss data transmission.	40

Figure 3. 6 Probability Density Function (Lognormal fitting) of the reference and estimated NO ₂ concentrations during time slots T5-T6 compared with the training data set (time slot T1). The TSKS-test results highlight the auto-detection skill of concept drift.	41
Figure 3. 7 Probability Density Function (Lognormal fitting) of the reference and estimated NO ₂ concentrations during time slots T7-T8 compared with the training data set (time slot T1). The TSKS-test results highlight the auto-detection skill of concept drift.	42
Figure 3. 8 Relative Expanded Uncertainties of all LCAQMSs computed in T2 and fitted with 8th degree polynomial.	43
Figure 3. 9 Relative Expanded Uncertainties in time slots [T1 - T4].	44
Figure 3. 10 Relative Expanded Uncertainties in time slots [T5 - T8].	45
Figure 3. 11 TSKS-Test Statistic results of NO ₂ MDL predictions and temperature of all devices when the MDL model is trained in T1 and T5. Concept Drift happen in T4 and go on in T5 that is used to recalibrate the model. This zone is called drift zone.	46
Figure 3. 12 Scheme of the proposed Diagnostic Add-On block.	47
Figure 4. 1 Hierarchical network for continuous calibration. (Picture from: Miskell et al., 2018).	50
Figure 4. 2 Evidence of concept drift highlighted on the co-location samples of the target and input variables.	51
Figure 4. 3 Data selection for calibration update, when a re-calibration request arrives from the concept drift detector.	52
Figure 4. 4 How the weights are used in the fitting process.	56
Figure 4. 5 The proposed stacking ensemble architecture.	59
Figure A.1. 1 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with data of T4 (Mixed scenario).	71
Figure A.1. 2 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with data of T4 (Mixed scenario).	72
Figure A.1. 3 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with data of T4 (Mixed scenario).	72
Figure A.1. 4 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with data of T4 (Mixed scenario).	73
Figure A.1. 5 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with data of T4 (Mixed scenario).	73
Figure A.1. 6 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with data of T4 (Mixed scenario).	74
Figure A.1. 7 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with data of T3 (Last scenario).	74
Figure A.1. 8 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with data of T3 (Last scenario).	75
Figure A.1. 9 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with data of T3 (Last scenario).	75

Figure A.1. 10 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with data of T4 (Mixed scenario).....	76
Figure A.1. 11 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with data of T4 (Mixed scenario).....	76
Figure A.1. 12 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with data of T4 (Mixed scenario).....	77
Figure A.1. 13 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with data of T3 (Last scenario).	77
Figure A.1. 14 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with data of T3 (Last scenario).	78
Figure A.1. 15 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with data of T3 (Last scenario).	78
Figure A.2. 1 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with global calibration model.....	79
Figure A.2. 2 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with global calibration model.....	79
Figure A.2. 3 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with global calibration model.....	80
Figure A.2. 4 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with global calibration model.....	80
Figure A.2. 5 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with global calibration model.....	81
Figure A.2. 6 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with global calibration model.....	81
Figure A.3. 1 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with the importance weighted calibration model.	82
Figure A.3. 2 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with the importance weighted calibration model.	82
Figure A.3. 3 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with the importance weighted calibration model.	83
Figure A.3. 4 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with the importance weighted calibration model.	83
Figure A.3. 5 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with the importance weighted calibration model.	84
Figure A.3. 6 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with the importance weighted calibration model.	84
Figure A.3. 7 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with the importance weighted calibration model.	85
Figure A.3. 8 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with the importance weighted calibration model.	85
Figure A.3. 9 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with the importance weighted calibration model.	86
Figure A.4. 1 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with the stacking ensemble calibration model.	87

Figure A.4. 2 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with the stacking ensemble calibration model. 87

Figure A.4. 3 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with the stacking ensemble calibration model 88

Figure A.4. 4 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with the stacking ensemble calibration model. 88

Figure A.4. 5 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with the stacking ensemble calibration model. 89

Figure A.4. 6 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with the stacking ensemble calibration model. 89

Figure A.4. 7 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with the stacking ensemble calibration model. 90

Figure A.4. 8 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with the stacking ensemble calibration model. 90

Figure A.4. 9 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with the stacking ensemble calibration model. 91

List of Tables

Table I.1 Area of representativeness of a monitoring station.....	xiv
Table 2.1 Recorded data and data losses.....	17
Table 2.2 Models performance with different choices for the training length (<i>L</i> , in weeks) for each node. Bold indicates best performance.....	19
Table 2.3 MAE performance of models with different choices for the training length (<i>L</i> , in weeks) for each node in minutely analysis.....	20
Table 2.4 First order characterization of recorded data.....	25
Table 3.1 Evaluation metrics.....	33
Table 4.1 MAE and MAPE results obtained after the data selection for model calibration update.....	53
Table 4.2 Summary of REU results for the “Mixed” scenario.....	54
Table 4.3 Summary of REU results for the “Last” scenario.....	55
Table 4.4 Metrics performance of the general calibration model.....	54
Table 4.5 Metrics performance of the importance weighting calibration model.....	57
Table 4.6 Summary of REU results for the general calibration model.....	58
Table 4.7 Summary of REU results for the importance weighting calibration model.....	58
Table 4.8 Metrics performance of the Stacking Ensemble model.....	59
Table 4.9 Summary of REU results for the stacking ensemble calibration model.....	59

Abstract

Air pollution is now well known to be one of the major causes of human and climate health issues. The global crisis related to COVID-19 pandemic has brought to the fore such theme. The importance of air quality has been rediscovered and counted among the main positive effects of lockdown. The spread of low-cost electrochemical sensors, joined with diffusion of the Internet of Things (IoT) technologies, will allow in the near future, the birth of a generation of air quality monitoring networks, characterized by the integration of regulatory grade analyzers and such IoT smart electrochemical and particulate multisensory devices. The former will provide a backbone of sparse but high reliable, high quality, measurements at a significant procurement and operational costs, while smart multisensory devices will provide high resolution and possibly redundant measurements with affordable costs and with reduced precision and accuracy. Consequently, high-resolution pollution maps will be provided, constituting an advanced informative support tool for institutional decision makers.

However, this paradigm shift in air quality monitoring, is currently hampered by a series of problems concerning the low-cost sensors, high fabrication variance and the dynamic nonstationary nature of the working environment where these devices have to operate; but the primer concern is related to the measurements data quality.

Field calibration, relying on statistical or machine learning models more generally, seems the only viable and feasible method to guarantee the short-term accuracy and precision of these systems. Although its robustness to long term deployment and so different environmental and pollution composition is criticized. Field calibration allows to expose, rapidly and cheaply, the sensors grabbing their response to a variety of (uncontrollable) conditions that are similar to the ones that will be encountered during operational life, in opposition to laboratory-based calibration that would need significant time and human efforts to achieve similar variability in controlled settings.

Addressing the long-awaited achievement of data quality objective (DQO), in our opinion, could be a turning point for the rapid large-scale diffusion of this technology, especially in smart city applications.

With this objective in mind, the present PhD research has been focused in the first part, in the assessment of the machine learning techniques for the calibration of low-cost air quality monitoring systems (LCAQMSs), comparing multivariate linear regression and neural networks. The purpose of this analysis was aimed at understanding whether a simpler technique is equally able to carry out acceptable performances in terms of data quality with respect to advanced but much more complex techniques. A mid-term experimental co-location campaign as well as a citizen science company have been performed for such kind of investigation, evidencing the effectiveness of the multivariate approach, both in fixed and mobile applications.

The extensive literature analysis executed has shown that most of the efforts of the scientific community operating in this research area was given to the inspection and assessment of the calibration models able to provide the best performances, while less emphasis is found looking for the answer to a simple question: *When does the sensor node need to be recalibrated?*

After the calibration phase, the LCAQMS will be subjected to performance degradation and forced to operate in conditions never seen before during the training phase. The outcomes will be bad quality measurements both in accuracy and precision. One of the phenomena that most influences this trend is the so-called Concept Drift. The awareness that the used model is no longer able to provide reliable data implies the risk to invalidate the model and to request a model update. Consequently, an original methodology based on the two-sample Kolmogorov–Smirnov test (TSKS test) is proposed to automatically detect the presence of the concept drift and a scheme of an add-on block based on the proposed approach is designed for the continuous monitoring of the metrological performance exhibited by the calibration model. As disposed by European directive, the relative expanded uncertainty (REU) is the paramount metric we will refer to. This functional block, in addition to monitoring the calibration model performance, is able to provide an alert to the user when a proper threshold is exceeded. Consequently, retraining or updating the calibration model ensuring compliance of the DQOs, is possible.

In the last part, different strategies have been analyzed to update the calibration model, trying to mitigate the effects of the concept drift in an air quality network operational scenario. Specifically, two alternative calibration models are taken into consideration: the general calibration model and the importance weighting calibration model. In some cases, both models have shown improvement of performances or matching those of the ad-hoc model, bringing the REU back to values in compliance with DQOs without requiring reference data. These models have also been used as the first layer in a stacking ensemble approach with the outcome of a further improving performance by requiring only the reference labels in the training process. The proposed approach guarantees the continuity of the data quality and extends the validity of the calibrations.

Introduction

Why air quality monitoring is important

The latest publication of the sixth assessment report on the climate changes drawn up by the Intergovernmental Panel on Climate Change (IPCC) of United Nations in addition to the alarm on problems linked to the global warming underlines also the growing worried about the consequent reduction of air quality and all the consequent risks for humans' health (Masson-Delmotte *et al.*, 2021). Although in recent decades the emissions of air pollutants have ample decreased, already in 2016, the World Health Organization (WHO) reported an estimation of about 4.2 million premature deaths worldwide ascribable to the outdoor air pollution both in cities and rural areas (WHO, 2021a). Emission of pollutants could be due to natural circumstances but nowadays are predominantly anthropogenic, in fact the combustion processes of fossil fuels and biomass to generate electrical or mechanical energy, have as outcomes a huge release in the atmosphere of various compounds like particulate matter (PM) and nitrogen dioxide (NO₂) among the most dangerous. In WHO Air Quality Guidelines (WHO AQG) 2021, it is reiterated that air pollution is a major global public health emergency, as evidenced by the statistics reporting how outdoor and household air pollution accounted for approximately 12% of all deaths in 2019 (WHO, 2021b, Hoffmann *et al.*, 2021). The recent European Environment Agency (EEA) online report "*Beating cancer - the role of Europe's environment*", claims how pollutants in the environment and in the workplace impact heavily on our health and in some cases cause the onset of cancer. Approximately 3 million of new diagnoses and 1.3 million deaths every year across Europe have been recorded. Unfortunately, not only cancer is attributable to polluted air, but also several other diseases like ischaemic heart disease, obstructive pulmonary disease, strokes, mental and neurological conditions, diabetes and more, as shown in the figure I.1 below. Recent studies have found associations between long-term exposure to particulate matter and leukaemia in adults and children. Furthermore, it must be added, that the incidence of the environmental factor on diseases is not equally distributed in Europe and in the rest of the world among the population

Introduction

groups (high densely populated country with a high rate of pollution and between different age groups).

Therefore, the only way to minimize such impact on people's health is a further drastic reduction of air pollutants at all levels, as stressed in the WHO AQG updated in September 2021. Indeed, with respect to 2005 WHO AQG, it is recommended reducing the annual mean concentration of PM_{2.5} from 10 µg/m³ to 5 µg/m³ and similarly for NO₂, whose limit pass from 40 µg/m³ to 10 µg/m³. Although the WHO AQG guidelines are not legally binding, the previous EU Action Plan: "Towards Zero Pollution for Air, Water and Soil", published just a few months before, goes right in this direction, as a cross-cutting objective contributing to the UN 2030 Agenda for Sustainable Development and part of the European Green Deal initiatives. The European Union (EU) has already adopted strict measures on air pollution with European Directive (EC, 2022) and now is involved in a review of these directives with the aim of better aligning with the most recent WHO AQG. At moment the commission adoption is planned for the third quarter 2022.

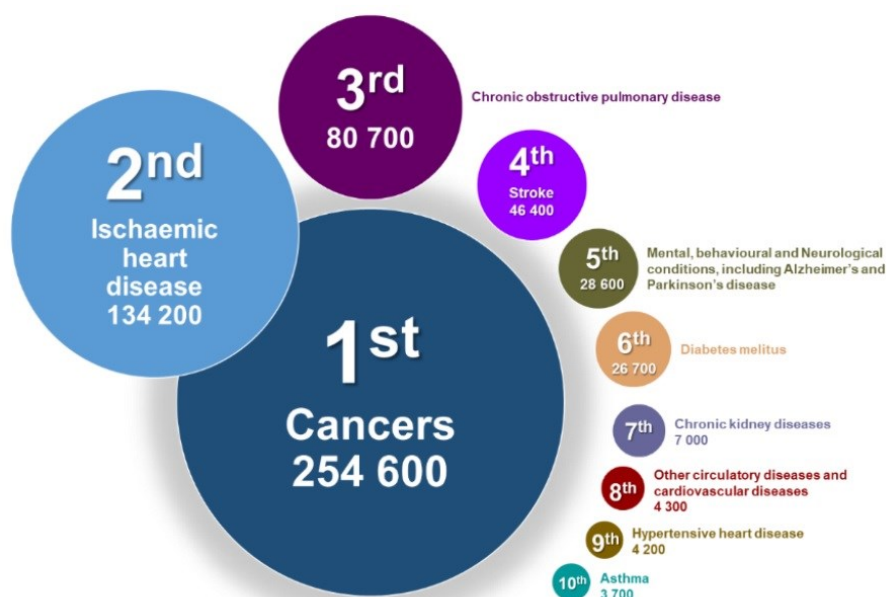


Figure I. 1 Top 10 noncommunicable diseases causing deaths attributable to the environment. (Picture from: EEA – Healthy environment, healthy lives, 2018 based on WHO (2016)).

Air Quality Monitoring networks

The only way to keep air quality under control is through continuous monitoring. In Italy, air quality monitoring activity is fulfilled by institutional agencies, whom are in charge for implementing the WHO AQG and/or EU directive, using fixed or mobile laboratory which require high instrumentation and management costs, therefore, covering inevitably a limited land area.

Fixed-site air quality monitoring stations are equipped with expensive measurement instrumentation to ensure high quality data strict routines of maintenance and calibration procedure are required moreover (Castell *et al.*, 2017). A distributed network of fixed stations is developed to meet the legal requirements for real-time air quality monitoring, but unfortunately relatively sparse, due to high management cost. Consequently, accurate data are available for only few locations, making it impossible to ensure widespread health citizen protection. For this reason, the fixed monitoring network is expanded with a mobile monitoring network, consisting of mobile vehicles within housed the same instruments as the fixed stations (Fig. I.2). The mobile solution extends slightly the geographical coverage but for a limited slot of time, having equally the drawback of high costs.



Figure I. 2 Typical fixed-site air quality monitoring station a); Mobile air quality monitoring station b).

All the information to support and to facilitate the assessments of air quality are assigned to the European Air Quality Monitoring Network (EUROAIRNET). The main goal of EUROAIRNET is to establish an air quality monitoring network with sufficient spatial coverage, representativeness, and data quality of every EU country. In addition, the different types of monitoring stations are also classified in its technical documents (EEA, 1999). Such monitoring stations are classified according to the following classification scheme in figure I.3.

Introduction

Type of station	Type of zone	Characterisation of zone
Traffic (T)	Urban (U)	Residential (R)
Industrial (I)	Suburban (S)	Commercial (C)
Background (B)	Rural (R)	Industrial (I)
		Agricultural (A)
		Natural (N)
		Res/Com (RC)
		Com/Ind (CI)
		Ind/Res (IR)
		Res/Com/Ind (RCI)
		Agri/Natural (AN)

Figure I. 3 Air quality monitoring stations subdivision by type and zone (Picture from Technical Report of EEA).

A “Traffic Station” is usually positioned near roads characterized by medium-high traffic level, meaning high level pollution. In industrial areas, where pollution due to industrial waste must be carefully monitored, we can find obviously “Industrial Station”. The “Background Stations”, on the other hand, are positioned so as not to be influenced by any source of pollutants, acting as a comparison with the previous ones. The differentiation in zones depends on the context in which they are installed.

The concentration measured is supposed the same along a geographical area called “area of representativeness” of the station (see range values in table I.1).

Table I.1 Area of representativeness of a monitoring station.

Station Class	Radius of area
Traffic station	Not applicable
Industrial stations	10 – 100 m
Background stations:	
- Urban background station	100 m – 1 km
- Near-city background station	1 – 5 km
- Regional stations	25 – 150 km
- Remote stations	200 – 500 km

Additional criteria for air monitoring stations installation are also linked to the number of inhabitants of a peculiar geographical area, as a matter of fact the largest number of stations are placed in large densely populated cities. Despite everything, due to too high costs, increasing the number of installations is economically impractical.

A solution to this problem should come from the introduction of the low-cost sensors. The spread of the IoT applications occurred in recent years, together with the pioneering research on electrochemical sensors and metal oxide technologies as well as on portable particulate matter devices allows the costs reduction of sure (an equivalent reference method developed with low-cost sensors technologies cost less than a tenth with respect to an institutional measurement instrumentation station) and at the same time to take advantage to collect data on larger spatial zone, giving to the user the ability to generate high spatio-temporal resolution map of pollution. These sensors mounted on a circuit board are connected via bus to a control unit, and together with the communication devices responsible for data management and transmission, represent a multi-sensor unit (hereinafter "node"). Nodes in a position to send data over internet could be called IoT measurement system as well. Multiple nodes or groups of nodes distributed over a geographical area that are organized for wireless communication create a Wireless Sensor Network (WSN). Users with such devices are free to share information about one or more monitored pollutants including personal exposure. The currently applied European Directive (EU, 2008) for fixed monitoring stations defines further air quality indicative measurements. The low-cost sensors, that are easily be mounted on vehicles, bikes or even worn, can play a significant part in air quality indicative measurements in fact, in recent years the number of science projects involving citizens and communities to monitor the air their breathe, has increased significantly.

The arising of the field low-cost sensors calibration problem

The prime drawback of low-cost sensor technology is the data quality, if compared to the accuracy and precision of the measurement instrumentation installed in fixed monitoring stations. The main challenge is reaching the data quality objectives (DQOs) established in European Directive and keep them over the operation time. There are several reasons for this behaviour: the transduction principle, aging, chemical interferences (cross-sensitivity) and environmental conditions (humidity and temperature). To get through this limitation, such sensors require frequent calibrations in order to provide more reliable measurement data. Usually, two types of calibrations procedure are performed: in laboratory and field calibration. In laboratory calibration, the node is placed in a chamber under reproducible and accurately controlled temperature and humidity conditions and a dilution system generates all the concentrations of the gases that the node will have to detect. Nevertheless, the laboratory calibration conditions can never be the same as those it will face in its real outside operational lifetime, so sadly the laboratory calibration procedure is not enough to guarantee "good in field data" (Castell *et al.*, 2017). This is the reason why field calibration has been proposed.

Introduction

The goal of any calibration process is to find the mathematical law f that relates sensor outputs and in case other physical parameters that influence the measurement of the sensor itself (environmental quantities), to the pollutant concentration value that we want to measure. A graphical overview of this concept is shown below in figure I.4.

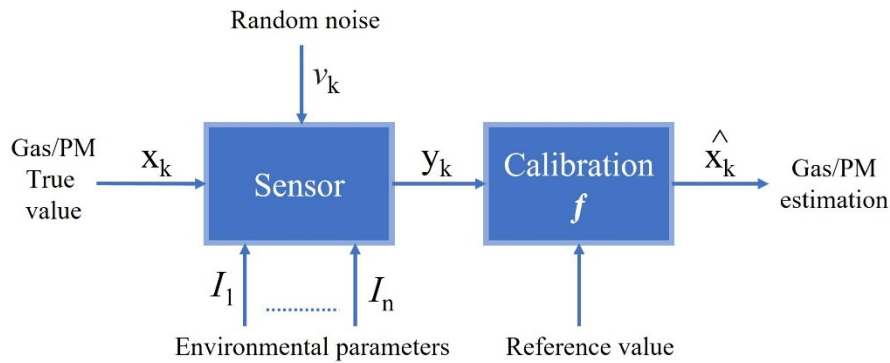


Figure I. 4 Graphical description of the sensor calibration process.

Field calibration involves a compulsory step, the so-called co-location, during which the node is positioned next to a reference station. In this way, is built the “training set” demanded for the determination of the calibration function f using statistical or machine learning techniques. The optimal calibration process is obtained if the estimation value equals the true value. The calibration models are designed to minimize the errors that can afflict the measurements of a sensor that operates in the field to get as close as possible to the quality of the measurements by more expensive and accurate instruments. Among the most common methods of low-cost sensors calibration there are Multiple Linear Regression (MLR) and various machine learning approach such as Neural Networks (NN).

Unsolved questions: Aims of thesis

The scientific community taking part in air quality monitoring with low-cost sensors, in the last decade has profuse considerable efforts into the search for an "optimal calibration function" employing artificial intelligence algorithms, even very complex ones as deep learning techniques. In the first part of this study, a comparison between the MLR and NN technique has been performed using the data obtained from a mid-term co-location campaign carried out in Portici (Naples – Italy) lasting about two months in winter 2020 in the context of the European Union project Air Heritage (www.uia-initiative.eu/en/uia-cities/portici). The node used is called MONICA from the Italian acronym of "MONItoraggio Cooperativo della qualità dell'Aria"

(whose translation is "Cooperative Air Quality Monitoring"), developed by ENEA (Italian National Agency for New Technologies, Energy and Sustainable Economic Development), which will be described in detail in the following sections. The objective of this analysis stems from the consideration that the main objectives in the application of low-cost sensors in air quality monitoring are the validation of new atmospheric models, the identification of pollutant hotspots, the generation of high-resolution pollutant maps and finally the assessment of personal exposure. Well, in order to obtain qualitatively reliable data from the MONICA device, the calibration function must be updated over time, or we must be sure that the device operates in a condition able to release reliable data. Using simpler techniques such as an MLR can be an advantage in such situations, truly updating the coefficients of the calibration function on the smartphone app is certainly easier than retraining a neural network. The obtained results prove that a MLR calibration model is adequate for these purposes. Another unsolved question not been addressed at present, is when the adopted calibration is no longer serviceable. The core of this thesis was prior to define a methodology that is efficient into detection the data quality degradation and enable a recalibration request. A functional add-on that gives added value to any device and that can practically monitor the performance of the calibration mode. A peculiarity missing in any device on the market to date. Moreover, this methodology has the advantage that it can be implemented both on the node and on the backend as a service. The last part of this manuscript is focused on an emergent topic: the global calibration models. Exploring if a generalized model has the potential to constitute a temporary alternative to recalibration.

Chapter 1

Low-Cost Air Quality Monitoring Systems (LCAQMS)

1.1 The new WHO air quality guidelines

In September 2021, WHO published the update of the global air quality guidelines providing the new recommended values for the six main atmospheric pollutants: particulate matter 2.5 and 10, ozone, nitrogen dioxide, sulphur dioxide and carbon monoxide (WHO, 2021b). The newly defined levels have been derived from the extensive scientific evidence available in the literature as well as widespread and extensive epidemiological studies. The basic idea is that reducing the levels of pollutants reduces both the number of pollution-related deaths and the number of people who could contract pollution-related diseases. The approved pollutants levels have been reduced compared to the old 2005 guidelines, as shown in the comparison table of the following figure 1.1.

Pollutant	Averaging Time	2005 AQGs	2021 AQGs
PM2.5 $\mu\text{g}/\text{m}^3$	Annual	10	5
	24-hour	25	15
PM10 $\mu\text{g}/\text{m}^3$	Annual	20	15
	24-hour	50	45
Ozone (O ₃) $\mu\text{g}/\text{m}^3$	Peak Season*+ 8-hour**	- 100	60 100
	Nitrogen dioxide (NO ₂) $\mu\text{g}/\text{m}^3$	Annual 24-hour*	40 -
Sulfur dioxide (SO ₂) $\mu\text{g}/\text{m}^3$	24-hour	20	40
Carbon monoxide (CO) mg/m^3	24-hour*	-	4

Figure 1. 1 New approved pollutant level in WHO AQG 2021 vs WHO AQG 2005 (Picture from: www.iqair.com).

Chapter 1

Although not binding, these guidelines represent a tool in the hands of institutional decision-makers capable of leading them to environmental policies aimed at reducing pollutants in order to achieve the objectives of the 2030 agenda for Sustainable Development defined by the United Nations. At the time of writing, the EU has launched actions and consultations for the revision of its outdated directives to align with these new guidelines, in fact on 26 October 2022, as part of the European Green Deal, the Commission proposed to revise the Ambient Air Quality Directives (EC, 2022).

For the sake of completeness, it should be noted that although the current pollutions levels are higher than those envisaged by the new guidelines, some countries have failed to keep the average concentrations of pollution below these legal limits in 2020 and have been subsequently fined. The EU Court of Justice in May 2022 has ascertained the systematic exceeding of the nitrogen dioxide limit value in all the cities areas under examination in Italy starting from the 2010, so Italy has been sanctioned (CJEU, 2022).

1.1.1 Particulate Matter

Particulate Matter (PM) are polluting particles present in the air that we breathe and can be organic or inorganic in nature. PMs are classified according to their size, which could determine a different level of harmfulness. In fact, more these particles are small plus they have the ability to penetrate the respiratory system. In particular, PM10 (diameter less than 10 μm) can be inhaled and penetrate into the upper respiratory tract, from the nose to the larynx. PM2.5 (diameter less than 2.5 μm) can be breathed in and pushed into the deepest part of the respiratory apparatus and even can enter in the blood.

There are two main sources of fine dust:

- ***Natural sources:*** forest fires, volcanic activity, dust, earth and sea salt raised by the wind (the so-called marine aerosol), pollen and spores, rock erosion;
- ***Anthropogenic sources:*** vehicular traffic, use of solid fuels for heating household (coal, wood and diesel fuel), residues from the wear of the road surface, brakes and car tires, industrial activity.

As mentioned above, there is a very close and quantitative relationship between high concentrations of atmospheric particulate matter and an increase in mortality, both in the short and long term. Conversely, when particulate concentrations are reduced (especially considering PM2.5) the relative mortality decreases and this is the reason for the new recommended values stated in WHO AQG 2021.

1.1.2 Nitrogen Oxides

Nitrogen oxides NO_x, among the most disturbing natural and anthropogenic pollutants, are essentially nitrogen oxide (NO) and nitrogen dioxide (NO₂). The term NO_x refers the sum of nitrogen monoxide and nitrogen dioxide (NO_x = NO + NO₂). The main source of NO_x emissions is the vehicular traffic; other sources are civil and industrial heating systems and energy production plants as well as a wide range of industrial processes. About 10% of NO once released into the atmosphere is transformed into NO₂ by the action of solar radiation (Sher, 1998).

NO is a primary pollutant generally formed by high temperature combustion processes. It is a gas with a limited toxicity, unlike NO₂.

NO₂ has a strong, pungent, irritating, odour. Nitrogen dioxide is a reddish-brown gas. It is responsible of the so-called photochemical smog, as the basis for the production of a variety of dangerous secondary pollutants such as ozone or nitric acid.

1.1.3 Ozone

Ozone (O₃) is very toxic for humans, irritating for all mucous membranes and continue exposure can cause cough, headache and even pulmonary edema. O₃ carries out a marked phytotoxic action against plant organisms, with immediately visible effects like leaf necrosis and less visible effects as enzymatic alterations and reduction of photosynthesis activity.

Ozone is a gas with a high oxidizing power, blue colour and a pungent odour. O₃ is formed in the atmosphere as a result of reactions favoured by solar radiation, in the presence of so-called precursor pollutants, especially nitrogen oxides (NO_x) and Volatile Organic Compounds (VOCs) which led to the formation of molecules consisting of three oxygen atoms (O₃). Its presence at ground level strongly depends on the meteorological conditions and therefore it is fluctuating both during the day and the seasons, in fact O₃ concentration increases when temperatures raise, (in spring and summer season). High temperature levels support the molecular oxygen dissociation and consequently the formation of ozone.

1.1.4 Carbon Monoxide

Carbon monoxide (CO) is an odourless and colourless gas that is formed from incomplete combustion of hydrocarbons present in fuels. It is a primary pollutant with a relatively long permanence time in the atmosphere (nearing four months) and with low chemical reactivity. The concentrations of this pollutant in the air are related to the traffic intensity in the measuring point, since in urban areas carbon monoxide is mainly emitted by motor vehicle. CO is considered as the reference tracer for this type of pollution throughout the

year. At high concentrations it is a powerful poison. The effects on humans are associated to the possibility of interfering with the transport of oxygen (formation of carboxyhemoglobin) to the tissues and in particular to the central nervous system.

1.2 Electrochemical gas sensor

The electrochemical gas sensors (ECs) used in MONICA device are manufactured by Alphasense™ (www.alphasense.com) and operate in amperometric mode using three electrodes. A generic EC practically generates a current that is linearly proportional to the fractional volume of the toxic gas. A special filter prevents the entry of dust and dirt allowing only the toxic gas (through the gas inlet) to reach the working electrode. The working electrode (WE) is also called sensing electrode as it is responsible for responding to the toxic gas with an oxidation reaction (in the case of CO, NO and SO₂ gases) or a reduction (for NO₂ and O₃ gases). Such reactions are aided by the presence of a catalyst on the surface optimizing the performance. Indeed, this electrode also allows the gas to come into contact with both an electro catalyst and an electrolyte in order to create a three-phase interface of gas, liquid and solid. This generates a current proportional to the incoming gas concentration. The sensor consists furthermore of two other electrodes: the counter electrode (CE) and the reference electrode (RE). Both have the same chemical composition of the working electrode. Lastly, each electrode is connected by means of a metal contact to the pins outside. Figure 1.2 shows the entire scheme of an electrochemical gas sensor.

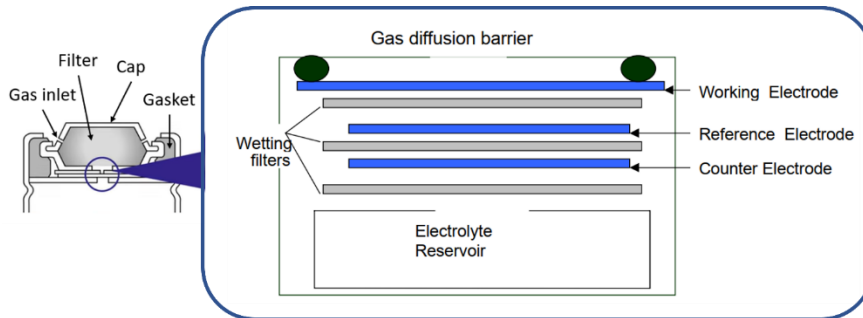


Figure 1. 2 Inside an electrochemical gas sensor.

As well as the normal Working, Reference and Counter electrodes, both B4 and A4 size Aphasense sensors include a 4th auxiliary electrode (AX), which is used to correct for zero current changes (Spinelle *et al.*, 2015). Since the redox reactions occur in pairs, if the working electrode oxidizes the incoming gas, then a reduction reaction takes place at the counter electrode

forcing a change of its potential. A suitable external potentiostatic circuit connected to the reference electrode anchors the working electrode at a fixed potential, at this point a potential difference is established between WE and CE in presence of the target gas. Therefore, the current induced by this potential difference is precisely the output of the sensor which will be proportional to the concentration of the incoming gas. To ensure that proportionality, the sensor has to work in the plateau region of the current-voltage characteristic curve, known as diffusion zone. Operating in diffusion zone the output current of the sensor is not critically sensitive to the applied potential and so the total sensor current is given by equation (1.1) where I is the sensor current generated across the electrolyte and electrode, k is a constant and CT is the concentration of the toxic gas.

$$I = k CT \quad (1.1)$$

This linear proportionality is valid for the whole range of concentrations (Alphasense, 2022a, Baron and Saffell, 2017).

1.2.1 Interfacing the ECs for use

As seen in the previous paragraph, a salient role in the operation of an EC take part by the potentiostatic circuit. This performs three premier functions: current measuring, control bias voltage and working electrode protection.

The output sensor current is measured with a single stage operational amplifier in transimpedance configuration. As previously stated, an EC must work in the diffusion zone to guarantee equation (1.1), well for many toxic gas species it has been found that whether a zero voltage is applied to WE with respect to RE this is assured. For all other species (NO for instance), the application of a bias voltage is required. Such bias voltage, which must be stable, is also provided by the potentiostatic circuit. It is normal practice to add a shorting FET for unbiased sensors so that the RE and WE are at the same potential when the instruments are switched off. This “zero bias” state ensures that when you switch the circuit back on, the sensor is ready immediately. If you do not use a shorting FET and leave the sensor open circuit when the circuit is off, the toxic gas sensor will take a few hours to stabilize when next switched on. For further details about a schematic of the Alphasense potentiostatic circuit consult the application note (Alphasense, 2022b).

The potentiostatic circuit is embedded on the Analog Front End board AFE-810-0020-00 manufactured by Alphasense (Alphasense, 2022c). The AFE board implements a low-noise circuit to optimise performance and it is designed for use with the A4 size air quality sensors. Connecting the AFE with a flat cable to the ADC converter of a micro-controller unit is possible recording air quality data immediately. The figure 1.3 shows the ECs and the AFE board as part of the MONICA node.

1.2.2 Elements affecting inherent sensors variability

Several external factors affect the output response of the EC sensors, among these the main ones are temperature and humidity.

The datasheet declares a sensitivity change (nA/ppm) from +0.1 up to +0.3%/K due to ambient temperature and also an increase of the sensor response time when temperature follows down under 10°C, affecting so the performance in harsh environment. When the humidity changes, the sensor shows current spikes that could be both positive and negative which effects decay in about 10 minutes. Similar consequences are also to be attributed to the atmospheric pressure change, but unlike humidity, its effect is negligible. To compensate these effects, the sensors must be calibrated appropriately correcting both temperature and humidity dependence. This is known as ad-hoc calibration. Another factor of variability to be aware is the variability between different batches of sensors even if the datasheets confirm a worst case error of $\pm 0.1\%/K$, related to the dependence in temperature, with 95% confidence level (Alphasense, 2022d). The inter-sensor variability will be taken into account when the emerging concept of general calibration model will be introduced.

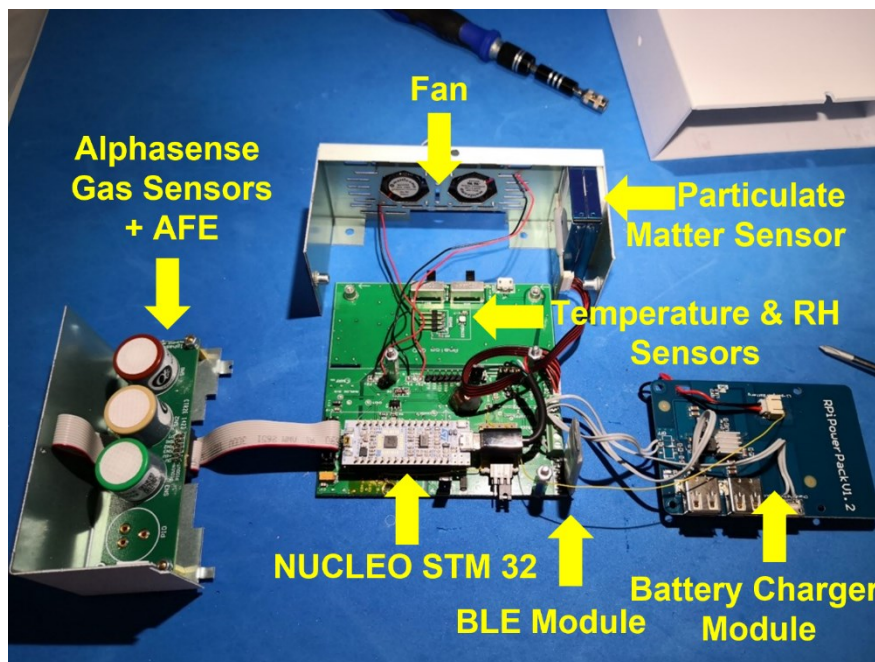


Figure 1.3 Exploded view of the MONICA node with Alphasense gas sensors, AFE and other paramount parts as well.

1.3 The MONICA node

The MONICA node is based on electrochemical sensors array using Alphasense A4 class sensor units, respectively targeted to Carbon Monoxide (CO-A4), Nitrogen Dioxide (NO₂-A43F) and Ozone (O₃-A431). Relative Humidity (RH) and Temperature (T) sensors complete the sensing array. Let call the version equipped with exclusively gas sensors MONICA 2.0 (Agresta *et al.*, 2017). The version that also includes Low-Cost Particulate Matter sensors (model type Plantower PMS-7003 able to measure PM10, PM2.5 and PM1) is named Monica 3.0 (De Vito *et al.*, 2021).

1.3.1 MONICA hardware architecture

The electrochemical gas sensors array is mounted on an analog front-end AFE 810-0020-00 provided by Alphasense that outputs the signals related to the concentration of monitored air gases. The Working Electrode (WE) and the Auxiliary Electrode (AE) signals of each sensors type are acquired and converted by a Nucleo LK432KC board from ST Microelectronics equipped with an ARM microcontroller with an integrated 12-bit Analog Digital Converter (ADC) for signals acquisition. Both temperature and relative humidity values are digitalized and acquired. In particular, AE readings may be used to partially correct for temperature interference affecting these sensors. The effect of temperature on the different electrode readings are different and temperature still affects their difference (WE-AE), due to their particular geometry and manufacturing difficulties. Two fans guarantee the minimum air flux to the sensors in order to follow the concentration dynamics and fosters the reactivity of the system. The power supply is provided to the entire node by a 3.7 V battery, a battery charger, and a step-up converter that boosts the voltage to 5 V. The battery has a capacity of 3800 mAh; when the node is driven in low power mode, it can stay in operation without recharge for more than 20 hours. MONICA sends a JSON packet containing all the sensors measurements at 15 samples/minute rate via Bluetooth.

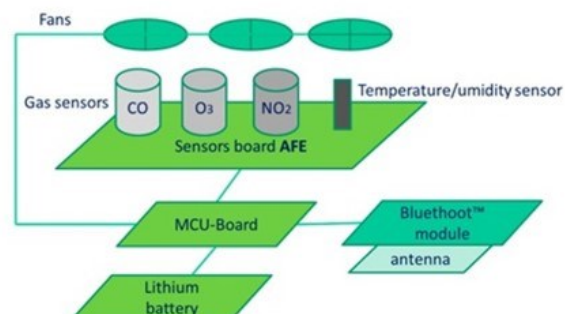


Figure 1. 4 Main functional blocks of a single MONICA 2.0 node.

The main improvement on MONICA 3.0 is the possibility to perform particulate concentration measurement with a Plantower PMS7003. It comes in a compact package that fits well in the node's case; the effective range of the sensor is 0 to 500 $\mu\text{g}/\text{m}^3$ and has a resolution of 1 $\mu\text{g}/\text{m}^3$. The last improvement concerns the Bluetooth transceiver; the new one is a Bluetooth low energy (BLE), which makes it possible to further reduce power consumption and enables communication with modern smartphones, which are adopting this technology as an interface to other devices.

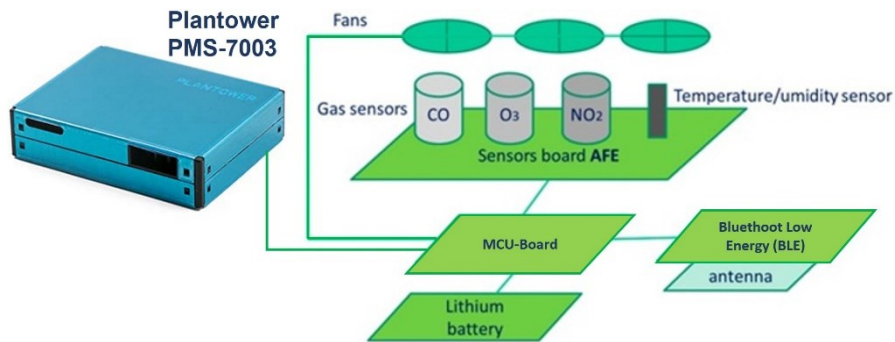


Figure 1. 5 Main functional blocks of a single MONICA 3.0 node.

1.3.2 IoT infrastructure

An IoT infrastructure is built around the MONICA device itself providing data communication, storage, processing, and visualization functionalities. MONICA gathered data are locally processed and sent to an ad-hoc backend in two different modes depending on the application setup: *stationary and mobile*. The stationary setup has been used during the co-location period with an institutional reference station needed to gather calibration data (training set). The mobile setup instead has been used for the citizen science campaigns.

During co-location, raw sensors data are captured and transmitted via Bluetooth/ or BLE to a Raspberry Pi Mod. 3B+ based datasink with Raspbian OS providing for local storage and WAN connectivity services through a mobile router wi-fi TP-Link M7650. Everything is managed by a python script running on the Raspberry Pi (Fig. 1.6).

At remote side, an ad-hoc IoT backend architecture relying on a contained NodeJS REST APIs server and MongoDB provides data inception, device management, storage, pre-processing and map based visualization functionalities (see figure 1.7 and 1.8).

In the mobile setup, data are sent to the backend through the use of an Android smartphone APP that completes the raw multisensory data tuple with position data gathered by the positioning service of the smartphone operative

system. Several activity screens guide the end user in the connection to the MONICA device, the initialization of a mobility session and its finalization, and the visualization of data gathered in past mobility sessions. In the MONICA 3.0 update the log in procedure both on mobile APP and web backend site has been developed using a professional service based on Auth0.

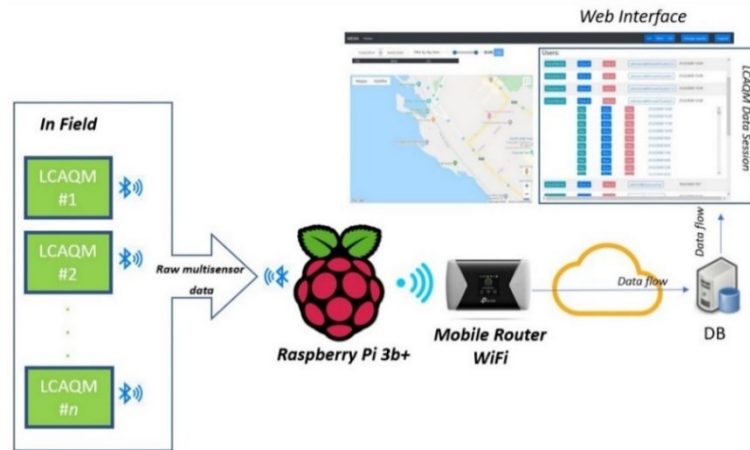


Figure 1. 6 Scheme of IoT architecture in stationary setup.

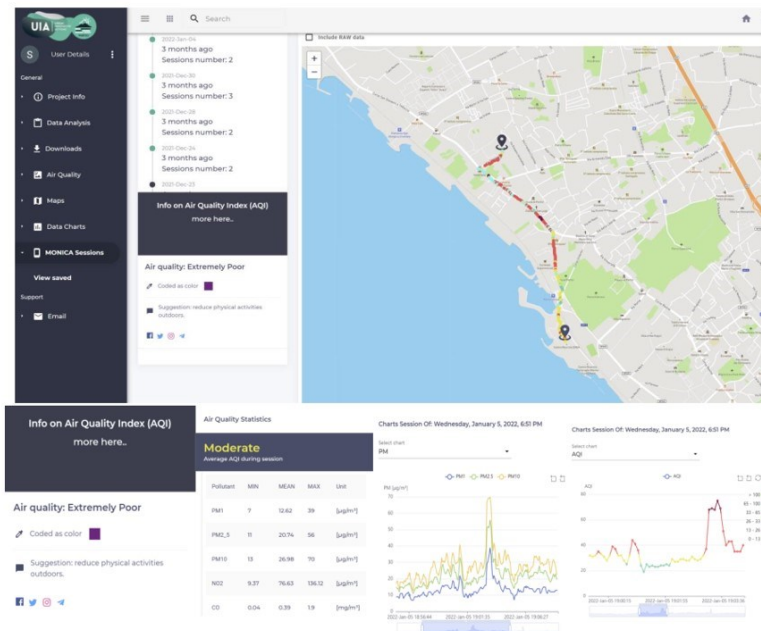


Figure 1. 7 Picture of a user session accessed through web service backend. PM plots are visible and also the user path on google maps.

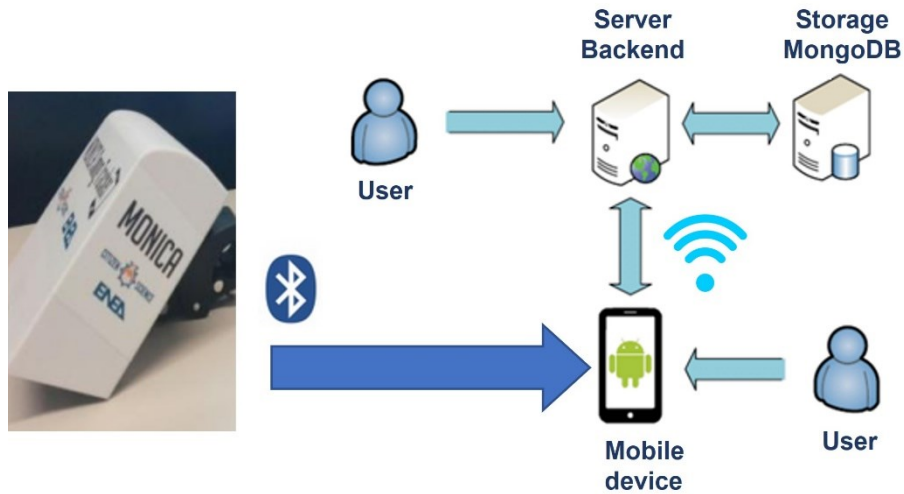


Figure 1. 8 *MONICA IoT architecture in mobile setup.*

1.4 Factory calibration and Laboratory characterization

As declared by the manufacturer, each sensor is tested before leaving the plant, in order to ensure that the specifications declared on the datasheets are respected. Traceability is also guaranteed and even a database with the sensitivities of each sensor tested is maintained. The so-called “*factory calibration*” is carried out through the following equation, which returns the concentration of gas under test in ppb:

$$gas_concentration = \frac{1}{S} [(Vwe_m - Vae_m) - (Vwe_0 - Vae_0)] \quad (1.2)$$

Where S is the sensor sensitivity, Vwe_m and Vae_m refers to WE and AX measured value, while Vwe_0 and Vae_0 indicate respectively measured value for both working electrode and auxiliary electrode for a zero offset (Mijling *et al.*, 2018, De Vito *et al.*, 2021).

For the proper study of gas sensors, it is necessary to be able to soak these sensors in an atmosphere of which the chemical composition, temperature and humidity can be modified with adequate precision and accuracy. Only in this way it is possible to correctly correlate the chemical-physical variations of the gas sensor to the variations of the traces of chemical compounds contained in the air. Normally this activity is defined as sensor characterization and usually performed in a volume test chamber.

Here is also briefly describe each mandatory steps to accomplish a sensor characterization as performed in ENEA gas sensor characterization laboratory (this is a state-of-art procedure, the same procedure that certified companies perform).

A 15 L large volume test chamber (LVTC, Fig. 1.9) is closed in an adjustable thermal box. The whole equipment constitutes a gas sensor characterization system (GSCS). In the LVTC, the air composition (humidity and chemical compound concentrations) is setup by using a certified mass flow meter. The accuracy of the gas chemical composition is ensured by the mixing of certified bottles (Rivoira SpA). For the accuracy on the nitrogen dioxide concentration, further validation is necessary by coupling the chamber gas output to a Teledyne T200 chemiluminescent total nitrogen oxide analyzer. Temperature and humidity are recorded with industrial sensors (LSI Pt100). The LVTC can sustain the calibration of several complete sensor systems at once. The calibration method consists in injecting in the inlet tube of the LVTC a constant flow of the target gas properly diluted at the maximum concentration (C_0) with humid synthetic air. The time-rising concentration $C(t)$ of the target gas is precisely predicted by the following exponential law that generally describes a transition between two steady states of a physical parameter under a time constant perturbation:

$$C(t) = C_0 \left(1 - e^{-\frac{t}{\tau}}\right) \quad (1.3)$$

The characteristic time (τ) is proportional to the free space inside the chamber and must be appropriately corrected when several sensors are inside the chamber.

The calibration procedure (run) consists of three time steps:

- 1) Synthetic air is injected for the unperturbed state recording of the sensor output (baseline);
- 2) The properly diluted gas target in the gas carrier is injected and the adsorbing phase of the sensor response is recorded;
- 3) The test chamber is washed in a constant flow of synthetic air while recording the desorbing phase of the sensor output.

With this procedure, it is possible to verify the sensor output behaviour during the adsorbing and desorbing phase of the chemical compound on the surface of the sensors. Sensing hysteresis or poisoning can be detected and measured.

The LVCT allowed us to place up to eight MONICA node (Fig. 1.9). A Raspberry Pi 3 with Raspbian and a Python script collected data via MONICA's parsers in the log files.

Sensor calibration is performed scanning the range from 0 up to 500 ppb for NO_2 and O_3 for instance at a controlled and constant temperature and humidity. For CO instead the range goes from 0 up to 5ppm. Figure 1.10a

Chapter 1

shows a graph of the time log for a sensor output during a calibration run with an injection of 5 ppm carbon monoxide. It is easy to distinguish the three steps of the calibration run; the red line underlines the adsorbing phase while the blue line shows the desorbing phase. As a result of the calibration run, a sensitivity curve was estimated by the sensor output log using a script in R language that synchronizes and correlates sensor output with the gas concentration. Once estimated, the sensitivity shown with a linear regression of the data (Figure 1.10b), can be used to explore the precision of the sensor output in the entire range of calibration. In this way, it is possible to estimate useful sensor parameters such as LOD (limit of detection), output linearity, precision, and accuracy (Massera *et al.*, 2020).



Figure 1. 9 LVTC with eight MONICA node under laboratory calibration procedure.

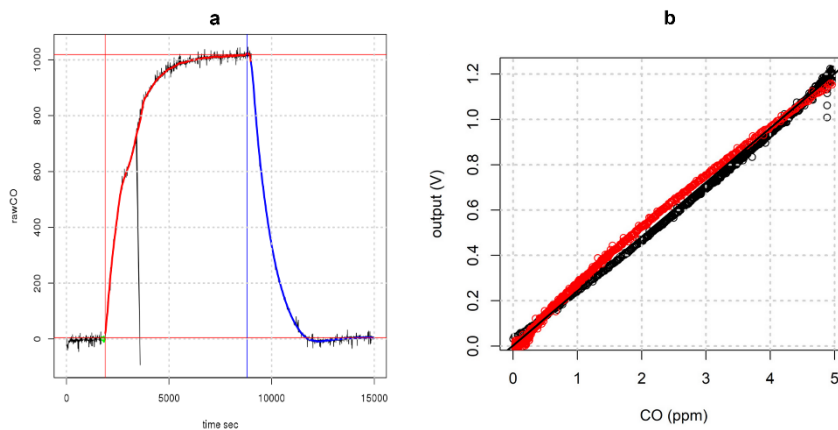


Figure 1. 10 a) Typical pulse with the three different phases of a calibration procedure (i.e., CO). b) Sensitivity curve and the linear regression for the sensor output in the range 0-5 ppm during adsorption (black dots) and desorption (red dots).

Chapter 2

Field Calibration of LCAQMS

2.1 Reasons for field calibration

One of the main factors that impacts the performance of low-cost sensors is the construction variability due to the manufacturing process. Identical sensors show different performances, despite the construction technology and belonging to the same platform. This become apparent already in laboratory test, where the conditions into test chamber are known and well defined, both in terms of gas that are injected (and therefore to be detected) and environmental conditions, i.e., temperature and humidity (Carotta *et al.*, 2001). A precise and rigorous analysis comparing laboratory and field tests is accessible in (Castell *et al.*, 2017). Authors find high correlations for all gaseous pollutants in the laboratory (the Pearson's correlation coefficient for NO₂ low-cost sensors is $r > 0.9$) when the sensors were tested under constant temperature and relative humidity conditions. Unfortunately, laboratory calibration alone is inadequate to deal with the unstable conditions of the operation field. Laboratory calibration cannot address for real world conditions and therefore a field calibration must be performed for each sensor individually. In fact, the same NO₂ low-cost sensors calibrated in the laboratory manifest a Pearson correlation index $r \leq 0.3$ in field calibration assessment. Additionally, calibration parameters may change over time depending on weather conditions and location, meaning that once nodes are deployed it will be difficult to determine whether the released data are compliance with DQOs. Therefore, it is necessary a severe evaluation of the low-cost sensor platforms under different environmental conditions. This not only makes clear the needfulness for field calibration, but also the need for special tools for the automatic detection of the performance decay of the calibration model, a topic that will be addressed by analyzing the concept drift problem in the next chapter.

2.2 Background to supervised machine learning algorithms

Two calibration methods were considered in this research project: Multivariate Linear Regression (MLR) and Neural Networks (NN). The goal is to identify the functional relationship that binds the input raw sensor data to gas concentration.

2.2.1 Multivariate Linear Regression

The essential assumption in linear regression is the linearity between input and output variables. Another fundamental assumption is that the input data follow a normal distribution (this simplification obviously has an impact on the model output). In field calibration of LCAQMS this means that for each sensors response is established the linearity of the raw sensor responses with reference measurement of the associated pollutant. Instead, the term multivariate suggest that the output variable Y is function of more input independent variables X_1, X_2, \dots, X_k and the overall mathematical law is due by the next equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad (2.1)$$

Where Y is the output variable (estimated gas concentration in our case), X_i the independent variables (features), β_i the coefficient of the multivariate regression model with β_0 the intercept and ε is the error of the model assumed to have a zero mean. Using the so-called “*fitlm*” function in MATLAB it is possible get the model coefficients β_i .

2.2.2 Neural Network

Neural networks (NN) are inspired by the human brain, imitating the way as biological neurons send signals each other (McCulloch and Pitts, 1943).

NN are composed of layers of nodes (neurons): an input layer, one or more hidden layers and an output layer (Fig. 2.1).

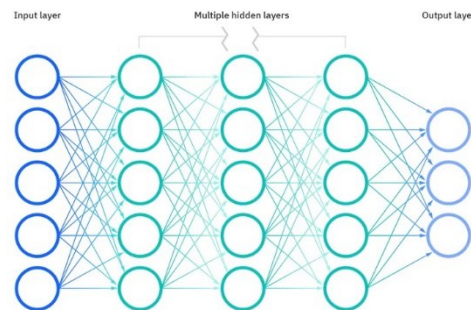


Figure 2.1 Functional layers of a neural network.

Each node connects to another and has an associated weight and threshold. If the output of any single node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed to the next level of the network. Such process is depicted in the figure below.

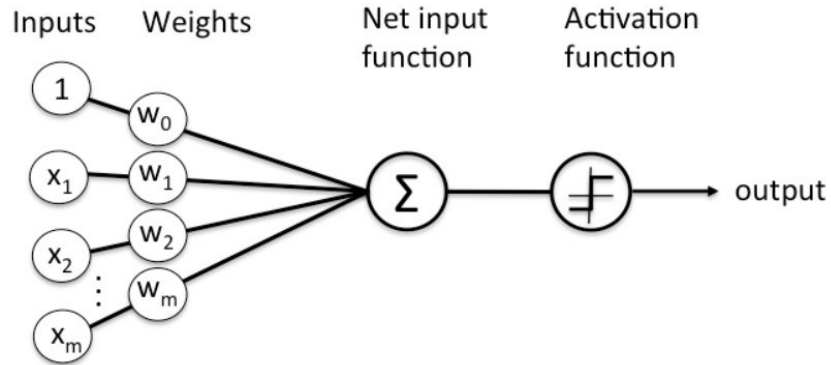


Figure 2.2 Diagram of node activation in a neural network.

Neural networks rely on training data (independently and identically distributed) to learn and improve their accuracy over time by searching for the right "weights". Deeper details about how a neural network works are available in (Bishop, 2006).

2.3 Evaluation Metrics

The use of machine learning methods in every application, but in our specific case in low-cost sensors calibration, need for evaluation metrics adequate to describe the quality of the algorithms applied both in term of fitting and predictive capability. Although the list of evaluation techniques and metrics is very extensive, here we will limit ourselves to using mostly relevant metrics in air quality systems performance assessment like the popular *coefficient of determination* R^2 and the related *Pearson correlation coefficient* r eventhough more emphasis will be caring in this study to the *Mean Absolute Error* (MAE [$\mu\text{g}/\text{m}^3$ for NO_2]), *Mean Absolute Percentage Error* (MAPE [%]) and *Relative Expanded Uncertainty* (REU) (EU, 2008, EC WG, 2010). In the following formulas y_i is the target value while \hat{y}_i is the predicted value.

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.2)$$

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2.3)$$

Details and formulas for the calculation of REU are covered in the next chapter.

2.4 Experimental framework

Included in the biggest challenges connected to the field calibration there are both choosing the length for the optimal collocation period (Esposito *et al.*, 2016) and the calibration model developing appropriate calibration strategies. With the data of the first mid-term co-location executed, we have been analyzed these issues.

2.4.1 Field calibration setup

Such field calibration experiment lasted two months long (January 02 to March 02, 2020) and four nodes named AQ6, AQ8, AQ11 and AQ12 have been deployed against a mobile regulatory grade analyser made available by the regional agency for environmental protection called ARPAC in the city of Portici, in the south of Italy. In Fig. 2.3 is detailed this co-location campaign. The following analysis, however, considers only on the winter campaign and focused on NO₂, considered one of most dangerous pollutants (Esposito *et al.*, 2020).



Figure 2.3 Field calibration scheme with a Google Earth view of the co-location and a photo of 4 MONICA 2.0 mounted on the roof of the mobile laboratory.

The recorded datasets consist of 1440 hours captured in a continuous sampling mode. Specifically, for each node, two datasets, with samples averaged at minute and hourly rate, have been built. These datasets contain averaged data from each of sensors embedded into the device, i.e., WE and AE raw sensors readings (mV) for NO₂, CO, O₃ targeted sensors plus T (°C)

and RH (%), joined to same time scale averaged data from a mobile ARPAC reference analyzer for NO_2 ($\mu\text{g}/\text{m}^3$), CO (mg/m^3), O_3 ($\mu\text{g}/\text{m}^3$). Table 2.1 resumes the acquired data from the 4 co-located nodes and data losses. Recorded data have been pre-processed, analysing the missing values, detecting the possible outlier's carrying out a correlation analysis.

Table 2.1 Recorded data and data losses.

MONICA node	Acquired data (hours/minutes)
AQ6	1432 h/83990 min
AQ8	1392 h/81854 min
AQ11	1422 h/81460 min
AQ12	1268 h/73832 min

The analysis of the ARPAC hourly validated data, collected during the co-location period shows a cyclical and long-term trend, according to the emission characteristics of the site. The significant decrease in average hourly NO_2 concentrations in the time series is due to the wind influence, with different intensity and direction (N-NW and N-NE). This allows the dispersion of pollutants of emission origin. In Fig. 2.4 is possible to see the NO_2 gas concentration distribution.

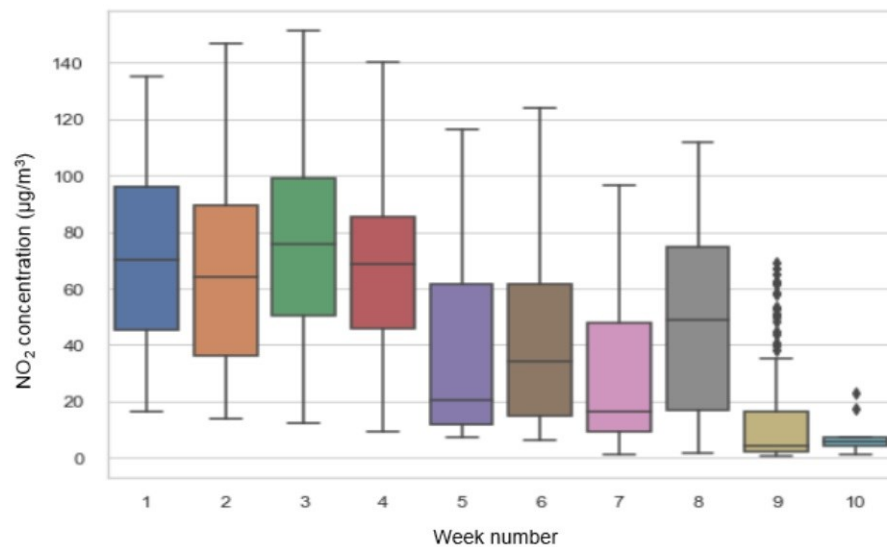


Figure 2.4 Box-plot representation of NO_2 gas concentration distribution along the entire co-location period.

2.5 Sensor calibration models

It is now common knowledge that chemical sensors array raw data needs to be processed by a calibration function to accurately and precisely estimate target gases concentration taking care of nonlinearities and interferent gases. Several and extensive on-field experiments along with theoretical results led to select two main calibration approaches, each of which has shown to be suitable in specific conditions: Multiple Linear Regression model (MLR) and Shallow Neural Network model (SNN).

Assuming that X is the input features vector and y the predicted value, the MLR model is the classic linear regression with multiple input features, mathematically expressed in equation (2.1). The selected nonlinear model is SNN, that has already proven very efficient for AQMS on field calibration. The analyzed SNN model is a three layers architecture, empirically equipped with three standard sigmoidal tangent neurons units in the hidden layer and a linear output layer. Automatic Bayesian Regularization (ABS) was used as training algorithm.

We focused on NO_2 hourly averaged concentration estimation problem using hourly averaged WE and AE sensors data of NO_2 , O_3 sensors plus T and RH data as inputs for the two calibration algorithms. The input matrix X , thus, consists of 6 features vectors as columns (WE_NO₂, AE_NO₂, WE_O₃, AE_O₃, T, RH) and the rows number depends on the training set length. So, the NO_2 concentration was calculated by using the following equation:

$$Y = \beta_1 + \beta_2 RH + \beta_3 T + \beta_4 AE_{NO_2} + \beta_5 WE_{NO_2} + \beta_6 AE_{O_3} + \beta_7 WE_{O_3} \quad (2.4)$$

where β_1 is the intercept, whereas β_2 up to β_7 are the regression coefficient for relative humidity (RH), temperature (Temp), auxiliary and working sensor signal of NO_2 (AE_{NO_2} , WE_{NO_2}) and auxiliary and working sensor signal of O_3 (AE_{O_3} , WE_{O_3}) respectively.

The two calibration algorithms have been tested using different choices of training lengths, meaning that calibration is performed in an ex-post mode by selecting for testing purposes only those samples that are temporally located after all the data used for training and validation purposes. This setting is the most adequate to simulate real conditions where nodes will be operated after the calibration took place. Details and results are reported in the following paragraph.

2.6 Experimental results

As previously mentioned, performance assessment experiments have been carried out using different training lengths combinations, aiming to the optimization of the involved parameters, furthermore models' performance were compared in order to select an optimal calibration strategy.

Table 2.2 captures the preliminary results of our experimentation. At a glance, it is possible to spot the different performance obtained by the different analysers. Irrespective of the calibration model and at each training length combination AQ8 node appears as the worst performing node. AQ12 node seems instead to express the best performance with respect to all the other nodes particularly when calibrated using an MLR approach. The SNN model performances are hampered when using data from second week; actually, without resorting to a validation set the learning process result in overtraining conditions that prevent the network to obtain good generalization capabilities. Generally, results obtained by MLR and SNN models appear similar with MLR keeping a limited edge on the performance obtained by SNN.

Table 2.2 Models performance with different choices for the training length (L , in weeks) for each node. Bold indicates best performance.

L	Mean Absolute Error (MAE) [$\mu\text{g}/\text{m}^3$]							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	11.7	7.94	21.94	23.36	8.20	7.78	12.23	6.55
2	7.53	7.70	25.64	16.78	10.07	9.51	8.82	6.92
3	8.89	7.73	19.48	13.30	10.09	8.86	8.33	6.49
4	8.74	7.56	11.71	12.63	10.24	9.88	7.08	6.31
5	7.98	7.63	13.15	11.37	9.6	9.65	5.79	5.15

L	Pearson Correlation Coefficient r							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	0.93	0.97	0.94	0.93	0.97	0.97	0.93	0.98
2	0.97	0.97	0.92	0.94	0.97	0.97	0.98	0.98
3	0.97	0.98	0.93	0.94	0.97	0.97	0.98	0.98
4	0.97	0.98	0.95	0.95	0.98	0.98	0.98	0.98
5	0.98	0.96	0.96	0.96	0.98	0.98	0.98	0.98

L	Coefficient of Determination R^2							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	0.79	0.91	0.47	0.41	0.91	0.92	0.78	0.94
2	0.91	0.90	0.22	0.62	0.85	0.88	0.88	0.92
3	0.88	0.89	0.49	0.74	0.86	0.88	0.89	0.92
4	0.87	0.88	0.77	0.75	0.84	0.84	0.91	0.93
5	0.88	0.88	0.75	0.81	0.87	0.87	0.94	0.95

Analyzing the performance indicators of Table 2.2 for results, it is clear that limited benefit could be obtained for using more than 3 weeks of data and that MLR and SNN held very similar results (let's not forget that increasing co-location time obviously means increasing deployment costs). The same trend is coming up from minutely analysis, as shown in table 2.3 where only the MAE metric is reported for the sake of brevity.

Table 2.3 MAE performance of models with different choices for the training length (L , in weeks) for each node in minutely analysis.

L	Mean Absolute Error (MAE) [$\mu\text{g}/\text{m}^3$]							
	AQ6		AQ8		AQ11		AQ12	
	NN	MLR	NN	MLR	NN	MLR	NN	MLR
1	11.16	10.44	16.07	19.83	10.20	9.72	11.72	11.12
2	11.51	12.01	20.59	22.01	10.70	10.61	10.76	12.53
3	12.91	11.59	19.71	19.21	11.90	10.28	11.93	10.92
4	12.07	11.23	16.17	17.12	10.84	10.48	10.48	10.28
5	10.57	10.87	14.75	16.05	9.77	9.86	9.91	9.78

We finally chose to select the MLR algorithm as the final calibration function for all of the devices. In fact, we used the entire dataset for training purposes, expecting a MAE for NO_2 estimation ranging from 6 to 12 $\mu\text{g}/\text{m}^3$ depending on the MONICA node. Applying a MLR model, moreover, it is a good and easy choice for embedding the resulting coefficient in the MONICA device-controlling Android APP used to perform following citizen science campaign for the personal exposure evaluation.

Figure 2.5 depicts the behaviour of SNN model-based estimations along with true concentrations for all the four nodes under calibration when using the first two weeks of data for training purposes and the third for test. It is easy to spot the sudden performance worsening attained at the end of the colocation period when concentration of the target pollutant is significantly lower than those encountered during training and validation period. Performance worsening are more evident for node AQ8 confirming its low capability to accurately estimate the target gas concentrations.

By focusing on figure 2.6, the underlying drivers of the poor performance of AQ8 node become evident with a significant bias error found along all the target concentration range and dramatic lack of precision when dealing with low target gas concentrations. A careful and continuous estimation of performance of the nodes can be helpful to rapidly detect poorly performing node and freeing its calibration spot to accommodate a new node. When benefitting from a continuous access to ground truth data significant differences in the linear fit parameters between node estimations may establish

a fast and robust way to determine anomalies in sensors operative performances.

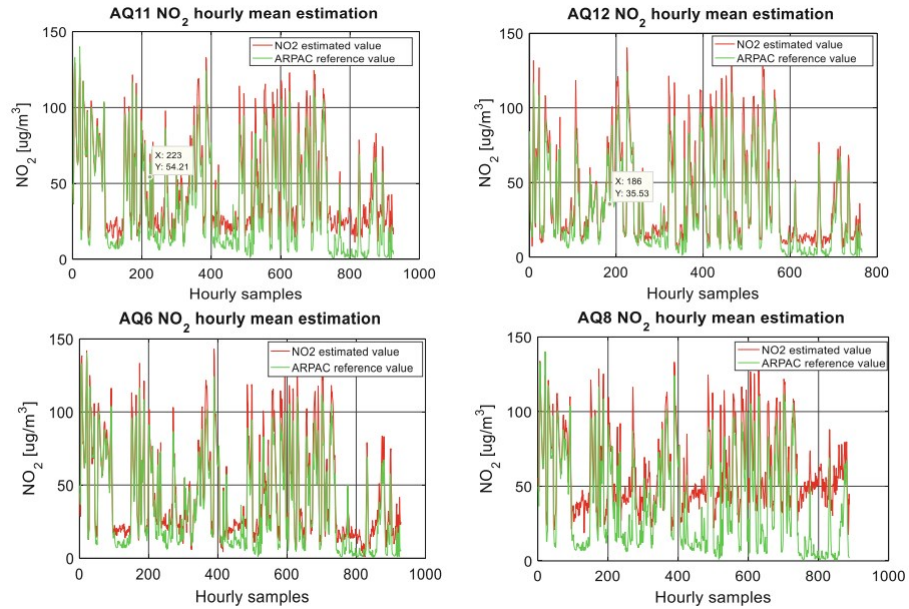


Figure 2.5 *NO₂ hourly concentration estimations versus target gas concentration starting from the 4th week of the co-location period.*

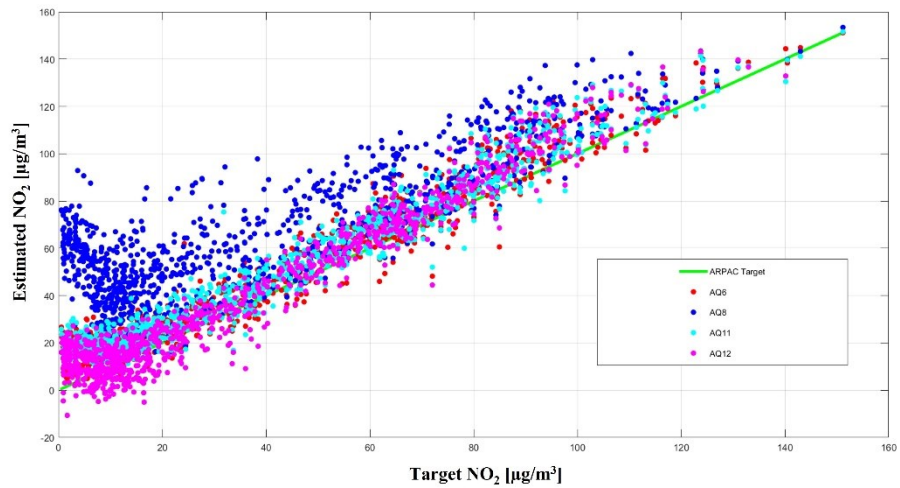


Figure 2.6 *NO₂ gas concentration estimation computed for each node, along the entire co-location period versus target gas concentration line. The differences among the four sensors performances become apparent when considering low true concentration of the target pollutant.*

2.7 Citizen science campaign in Portici during COVID 19 Phase 2

In order to assess the mid-term performance calibration procedures and test whether a simple MLR model is capable in detecting pollution hotspots, a citizen science campaign is performed embedding in an Android APP the coefficients of the model. In that way volunteers can see, on their own smartphone, the level of pollution observed during each session and at the same time send and store all data on backend.

2.7.1 Campaign details

During COVID 19 outbreak in Italy, government has decided to impose strict mobility limitations that have significantly and positively affected the Air Quality. In the so-called phase 1, only a fraction of commuters, belonging to strategic production companies, were allowed to physically reach the working place. All the others were forced to use smart working strategies. All the shops, except those belonging to food chain, were shut down. Figure 2.7 shows the NO_2 trends in the city of Portici as measured by the ARPAC mobile laboratory. A dramatic reduction of NO_x pollution levels has been observed starting from March 10th, the phase 1 start date. PM levels though, show a less pronounced reduction trend showing to be less affected by the reduction in car mobility emissions (Figure 2.8 for $\text{PM}_{2.5}$ fraction).

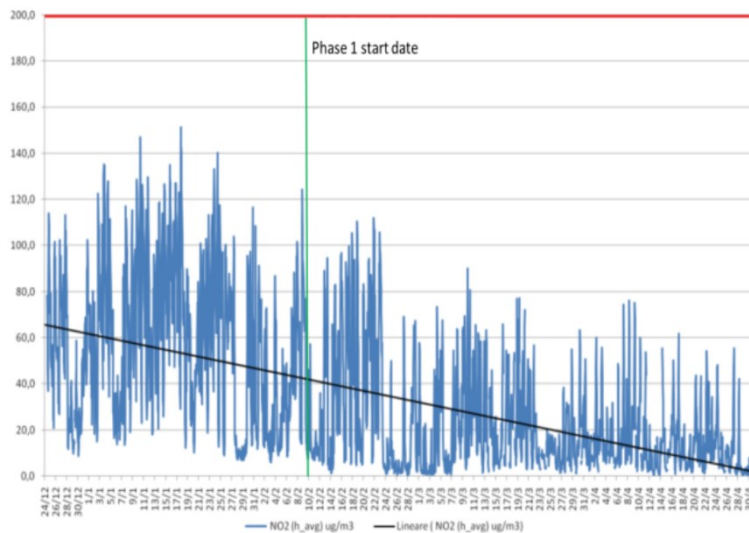


Figure 2.7 NO_2 gas concentration as measured by ARPAC mobile laboratory within the city of Portici and specifically in winter 2020 co-location at the Waterfront building site. Linear trend is highlighted in black.

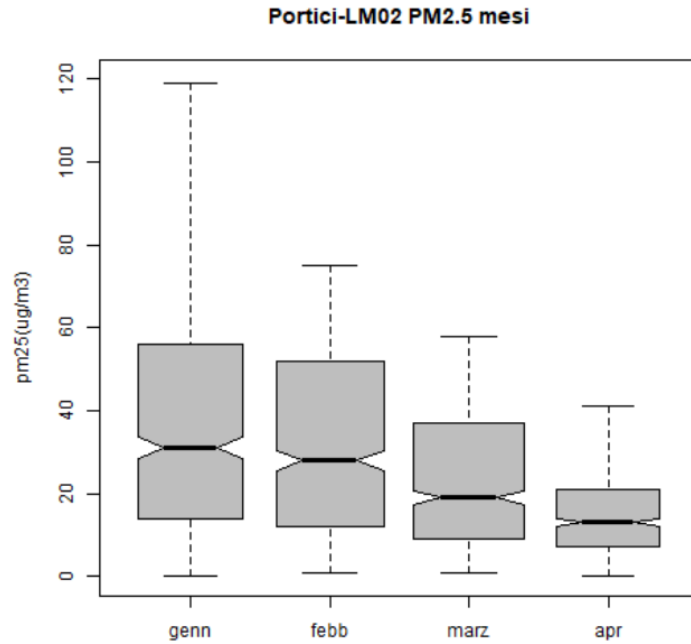


Figure 2.8 Monthly boxplot of the daily average concentrations of PM2.5 measured by ARPAC mobile Lab at Waterfront site in Portici.



Figure 2.9 The four pre-set monitoring paths (red, blue, green, orange) along with the mobile laboratory location (blue dot).

Starting from May 4th, Phase 2 began, carrying a relaxation of the mobility limitations. Though schools were still closed and smart working remained the

preferred working model, non-food shops were allowed to reopen and non-delayable tasks allowed workers to physically reach the working place. This forcibly simulated conditions in which population improved their mobility strategy toward more sustainable approaches. In these conditions, 6 volunteers citizens, belonging to local associations in the city of Portici were selected to use the calibrated MONICA 2.0 devices to monitor air quality according to a specific proposed monitoring scheme (Fig. 2.9). This implied a minimum of 1hr cumulative monitoring sessions duration each working day by feet following one of 4 different paths with one of the 4 calibrated devices that was assigned to single volunteers on weekly basis. Aside from technical difficulties, only four of the foreseen totals 60 (15x4) monitoring slots have been deserted.

2.7.2 Air Quality Geo-Mapping algorithm

A pervasive monitoring of the air quality by using mobile sensing devices as MONICA returns very large sample point datasets. For analysing their spatial pattern, in geostatistics air pollution distribution maps can be generated using deterministic and/or stochastic models (Brindha *et al.*, 2023). Among deterministic interpolation methods, Inverse Distance Weighted (IDW) interpolation is well-suited to be used with large pollution level sample datasets (Shepard, 1968, Du, 2020). Moreover, it allows to model properly the air pollution phenomenon driven by local variation, which it captures by defining an adequate search neighbourhood. This method explicitly assumes that phenomena that are close to one another, subject to the choice of an appropriate measure function, are more alike than those that are farther apart. To predict a value for any unmeasured location, IDW uses the measured values surrounding the prediction location. The measured values closest to the prediction location have more influence on the predicted value than those farther away. The map calculated depends on the selection of the power value p for inverse distance measurements weighting and the search neighbourhood strategy (*circle or ellipse*). IDW is an exact interpolator, where the maximum and minimum values in the interpolated surface can only occur at sample points, and the output surface is sensitive to clustering and the presence of outliers. IDW was actually used to compute an average interpolation of pollutant concentrations on a predetermined grid by applying the opportunistic measurement taken in a particular time slot using:

$$c(x, y) = \frac{\sum_1^N w_i c(x_i, y_i)}{\sum_1^N w_i} \quad (2.5)$$

Where $c(x, y)$ is the concentration at interpolated location $p = (x, y)$, the concentration at measurement points $p_i = (x_i, y_i)$ is $c(x_i, y_i)$, while w_i is the weight due by:

$$w_i = \frac{1}{d(p,p_i)^k} \quad (2.6)$$

With d the Euclidean distance with $k = 2$. At an interpolating position, IDW uses the actual concentration recordings. A preliminary step is undertaken to average all measurements that took place within a specific grid cell. The IDW surface grid resolution was 10m.

2.7.3 Citizen science campaign results

Table 2.4 captures a first level statistical characterization of the monitored values. For CO and NO₂ results are compatible with the expected increase in pollutants concentrations with respect to phase 1 measurement, due to the slow restart of the productive activity in the area due to phase 2 regulatory framework. Ozone is keeping similar values to the ones recorded during the last days of phase 1. Figure 2.10 actually shows the measurement density pattern along the predetermined paths within the city urban boundary. In particular some areas appear overrepresented, this is mainly due to the different length of the proposed paths that lead to multiple shots taken during multiple passages in the same day. Care should be taken in evaluating underrepresented areas (darkest colours) that will suffer from temporal variance dependence, potentially leading to no representative results in the resulting IDW averaged spatial patterns.

Table 2.4 First order characterization of recorded data.

	First order statistics	
	Average	Standard deviation
CO (mg/m ³)	0.44	0.64
NO ₂ (µg/m ³)	40.0	37.1
O ₃ (µg/m ³)	76.2	34.3

Calibrated data featuring measured concentrations were fused to build Inverse Distance Weighting maps. Figures 2.11-2.13 show the resulting pollution patterns. Specifically, figure 2.11 shows the average concentration patterns of CO as monitored during the campaign from all the volunteers in the respective hours of the day. These are characterized by localized hotspots near main crossroads as well as in areas that are subjected to heavy car traffic. However, an unforeseen hotspot emerged confirming the unprecedented resolution power of cooperative mobile monitoring. If confirmed by ad-hoc measurements campaigns, this could lead to the development of fact based remediation policies by the main urban administration entity. NO₂ pattern (Fig. 2.12) analysis basically confirms the hotspots identified by CO pattern analysis, however some of the most polluted areas are characterized by values

Chapter 2

that come closer to regulatory thresholds with respect to measured average CO concentration values. Ozone IDW averaged values show a lower spatial variance but are relatively closer to regulatory thresholds and locally overcome them. While this behaviour is common in the summer season in the monitored area, this results call for a closer analysis of the main drivers (Fig. 2.13).

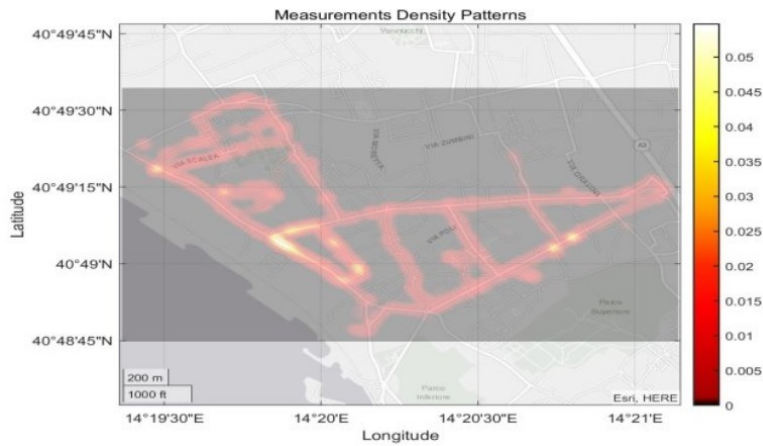


Figure 2.10 Measurements density plot shows slightly uneven density patterns and unforeseen measurements paths.

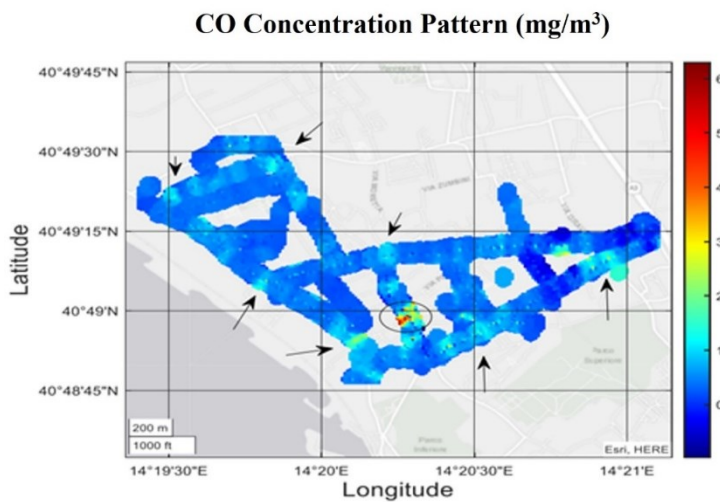


Figure 2.11 IDW averaged CO concentration pattern is characterized by localized hotspots near main crossroads or streets characterized by heavy traffic load (arrows). Unforeseen hotspots have also arised prompting for ad-hoc measurements campaigns (ellipse).

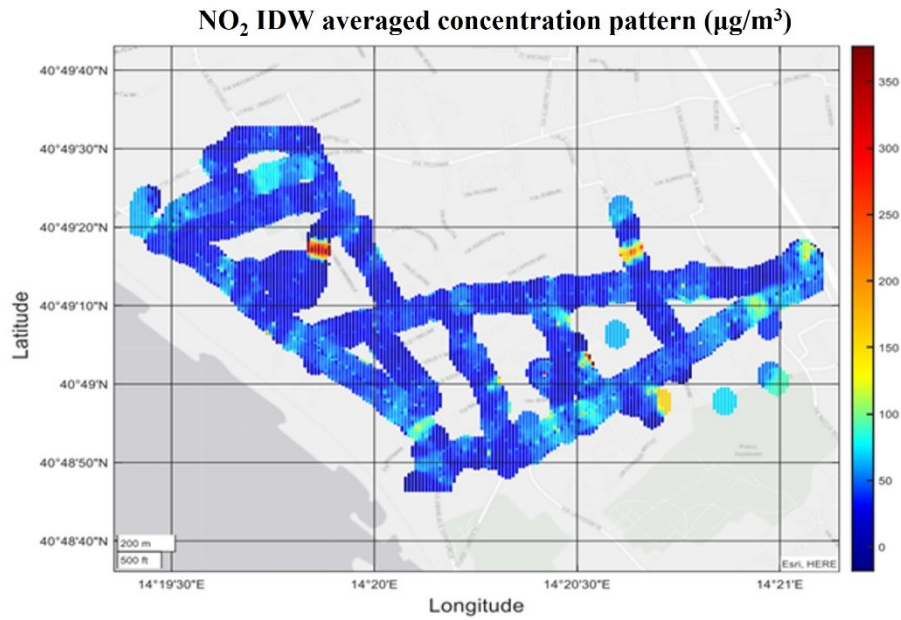


Figure 2.12 IDW averaged NO₂ concentration pattern.

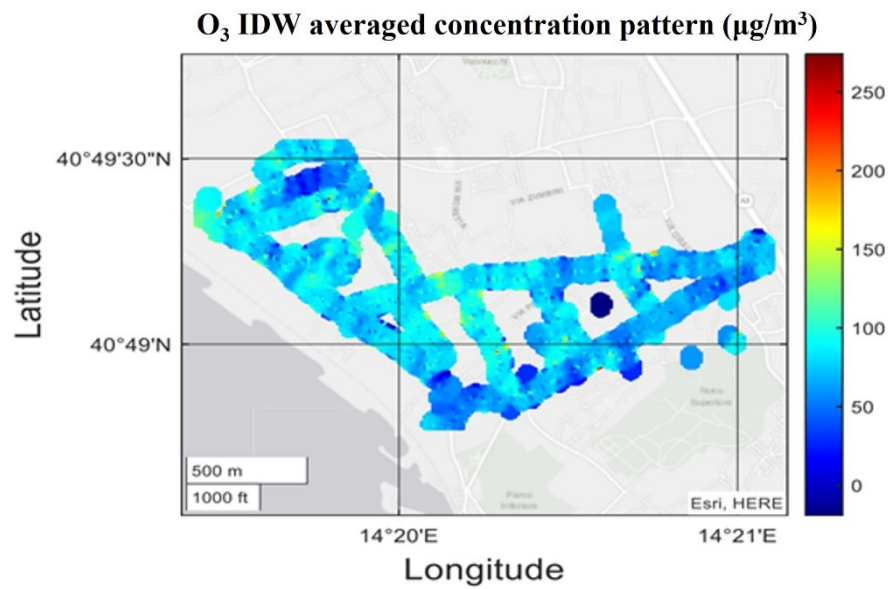


Figure 2.13 IDW averaged O₃ concentration pattern shows generally lower spatial variance with average values that reach or overcome the regulatory threshold.

2.8 Global Calibration Methodology

As we have seen so far, field calibration is a compulsory step to gain satisfactory data quality level in low-cost sensors technology over medium term. Nevertheless, for large deployments, especially in future smart city applications, field calibration represents a non-negligible cost. One of the emerging approaches to achieve cost reduction is the assessment and evaluation of generalized models. Well, in this paragraph we report some of the most promising works in this area.

A generalized model can be assimilated to the model of a virtual sensor that can be applied effectively to all copies of the same sensor. However, since two "identical" items do not exist in the sensor manufacturing process, the approach of (Malings *et al.*, 2019) to reduce this sensor-to-sensor variability is in the exploitation of the overall information gathering during field collocation, building a training set employing the median between the values released by the various sensors located in the same batch. Overall step for build the training set of a generalized median model are reproduced in figure 2.14. Thus, the constructed training set has the peculiarity of a better generalization about the inherent variability between the instruments, but in order to be able to carry out an exhaustive and rigorous performance evaluation, a comparison should be carried out with a fixed training algorithm. For this reason, in our work, comparisons will be made with the same training algorithm, a multivariate regression model. Another interesting property of the median generalized model is the robustness against the relocation to other sites problem. This is even more pronounced for CO sensors, but also a minor error in the long run (one year or more) when the generalized model is built with sensors that are about the same age. These promising results are related to electrochemical gas sensors.

A different approach is the one proposed by (Miquel-Ibarz *et al.*, 2022) whose ground idea is that there is a common hidden structure between the responses of multiple units of the same sensor model and the gas concentrations. Well, finding this common structure means discovering a model that is suited on all sensors. This procedure has already been passed through by (Solórzano *et al.*, 2018) for a classification task, while Miquel-Ibarz extends it to a regression task, but both works are applied to Metal-oxide gas sensor (MOX). Among the multilinear regression techniques, the Partial Least Squares is certainly one of the most used and popular (especially in chemometrics) for its ability to deal with problems with high dimensionality data. A PLS variant able of extracting the maximum information that reflects the maximum variation in the data set by reducing the functional dependence between the predictor variables (X) and responses (Y) is the Orthogonal-PLS (Trygg and Wold, 2002). Finding the hidden structure means finding a relationship between the so-called latent variables. It is clear that the concept of latent variables derives from the even more well-known Principal

Component Analysis (PCA) technique and in our applicative scenario suggests that there are some robust features associated to the fabrication variance.

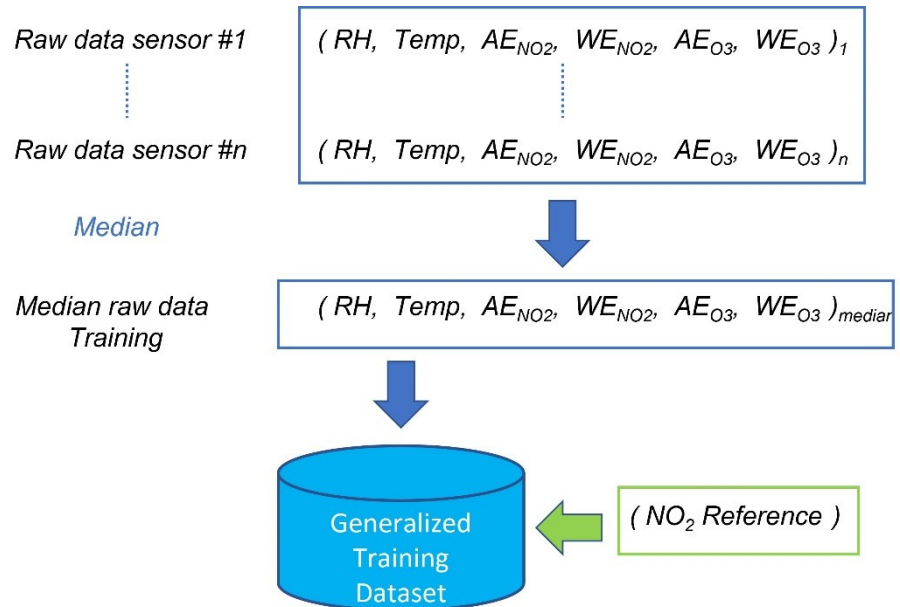


Figure 2.14 Procedure for creating the training set of the general calibration model.

Chapter 2

Chapter 3

Influence of Concept Drift on Metrological Performance of Low-Cost NO₂ Sensors

3.1 A neglected phenomenon in field calibration: The concept drift

One discussed point that still grips the massive diffusion of these systems is the low precision and accuracy, as the fact that no exist standardized protocols for the calibration process yet. The European directive disposes to use the relative expanded uncertainty (REU) to evaluate the equivalence of LCAQMS measurements with respect to the reference methods and consequently the compliance with DQOs (EC, 2008, EC WG, 2010). By now, during the process of the on-field calibration, machine learning (ML) algorithms are normally used, in fact in the last decade, a lot of emphasis was given to the search and analysis of the calibration models able to provide the best performances. An extended list is available in (Concas *et al.*, 2021). Whatever ML model is used, both the simplest and the most complex, all are based on the common assumption that training and test distributions are independently and identically distributed (i.i.d. assumption). In practical scenario, because of dynamic environments, noise, and other factors, this assumption may fail (Ditzler *et al.*, 2015).

In such cases, the model operates under different distributions of training and test set bringing blunders in previsions. The calibration model tends to become inexorably unreliable. Although over the years, researchers have given different names to the same subject, this phenomenon is well known in ML field as “*Concept Drift*” (Quinonero-Candela *et al.*, 2008, Lu *et al.*, 2019).

Such topic has acquired an ever more due to disruptive impact it could have on several application sectors (Gemaque *et al.*, 2019). In fact, issues ascribed to concept drift only recently have been relieved in low-cost air quality monitoring community. Different distributions of target gas, “interfering”

gases, and environmental changing variables during the phases of training and test play negatively on the quality of the model’s predictions. A theoretical and quantitative basis is outlined in (De Vito *et al.*, 2020). If the latter is essentially a statistical approach, on the other side, a predictive maintenance technique has been also investigated (Tancev, 2021). Anyway, both in fixed and mobile applications, the LCAQMSs will come across different operating environments. In such scenario, it is evident the probability that a measuring node will operate under the concept drift is very high. The majority approach found in the literature is the search for the so-said “optimal model,” a calibration model that it is hoped to have good performances for as long as possible time and is able to be up against each variable changing.

Instead, we believe that an adaptive approach is more suitable and effective in dynamic environmental spaces. Therefore, the proposal that we are going to investigate explores the possibility of identifying the “changing conditions” experimented by the instrumentation (i.e., the detection of the concept drift) and then eventually to adapt and/or to retrain the calibration model (as depicted in Fig. 3.1). To this aim, equipping an LCAQMS with an add-on capable of detecting the concept drift is paramount.

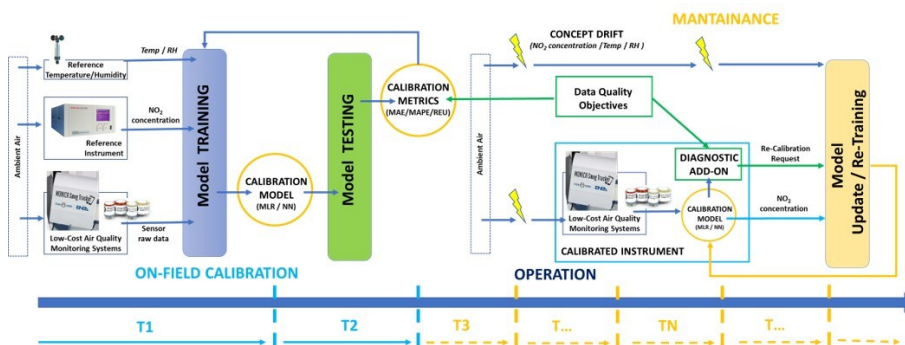


Figure 3. 1 Typical life cycle of LCAQMS: on-field calibration, instrument operation, and maintenance.

For all that reasons, a first step will be an evaluation of how the presence of the concept drift affects the performances of the calibration model. The computation and analysis of the REU is carried out working on measurement data that have been gathered during a middle-term experimental campaigns unrolled in Portici (Naples, Italy) in the domain of the European Union project AIR HERITAGE during winter 2020, already described in previous chapters.

3.2 Calibration metrics and data quality requirements

Many metrics have been suggested in the scientific literature regarding the performance analysis and comparison between ML approaches applied to the

on-field calibration of LCAQMS. Table 3.1 reports the most adopted metrics in terms of description and corresponding formulas (y_i is the target value while \hat{y}_i is the predicted value) that have been first considered to find out if there are any clues attributable to the existence of a concept drift. However, evidence of changing conditions could be masked in MAE and MAPE; therefore, a deeper statistical analysis will be conducted with reference to the training and test data in the following sections.

Table 3.1 Evaluation metrics.

Symbol	Description	Formula
MAE	Mean Absolute Error	$\frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i $
MAPE	Mean Absolute Percentage Error	$\frac{100}{n} \sum_{i=1}^n \left \frac{y_i - \hat{y}_i}{y_i} \right $

A LCAQMS to be used as pollution monitoring system for indicative measures must be compliance with DQO defined in the European Directive (EC, 2008), which in case of NO₂ establishes the REU limit at 25%. The Relative Expanded Uncertainty U_r should be estimated using Equation (3.1) and (3.2)

$$U_r(y_i) = \frac{2 \left(\frac{RSS}{n-2} - u^2(x_i) + [b_0 + (b_1 - 1) x_i]^2 \right)^{(1/2)}}{y_i} \quad (3.1)$$

$$RSS = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \quad (3.2)$$

where n is the number of measurements, x_i and y_i , for $i = 1, \dots, n$, refer to the reference and low-cost system measurements, correspondingly. RSS represents the sum of (relative) residuals calculated making use of equation (3.2) and $u^2(x_i)$ is the random uncertainty of the standard method (EC WG, 2010).

The coefficients b_0 and b_1 are the intercept and the slope of the orthogonal regression, typically evaluated with the procedure described in Annex B of (EC WG, 2010).

3.3 Methodology

3.3.1 Concept Drift

ML models are based on the hypothesis that the probability distribution of test data matches training distribution data, but if something changes the

model will fail giving bad quality values to the user. The described phenomenon is known as concept drift and it plays a primary role in the performance evaluation of a ML model, although metrics such as those introduced in Table 1 are usually only adopted. In agreement with (Lu *et al.*, 2019) there is a concept drift when the following condition holds:

$$\exists t : P_t(X, Y) \neq P_{t+1}(X, Y) \quad (3.3)$$

where $P_t(X, Y)$ is the joint probability of feature vector (X) and model output (Y) at instant time t, while obviously $P_{t+1}(X, Y)$ is the same but referred to instant time $t + 1$. Reporting this formalism in Air Quality Monitoring we will refer at t as the training data window, while to $t + 1$ as the test window. Splitting the joint probability in two contributes it is possible rewrite the relation (3.3) as in (3.4):

$$P_{training}(X)P_{training}(Y|X) \neq P_{test}(X)P_{test}(Y|X) \quad (3.4)$$

meaning that the concept drift is associated both to $P_t(X)$ and to $P_t(Y|X)$. For this reason, it is linked to the notions of covariate shift $P_t(X)$ and prior probability shift $P_t(Y|X)$. Moreover, as time evolving, the drift could be following different pattern types which go by the name of abrupt, incremental and gradual drift (Losing *et al.*, 2016). When one or more types of drift happen, the model predictions move further away from the true value. The model error could increase up to a such unacceptable levels enforcing the user to "invalidate" the model.

Turning to the metrological aspects, the temperature may be considered among the most relevant input variable influencing the output from calibrated electrochemical sensors, so much that temperature correction algorithms have been also proposed in the literature (Wei *et al.*, 2020). Thus, it will be treated in the covariate shift evaluation. Instead, in the prior probability shift evaluation the reference nitrogen dioxide concentration will be examined. The usage of both variables has been already suggested in the dissimilarity assessment of a multivariate approach reported in (De Vito *et al.*, 2020). The NO₂ reference values are handled only to demonstrate the presence or absence of the concept drift. Since the reference values are not available in operating conditions, it will be investigated the auto-detection of the concept drift directly from the model predictions.

3.3.2 Statistical tools for detecting the Concept Drift Events

To evaluate if the training and test samples come from the same distribution, the Two-Sample Kolmogorov-Smirnov test (TSKS-test) may be adopted. Indeed, the TSKS-test is a non-parametric statistic test that does not

imply any assumption on the type of data distribution unlike t-Student or Chi-square test.

Computing the maximum difference D between the cumulative density functions of both training and test:

$$D = \max_x |F_{training}(x) - F_{test}(x)| \quad (3.5)$$

and indicating with H_0 the null hypothesis when samples are from the same continuous distribution:

$$H_0 : F_{training}(x) = F_{test}(x), \forall x$$

if $D > D_{critical}$ according to the Kolmogorov-Smirnov distribution at a level of confidence α , then the null hypothesis H_0 is rejected (Rabanser *et al.*, 2019, Raab *et al.*, 2020).

3.3.3 Estimation of Relative Expanded Uncertainty

When a LCAQMS is calibrated by means of a ML model, we are willing to accept a minimum error during the training phase and surely, at this stage, the i.i.d. assumption is respected.

An analogous condition (i.e., the LCAQMS is characterized by an acceptable accuracy, similarly to the reference instrument) can be highlighted by the Relative Expanded Uncertainty (Walker and Schneider, 2020) when the following conditions hold:

$$b_0 \rightarrow 0 \text{ and } b_1 \rightarrow 1; \quad (3.6)$$

resulting that $y_i \approx x_i$ and entailing that also their distributions are equal.

The conclusion is that to ensure compliance with the DQOs, it is necessary to work in these conditions. In a such scenario it is compulsory give to an LCAQMS the ability to detect the concept drift, and this could be the right way to achieve the goal of compliance with European Directive. However, a change in training and test distributions is neither necessarily index of a dramatic drop in model performance, nor a good reason to invalidate the model. It is required set up an appropriate threshold beyond which the change of distributions negatively affects the estimated REU and consequently the whole model performance. Final user must be aware that the LCAQMS is not able to be trusted or relied on in this operating context. The succeeding proposed approach goes to this direction.

3.3.4 The proposed approach

The objective of the present study is evaluating how the concept drift affects the metrological performance of the LCAQMSs rather than to compare

different ML calibration models (topic covered by an extensive scientific literature). Indeed, both the MLR and NN model have been proved to be effective for carrying out the on-field calibration of low-cost gas sensors and make the corresponding instruments useful for achieving the qualitative measurements suggested by international legislation. On the other hand, the Quality Control of these instruments is yet a topic of research (i.e., the estimation of the time intervals and operating conditions during/under which the performed on-field calibration is still valid as well as the proposition of suitable strategies of re-calibration or model update).

Thus, the following methodology based on a data-driven approach is proposed to monitor the validity of the on-field calibration and consequently perform the Instrument Maintenance of the LCAQMSs. In detail, it consists of the following three steps:

- 1) Detection of the Concept Drift events, through the continuous application of the TSKS-test to the distributions resulting from the measurements during the co-location adopted respectively for calibration and LCAQMS operation (on the same window time).
- 2) computation and analysis of the REU: we will indicate a REU plot completely above the 25% level with the label FAIL against the label PASS, that indicates the significant presence of any REU plot points below.
- 3) evaluation of a suitable threshold (using the TSKS-test statistic obtained correspondingly to the label PASS) in order to confirm the influence of the concept drift events on the metrological precision of the calibration model.

A heuristic based on the previous threshold should allow the scheme of an add-on block to be drawn and embedded easily in the hardware architecture of an LCAQMS or on the backend side, able to guarantee the continuous monitoring of the model performance and promptly alert for the necessity of the model recalibration.

3.4 Results and discussion

The validation of the proposed methodology is detailed in the following subsections with reference to a mid-term measurement campaign. Indeed, after an overview of the experimental setting and the performed sensor calibration, the main issues of the methodological steps arising from the critical observation of the experimental data are discussed about: *i*) the detection of possible concept drift causes (through the application of statistical analysis); *ii*) the quantification of the corresponding influence on metrological performance (based on the estimation of the expanded uncertainty); and *iii*) the detection of significant concept drift events from the analysis the sensor

system output (based on the correlation between the statistical test results about the observed data distributions and the estimated precision of the calibration models).

Finally, the implementation of the methodology aiming to achieve the self-capability of detecting the concept shift by the low-cost sensor system is suggested.

3.4.1 Experimental Data and MLR Calibration

Four LCAQMS denominated AQ6, AQ8, AQ11 and AQ12 (MONICA 2.0) were co-located with the reference mobile laboratory provided by ARPAC during two-months measurement campaign in Portici (40°49'18.0"N 14°19'27.6"E) started on 2 January 2020 and finished on 2 March 2020. In detail, the reference instrument is the chemiluminescence NO-NO₂-NO_x Analyzer Model Type 42i by Thermo Fisher Scientific Inc. certified according to the Standards UNI-EN 14211:2012 and US-EPA RFNA-1289-074 and characterized by an average relative expanded uncertainty REU_{ref} equal to 7.9% for hourly measurements in laboratory whereas, the AQ8 prototype was not taken into consideration as it was out of order and broken during the next summer collocation in 2020.

To evaluate the performance evolution of the calibration model and eventually highlight the occurrence of the concept drift, the time series was divided in eight consecutive slot times, each one lasting 1 week. The last days data have not been considered because characterized by regional lockdown due COVID-19 pandemic emergency. In such way almost the whole dataset has been exploited, as shown in figure 3.2.

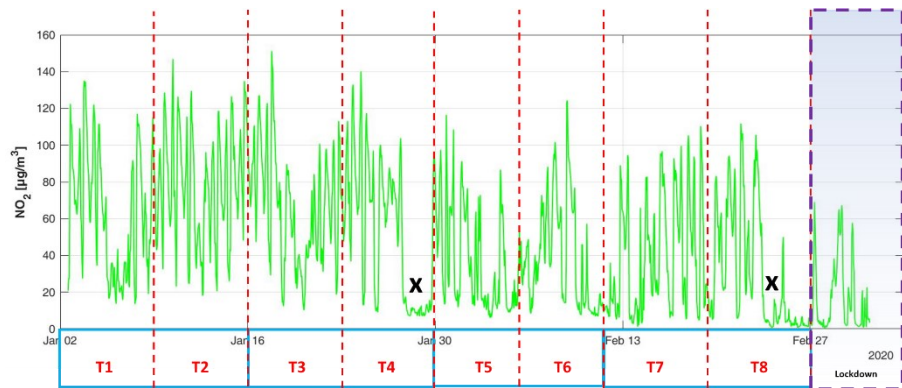


Figure 3. 2 NO₂ concentration (reference station) during the co-location period and time slot partition. Two abrupt changes in time series are marked with x.

Chapter 3

The hourly measurement data corresponding to the first week have been adopted for estimating the MLR calibration model for each one of the three considered LCAQMSs about the NO₂ concentration according to Eq. (2.4). Satisfying models have been achieved concerning with the corresponding accuracy in terms of the traditional metrics (MAE and MAPE), whose trends during the measurement campaign are reported in figure 3.3. Indeed, for each sensor system both the metrics are characterized by low levels (ever lower than 5 µg/m³ and 10% respectively) during the second week (i.e., the first period of the time series considered as testing set) and comparable with the accuracy of the models estimated correspondingly to the measurement data during the first week of co-location (i.e., the training set). Further details of the MLR calibration models achieved are concerned with the behaviour exhibited at very low pollutant concentration (expected on the basis of the corresponding scientific literature). Then, the estimated negative values with AQ6 and AQ12 from Eq.(2.4), corresponding to at very low reference values (< 20 µg/m³), were substituted with the respective medium reference value, instead the only one negative value of estimated NO₂ concentration with AQ11 has been removed. Furthermore, the regression model of each LCAQMS contains only independent variables that are statistically significant.

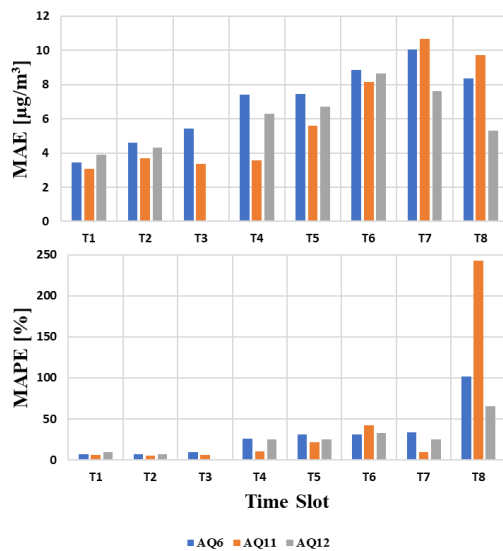


Figure 3. 3 MAE and MAPE of examined devices in the respective slot time. While a smooth trend is visible from T1 up to T4, in range T5 - T8 the increasing values of both the metrics point out a worst model performance.

A first indication about the possible presence of concept drift occurred during the mid-term campaign is already observable in the graphs of MAE and MAPE (Fig. 3.3 where you may see increasing metric values during the second month of the co-location), but above all from the boxplots of figure 3.4, where stands out clearly a double operation regime about the NO₂ concentration measured by the reference station: the slot times T1 - T4 with relatively high concentrations and a second part T5 - T8 characterized by lower concentrations, especially at the end of T8. The same conclusions are also valid for the measured ambient air temperature.

3.4.2 Validation of the proposed methodology

As first step, the TSKS-test is suggested to be performed by setting the first co-location week data as training and the following weeks data as test. The null hypothesis H_0 refers to both samples come from the same distribution at the 5% significance level and the result of the test is $h = 0$, otherwise if the hypothesis H_0 is rejected $h = 1$. Such analysis has been implemented in MATLAB® environment. The results of the TSKS-test on institutional validated nitrogen dioxide reference values confirm the presence of concept drift. Precisely, the computed h values show how the distributions between the training and tests phases are already different in the second week, while they are equal in the third one. From T4 onwards the distributions may be considered different. Evidence of this trend is depicted in figure 3.5 - 3.7, where the NO₂ reference concentration measured during each time slot has been fitted with a Lognormal distribution (De Vito *et al.*, 2020). At same time, the calculation procedure was carried out on temperature, highlighting how the model is also affected by covariate shift.

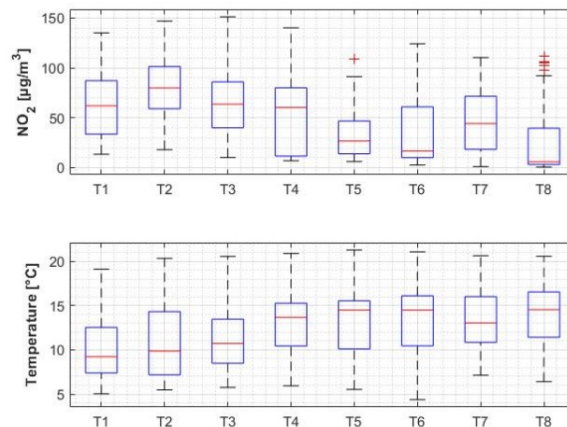


Figure 3. 4 Boxplot of NO₂ concentrations measured by reference station and temperature during the co-location period in each time slot.

Chapter 3

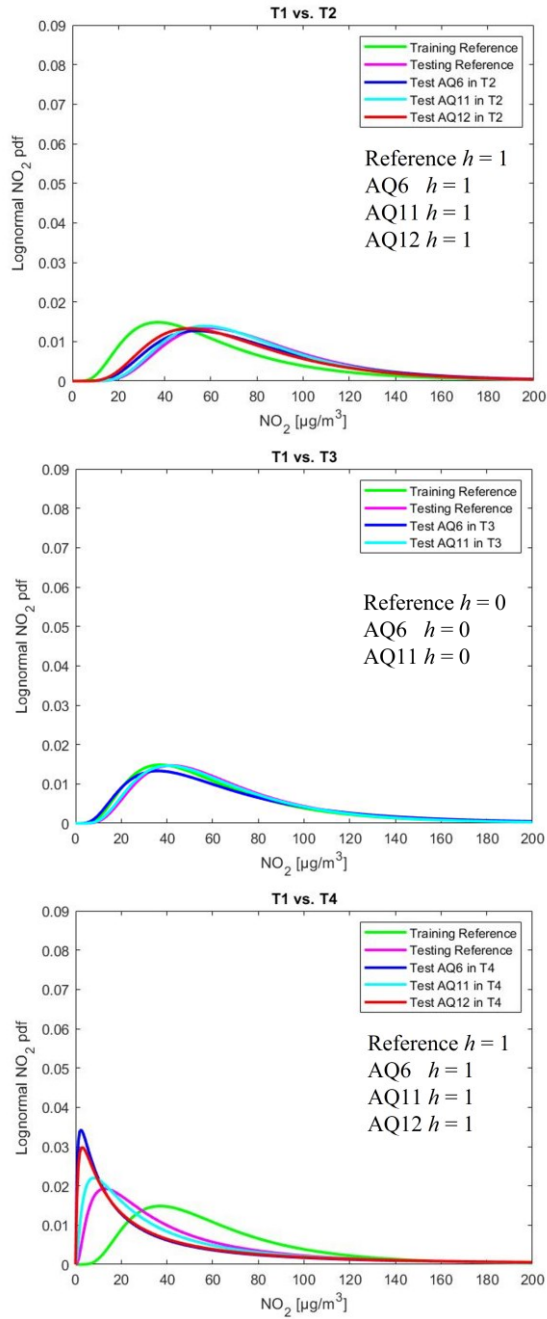


Figure 3. 5 Probability Density Function (Lognormal fitting) of the reference and estimated NO₂ concentrations during time slots T2-T4 compared with the training data set (time slot T1). In T3 there is no AQ12 data due to loss data transmission.

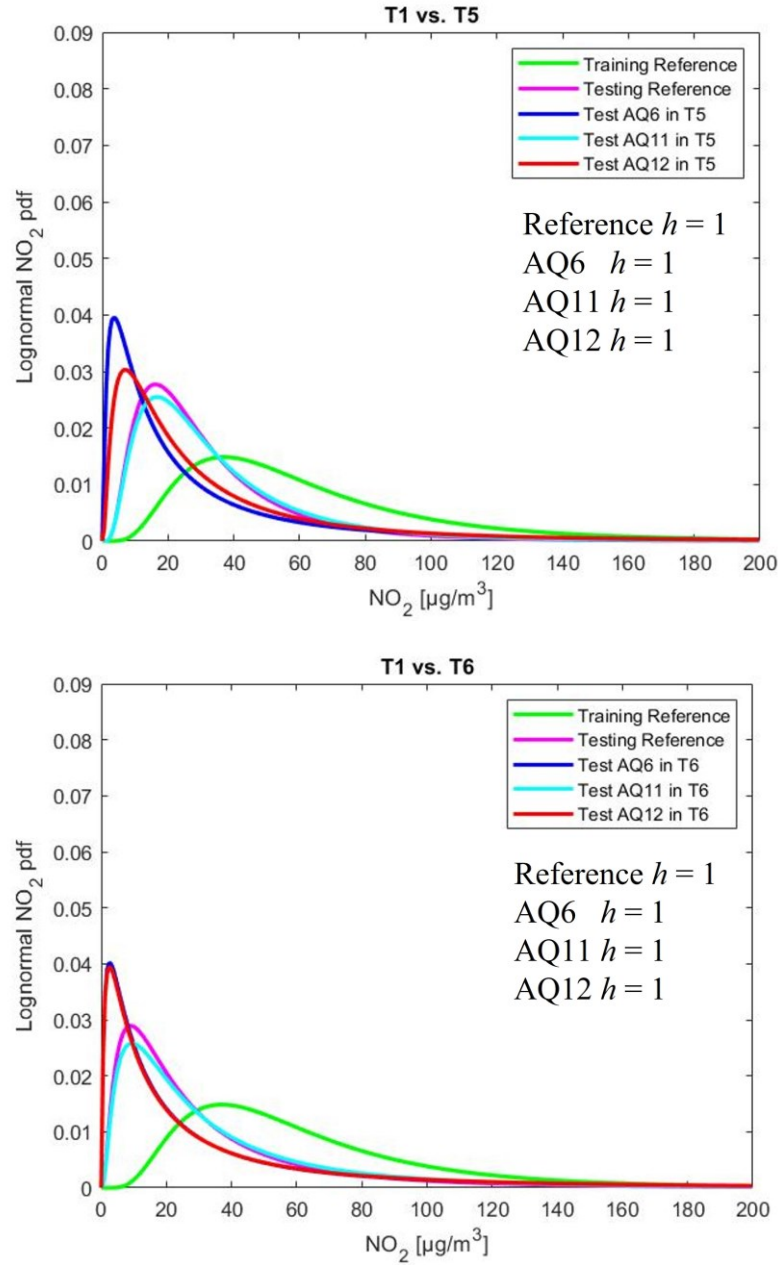


Figure 3. 6 Probability Density Function (Lognormal fitting) of the reference and estimated NO₂ concentrations during time slots T5-T6 compared with the training data set (time slot T1). The TSKS-test results highlight the auto-detection skill of concept drift.

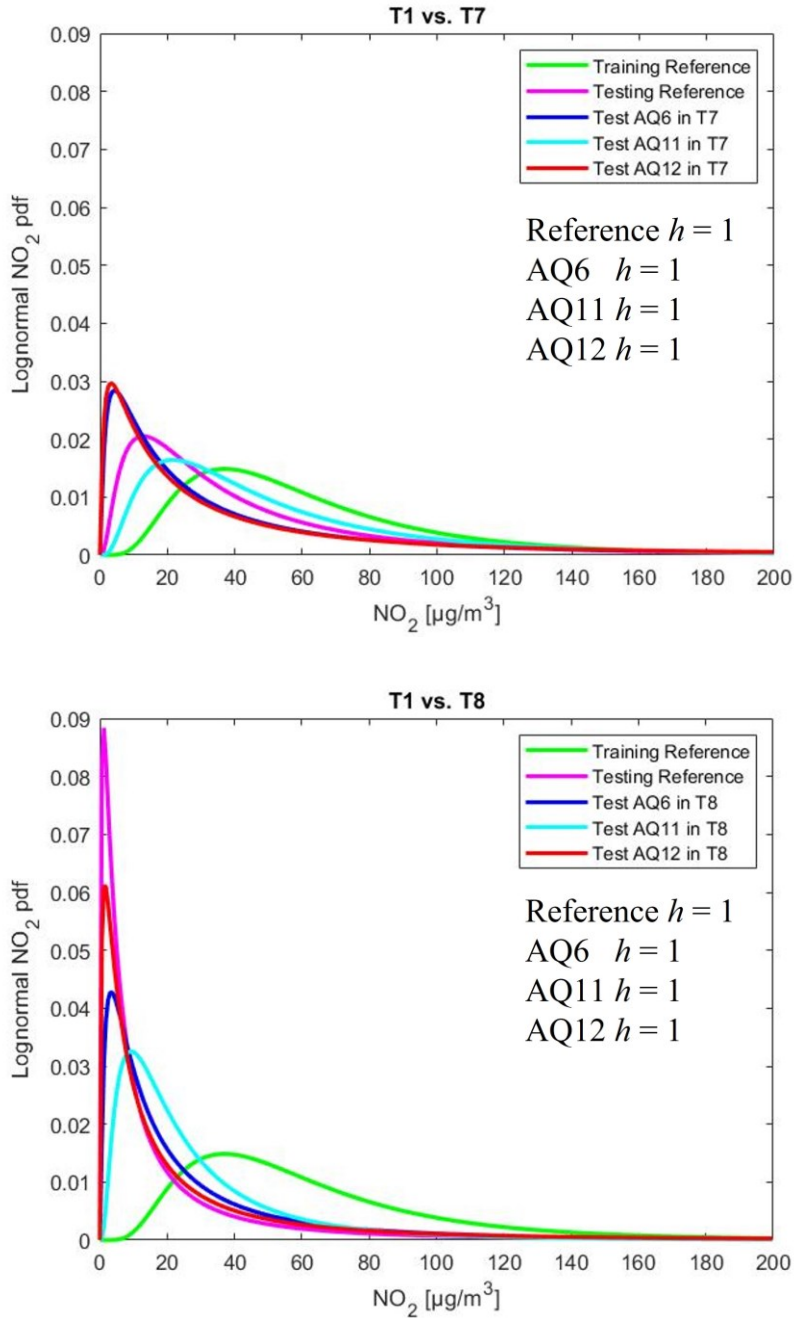


Figure 3. 7 Probability Density Function (Lognormal fitting) of the reference and estimated NO_2 concentrations during time slots T7-T8 compared with the training data set (time slot T1). The TSKS-test results highlight the auto-detection skill of concept drift.

As second step, the REU has been evaluated for all device under analysis based on equations (3.1) and (3.2). All used data for $u^2(xi)$ were provided and processed by the institutional regional authority according to the EN 14211:2012.

As you can see in the plot of figure 3.8, the estimated REU in T2 satisfies the European Directive requirements for the indicative measurements going down below the threshold of 25% imposed for the nitrogen dioxide. Such plot makes clear the poor ability of an MLR model in estimating very low concentration (Zimmerman *et al.*, 2018), although AQ11 exhibits very good performance till 33 $\mu\text{g}/\text{m}^3$.

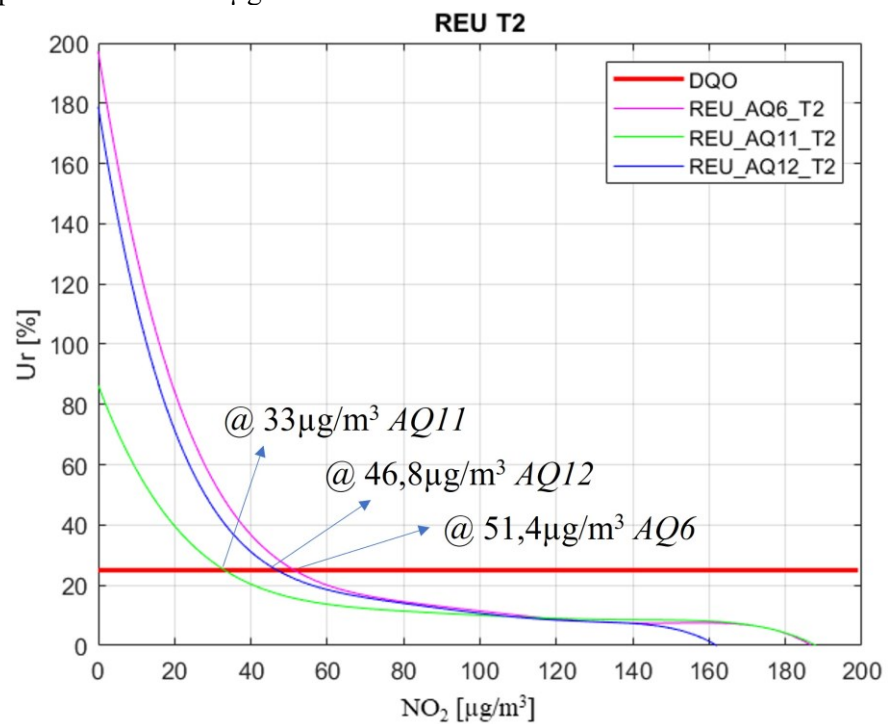


Figure 3. 8 Relative Expanded Uncertainties of all LCAQMSs computed in T2 and fitted with 8th degree polynomial.

During the field co-location the devices stay nearby a regulatory station, so ground truth data are available, hence computing the REU in T2 (or in the whole period when the reference data are available), could be an effective anomaly detection procedure for all LCAQMSs.

Moreover, the analysis of the REU plots could be a deployment cost-reduction technique if it is applied to choose a set of better devices which will be taken into account in development of a general calibration model (Malings *et al.*, 2019). Another noteworthy application is the node-to-node field calibration: the analysis of the REU plots could be the fundamental procedure for the “golden nodes selection” (Kizel *et al.*, 2018).

Chapter 3

However, for the purposes of the present work, the most important experimental observation is that, on despite of the change in distribution identified by the TSKS-test in T2, the REU computation claims the goodness of the model. It means that a “critical threshold” has not been probably exceeded.

Looking at figure 3.9, identical findings characterize the time slot T4, although an abrupt change in NO_2 concentrations is tangible (Fig. 3.2). Exactly this abrupt change marks the beginning of the concept drift in NO_2 distribution from T5 on ahead (see Fig. 3.10). The Relative Expanded Uncertainty during time slots T5-T8 is de facto above the red line of 25% threshold (apart few occasional points right on the line) indicating how the concept drift negatively affects the metrological performance of the sensor systems.

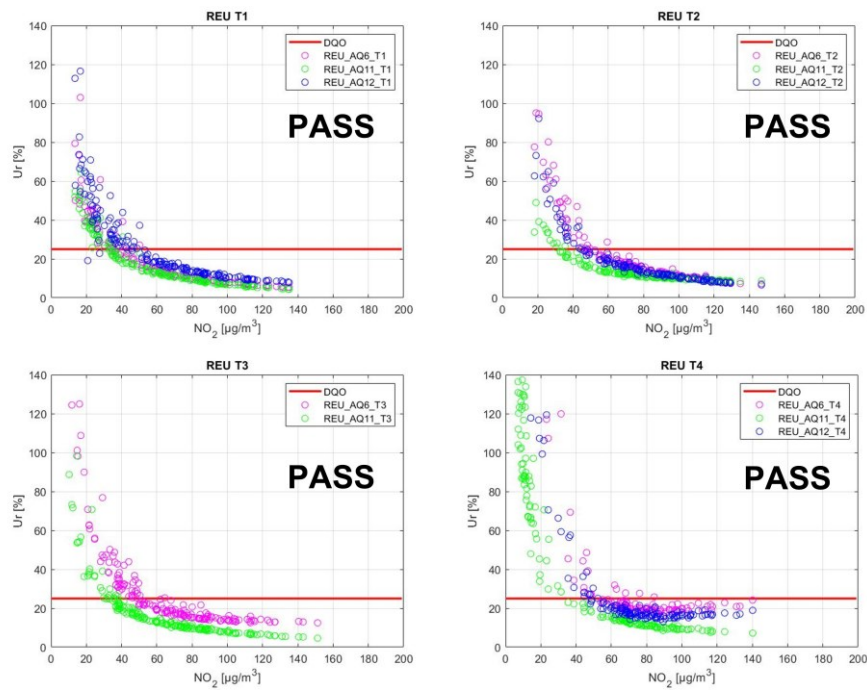


Figure 3. 9 Relative Expanded Uncertainties in time slots [T1 - T4].

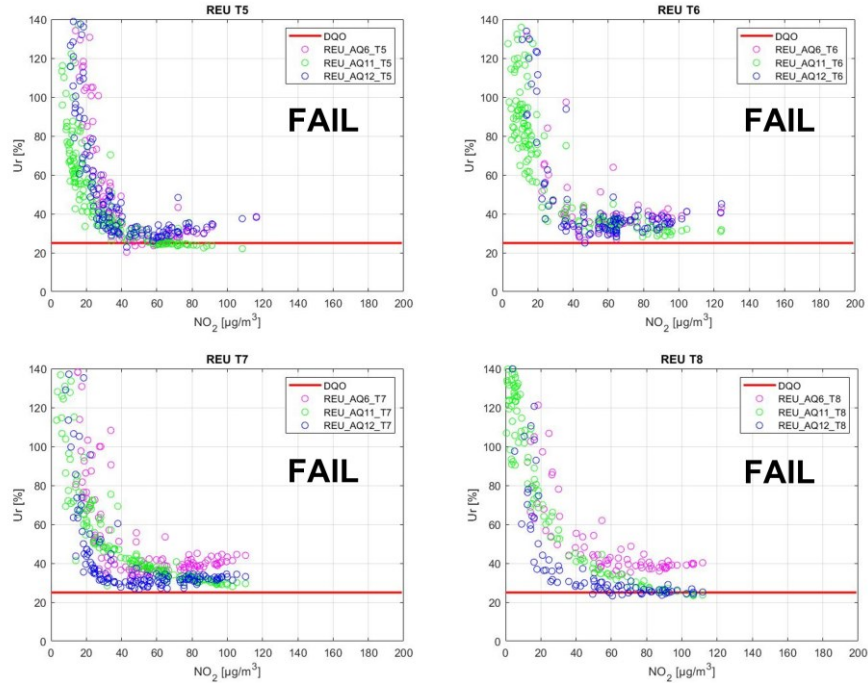


Figure 3. 10 *Relative Expanded Uncertainties in time slots [T5 - T8].*

Previous outcomes underline after all that the MLR model works fine at relatively high concentrations but not at low concentrations or when the concept drift takes place. Thus, as third methodological step, this peculiarity may be exploited for the self-detection of the concept drift. The TSKS-test has been also performed by considering the data distribution resulting from the measurements by the sensor systems (AQ6, AQ11, AQ12). The same h values have been achieved with respect to the case of the reference data distribution, thus proving the concept shift auto-detection skill (see Fig. 3.2 and Fig. 3.6).

It is significant to underline how the distributions of the three devices are close together in time slot T2 and T3: the conditions are like those seen during the training set slot and the model is working in stationary conditions. Then, the distributions move away from each other and from the reference one when the concept shift happen. Thus, the condition of T2 and T3 is certainly the most appropriate to set a threshold effective to ensure that the model works in optimal conditions and therefore with a REU that falls within the limits of the EU legislation.

For an optimal threshold evaluation, we will not consider T4 and T5 time slots because this is a boundary zone (drift zone), so that T5 time slot may be adopted to recalibrate. In such a way there will be a strong separation between the two operating stable areas: first and after the concept shift. Moreover, the recalibration will bring the REU to Pass.

An optimal threshold of 0.3 has been chosen by means of results obtained from the TSKS-test statistic D for each couple of training/test data set about the measured NO_2 concentration and the reference temperature, that fall in REU PASS area, as reported in figure 3.11. In detail, the TSKS-test statistic D represents the maximum difference between the empirical Cumulative Distribution Functions (CDFs) estimated from the distributions under test. In both the plots of figure 3.11, are reported two curves that represent the trend of the test statistic during the time slots. The blue curves refer to the distribution of the measured NO_2 concentration and reference temperature, when the calibration model resulting from T1 week is considered (thus, the corresponding test statistic D is null). The yellow curves are achieved when the adopted calibration model is trained according to the time slot T5 data.

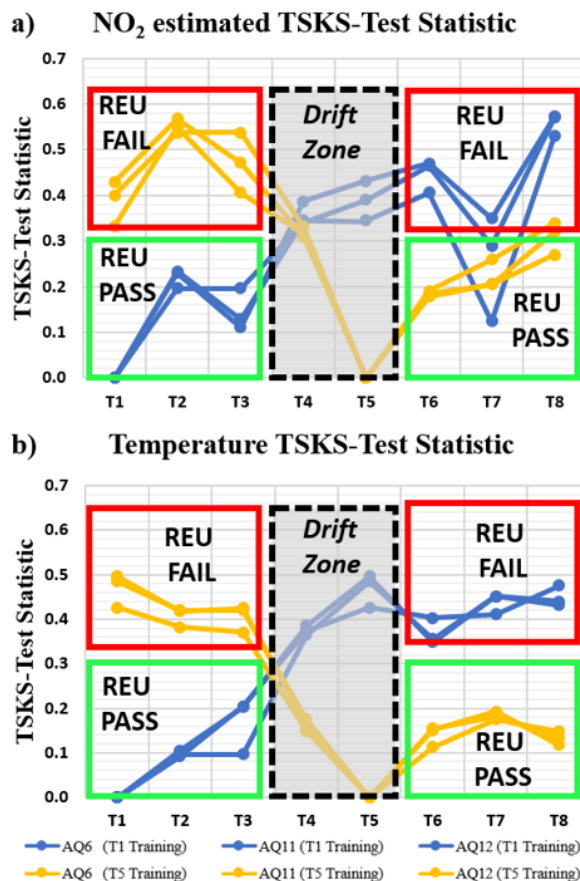


Figure 3. 11 TSKS-Test Statistic results of NO_2 MDL predictions and temperature of all devices when the MDL model is trained in T1 and T5. Concept Drift happen in T4 and go on in T5 that is used to recalibrate the model. This zone is called drift zone.

From the analysis of the plots, the threshold equal to 0.3 seems to be optimal because both the test statistic values computed for the NO₂ concentration and the reference temperature are not exceeded when the corresponding distributions are not influenced by the concept shift (in terms of REU).

The threshold 0.3 is just exceeded in T8 (figure 3.11a) when the model is recalibrated in T5 although this area is labelled with REU PASS. This situation is due to the second abrupt change present in T8 about the reference pollutant concentration with respect to the previous time slots (see Fig.3.2 and Fig. 3.6). Furthermore, the same models obtained recalibrating in time slot T5 are also applied in the intervals T1-T4 preceding the concept drift (see yellow plots) as a further verification of the effectiveness of the threshold value.

3.4.3 Auto-Detection of the Concept Shift

Based on the previous observations, a heuristic may be achieved to automate the continuous monitoring of the performance degradation about the calibration model. We will trigger a recalibration request when the TSKS-test statistic (D) of at least one input variable (estimated NO₂ concentration or temperature) is greater than 0.3. A picture of the diagnostic add-on scheme is reported in figure 3.12.

Embed this diagnostic block in the existing architecture of an LCAQMS could give to the final user better information represented by the awareness of the quality about the released data. If a recalibration request is sent, the user could decide to recalibrate if reliable data are accessible (as an example from the reference station nearby the considered LCAQMS).

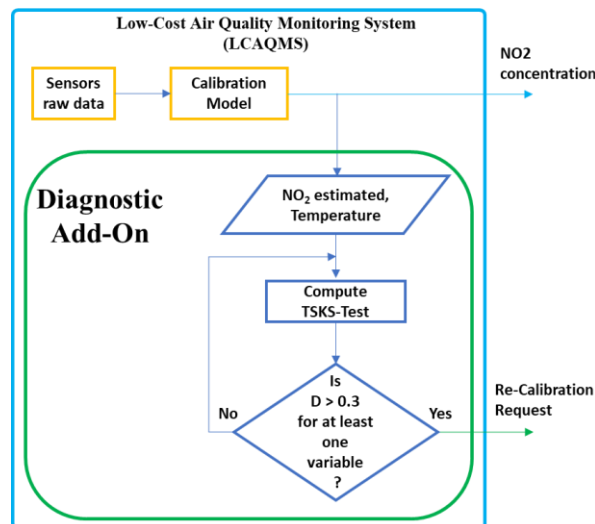


Figure 3. 12 Scheme of the proposed Diagnostic Add-On block.

3.5 Concluding remarks

This chapter has dealt with the proposal and experimental validation of an advanced statistical methodology aiming to improve the Quality Assessment and Quality Control (QA/QC) of the instrumentation actually proposed for the distributed monitoring of ambient air quality. It addresses two of main open points which still hinder the spread of Low-Cost Air Quality Monitoring Systems: the compliance with DQO of European Directive and the time identification of the necessary recalibration. As concerns the data quality, the influence of the concept drift on the Relative Expanded Uncertainty evaluation has been investigated proving how both the metric and the concept drift are the preeminent factors to be considered for the quality assessment of the calibration models based on Machine Learning. Moreover, applying Two-Sample Kolmogorov-Smirnov test has been proven an effective tool for the Quality Control of the instrumentation both in detecting concept drift events and in setting an appropriate threshold beyond which the recalibration request should be sent. The proposed methodology can be integrated into the current architecture of an LCAQMS either as a software block or as a module on the backend side of an Internet of Things platform to expanding the institutional air quality monitoring network. The experimental analysis carried out has shown how the concept drift can appear even a month after a field co-location. Because performing a second co-location after such a short time is not feasible in real scenario, interesting directions of the present research should be focused on the development of remote and cheaper calibration/recalibration techniques.

Chapter 4

Strategies for Concept Drift Mitigation in Low-Cost Air Quality Networks

4.1 Calibration update triggered via concept drift detector

In the previous chapter it was highlighted how the concept drift is one of the prime causes concerning losses in data quality, so a functional Add-On (based on a statistical approach) has been proposed. Such block is qualified in forwarding an alert and/or a re-calibration request to the user or to the low-cost air quality monitoring network administrator. Aim of this final chapter is pursue an effective and efficient model update with the goal to mitigate concept drift effects, bringing back data quality level at the permitted grade by the European directive. This objective is within reach whether reference data are accessible, otherwise challenging without.

The diagnostic Add-On block allows to trigger a calibration update procedure. Now the question is: *what are the effective calibration update strategies to adopt?*

4.1.2 Remote calibration

With the term *Remote Calibration*, the researchers indicate continuous re-calibration schemes relying on reference data from remote stations exploiting particular conditions and hypothesis.

Since 2005 different teams involved in air quality monitoring using low-cost sensors technologies, have suggested to exploit data coming from remote regulatory grade station, in low spatial variance conditions, to correct for drifting network of low-cost systems. It is reasonable to suppose a uniform distribution of the pollutant concentration in a restricted geographical area and therefore thanks to data from a nearby regulatory station, it is possible to correct the baseline of the gas sensor response (Tsujita *et al.*, 2015).

Chapter 4

Hierarchical networks including "golden" reference stations (proxies) on top have also been proposed as a solution to the problem of continuous calibration of LCAQMSs. Usually, regulatory instrumentations are on top in hierarchy, while also some well-calibrated low-cost nodes are entered in the intermediate positions. In such approach the proxy measurements are used to adjust the calibration by matching mean and standard deviations between the measures of the proxy itself and low-cost sensor data over a fixed time window (Miskell *et al.*, 2018). Based on this methodology the Aeroqual company (www.aeroqual.com) has developed the virtual calibration service for its own products.

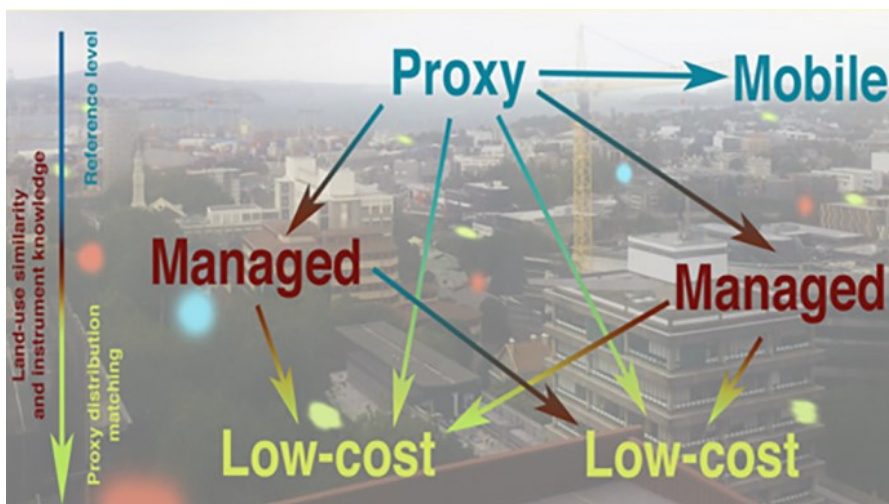


Figure 4. 1 Hierarchical network for continuous calibration. (Picture from: Miskell *et al.*, 2018).

Continuous calibration strategies exploiting remote data are promising, in fact hourly average data pollution are published by regulatory institutions for the convenience of the citizen and of other users. This means that such data are reachable for everyone through API REST services. Unlike MOMA, this is the name of the virtual calibration service distributed by Aeroqual, which updates the calibration at fixed time intervals, we take the advantage from this opportunity to explore a new road: the calibration update triggered via a concept drift detector.

This issue will be addressed in three steps: *i*) understand which data use for model update *ii*) explore and validation of the general calibration model and importance weighting calibration model (base learners or weak learners) in the presence of concept drift *iii*) attempting to improve the pollution estimations data quality with stacking ensemble technique in air quality networks scenario.

4.1.3 Reference data selection for calibration update in presence of Concept Drift

The data of the closest regional station could be used as label to update the calibration model after a concept drift detection. In the case of our interest, however, it is essential to know which data to select so that the new concept is learned in the fitting process. Taking advantage from the winter 2020 collocation dataset, characterized by the presence of the concept drift, we tackle the problem of concept drift handling trying to understand which data contain the useful information to update the calibration model incorporating the “*new concept*”. In figure 4.2 it is possible to see a clear graphic representation of the concept drift existing in the dataset under analysis. The green circles represent the “*concept*” that characterizes the target variable and the temperature during the collocation period and therefore during the training process in T1, while the red circles are representative of the “*new concept*” that characterizes the test set, that is the interval from T5 time slot onwards.

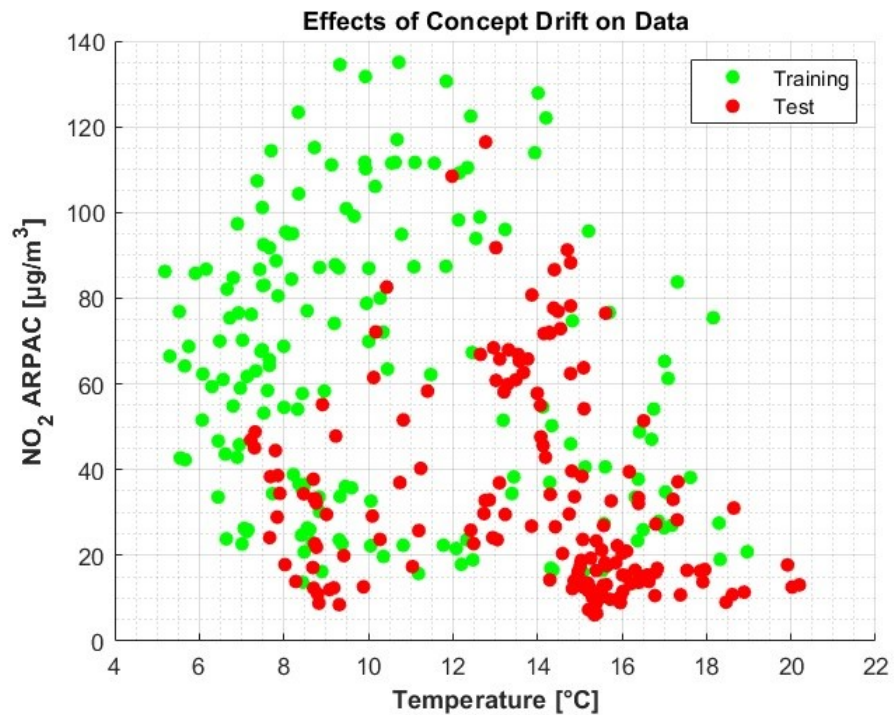


Figure 4. 2 Evidence of concept drift highlighted on the co-location samples of the target and input variables.

Three possibility are investigated: the data preceding the concept drift alert (called “Last”), the data subsequent the concept drift alert (“Next”) or part of

both (“Mixed”) (Baier *et al.*, 2018). These circumstances are schematized in the figure 4.3, in which the slot time T3 was identified as “Last”, T4 represents the “Mixed” data but at the same time it contains the time instant t_0 when the alert signal is released by the concept drift detector triggering the calibration update request, and finally T5 encloses “Next” data. As emerged from the analysis carried out in preceding chapter, the quality of the model's predictions deteriorates from the T5 interval onwards, therefore our attention will focus on trying to restore a compliant data quality in these time intervals. However, it was decided to exclude T8 slot time since, as visible in figure 3.2, it is subject to a further abrupt drift and the used dataset is not long enough to enable us to analyse even this situation. The following tests are fulfilled by updating the model using on hand reference data (in the proper time windows) and then evaluating the performance of the new model in T5 - T6 -T7.

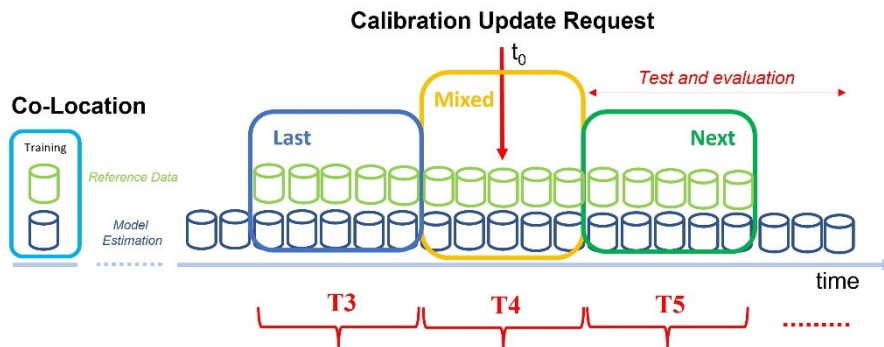


Figure 4. 3 Data selection for calibration update, when a re-calibration request arrives from the concept drift detector.

Table 4.1 collects the results of this analysis (remember that AQ12 dataset has missing records in T3). Looking at the values highlighted in bold representing the best performances, it easy arrive at the conclusion that the “Next” approach is to be preferred over the others. This was foreseeable since it is after the alert that the new operating conditions fully manifest themselves. However, in order to learn the new operating domain, it is necessary to wait for the time necessary to acquire a certain number of samples to train the new model and this means continuing in the meantime to invalidate the data released by the node (recall that the REU is PASS up to T4). The drawback due to delay time to get reference data before to have a running updated model must be remarked. It would be possible to wait for the reference data of T5 and then update the model surely embedding the new concept, thus allowing good performances in T6 and T7, but in this way we would have lost the continuity of the data quality in T5 which was labelled as FAIL. The goal is achieving the continuity in data quality wishing the REU in T5 – T6 – T7 be equal PASS. So, it would be helpful that the “Last” and/or “Mixed” approaches present acceptable value of REU.

Table 4.1 MAE and MAPE results obtained after the data selection for model calibration update.

L	AQ6					
	Last		Mixed		Next	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
T3	4.29	7.99	-	-	-	-
T4	5.37	14.87	4.56	13.93	-	-
T5	8.14	32.35	7.37	32.33	2.94	12.50
T6	6.93	32.44	6.07	29.11	4.56	20.06
T7	8.07	27.74	6.68	32.94	4.45	29.02

L	AQ11					
	Last		Mixed		Next	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
T3	3.27	6.96	-	-	-	-
T4	4.46	20.33	3.40	10.37	-	-
T5	5.08	19.97	6.63	31.77	2.79	11.45
T6	7.72	36.95	7.38	47.42	3.74	23.85
T7	10.70	41.42	9.05	47.00	3.49	18.60

L	AQ12					
	Last		Mixed		Next	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
T3	-	-	-	-	-	-
T4	-	-	4.43	12.86	-	-
T5	-	-	5.31	23.24	3.01	13.44
T6	-	-	5.43	26.46	3.95	15.71
T7	-	-	5.58	28.05	3.84	27.47

Remember that in the previous chapter it was verified as MAE and MAPE are metrics not adequate in ensuring the compliance with the European Directive if compared versus the REU. For this reason, the REU has been calculated both in the “Last” and “Mixed” scenario for all devices under test. From here on, refer to the appendices for all the graphs of the REUs of different scenarios.

Let's start with AQ12 device where only the Mixed approach can be considered. Well, observing the graphs of the REU from T5 to T7 it can be seen that the REU drops below 25% from 55 $\mu\text{g}/\text{m}^3$ onwards in T5 and T6 (figure A.1.1 and A.1.2), while from 51 $\mu\text{g}/\text{m}^3$ in T7 (figure A.1.3). Surely such a situation is to be preferred since it allows the continuity of data quality

extending the PASS of the REU for 3 weeks more. Regarding AQ11, the mixed approach works similarly to AQ12 but the inherent variability of the device negatively affects model performance, meaning it fails to deal the effects of concept drift. In fact, it can be seen that the intersections with the DQO thresholds line at 25% are higher, respectively $60 \mu\text{g}/\text{m}^3$ in T5 (figure A.1.4) and $80 \mu\text{g}/\text{m}^3$ in T6 and T7 (figure A.1.5 and A.1.6). AQ11 in “last” approach instead, works only in T5 (figure A.1.7), while it does not reach the desired levels in the T6 and T7 (figure A.1.8 and A.1.9). This means that the intrinsic variability of AQ11 and the operating conditions in T3 used for the fitting of the new model are too different from those present after the concept drift. The AQ6 device in mixed approach does not work in T5, while in T6 and T7 the goal is reached at 65 and $70 \mu\text{g}/\text{m}^3$ respectively (figures A.1.10 – A.1.12). The last scenario for AQ6 essentially does not work except in T6 but at high pollution level ($80 \mu\text{g}/\text{m}^3$ see figures A.1.13 – A.1.15).

In order to emphasize the ability in the mitigation of the concept drift effects (meaning accurate and precise NO_2 estimations) the tables 4.2 and 4.3 have been reported. Both tables show, for each slot time and for each device, the assigned REU labels (PASS and/or FAIL) as well as the value corresponding to the intersection with the DQO threshold line at 25%. Obviously, the missing values are related to the fact that the REU plot fails to reach the minimum threshold of 25%.

Table 4.2 Summary of REU results for the “Mixed” scenario.

<i>Mixed</i>						
T5		T6		T7		
REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	
AQ6	FAIL	-	PASS	65	PASS	70
AQ11	PASS	60	PASS	80	PASS	80
AQ12	PASS	55	PASS	55	PASS	51

Table 4.3 Summary of REU results for the “Last” scenario.

<i>Last</i>						
T5		T6		T7		
REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	
AQ6	FAIL	-	PASS	80	FAIL	-
AQ11	PASS	55	FAIL	-	FAIL	-

From the analysis carried out, it can be concluded that all in all, if there is the possibility of drawing on the data of a regulatory reference station, it is possible to update the calibration model after the trigger signal using the “Mixed” approach, i.e., selecting part of the reference data before the concept drift alert and the remainder after that.

4.2 Calibration update without reference data

In this part of the research project a very ambitious task is tried to tackle that is the updating of the calibration model without drawing on the reference data. Nevertheless, bringing back the data quality at the regulatory level remains the mandatory goal, implying the mitigation of the effects concern to the concept drift. The key idea is to exploit as much as possible the information content of the co-location data. Making use of co-location dataset two new calibration models will be created and assessed in order to understand ones is suitable for the purpose. The following techniques will be explored: General (or also said Global) calibration and Importance weighting.

4.2.1 General calibration model

This methodology, already described in paragraph 2.8, has been introduced in recent years as an attempt to reduce the calibration costs. It consists in identifying and applying a general calibration model to all the nodes involved in the network, thus avoiding the need for additional ad-hoc calibrations. As we know, the output of each electrochemical gas sensor is the voltage (mV) measured at the working and auxiliary electrodes that is representative of the measured gas concentration. Well, if there are n sensors placed in co-location, then the median between these n values is evaluated. The same goes for temperature and humidity values. The set of the medians of all single quantities involved in model creation constitutes the training set which the generalized model is trained on (Malings *et al.*, 2019). A similar approach is capable of incorporating the inherent variability of each individual sensors into a single model. Follonosa *et al.*, 2016 investigated the use of a generalized model to contrast the effects of concept drift on MOXs. Taking inspiration from these works, the same approach will be applied to the electrochemical gas sensors (NO₂) in our case study.

Table 4.4 Metrics performance of the general calibration model.

L	Global Calibration Model					
	AQ6		AQ11		AQ12	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
T5	4.64	20.12	25.25	148	6.38	23.40
T6	5.82	25.43	30.96	247	7.04	25.63
T7	7.27	35.28	26.91	198	7.32	28.60

The global calibration model works well for the AQ6 and AQ12 devices (results of reported MAE and MAPE metrics in table 4.4) while it is unusable for AQ11. The AQ11 intrinsic node variability makes this instrument too

different from the others, therefore, a global model built in this way it is unable to embed it. Future research projects could investigate this aspect and evaluate whether it is better to build more global models but among similar devices by means of clustering (Smith *et al.*, 2019). As previously explained, it is essential to evaluate the REU to understand if the global model is able to mitigate positively the effects of concept drift. Obviously, in this case we evaluate the REU only for AQ6 and AQ12 nodes. Specifically, the comparison will be made between the REU obtained with the global model in the same time slot with that one's obtained with ad-hoc model calibrated in T1 and therefore subject to concept drift as detailed in chapter 3. Applying the global model to AQ6, the REU plot drops below 25% at 45 $\mu\text{g}/\text{m}^3$, suggesting that global calibration model is efficient into the mitigation of the concept drift consequences (Figure A.2.1). This value shifts towards at higher levels of concentration in T6 and T7 at 63 $\mu\text{g}/\text{m}^3$ and 80 $\mu\text{g}/\text{m}^3$ respectively (Figures A.2.2 – A.2.3). Although the REU does not fall within a considerable range of values below 25% in T6 and T7, it must be said that such situation presents better results than the ad-hoc model. A first applicative scenario, therefore, would be to use the global model for AQ6 from T5 onwards instead continuing with the ad-hoc one. For AQ12 instead the global model matches the ad-hoc model performance (Figures A.2.4 – A.2.6). The same result has been obtained for PMs in terms of MAE (De Vito et al, 2022).

4.2.2 Importance weighting calibration model

The idea of this approach is to “weigh” the samples of the test set in order to “match” the distribution used during the training. Once the weights are obtained, these will be applied in the training process obtaining a new calibration model (Sugiyama et al., 2007).

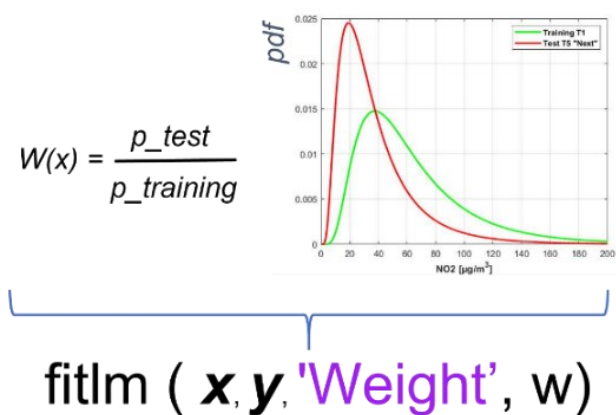


Figure 4. 4 How the weights are used in the fitting process.

The importance of a sample (the “weight”) is calculated as the ratio between the probability density functions of test and training set (figure 4.4). If such ratio $w(x)$ is equals to one means that the sample has the same importance both in test set and in the training set, while if $w(x) > 1$ the considered sample is much more important in describing the test set rather than the training set. In the present analysis the application of the importance weighting calibration model has been limited to the target variable only. Moreover, taking advantages of autodetection feature previously established, the practicability of obtaining the weight by the ad-hoc model prediction in T4 characterized by REU PASS has been explored.

As can be seen from table 4.5 re-weighting the target variable in T4, a considerable improvement is obtained in the performance of MAE and MAPE for the AQ11 device.

Table 4.5 Metrics performance of the importance weighting calibration model.

L	Importance Weighting Calibration Model					
	AQ6		AQ11		AQ12	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
T5	7.31	25.61	4.38	15.21	7.91	30.79
T6	8.60	31.05	6.68	30.70	8.50	30.80
T7	10.93	34.34	8.94	32.04	9.43	39.40

In fact, looking at the REU plot of the device AQ11 in T5, the intersection with the 25% value has been observed at $40 \mu\text{g}/\text{m}^3$ (figure A.3.4), while in T6 and T7 it is only possible to brush the line of the DQOs, although the importance weighted calibration model performs better than the ad-hoc one (figures A.3.5 – A.3.6).

As far as AQ6 and AQ12, although the metric outcomes are acceptable, the REU plot equals those of the ad-hoc model in some slot times, while in some other cases are even worse (figures A.3.1 – A.3.3; figures A.3.7 – A.3.9).

Summarizing the application of the importance weighted calibration model brings the AQ11 device in T5 back to the allowed REU values.

4.2.3 Extending the calibration validity

An important consideration can be made looking at the results reported in the tables 4.6 and 4.7 that summarize the REU achievements in the application of the general calibration model and importance weighting calibration model. Thus, both models if applied on the AQ6 and AQ11 devices are able to deal with the concept drift in T5, therefore switching from the ad-hoc model to the respective models that have a REU equal to PASS in T5 would allow extending the validity of the calibration for an additional slot time (1 week)

ensuring compliance of data quality moreover without requiring access to reference data. Furthermore, the general calibration model applied on AQ6 exhibits a REU PASS for all the remaining time slots up to T7, extending the calibration for a total of three weeks.

Table 4.6 Summary of REU results for the general calibration model.

<i>General Calibration Model</i>						
	T5		T6		T7	
	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]
AQ6	PASS	45	PASS	63	PASS	80
AQ11	FAIL	-	FAIL	-	FAIL	-
AQ12	FAIL	-	FAIL	-	FAIL	-

Table 4.7 Summary of REU results for the importance weighting calibration model.

<i>Importance Weighting Calibration Model</i>						
	T5		T6		T7	
	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]
AQ6	FAIL	-	FAIL	-	FAIL	-
AQ11	PASS	40	FAIL	-	FAIL	-
AQ12	FAIL	-	FAIL	-	FAIL	-

4.3 Handling concept drift with Stacking Ensemble model

In the earlier analysed “Mixed” scenario, it was seen that by recalibrating the nodes with reference data grasped from the network in T4, the effects of the concept drift are really attenuated. In some cases, however, this beneficial effect is manifested only at high concentration values (i.e., $80 \mu\text{g}/\text{m}^3$). In some other cases this positive effect is not revealed.

Recently, the stacking ensemble technique has been successfully also applied in domain adaptation under concept drift, due to its ability in reducing the deviation and variance in neural networks producing robust predictions (Yuan *et al.*, 2022), as well as also in air quality sensor networks (Bagkis *et al.*, 2022).

Stacking ensemble consists in combining the outputs of several models produced by different algorithms (generally called as basic learner or also weak models) in order to increase the total accuracy and increase the generalization of the basic learners. The estimations of the basic learners are combined in a second level, called meta-learner which, in the case of regression problems, can be a simple linear regression. In our case study, instead of implementing additional algorithms such as base learner, the estimations from the general calibration model and the importance weighting

calibration model have been exploited. It must be underline that both used base learners have the peculiarity of only make the best use of the co-location data. Figure 4.5 describes the detail of the architectural scheme. The reference data (labels) for the meta-learner training phase will be requested from the network in the T4 interval similarly already done for the "Mixed" approach. Likewise, to the cases treated above, the table with the MAE and MAPE performance as well as the comparison plots of the REUs (figures A.4.1 – A.4.9) in the individual time slots are listed below.

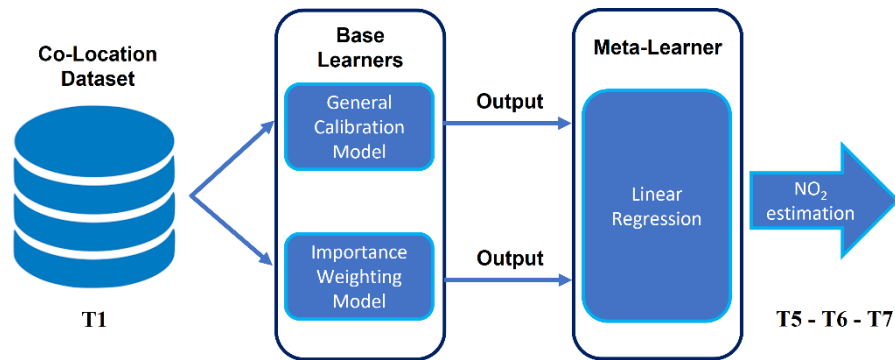


Figure 4.5 The proposed stacking ensemble architecture.

Table 4.8 Metrics performance of the Stacking Ensemble model.

<i>Stacking Ensemble Model</i>						
	AQ6		AQ11		AQ12	
	MAE	MAPE	MAE	MAPE	MAE	MAPE
T5	4.57	19.19	4.69	20.27	4.89	20.92
T6	5.58	24.61	6.53	36.91	4.88	21.02
T7	6.96	31.15	8.72	41.98	4.75	25.83

Table 4.9 Summary of REU results for the stacking ensemble calibration model.

<i>Stacking Ensemble Calibration Model</i>						
	T5		T6		T7	
	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]	REU	Value [$\mu\text{g}/\text{m}^3$]
AQ6	PASS	44	PASS	56	PASS	65
AQ11	PASS	44	PASS	70	PASS	76
AQ12	PASS	51	PASS	54	PASS	45

The inference to be drawn comparing the outcomes of table 4.9 with those of the table 4.2 is that the application of the stacking ensemble technique

Chapter 4

produces a lowering of the value in which the REU reaches 25%, making the mitigation of the concept drift effects in practice more robust in all analysed cases. May not be excluded that the use of further even more complex base-learners algorithms, could contribute to any additional reducing.

The proposed solution, although not yet an optimal solution, anyhow constitutes a good compromise, since it guarantees the continuity of the data quality even after the detection of the concept drift. The alternative would be to learn the new concept in T5, but there is a negative side due to delay in waiting for the reference data in T5. In such situation, the implication of losing the continuity of data quality right in all the T5 slot time it's obvious.

Operating within an air quality network the implementation of the proposed stacking ensemble approach requires only the reference data in T4. It is good thing underline a further advantage of the introduction of the concept drift detector add-on into the network, i.e., the alert could be used by the network administrator to search and/or request the reference data needed for node recalibration. Reference data that in a network could come from nearby regulatory stations or from nearby "Golden" node

Conclusion

In this research activity a further step towards achieving of the data quality objectives in the measurements of environmental pollutants through machine learning calibrated low-cost gas sensors has been attempted, with the hope of accelerating the spread of this technology in smart city applications. The study carried out has taken into consideration the overall data chain value, from the generation of the data up to its sharing with users, citizens and the administrator of the air quality monitoring network.

The first part of the experimental activity was dedicated to the data collection through co-location campaigns in order to create the calibration models of the sensors. Then an assessment of both multilinear regression model and neural networks model has been performed. The achieved outcomes have shown that a multilinear regression is able to provide reliable data quality in fixed application but also in mobile applications, managing to detect pollution hotspots and offering high resolution pollutants maps.

The quality of the data of the NO₂ estimation was the goal of this thesis, therefore more attention was given to a yet neglected phenomenon in such context, namely concept drift. Since the concept drift negatively affects the estimations provided by the calibration model trained and obtained from the initial co-location, an automatic procedure based on the Kolmogorov–Smirnov test has been proposed to identify when a model update is necessary. The usefulness of inserting a concept drift detection block, either embedded on the node or on the backend, has been proposed for the first time in low-cost air quality monitoring background. The proposed technique is obviously a preliminary approach since countless other techniques are present in the literature to detect the concept drift and therefore new scenarios are already opening up for future research activities such as, the evaluation of other techniques in order to find an optimal configuration as well as moving on to the online streaming rather than batch data as here has been discussed.

Once the problem of the deterioration in the measurements quality following the occurrence of a concept drift has been raised, the subsequent problem is how to update the calibration model and get back compliance of

Conclusion

DQOs. This topic is known in the literature among the community of researchers involved in machine learning as "concept drift adaptation", a research area still in an early stage. In fact, the crucial challenges for concept drift adaptation are related to the useful number of samples for updating and/or retraining the model, whether reference data are available or not and how to deal with the different types of concept drift. Well, an attempt has been achieved to address this issue in the context of low-cost air quality monitoring by exclusively exploiting the data available in co-location, exploring and combining different techniques.

A first step was understood which data were useful to update the model when a concept drift detector is used. The analysis carried out highlighted how taking the reference data straddling the alert ("Mixed" scenario) leads to good results in terms of REU (< 25%) but in some cases at too high concentration values (about 80 $\mu\text{g}/\text{m}^3$). The general calibration model and the importance weighting calibration model show promising results in the mitigation of the concept drift effects by allowing an extension of the calibration soundness. In this way, switching off the ad-hoc model and switching on the general calibration model for instance, the node is able to release reliable data.

Another advantage of the introduction of a concept drift detector block is also the possibility to alert the administrator of the air quality monitoring network for the model update request of a node and activate nearby nodes or looking for nearby regulatory station for reference data collection.

A further improvement in reduction of the concentration value at which the REU is equal to 25% has been obtained by inserting the general calibration model and the importance weighting calibration model in the first layer of a stacking ensemble and relying on the reference data in mixed approach.

Even though the proposed methodology offers encouraging results, future research activities could be undertaken to test other models in the first layer of the stacking ensemble in order to increase performance, but also the assessment and validation of the methodology using longer datasets with a greater number of devices. Another major research area which needs to be investigated is linked to the issue of the missing data due to faults of the sensors or transmission losses such it affects the data quality as well as the creation of high-resolution air quality maps (geostatistics). In that case both distributed network approach and the evaluation of multiple machine learning models should be explored (Lay-Ekuakille and Trotta, 2011).

Finally, an important aspect to remark is the feasibility of the implementation on the backend side of the platform of an air quality network of the proposed approach. This might deliver a continuous calibration service like the one offered by Aeroqual. Really in recent years, the issue of evaluate and monitor machine learning models from validation to production has come to the fore, so that numerous start-ups have implemented their own platform and have begun to offer services in many application areas. The low-cost air quality monitoring network could be one of such applications.

Reference

Agresta *et al.*, 2017 Annalisa Agresta, Saverio De Vito, Fabrizio Formisano, Ettore Massera, Elena Esposito, Maria Salvato, Grazia Fattoruso, Girolamo Di Francia,
“Cooperative Air Quality Sensing with Crowdfunded Mobile Chemical Multisensor Devices”
Proceedings Eurosensors 2017 Conference, 1(4), 602;
<https://doi.org/10.3390/proceedings1040602>

Aimé Lay-Ekuakille and Amerigo Trotta, 2011
“Predicting VOC Concentration Measurements: Cognitive Approach for Sensor Networks”
IEEE Sensors Journal, vol. 11, no. 11, pp. 3023-3030, Nov. 2011;
<https://doi.org/10.1109/JSEN.2011.2143705>

Alphasense, 2022a
Application notes AAN 104
Accessed: Dec. 06, 2022. [Online].
<https://www.alphasense.com/downloads/application-notes/>

Alphasense, 2022b
Application notes AAN 105
Accessed: Dec. 06, 2022. [Online].
<https://www.alphasense.com/downloads/application-notes/>

Alphasense, 2022c
AFE Sensor Board (Datasheet)
Accessed: Dec. 06, 2022. [Online].
<https://www.alphasense.com/downloads/>

Alphasense, 2022d
Application notes AAN 110
Accessed: Dec. 06, 2022. [Online].
<https://www.alphasense.com/downloads/application-notes/>

Reference

- Bagkis *et al.*, 2022 Evangelos Bagkis, Theodosios Kassandra and Kostas Karatzas,
“Learning Calibration Functions on the Fly: Hybrid Batch Online Stacking Ensembles for the Calibration of Low-Cost Air Quality Sensor Networks in the Presence of Concept Drift.”
Atmosphere 2022, 13, 416.
<https://doi.org/10.3390/atmos13030416>
- Baron and Saffell, 2017 Ronan Baron and John Saffell,
“Amperometric Gas Sensors as a Low Cost Emerging Technology Platform for Air Quality Monitoring Applications: A Review”,
ACS Sensors 2017 2 (11), 1553-1566
<https://doi.org/10.1021/acssensors.7b00620>
- Bishop, 2006
“Pattern Recognition and Machine Learning”,
Springer, New York, USA, 2006
ISBN:978-0-387-31073-2
- Brindha *et al.*, 2023, Majid Taie Semiromi, Lamine Boumaiza and Subham Mukherjee,
“Comparing Deterministic and Stochastic Methods in Geospatial Analysis of Groundwater Fluoride Concentration”,
Water **2023**, 15, 1707
<https://doi.org/10.3390/w15091707>
- Carotta *et al.*, 2017, Maria Cristina Carotta, Giuliano Martinelli, Luigi Crema, Cesare Malagù, Marco Merli, Giovanna Ghiotti, Enrico Traversa
“Nanostructured thick-film gas sensors for atmospheric pollutant monitoring: quantitative analysis on field tests”
Sensors Actuators B Chem., 76 (2001), pp. 336-343
[https://doi.org/10.1016/S0925-4005\(01\)00620-7](https://doi.org/10.1016/S0925-4005(01)00620-7)
- Castell *et al.*, 2017, Nuria Castell, Franck R. Dauge, Philipp Schneider, Matthias Vogt, Uri Lerner, Barak Fishbain, David Broday, Alena Bartonova,
“Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?”
Environment International, Volume 99, 2017, Pages 293-302,
<https://doi.org/10.1016/j.envint.2016.12.007>
- CJEU, 2022 Court of Justice of the European Union (CJEU)
PETI-CM-738526_EN

- Accessed: Dec. 06, 2022. [Online].
Available: https://www.europarl.europa.eu/doceo/document/PETI-CM-738526_EN.pdf
- Concas *et al.*, 2021 Francesco Concas, Julien Mineraud, Eemil Lagerspetz, Samu Varjonen, Xiaoli Liu, Kai Puolamäki, Petteri Nurmi, Sasu Tarkoma
“Low-cost outdoor air quality monitoring and sensor calibration: A survey and critical analysis”
ACM Trans. Sensor Netw., vol. 17, no. 2, pp. 1–44, May 2021,
Available: <https://doi.org/10.1145/3446005>
- De Vito *et al.*, 2020 Saverio De Vito, Elena Esposito, Nuria Castell, Philipp Schneider, Alena Bartonova
“On the robustness of field calibration for smart air quality monitors”
Sens. Actuators B, Chem., vol. 310, May 2020, Art. no. 127869,
doi: <https://doi.org/10.1016/j.snb.2020.127869>
- De Vito *et al.*, 2021
“Crowdsensing IoT Architecture for Pervasive Air Quality and Exposome Monitoring: Design, Development, Calibration, and Long-Term Validation”,
Sensors 21, no. 15: 5219.
<https://doi.org/10.3390/s21155219>
- Ditzler *et al.*, 2015 Gregory Ditzler, Manuel Roveri, Cesare Alippi, Robi Polikar
“Learning in nonstationary environments: A survey,”
IEEE Comput. Intell. Mag., vol. 10, no. 4, pp. 12–25, Nov. 2015,
Available: <https://doi.org/10.1109/MCI.2015.2471196>
- Du, 2020
“IAM–Interpolation and Aggregation on the Move: Collaborative Crowdsensing for Spatio-temporal Phenomena”.
In Proceedings of the MobiQuitous 2020–EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, Virtual, Germany, 31 December 2020.
Available: <https://doi.org/10.1145/3448891.3448918>
- EC European Commission, 2008
Directive 2008/50/EC of the European Parliament and the Council of 21 May 2008 on Ambient Air Quality and Cleaner Air for Europe
Accessed: Nov. 10, 2022. [Online].
<https://eur-lex.europa.eu/eli/dir/2008/50/oj>

Reference

- EC European Commission, 2022a
Air Quality
Accessed: Nov. 10, 2022. [Online].
Available: https://environment.ec.europa.eu/topics/air/air-quality_en
- EC European Commission, 2022b
Proposal for a revision of the Ambient Air Quality Directives
Accessed: Dec. 06, 2022. [Online].
Available: https://environment.ec.europa.eu/all-publications_en
- EC WG, 2010
“Guide to the Demonstration of Equivalence of Ambient Air Monitoring Methods”,
Report by *EC Working Group on Guidance*
Accessed: Nov. 11, 2022. [Online].
https://environment.ec.europa.eu/topics/air/airquality/assessment_en
- EEA European Environment Agency, 1999
Technical Report n.12 “Criteria for EUROAIRNET The EEA Air Quality Monitoring and Information Network”,
Accessed: Nov. 10, 2022. [Online].
Available: <https://www.eea.europa.eu/publications/TEC12>
- Esposito *et al.*, 2016, E. Esposito, S. De Vito, M. Salvato, V. Bright, R.L. Jones, O. Popoola,
“Dynamic neural network architectures for on field stochastic calibration of indicative low cost air quality sensing systems”,
Sens. Actuators B Chem., 231 (2016), pp. 701-713
Available: <https://doi.org/10.1016/j.snb.2016.03.038>
- Esposito *et al.* 2020 E. Esposito, G. D’Elia, S. Ferlito, A. Del Giudice, G. Fattoruso, P. D’Auria, S. De Vito, G. Di Francia
Optimal Field Calibration of Multiple IoT Low Cost Air Quality Monitors: Setup and Results.
Computational Science and Its Applications – ICCSA 2020. Lecture Notes in Computer Science, vol 12253. Springer,
https://doi.org/10.1007/978-3-030-58814-4_57
- Gemaque *et al.*, 2019 Rosana Noronha Gemaque, Albert França Josuá Costa, Rafael Giusti, Eulanda Miranda dos Santos
“An overview of unsupervised drift detection methods”
WIREs Data Mining Knowl. Discovery, vol. 10, no. 6, p. 1381, Nov. 2020,
doi: <https://doi.org/10.1002/widm.1381>

- Hoffmann *et al.*, 2021
“WHO Air Quality Guidelines 2021–Aiming for Healthier Air for all: A Joint Statement by Medical, Public Health, Scientific Societies and Patient Representative Organisations”
Int J Public Health, 66:1604465.
<https://doi.org/10.3389/ijph.2021.1604465>
- Kizel *et al.*, 2018 Fadi Kizel, Yael Etzion, Rakefet Shafran-Nathan, Ilan Levy, Barak Fishbain, Alena Bartonova, David M. Broday
“Node-to-node field calibration of wireless distributed air pollution sensor network”
Environ. Pollut., vol. 233, pp. 900–909, Feb. 2018,
doi: <https://doi.org/10.1016/j.envpol.2017.09.042>
- Losing *et al.*, 2016 Viktor Losing, Barbara Hammer, Heiko Wersing
“KNN classifier with self-adjusting memory for heterogeneous concept drift”
in Proc. IEEE 16th Int. Conf. Data Mining (ICDM), Dec. 2016, 291–300,
doi: <https://doi.org/10.1109/ICDM.2016.0040>
- Lu *et al.*, 2019 Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, Guangquan Zhang
“Learning under concept drift: A review”
IEEE Trans. Knowl. Data Eng., vol. 31, no. 12, 2346–2363, Dec. 2019,
doi: <https://doi.org/10.1109/TKDE.2018.2876857>
- Malings *et al.*, 2019, Carl Malings, Rebecca Tanzer, Aliaksei Hauryliuk, Srinivasa P. N. Kumar, Naomi Zimmerman, Levent B. Kara, Albert A. Presto, R. Subramanian
“Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring”,
Atmos. Meas. Tech., 12, 903–920,
Available: <https://doi.org/10.5194/amt-12-903-2019>
- Massera *et al.*, 2020 E. Massera, B. Alfano, M. L. Miglietta, T. Polichetti, S. De Vito, F. Formisano, G. Di Francia, P. Delli Veneri
Rapid Parallel Calibration for Environmental Bulky Gas Sensor Systems.
Sensors and Microsystems. AISEM 2019. Lecture Notes in Electrical Engineering, vol 629. Springer,
https://doi.org/10.1007/978-3-030-37558-4_34

Reference

Masson-Delmotte *et al.*,

“Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change,”

IPCC, Geneva, Switzerland, Tech. Rep., 2021.

Accessed: Sep. 18, 2021. [Online].

Available: <https://www.ipcc.ch/report/ar6/wg1/#FullReport>

McCulloch and Pitts, 1943

“A Logical Calculus of Ideas Immanent in Nervous Activity”,

Bulletin of Mathematical Biophysics. 5 (4): 115–133.

<https://doi.org/10.1007/BF02478259>

Mijling *et al.*, 2018 Bas Mijling, Qijun Jiang, Dave de Jonge,

Stefano Bocconi,

Field calibration of electrochemical NO₂ sensors in a citizen science context,

Atmos. Meas. Tech., 11, 1297–1312, 2018

<https://doi.org/10.5194/amt-11-1297-2018>

Miquel-Ibarz *et al.*, 2022, Albert Miquel-Ibarz, Javier Burgués,

Santiago Marco

“Global calibration models for temperature-modulated metal oxide gas sensors: A strategy to reduce calibration costs”,

Sensors and Actuators B: Chemical, Volume 350, 2022,

Available: <https://doi.org/10.1016/j.snb.2021.130769>

Miskell *et al.*, 2018 Georgia Miskell, Jennifer A. Salmond, and David E.

Williams,

Solution to the Problem of Calibration of Low-Cost Air Quality Measurement Sensors in Networks,

ACS Sens. 2018, 3, 4, 832–843

<https://doi.org/10.1021/acssensors.8b00074>

Quinonero-Candela *et al.*, 2008, Joaquin Quiñonero-Candela, Masashi

Sugiyama, Anton Schwaighofer, Neil D. Lawrence

“Dataset Shift in Machine Learning”.

Cambridge, MA, USA: *MIT Press*, 2008,

doi: <https://doi.org/10.7551/mitpress/9780262170055.001.0001>

Raab *et al.*, 2020 Christoph Raab, Moritz Heusinger, Frank-Michael Schleif

“Reactive soft prototype computing for concept drift streams”

Neurocomputing, vol. 416, pp. 340–351, Nov. 2020,

doi: <https://doi.org/10.1016/j.neucom.2019.11.111>

- Rabanser *et al.*, 2019, Stephan Rabanser, Stephan Günnemann, Zachary C. Lipton
“Failing loudly: An empirical study of methods for detecting dataset shift”
in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 1394–1406.
- Shepard, 1968,
“A two-dimensional interpolation function for irregularly-spaced data”,
In *Proceedings of the 1968 23rd ACM National Conference*, New York, NY, USA, 27–29 August 1968; pp. 517–524.
Available: <https://doi.org/10.1145/800186.810616>
- Sher, 1998 Eran Sher,
Environmental Aspects of Air Pollution,
Academic Press, 1998, Pages 27-41, ISBN 9780126398557
<https://doi.org/10.1016/B978-012639855-7/50041-7>
- Solórzano *et al.*, 2018, Ana Solórzano, Raquel Rodríguez-Pérez, Marta Padilla, Thorsten Graunke, Luis Fernandez, Santiago Marco, Jordi Fonollosa,
“Multi-unit calibration rejects inherent device variability of chemical sensor arrays”,
Sens. Actuators B Chem., 265 (2018), pp. 142-154,
Available: <https://doi.org/10.1016/j.snb.2018.02.188>
- Spinelle *et al.*, 2015 Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, Fausto Bonavitacola,
Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide,
Sensors and Actuators B: Chemical, Volume 215, 2015, Pages 249-257,
<https://doi.org/10.1016/j.snb.2015.03.031>
- Tancev, 2021
“Relevance of drift components and unit-to-unit variability in the predictive maintenance of low-cost electrochemical sensor systems in air quality monitoring”
Sensors, vol. 21, no. 9, p. 3298, 2021,
doi: <https://doi.org/10.3390/s21093298>
- Trygg and Wold, 2002
“Orthogonal Projections to Latent Structures (O-PLS)”,
Journal of Chemometrics, 16, 119-128.
Available: <http://dx.doi.org/10.1002/cem.695>

Reference

Walker and Schneider, 2020

“A study of the relative expanded uncertainty formula for comparing low-cost sensor and reference measurements”
NILU, Kjeller, Norway, Tech. Rep., 2020

Wei *et al.*, 2020 Peng Wei, Li Sun, Abhishek Anand, Qing Zhang, Zong Huixin, Zhiqiang Deng, Ying Wang, Zhi Ning,

Development and evaluation of a robust temperature sensitive algorithm for long term NO₂ gas sensor network data correction,
Atmos. Environ., 230 (2020), Article 117509,
<https://doi.org/10.1016/j.atmosenv.2020.117509>

WHO World Health Organization, 2021a

“Ambient (outdoor) air pollution report”

Accessed: Nov. 10, 2022. [Online].

Available: [https://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health)

WHO World Health Organization, 2021b

“World Health Organization. (2021). WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulphur dioxide and carbon monoxide.”

Accessed: Nov. 10, 2022. [Online].

Available: <https://apps.who.int/iris/handle/10665/345329>

Yuan *et al.*, 2022 Liheng Yuan, Heng Li, Beihao Xia, Cuiying Gao, Mingyue Liu, Wei Yuan, Xinge You

“Recent Advances in Concept Drift Adaptation Methods for Deep Learning”,

Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)

<https://doi.org/10.24963/ijcai.2022/788>

Zimmerman *et al.*, 2018

“A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring”

Atmos. Meas. Techn., vol. 11, no. 1, pp. 291–313, 2018,

doi: <https://doi.org/10.5194/amt-11-291-2018>

Appendix A

A.1 REU plots of reference data selection for calibration update

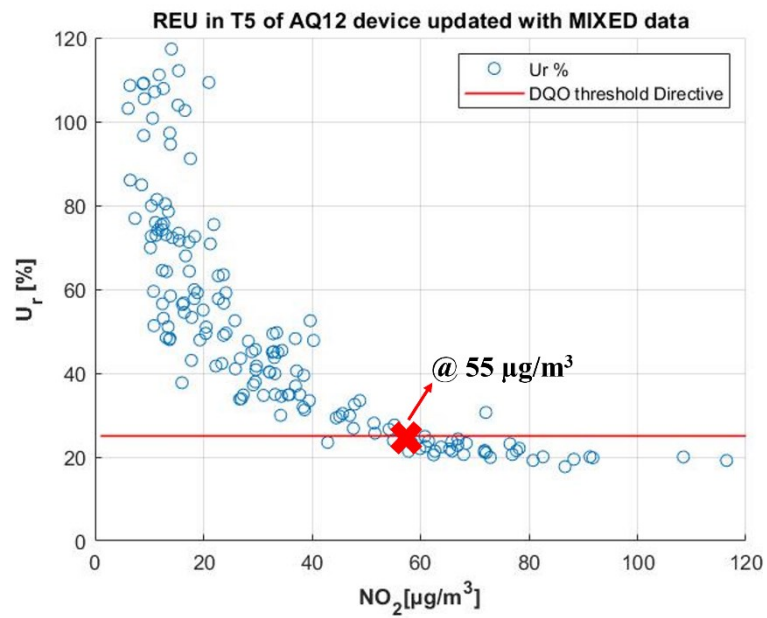


Figure A.1. 1 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with data of T4 (Mixed scenario).

Appendix

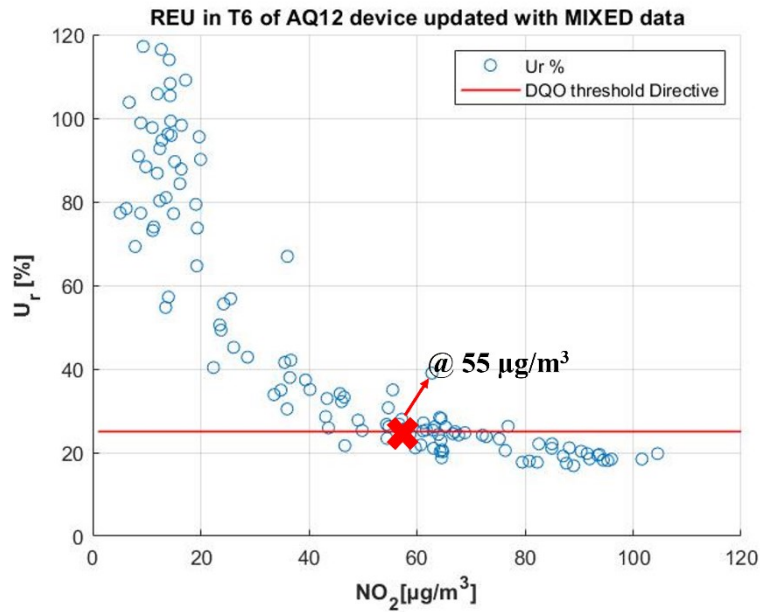


Figure A.1. 2 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with data of T4 (Mixed scenario).

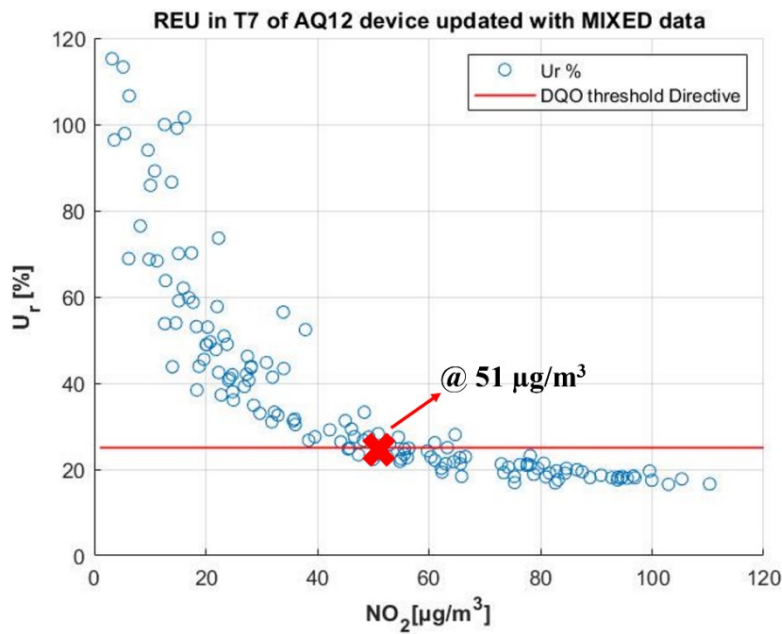


Figure A.1. 3 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with data of T4 (Mixed scenario).

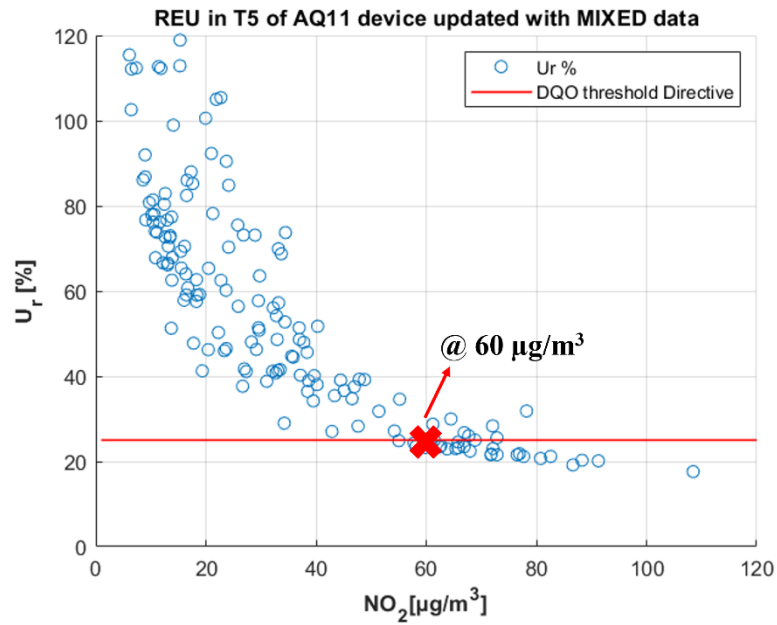


Figure A.1. 4 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with data of T4 (Mixed scenario).

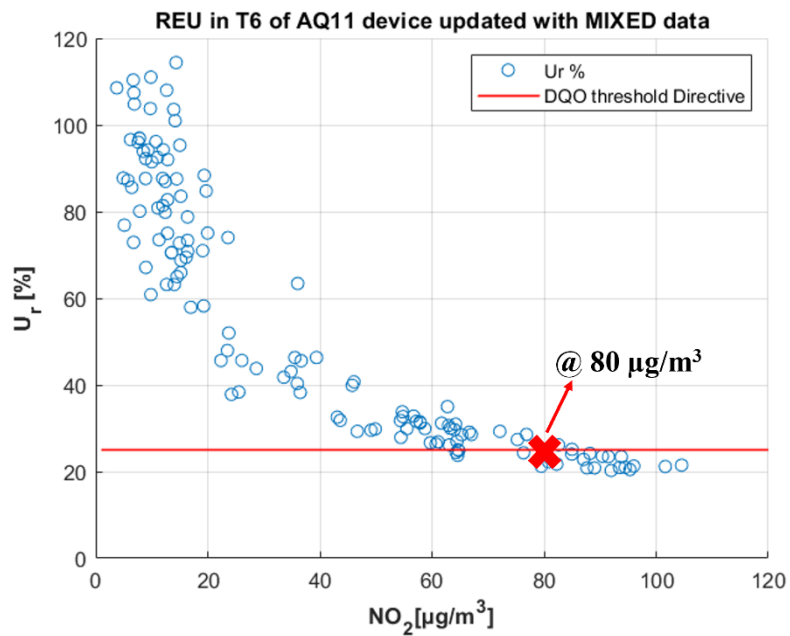


Figure A.1. 5 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with data of T4 (Mixed scenario).

Appendix

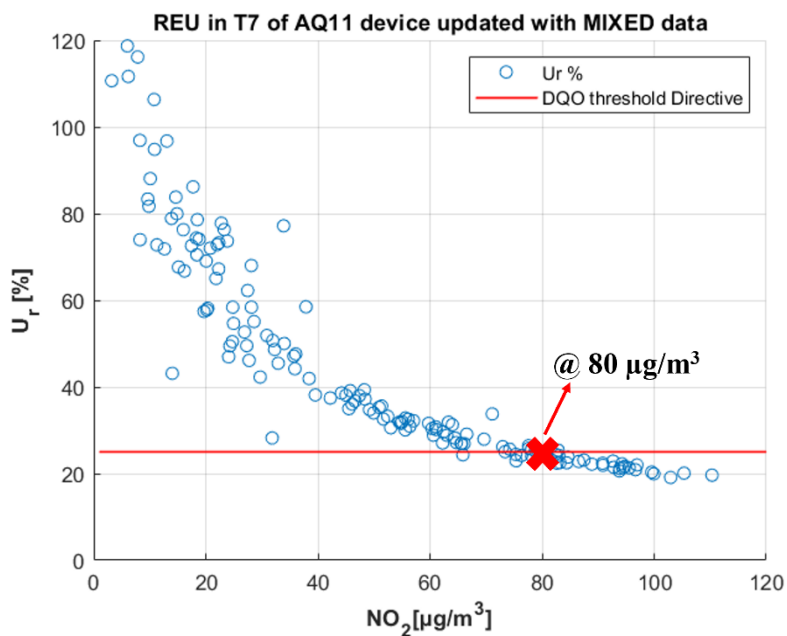


Figure A.1. 6 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with data of T4 (Mixed scenario).

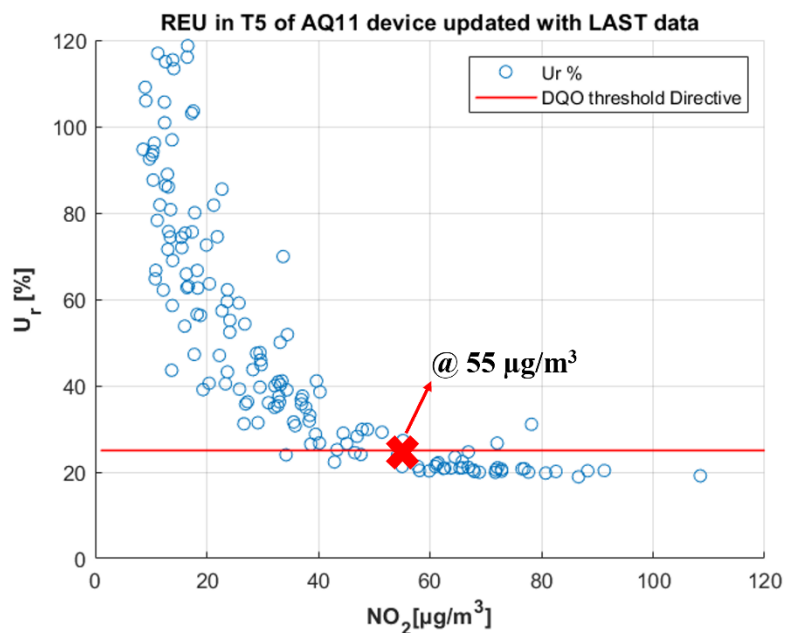


Figure A.1. 7 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with data of T3 (Last scenario).

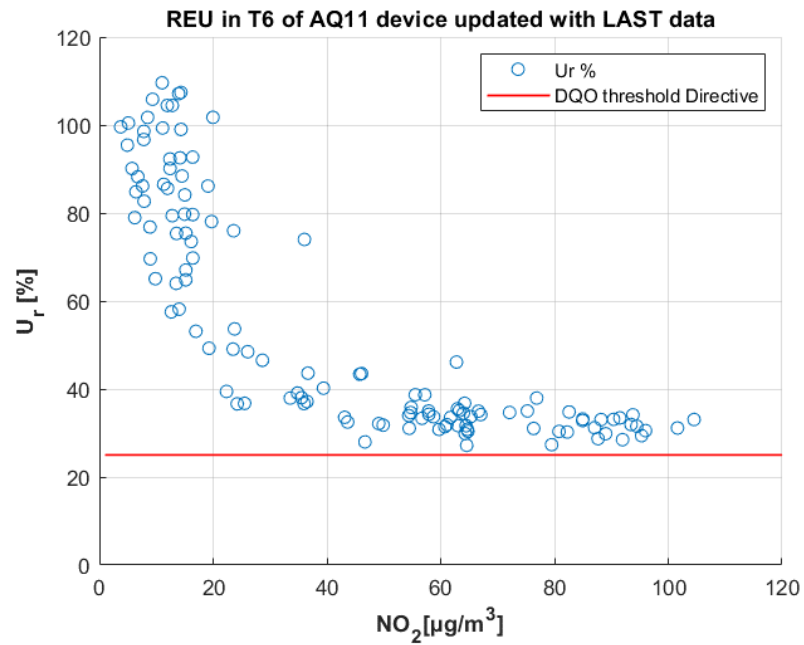


Figure A.1. 8 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with data of T3 (Last scenario).

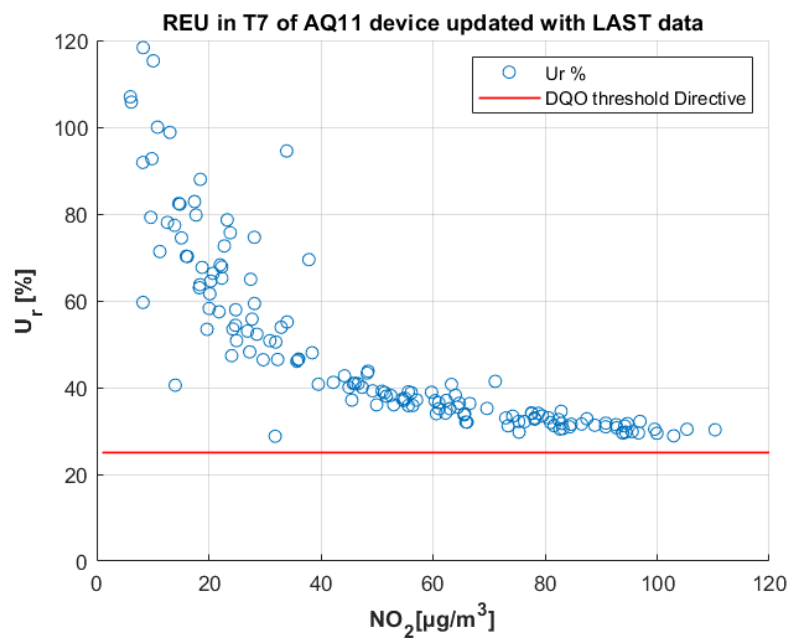


Figure A.1. 9 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with data of T3 (Last scenario).

Appendix

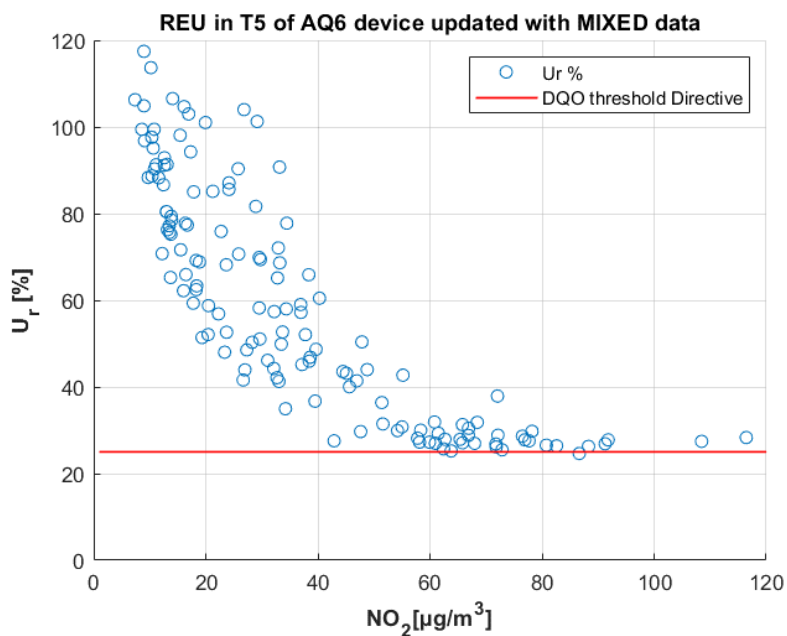


Figure A.1. 10 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with data of T4 (Mixed scenario).

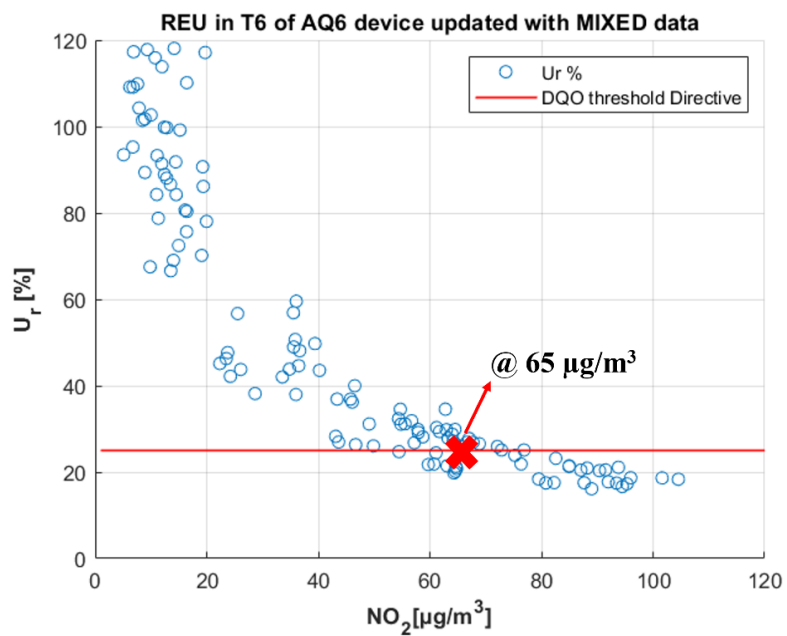


Figure A.1. 11 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with data of T4 (Mixed scenario).

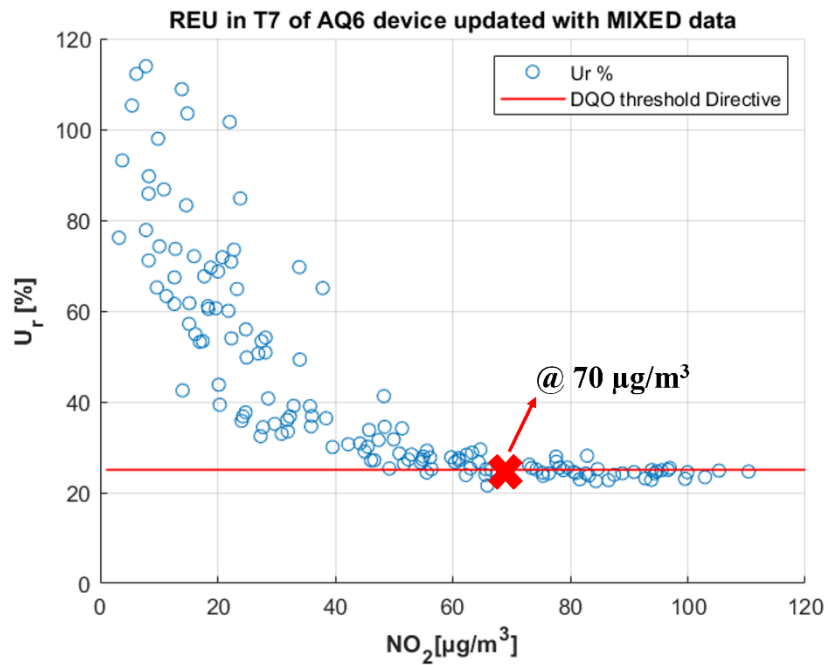


Figure A.1. 12 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with data of T4 (Mixed scenario).

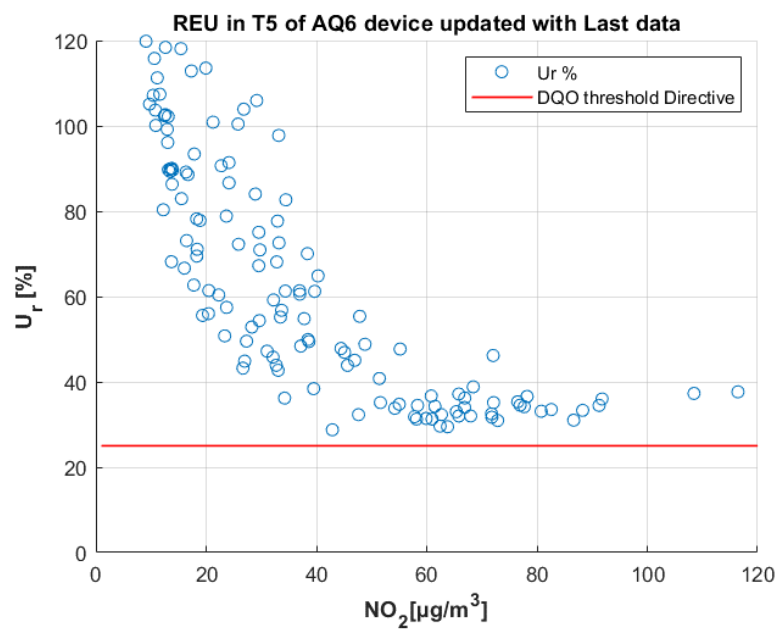


Figure A.1. 13 Plot of Relative Expanded Uncertainties in T5 when AQ6 is re-calibrated with data of T3 (Last scenario).

Appendix

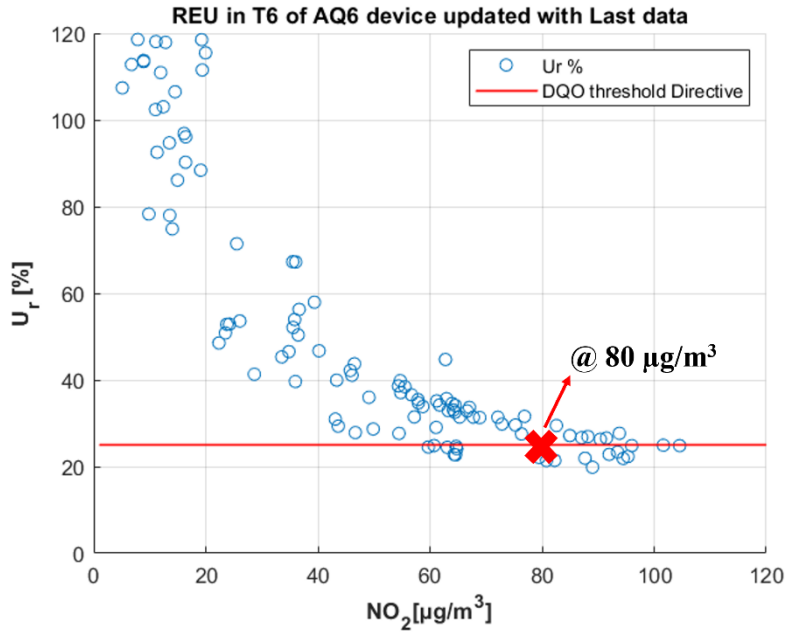


Figure A.1. 14 Plot of Relative Expanded Uncertainties in T6 when AQ6 is re-calibrated with data of T3 (Last scenario).

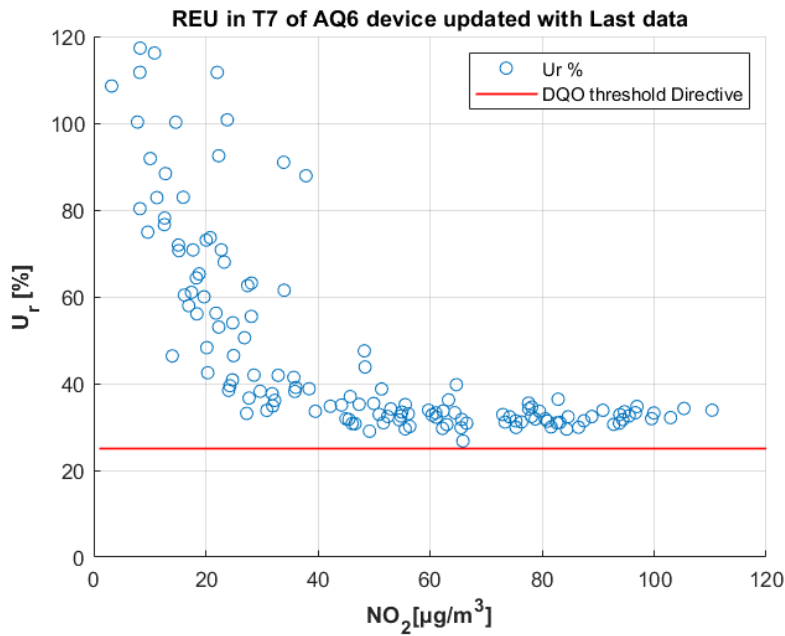


Figure A.1. 15 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with data of T3 (Last scenario).

A.2 REU plots of general calibration model

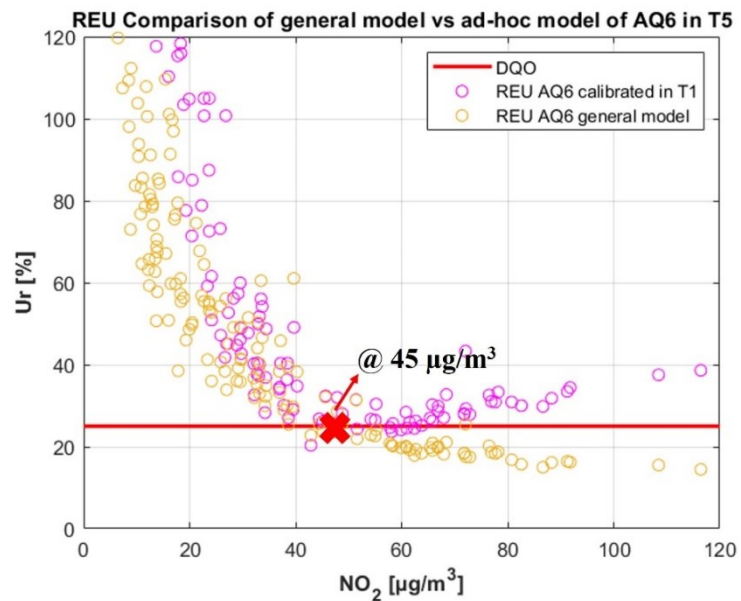


Figure A.2. 1 Plot of Relative Expanded Uncertainties in T5 when AQ6 is recalibrated with global calibration model.

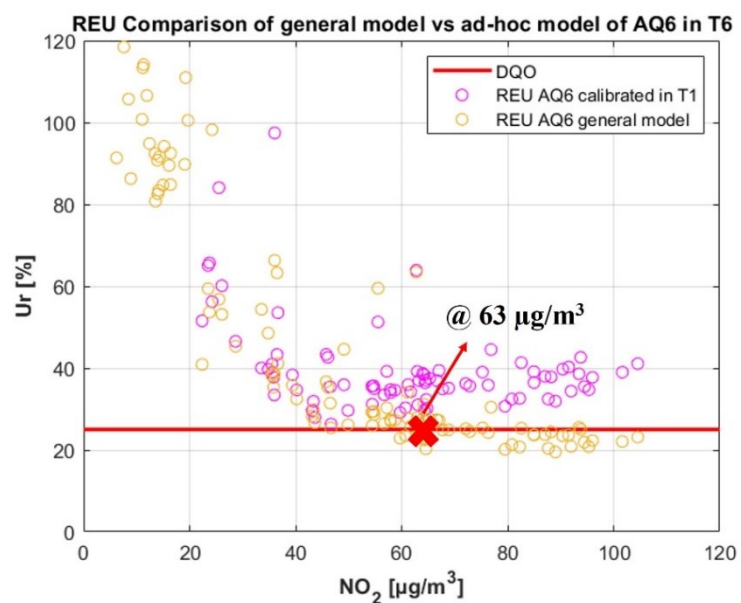


Figure A.2. 2 Plot of Relative Expanded Uncertainties in T6 when AQ6 is recalibrated with global calibration model.

Appendix

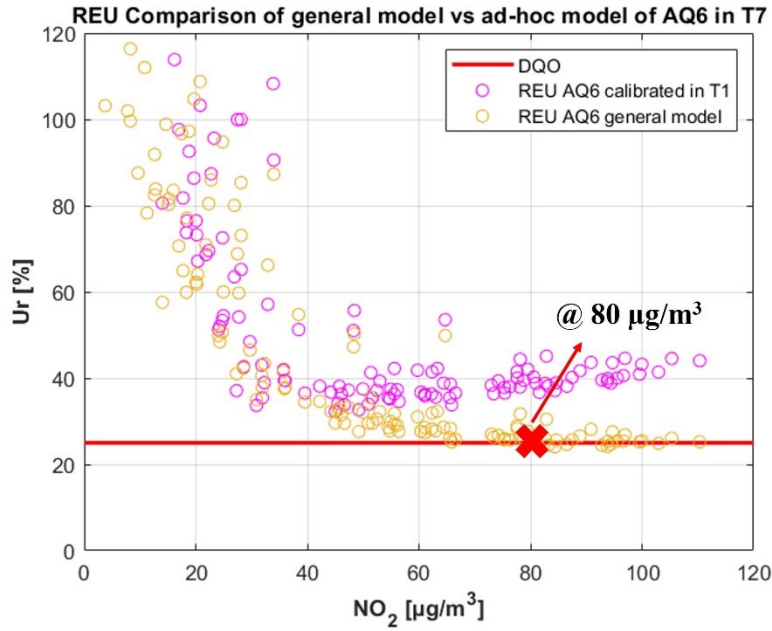


Figure A.2. 3 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with global calibration model.

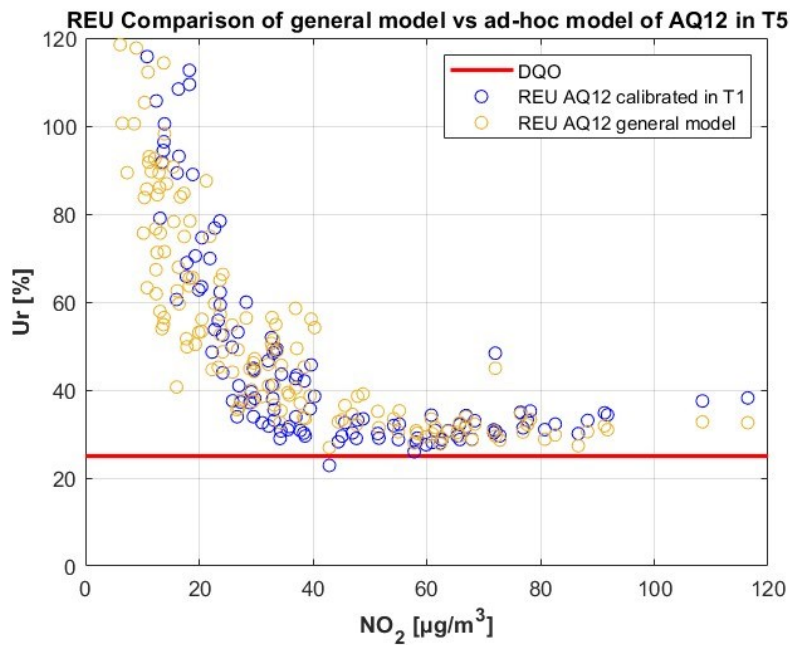


Figure A.2. 4 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with global calibration model.

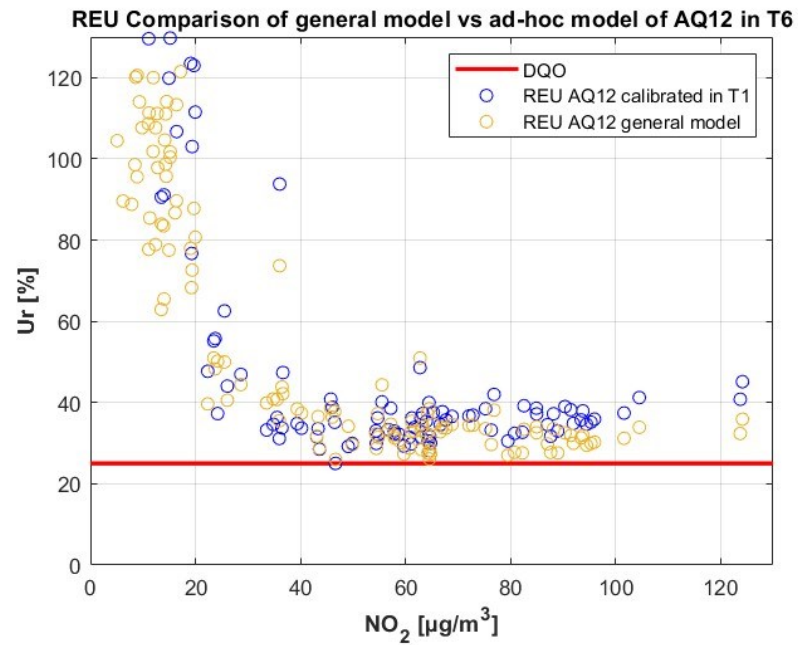


Figure A.2. 5 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with global calibration model.

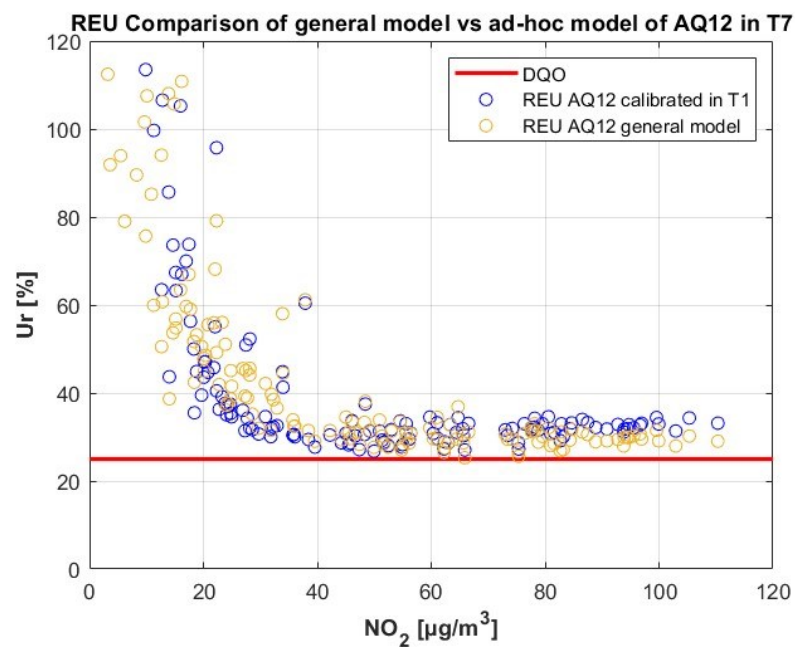


Figure A.2. 6 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with global calibration model.

A.3 REU plots of importance weighting calibration model

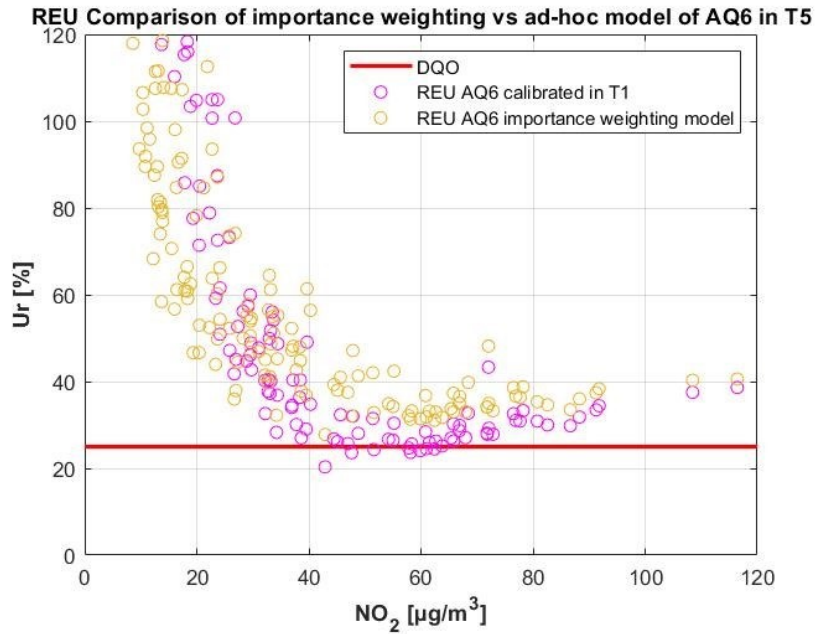


Figure A.3. 1 Plot of Relative Expanded Uncertainties in T5 when AQ6 is recalibrated with the importance weighted calibration model.

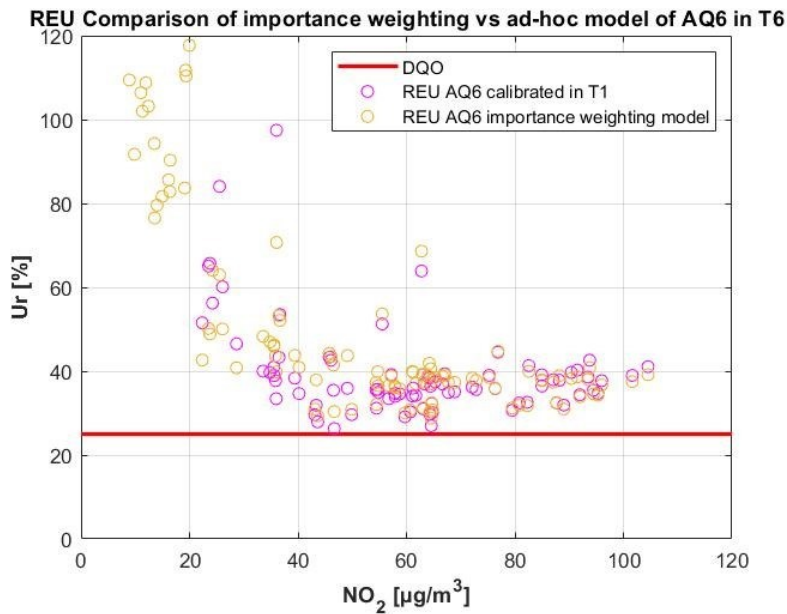


Figure A.3. 2 Plot of Relative Expanded Uncertainties in T6 when AQ6 is recalibrated with the importance weighted calibration model.

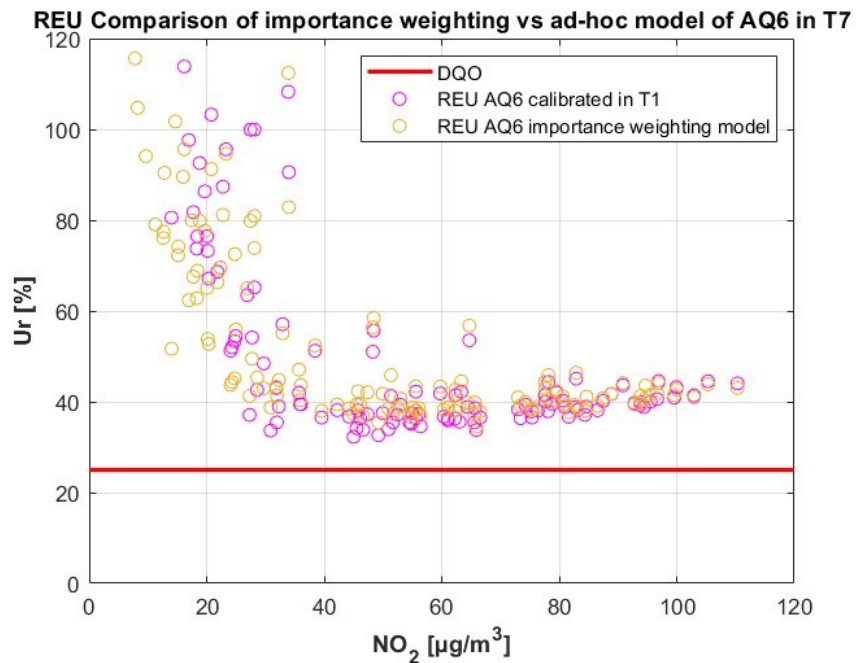


Figure A.3. 3 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with the importance weighted calibration model.

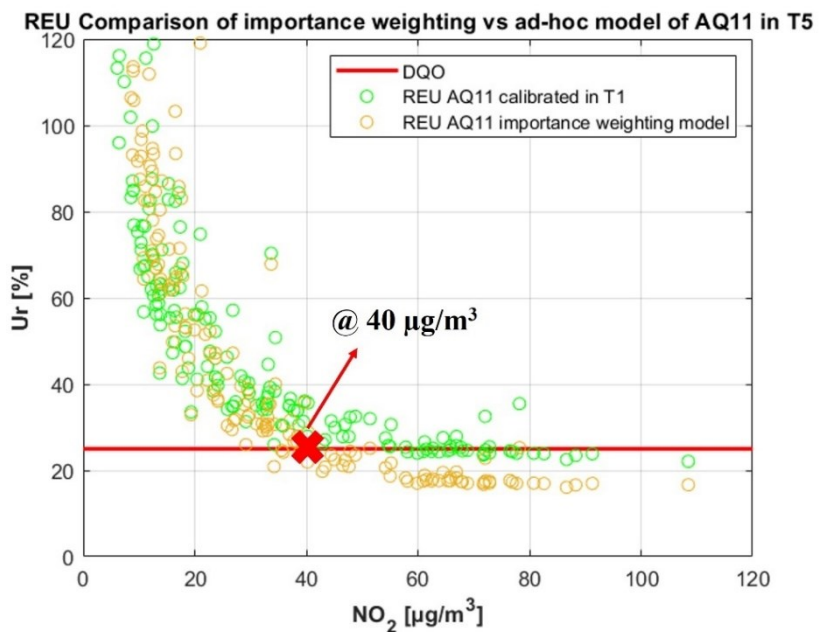


Figure A.3. 4 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with the importance weighted calibration model.

Appendix

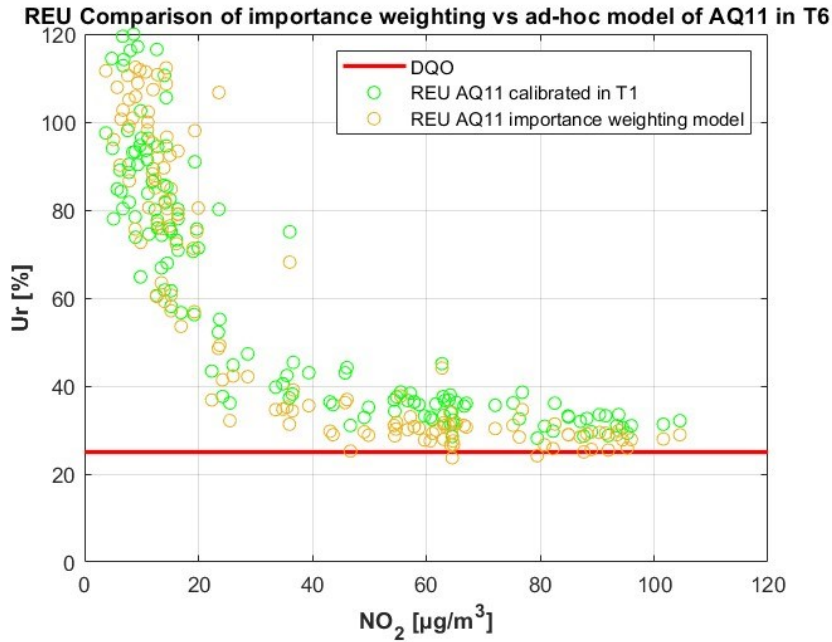


Figure A.3. 5 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with the importance weighted calibration model.

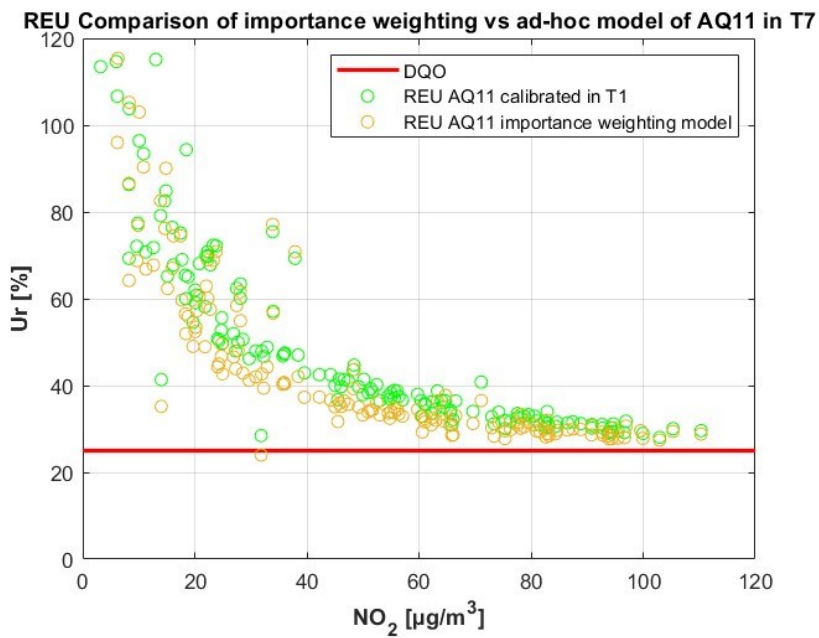


Figure A.3. 6 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with the importance weighted calibration model.

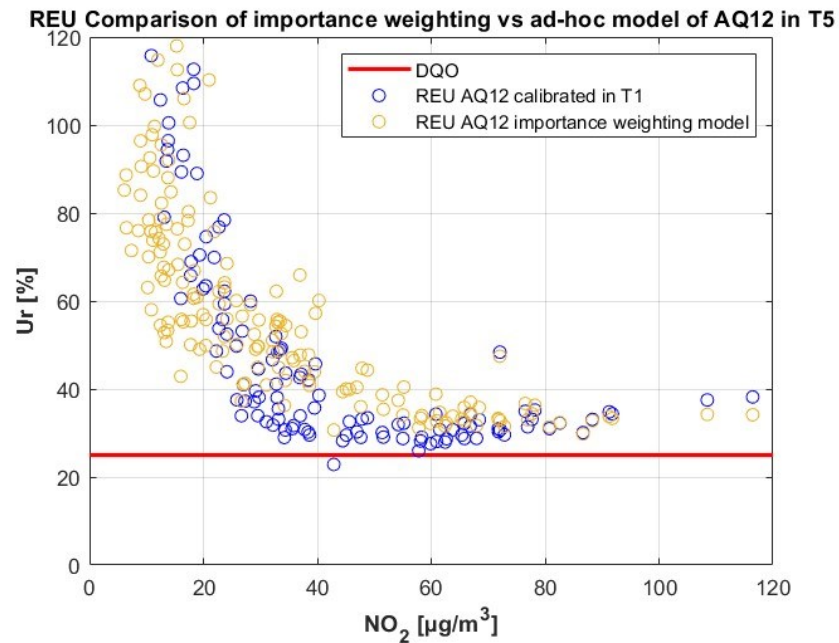


Figure A.3. 7 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with the importance weighted calibration model.

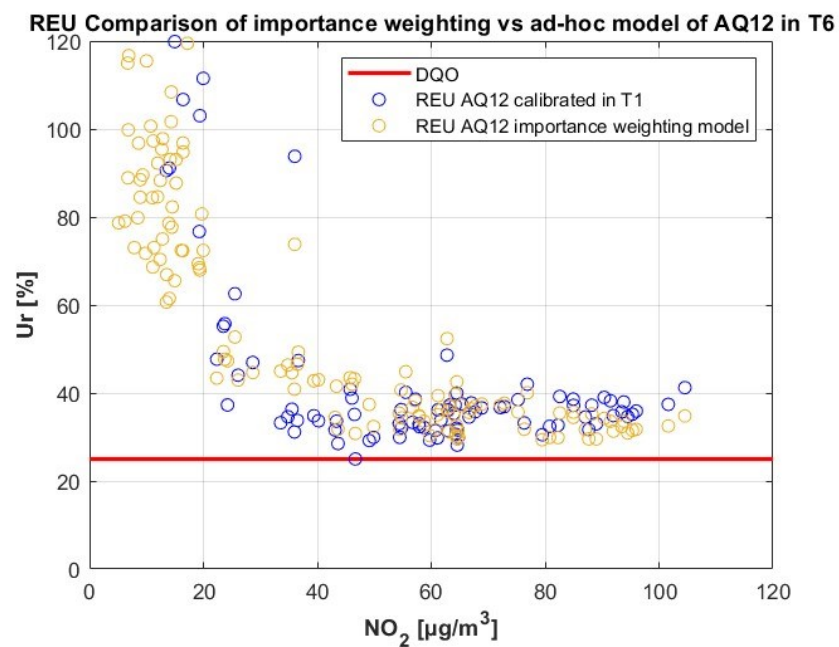


Figure A.3. 8 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with the importance weighted calibration model.

Appendix

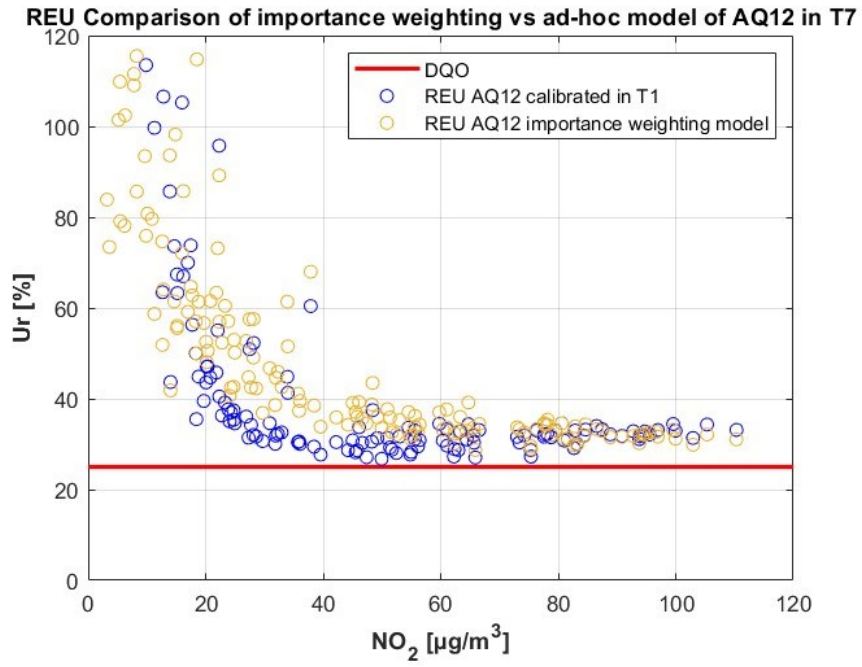


Figure A.3. 9 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with the importance weighted calibration model.

A.4 REU plots of stacking ensemble calibration model

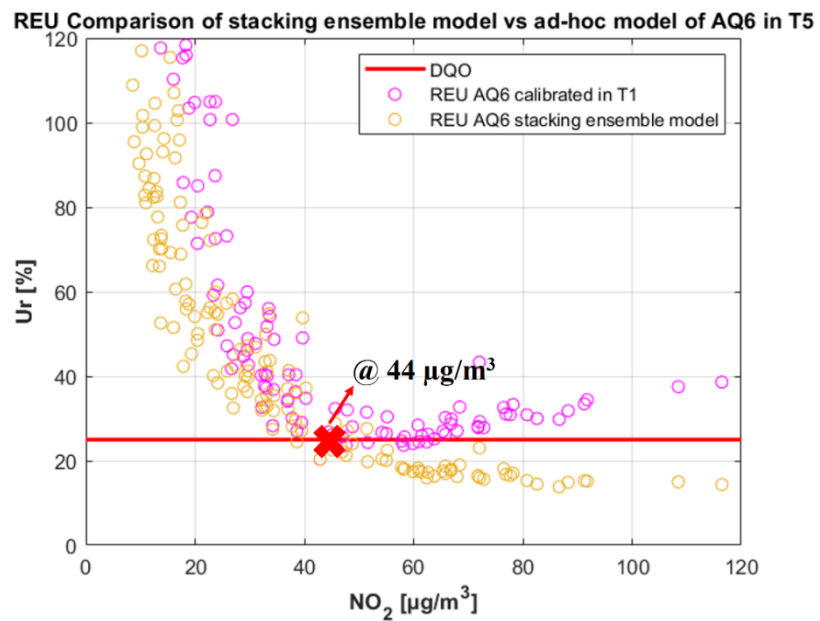


Figure A.4. 1 Plot of Relative Expanded Uncertainties in T5 when AQ6 is recalibrated with the stacking ensemble calibration model.

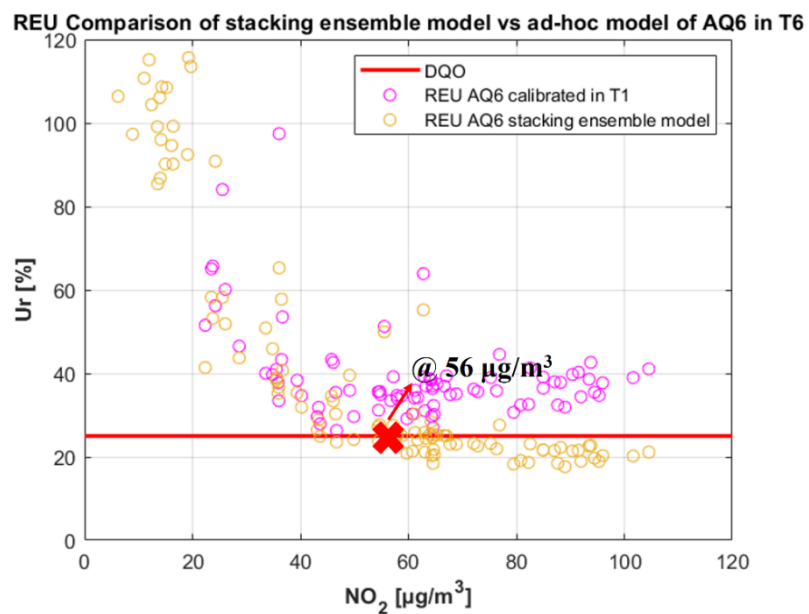


Figure A.4. 2 Plot of Relative Expanded Uncertainties in T6 when AQ6 is recalibrated with the stacking ensemble calibration model.

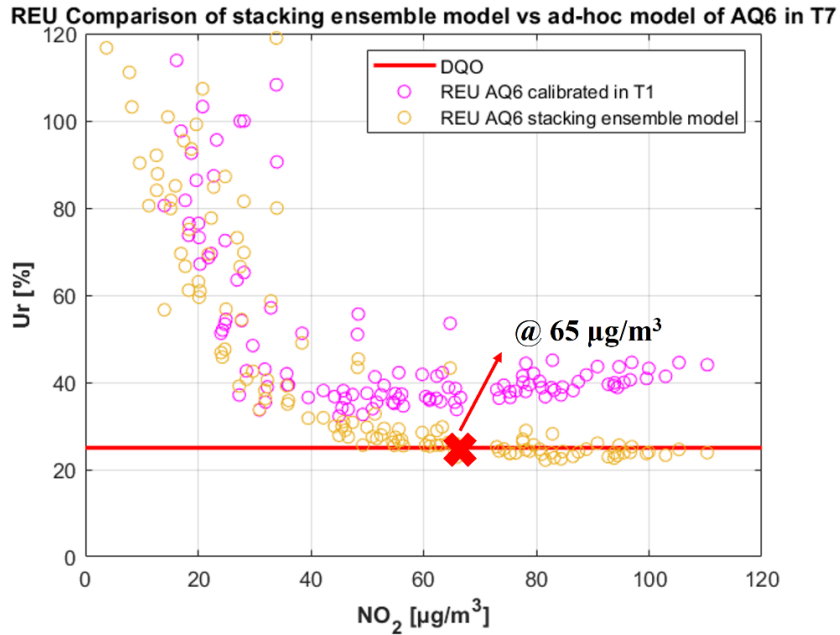


Figure A.4. 3 Plot of Relative Expanded Uncertainties in T7 when AQ6 is re-calibrated with the stacking ensemble calibration model

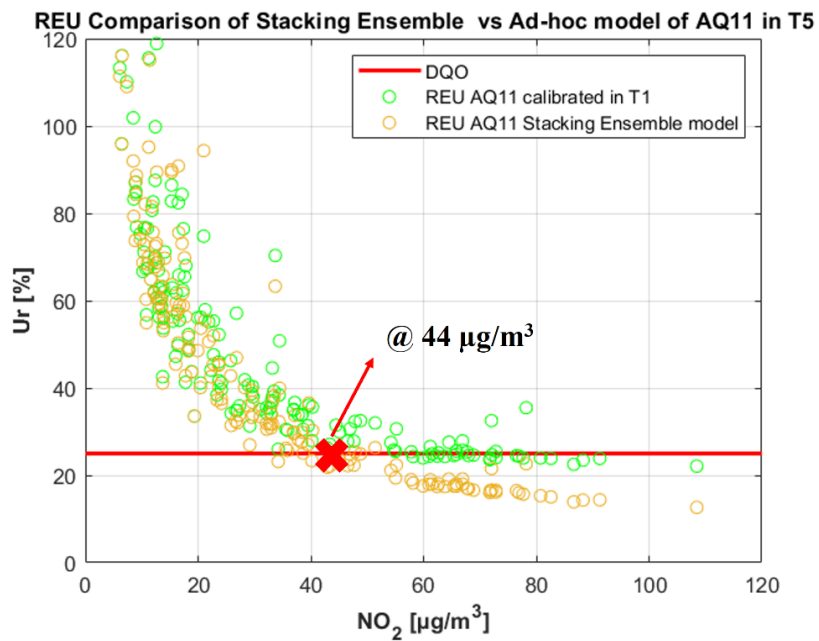


Figure A.4. 4 Plot of Relative Expanded Uncertainties in T5 when AQ11 is re-calibrated with the stacking ensemble calibration model.

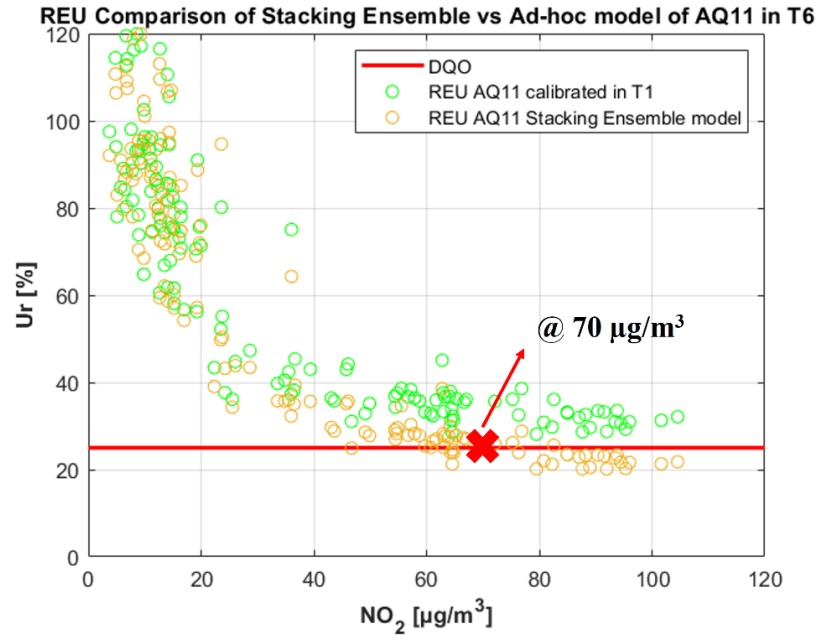


Figure A.4. 5 Plot of Relative Expanded Uncertainties in T6 when AQ11 is re-calibrated with the stacking ensemble calibration model.

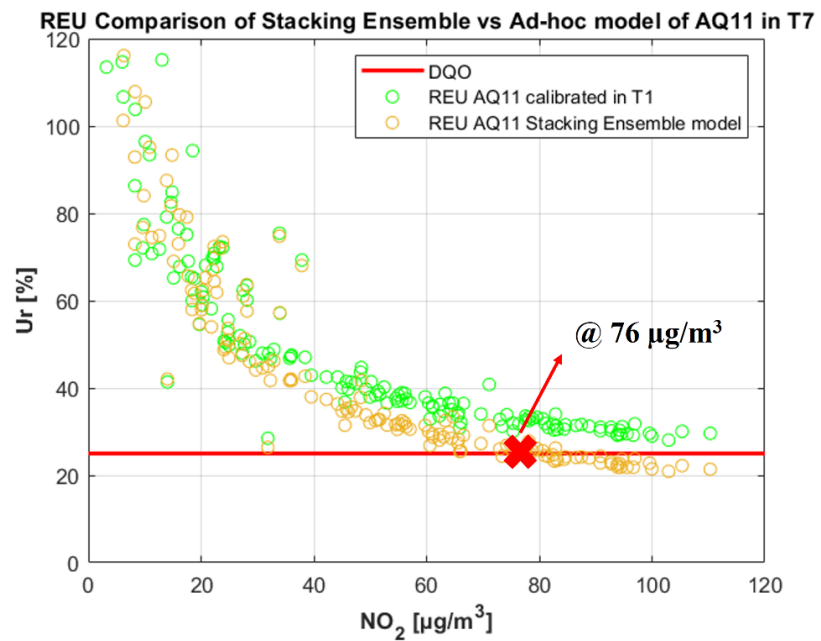


Figure A.4. 6 Plot of Relative Expanded Uncertainties in T7 when AQ11 is re-calibrated with the stacking ensemble calibration model.

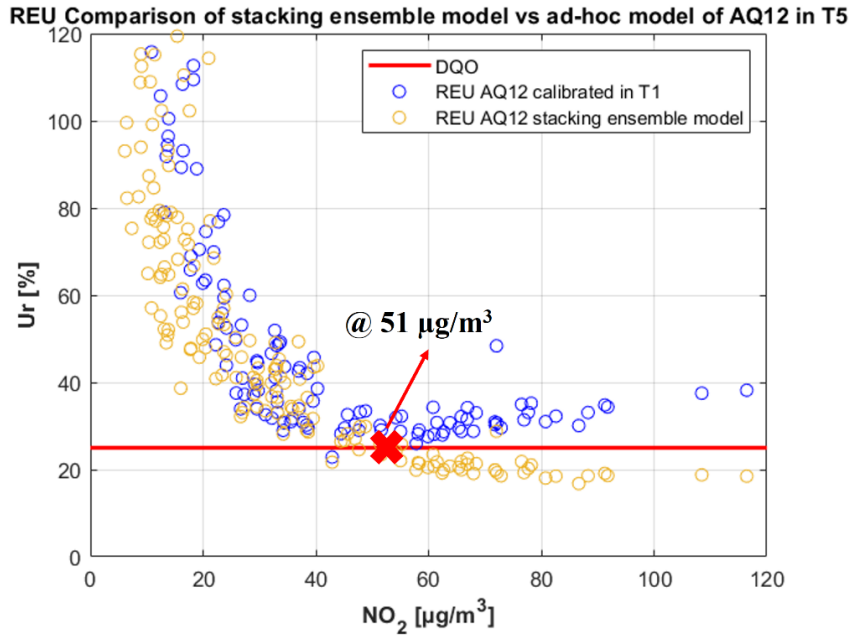


Figure A.4. 7 Plot of Relative Expanded Uncertainties in T5 when AQ12 is re-calibrated with the stacking ensemble calibration model.

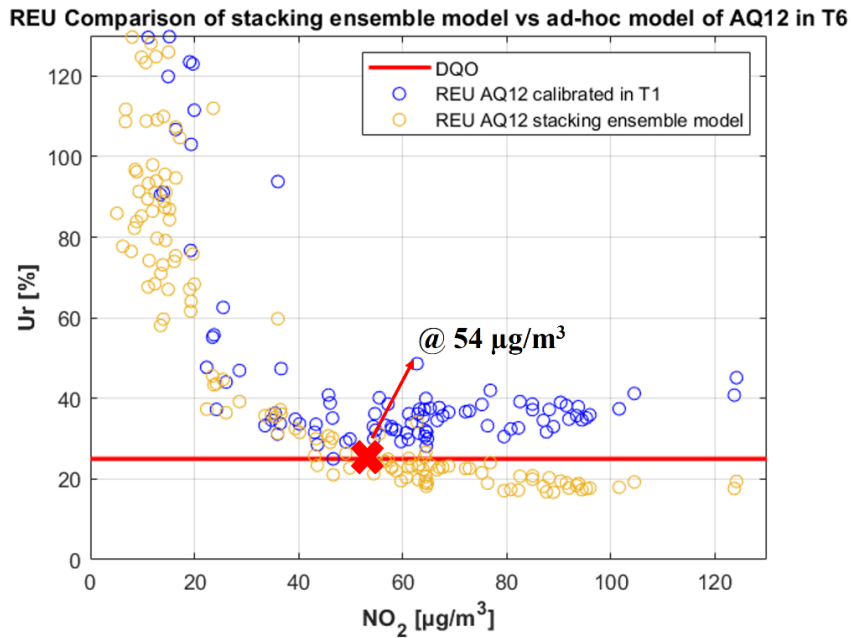


Figure A.4. 8 Plot of Relative Expanded Uncertainties in T6 when AQ12 is re-calibrated with the stacking ensemble calibration model.

REU Comparison of stacking ensemble model vs ad-hoc model of AQ12 in T7

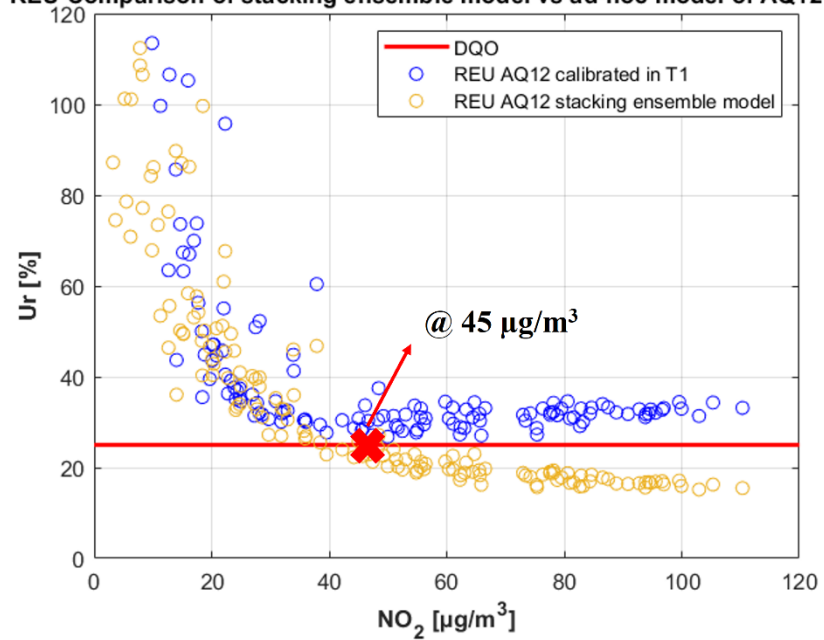


Figure A.4. 9 Plot of Relative Expanded Uncertainties in T7 when AQ12 is re-calibrated with the stacking ensemble calibration model.