On the Robustness of Field calibration for Smart air quality monitors

Article · February 2020

DOI: 10.1016/j.snb.2020.127869

CITATIONS
26

READS
344

5 authors, including:

Saverio De Vito
ENEA
137 PUBLICATIONS **1,603** CITATIONS
SEE PROFILE

Elena Esposito
ENEA, Portici, Italy
48 PUBLICATIONS **686** CITATIONS
SEE PROFILE

Philipp Schneider
Norwegian Institute for Air Research
75 PUBLICATIONS **3,326** CITATIONS
SEE PROFILE

Alena Bartonova
Norwegian Institute for Air Research
177 PUBLICATIONS **3,585** CITATIONS
SEE PROFILE

Published in Sens. and Act. B: Chemical, Vol. 310, 2020, 127869, ISSN 0925-4005, doi.org/10.1016/j.snb.2020.127869.

On the Robustness of *Field calibration* for smart air quality monitors

Saverio De Vito[1], Elena Esposito[1], Nuria Castell[2], Philipp Schneider[2], A. Bartonova[2]

[1]Smart Network Division, ENEA, P. le E. Fermi, 1, 80055, Portici, Naples
[2]NILU - Norwegian Institute for Air Research, N-2007 Kjeller, Norway

Abstract.

The robustness of field calibrated Air Quality Multisensors (AQM) performances to long term and/or mobile operation is still debated. Though accuracy generally exceeds the one of laboratory calibrations models, experimental results show that field calibration models cannot sustain optimal field performances due to *changes* occurring in operative conditions. Among them, the relocation of calibrated multi-sensors platforms and sensors drift are considered as the most relevant.

In this work, we want to provide an answer to the general issue of field calibration robustness analysing theoretical foundations and providing tools for determining the main drivers affecting performances. In particular, by leveraging on probability distribution of target and interferents gas as well as environmental variables, measures of dissimilarity between calibration and operative phase conditions are provided to quantitatively capture the occurring change. A 6 months multiple nodes dataset including node relocations events in several sites, have been processed for deriving nonlinear multivariate field calibrations whose robustness to changing conditions have been analysed. Kullback-Leibler, Euclidean and Hellinger dissimilarity measurements have been correlated with recorded performance degradation. Results show that quantifying relevant factors probability distribution changes allows to explain and predict performances of in field data driven calibration models. They also highlights the role of concept drifts in explaining field performances ameliorating our capability to select optimal conditions in which a field calibration should be derived. Finally, smart air quality monitors could now autonomously detect the need for re-calibration.

Keywords. Field Calibration robustness; Mobile Air Quality multi-sensor platforms; Sensors relocation; Concept drift; Learning in Dynamic Environments.

1. Introduction

Smart air quality monitors, mostly based on solid state micro-sensors, are receiving

an increasing interest from the general public. This is due primarily to their capability to act as pervasive sentinels enabling to assess personal exposure to air pollutants (see [1-2,26]), a primary concern for general population and especially for those who live in urban environments. Actually, the fraction of total exposome due to air pollutants is deemed as extremely relevant in determining the incidence rate of several severe illnesses like strokes, cardiovascular diseases, asthma, COPD, etc. [3,24,25]. Aside from wearable/portable systems, fixed multi-sensor platforms will also have a significant role in augmenting the information to the citizen and policy maker. Should we be able to harmonize data provided by such different systems and classical model approaches, we would be capable to build high space and time resolution maps, that will allow the estimation of air quality for exposome monitoring, nowcasting applications and predictive scenarios with unprecedented accuracy and hence, value [4,5].

Unfortunately, most smart AQM systems are nowadays sold without any guarantee on data accuracy and precision [6]. Moreover, raw data are usually not available for post processing.

It is well known that chemical as well as particulate matter sensors are affected by long term drift, interferents and non-linearities. The latter requires laboratory calibration procedures to use several set points to completely describe sensors response curve. Lack of selectivity cause several unwanted factors, like non-target gases and environmental conditions, to elicit a significant response negatively interfering with the response to the nominal target. The set of target gases, non target gas and environmental interferents, are hereafter collectively referred to as *forcers*.

To correct for interference, *multiple regression* schemes are employed [6]. However, the number of forcers (non-linearly) influencing sensor responses is usually so significant that a complete lab based calibration procedure appears simply unfeasible. Moreover, lack of reproducibility in sensors fabrication requires ad-hoc calibration for each multisensors multiplying the overall costs while calibration transfer technology is still to provide evidence of scalability. Simplified calibration procedures retain a high cost with significant risks of badly performing under field conditions. Schneider et al. [4] showed how a lab calibrated device whose calibration model was developed taking into account temperature dependence correction was still negatively impacted by different weather conditions encountered in the field.

Supplementing lab based calibration, *in Field calibration* models may be obtained by jointly analysing multi-sensors data and pollutants concentrations as measured by collocated reference analysers. This methodology offers a viable solution for obtaining accurate pollutant concentration predictions allowing to compute models that well describe the field behavior of the multi-sensors. The results are so impressive that Borrego et al. [7] have shown that, at least for short term periods, the calibrated multi-sensors can already meet the *data quality objectives* (DQO), set by European Union in the 2008 directive, actually qualifying the sampled data for supplementing the sparse regulatory AQ monitoring network. However, the robustness of field calibration is still debated.

In facts, multiple factors can negatively affect multi-sensors in field calibrations performances. Changes in environmental conditions and local atmosphere composition either of natural or anthropogenic origin along with sensors drift, challenge their robustness to effective field deployment. This is particularly evident for *non-parametric* calibration models.

Introducing the main point of our work, to cope with field conditions variability and cross sensitivities a calibration model should be ideally trained with enough data to span all the meaningful configurations of forcers values[1]. In this way, an automated learner may enjoy a complete knowledge on the influences of all the relevant factors on sensor response and the full calibration function may be learnt. If provided with a limited number of samples, on the contrary, the resulting model may be incomplete. Moreover, it may learn false, or at least temporary, correlations among forcers and sensors responses due to special conditions occurring in space and time. The resulting model will not be able to survive to significant changes in the relevant conditions. Unfortunately, these often occur in time, at least on a seasonal basis (e.g. weather conditions) or due to changing human activities. In the space dimension, moving away from calibration site, conditions could severely change due, for example, to dramatically different car traffic conditions.

Loss of accuracy due to model incompleteness is related with the so called "concept drift" effects, as opposed to sensor drift effects due to sensor ageing and poisoning. Recently some authors, including De Vito [8,9,15], Zimmermann et al. [10], Cross et al. [11], Spinelle et al. [12], Hagan [14], although advocating for the use of field multivariate calibration warn the practitioners about some of these possible effects. De Vito et al. showed evidence of long term drift effects recovering while proceeding towards the season in which their multi-sensor was originally field calibrated [8], thus highlighting cyclostationarity in the performances. Spinelle et al., warned against false or local correlations and advised not to use sensors targeted to species (proxies) that show high correlation with the target gas by operating an accurate preliminary feature selection scheme [12]. That correlation obviously depends on atmosphere composition and may disappear or change in intensity when moving away from calibration site. Cross et al identified temperature as a major interferents for Alphasense electrochemical sensors with respect to humidity and argued that environmental variables range changes may be the primary source of calibration errors with respect to sensor drift itself during extended field operations [11].

Most of these studies are conducted using a single or a few multi-sensor devices and hence their generalization capability could be limited. The field knowledge could be improved by deploying several devices in field for multi-seasonal periods and collectively analysing the devices performances over time. Cordero et al. [22], for example, attempted to develop a multilinear regression framework for field calibrating several AqMesh Pods [23] and evaluating the significance of non-target and interferent gases with state-of-the-art statistical tests. They account for a limited influence of temperature but these results may be due to the non-linearity of

[1] Along the paper, we will use the term *manifold* to refer to the multidimensional space subset containing forcers values configurations.

temperature influence reported in the adopted sensors datasheets [21] or to limited operational time.

Field calibration robustness is often linked by practitioners to location issues. Few studies, however, analyse the performances when sensor nodes are operated elsewhere with respect to its field calibration site. During spring and summer 2017, our research group investigated the location dependence of the performances of field calibrations in the city of Oslo [13]. Surprisingly, we didn't observe significant differences among the performance of relocated multi-sensors and the ones that continued to operate at calibration site. Instead, seasonal changes in mean values of forcers appeared to negatively affect performances of all the nodes, irrespective of their location. Later the same year, Hagan et al. highlighted the relevance of the target gas range change in determining the performance of a field calibration scheme [14]. During 2018, Casey and Hannigan obtained results that shown how pollutant emissions composition changes caused performance degradation in field calibrated multi-sensors devices [17]. Masey et al. [18], studied the temporal performance behaviour of field calibrated again generally attributing to sensors drift and concentrations range changes the loss of calibration quality. They propose to gradually enlarge the calibration dataset with recurrent colocation measurements period. Although very effective, this approach will boost calibration costs and pose a risk of *catastrophic interference* (or even *catastrophic forgetting* if sequentially learning) effects, often reported when learning in nonstationary environments [19,20].

These works clearly highlighted the role of concept drifts in determining performances of field calibrated AQ monitors, however to the best of our knowledge, none of them analyzed the theoretical foundations of the underlying phenomena, nor they attempted to quantitatively investigate their drivers. At the same time, they advocated for further experiments to be conducted in order to clarify the robustness of field calibration strategy. It is indeed clear that we need to set up a new theoretical reference model to encode these aspects. This knowledge is, in fact, would be of paramount importance, especially for developing efficient calibration strategies for mobile or fixed long term deployments at the base of future dense air quality monitoring networks and exposome monitors. As a consequence, in this work we will focus on the following objectives:

a) Introducing a theoretical model for quantitatively analysing and predicting the robustness of a calibration model in nonstationary environments identifying the main drivers of performance. In our view, the *manifold* spanned by calibration data points becomes crucial in determining the calibration function robustness to changing field conditions. In particular, we will analyse the differences between the statistical distribution of relevant forcers during calibration learning and field operation to explain the overall performances. We formulate the hypothesis that the distribution difference consistently and negatively correlates with the calibration performance. This knowledge will consequently help to understand the best conditions for field calibration sites selection. As a further result, this will allow the multisensor to assess its current accuracy and anticipate the need for recalibration.

b) Testing the introduced model and hypothesis using a dataset featuring multiple

devices operated in different locations during several months with the constant availability of ground truth regulatory grade data about pollutants concentrations and weather conditions. Target gas and a primary interferent will be chosen to investigate the relationship between distribution changes and performance indicators.
In the following sections, background and the experimental setting are described. Specifically, we first introduce the experimental campaign upon which we developed our preliminary analysis along with a motivational review of the most related works. Theoretical framework and methodology description follow right after. Results are detailed and discussed in Section 4. Finally, conclusions are drawn in Section 5.

2. Background

2.1. Dataset: Multisensor device and deployment campaign description

The dataset analysed in this work has been extracted from a recording campaign conducted by NILU using 24 AQMesh multi-sensor platforms [4,23]. These AQMesh multi-sensor devices, deployed in Oslo for more than 6 months, were equipped with NO, NO_2, O_3, CO, $PM_{2.5}$ and PM_{10} sensors plus Temperature (T) and Relative Humidity (RH) sensors. Specifically, all gas sensing capabilities were based on Alphasense Electrochemical sensors. Alphasense electrochemical sensors are reported by several studies to be among the best performing in field conditions [7], although affected by slow response [9] and interference by temperature [11] and non-target gases [21]. Datasheets are available at [21]. AQMesh pods are sold with a factory derived calibration that partially correct for cross-sensitivities although temperature interference can still be significant as shown in [4]. In this work, however, we field calibrated the pods against co-located reference stations for optimal performances.

Actually, the 24 platforms were co-located at the reference AQM station of Kirkeveien road from April to June 2015. Then, a subset of them was relocated to 3 other different regulatory monitoring stations in Oslo. In particular, 5 pods were relocated to Akebergveien Rd, 4 pods to Manglerud Rd and 4 pods to Alnabru Rd. (Fig. 1). The latter are not used in this study due to a critical malfunction. Kierkeveien and Manglerud stations are located close to highly busy traffic while the Akebergveien station is located at intersection of two streets with low to medium traffic. More detailed information about sensors deployments and air pollution local assessment are available in [4]. This setting allowed all the pods to be collocated with at least a reference station along the entire deployment duration. To us, that meant being able to have access to true, or better saying to high accuracy, concentrations levels from a regulatory grade station, all the time. Reference data have been, in facts, exploited for supervised calibration learning. Note that while NO_2 reference data were available in all the reference stations, O_3 reference data were available only at Kierkevein rd. station.

Furthermore, several pods experimented sustained data losses. For our investigations, we have taken into account only 12 of the total 24 pods, actually the ones for which

more data were available. To balance spatial composition, 4 of the nodes were chosen among the ones that remained in Kirkeveien, 4 nodes were chosen among the ones that were relocated from Kirkeveien to Akebergveien and 4 were chosen among the ones that were relocated to Manglerud Rd. (for a summary of the considered AQMesh nodes, see Table 1.)

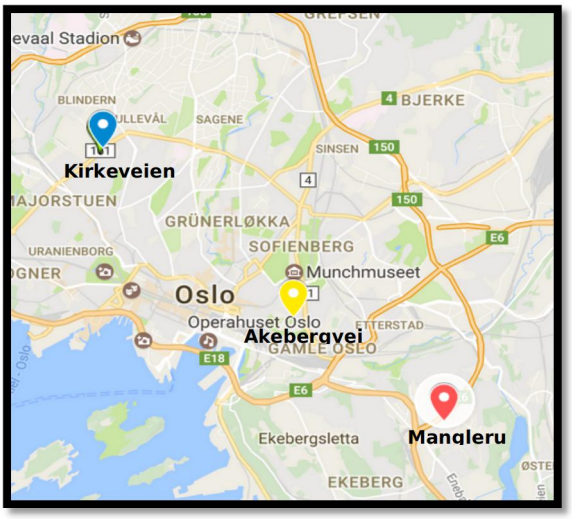


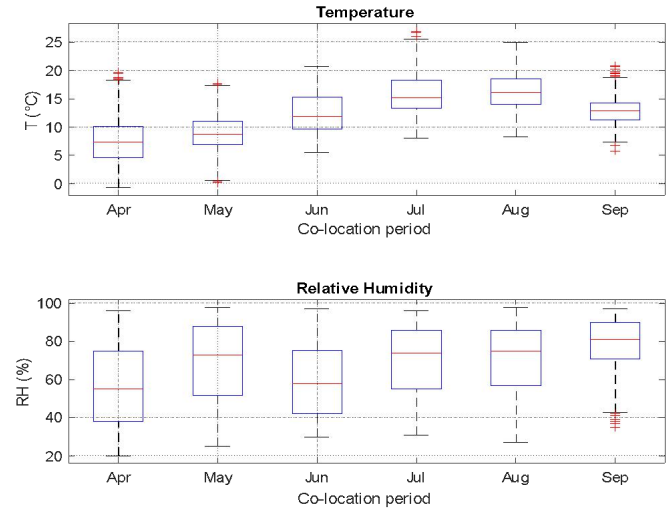
Figure 1: Sensor nodes (Pods) and Reference stations deployment locations in the city of Oslo.

Platform	Monitored Species/parameters	Final Location (starting from July 2017)	No. of recorded samples (hourly averages)
715150	CO, NO, NO_2, O_3, $PM_{2.5}$, PM_{10}, T, RH	Kirkeveien rd., Stationary pod	3061 hrs
764150	CO, NO, NO_2, O_3, $PM_{2.5}$, PM_{10}, T, RH	Kirkeveien rd., Stationary pod	2924 hrs
785150	CO, NO, NO_2, O_3, $PM_{2.5}$, PM_{10}, T, RH	Kirkeveien rd., Stationary pod	2893 hrs
849150	CO, NO, NO_2, O_3, $PM_{2.5}$, PM_{10}, T, RH	Kirkeveien rd., Stationary pod	2915 hrs
712150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Akebergveien Rd.	1151 hrs
743150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Akebergveien Rd.	2764 hrs
828150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Akebergveien Rd.	2930 hrs
850150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Akebergveien Rd.	2974 hrs
718150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Manglerud Rd.	3119 hrs
737150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Manglerud Rd.	2905 hrs
751150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Manglerud Rd.	2833 hrs
856150	NO, NO_2, $PM_{2.5}$, PM_{10}, T, RH	Relocated to Manglerud Rd.	3064 hrs

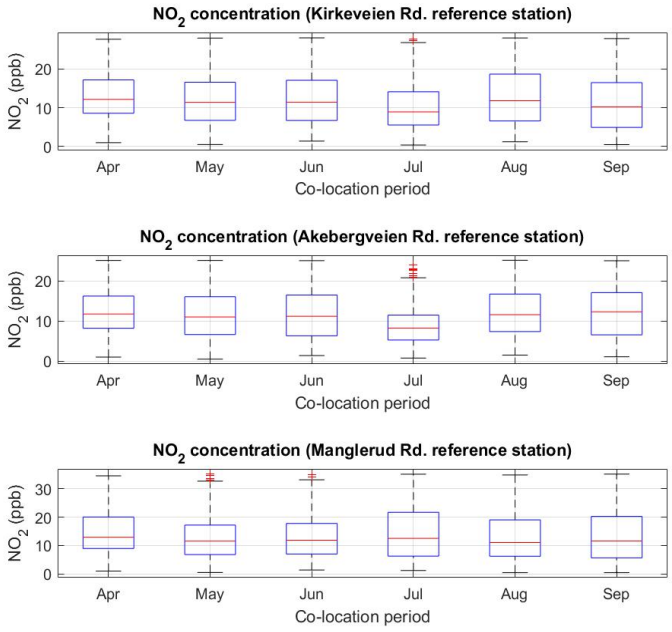
Table 1: Summary of the AQMesh nodes, along with their composition and positioning strategy along the dataset, used in the evaluations.

2.2. Motivation and Related Works

The motivation for performance drivers assessment in field calibration comes from earlier studies that revealed difficult to interpret. In a preliminary investigation, we calibrated AQMesh multi-sensor nodes (pods) using field sensors data and colocated regulatory grade station data recorded during the previously described campaign. Shallow neural networks were trained with these data to serve as a calibration function [13]. In particular, pods were calibrated against NO_2 using data recorded during their localization in Kirkeveien rd. (April 13rd to May, 31st) in Oslo. Performance were assessed using data recorded afterwise (4 10-days periods from June, 1st to Sept. 10) when a subset of the pods were relocated in different streets (Mangerud rd. and Akeberg rd.) in colocation with a different regulatory grade instrument. The obtained results have shown a complex picture that did not allow for a straightforward interpretation. Our investigations showed that, regardless of the position of the pods, a similar behaviour can be observed, i.e. a collective deterioration of the performances in mid-summer time. Actually, we observed that both the stationary and the relocated nodes suffered from a worsening of correlation index between sensor estimations and true concentrations of NO_2 as provided by co-located reference stations. Surprisingly, stationary pods, that stayed at Kirkeveien rd. for the whole data collection period, did not show the best relative performances. We observed that, during the Oslo summer weeks, the average temperatures were significantly higher than during the training phase (see Fig.2a,b). In the same period, NO_2 concentration median slightly decreased in Manglerud rd. and in Kierkeveien rd. sites (see Fig.2c,d,e), due to a decrease of anthropogenic emissions (i.e. car traffic) during summer (see [4]). During the last weeks of the deployment, temperatures lowered approaching the levels recorded during training phase and performance improved accordingly. We graphically compared the temperature changes with observed performances showing an interesting correlation among performance worsening and temperature increase [13].



(a,b)



(c,d,e)

Figure 2: (a,b) Box plot of hourly averaged environmental parameters (T, RH) recordings during the colocation period in OSLO (Blindern City official weather station data). The Pearson coefficient between the two variables computed along the entire deployment period was found to be negative (r = -0.36). (c,d,e) Box plot of hourly averaged NO2 concentrations recorded by the reference analyzers in the three relevant deployment site.

These preliminary analysis supported a primary role for natural and anthropogenic forcers distribution changes (concept drift) in determining the performance of calibrated nodes subjected to long term deployments and/or nodes relocation. Simple relocation, obviously, does not necessarily lead to performance degradation, unless there is a significant change of relevant conditions. These, in turn, depending on the deployment time scale, may prevail on sensors drift as a major cause of accuracy losses.

In fact, Esposito et al. [9] in 2016 (single site 5-months deployment, NO2 targeted

ANN model) and, more extensively, Hagan et al. [14] in 2018 (multisite 8-month deployment, SO2 targeted kNN model) have empirically shown that non-linear non-parametric calibration functions are challenged by extrapolation tasks, i.e. when they are used to estimate target concentrations outside the ***range*** encountered in calibration set. Casey and Hannigan reported different scales performance degradation of field calibrated Ozone and CO2 targeted multi-sensor devices when operated at different locations or different seasons with respect to calibration sites. Supported by their results, they argued how performances can be affected by the different ***composition*** of the pollution sources (car traffic in urban areas, oil and gas processing in industrial areas) occurring in different locations and by generalization capabilities of the concerned algorithm (linear or non-linear models) [17]. However, multiple regression approaches will typically embed[2] interferents response model and, hence, their performance will be also affected by (possibly multiple) ***interferents*** ranges mismatches as also suggested in [14]. As a result, we are no more concerned of target range mismatch alone but, more generally, of mismatch occurring between delimiting ***surfaces of calibration and operative manifolds*** for all forcers.

Furthermore, while adequately matching or just including the operative envelope, we may still fail to adequately cover the (hyper-)volume subsets depicting actual operative relationship among forcers values as a result we may not obtain sufficient knowledge of the shape of calibration function in that relevant volumes. Generalizing, a small number of training points located in strongly nonlinear regions of the multi-sensors response model may be insufficient for learning an accurate description of its true curvature. The resulting non parametric model will hence be also challenged by interpolation conditions: whenever operating conditions will limit the forcers distribution to a restricted nonlinear volume that is insufficiently represented in the training set, the model will be inaccurate and this will lead to poor predictions. The training manifold ***density*** is hence, locally relevant.

With this specific background, we propose to quantitatively assess the loss of performance of a calibration scheme specifically due to concept drift leveraging on *forcers' statistical distribution changes*.

In our opinion, multivariate distribution dissimilarity analysis will provide a more complete picture of the performance drivers helping improving robustness of field calibration procedures. More specifically, concerning training and operative conditions, similar multivariate forcers probability distribution functions should lead, to best predictions; large differences instead should lead, as a sufficient condition, to poor predictions. In our hypothesis, nonlinearities coverage and adequate range extension is the key to the selection of a calibration site (and time) that will lead to a robust calibration algorithm.

[2] either implicitly or explicitly

3. Methods

3.1. Distributions dissimilarity tools theoretical and applicative framework

In order to quantitatively check the relevance of the distribution similarity hypothesis, it is paramount to select a set of tools that can appropriately describe the similarities and summarize the differences among two probability distribution functions detecting the change. In this regard, tenths of different operators have been proposed in the last decades. A taxonomic attempt was made available by Cha in [34]. Among the most interesting operators are those belonging to *f-divergences [30,31].* The latter is a flexible family of non-negative measures of dissimilarities comparing the two probability distribution functions, say *P(x)* and *Q(x)*, along their entire support, specifically weighing their odds ratio by the use of a specific function *f*:

$$D_f(P \parallel Q) = \int_{\Omega} f\left(\frac{p(x)}{q(x)}\right) q(x)\, d\mu(x) \tag{1}$$

They are particularly suited for our purposes since they make no peculiar assumptions on original data distributions and hence can be used for forcers belonging to different and/or unspecified distribution.

Different choices for function *f* generates the different operators. Specifically, substituting

$$f(x) = xlog(x) \tag{2}$$

in (1) generates the *Kullback-Leibler divergence* (KL) operator:

$$D_{KL}(P||Q) = \int P(x)\log\left(\frac{P(x)}{Q(x)}\right)\mathrm{dx} \tag{3}$$

KL divergence provides an *asymmetric* measure of how one probability distribution differs from a second, expected probability distribution. Like any f-divergence, it is non-negative and take zero value only when *P(x)* equal *Q(x) almost everywhere*. KL divergence is usually considered as a measure of distance between probability distributions, though it does not exactly defines a geometrical *metric*. In fact, symmetry property and triangle inequality, mandatory requirements in this case, do not hold true for the KL divergence. More precisely, it defines a dissimilarity measure. It is widely used in information theory and machine learning, where it is used to compute the information gain achieved if *Q* distribution is used to estimate *P* distribution.

Hellinger distance (see [32]), *H(P,Q)* which is also an *f-divergence,* is generated by :

$$f(x) = \left(\sqrt{x} - 1\right)^2 \quad (4)$$

and is hence defined as:

$$H^2(P||Q) = \frac{1}{2}\int \left(\sqrt{P(x)} - \sqrt{Q(x)}\right)^2 dx \quad (5)$$

Differently from KL divergence, Hellinger divergence also defines a metric, properly accounting for a distance. It is bound to take value in [0, 1] range with 1 representing two completely different distributions. Hellinger distance have been previously proposed for tackling machine learning challenges when operating in dynamic environments; Ditzler and Polikar actually used univariate Hellinger distance between single feature distributions for their concept drift detector in nonstationary environments [32].

A Minkowsky L_p distance complete the set of analysed dissimilarity indicators. In particular, we use the classical Euclidean distance defined as

$$D_{Euc}(P||Q) = \int (P(x) - Q(x))^2 \, dx \quad (6)$$

Summarizing, three different measures of dissimilarity are taken into account. The first two (KL, Hellinger) differs from the facts that only the second properly accounts for a distance being also value-bound. Being asymmetric, KL divergence favour conditions in which *Q(x)* is wider than *P(x)* , Hellinger distance, instead treats the two distribution equally, penalizing all differences. Finally, we also exploited the simple Euclidean distance, which is not a divergence, but returns a geometrical difference assessment which is more sensible to localized differences. The combined usage of these three different mathematical operators (see fig. 3) will contribute to a more complete assessment of our hypothesis.

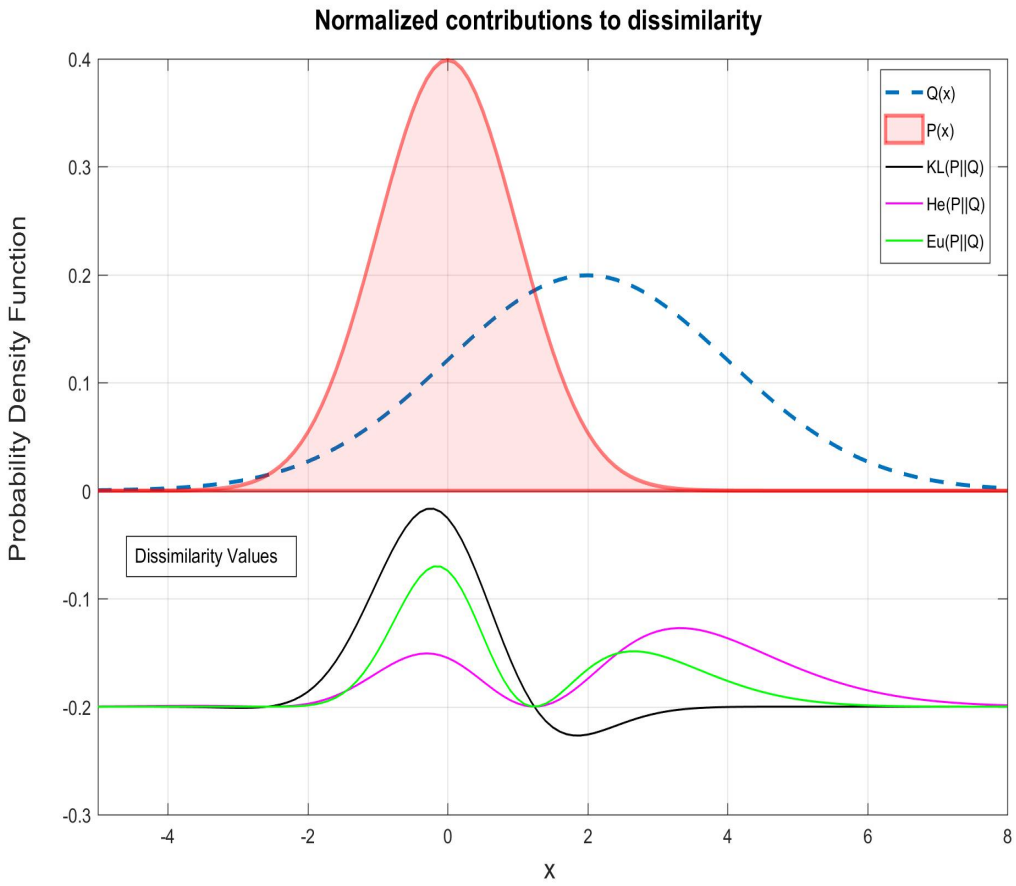


Figure 3: Integrand values of KL, Hellinger and Euclidean integral distance when comparing N(0,1) and N(2,2)gaussian probability density functions. Note that values have been biased and rescaled to improve picture readability and favour comparisons.

In the following, the relevant calibration time forcers probability distribution function substitute P in (3,5,6) and the test time one, in turn, substitutes Q. As an example, for KL divergence we will obtain:

$$d_y = D_{KL}(y_{Calibration}||y_{test}) \int_x^{\square} y_{Calibration}(x) log\left(\frac{y_{Calibration}(x)}{y_{test}(x)}\right) dx \quad (7)$$

with x being, using a simplified notation, a vector scanning the joint support of the calibration and test probability distribution of any relevant forcers subset, respectively $y_{Calibration}$ and y_{Test}. In this way, using KL divergence, we would actually compute the information lost when we approximate the actual forcers concentration distribution (e.g. NO_2 or T) as would be recorded on a test set using the one recorded on training set.

Obviously, for any multi-target/multi-sensors system, forcers set should encompass all target gases and all variables that are proven or suspected to be an interferent to any of the sensors response.

Given the dynamic nature or sensors systems, in certain cases, the derivative of a

variable may be considered among the relevant forcers. A specific example is given by temperature derivative that have been suggested as relevant in affecting electrochemical sensor responses [9,11]. Pang et al. [27], also observed short-term interferences by stepwise RH level transients, induced in laboratory conditions in electrochemical sensors. However, a complete analysis of all sensors known interferents impact is outside the scope of this work. Here, we restrict the analysis to dissimilarities of probability distributions over manifolds described by a subset of forcers. These include NO_2 as a target gas which reference readings are always available along the dataset, regardless of the location and the time, plus temperature. The latter, as above mentioned, is considered the main environmental interferent for Alphasense EC sensors under analysis. We currently do not include Ozone, a non target gas that proven a relevant interferent for the concerned NO_2 sensor.
Of course, comparing forcers joint probability distributions can provide for a more complete picture of the distributions changes. For these reasons, in our analysis, we also included joint probability distributions. In the following, we will hence substitute y in (6) with the discrete probability distribution function of NO_2 concentration, namely $p(NO_2)$, temperature, namely $p(T)$ and their joint probability distribution function, $p(NO_2,T)$.
In the next sections we will correlate the value of these dissimilarity indicators with nodes concentration estimation performance indices. In this way we want to show the dependence of the field calibration effectiveness on the differences between the multivariate forcers distribution on calibration and operative (test) manifolds.

3.2. Multisensors performance assessment procedure

In this contribution, we choose to calibrate the multi-sensor nodes using the Support Vector Regressor (SVR) as a prototype calibration function (see [15]). This structural risk minimization rooted machine learning architecture is characterized by good generalization properties and sparseness in knowledge representation [16,29]. SVRs also allow to reduce the uncertainty sources in the model training procedures associated to neural networks due to random weights initialization procedure.
According to the choice in the previous chapter, the multisensors nodes are specifically calibrated against NO2. NO_2 sensors, Temperature, and RH sensors readings are used as SVR inputs. Note that including environmental sensor readings allow for a correction of their influence on EC sensors response.
Concentration estimations performances are assessed through selecting multiple *calibration* and *test* sets along the entire deployment time. Precisely, for each pod, we divide the time domain into different timewise subsets of fixed length (180 hourly samples). Afterwards, we scan the pod dataset serially selecting two consecutives subsets for their use in the calibration procedure. As such, each calibration set

includes 360 samples (15 days). Two third of *calibration set* samples have been selected, by interleaving, for actual training (*training set*) while the other third (*validation set)* have been used for implementing an hyperparameters selection procedure. In particular, SVR model hyper parameters selection have been implemented by a *grid search* optimization procedure. Table 2 reports which hyperparameters have been actually taken into account, along with the corresponding spanned value range.

	Kernel Function	Box Constraint	Kernel Scale	Epsilon
Selected range	*Radial basis function*	*(2^0, 2^1, .., 2^5)*	*(2^0, 2^1, .., 2^5)*	*(0.1: 0.1: 3)*

Table 2: SVR hyperparameters optimization ranges.

As such, each training and optimization procedure required 6x6x30=1,080 evaluations accounting for the different hyperparameter combinations aiming to minimize errors on validation set.For each calibration set choice, we cycle through all the remaining timewise subsets, one at a time, selecting them for testing and final performance assessments (see Fig. 4); each test set is hence built up by 180 samples (about one week). In this way, we can explore the entire space of all possible relative combinations of times of the year and mutual location of the single pods during the calibration and respective test period. We are hence generating realistic conditions in terms of manifolds and joint probability distribution of forcers reflecting what could happen in real world conditions for both *fixed* and *mobile* applications. On average, every pod generated 90 different calibration/test set pairs with a minimum of 12 pairs, a maximum of 120 pairs and only one pod generating less than 90 pairs (pod #712150). Arithmetically, we computed a total number of 1,116,400 SVR training sessions.

Mean Absolute Error (MAE) has been selected as performance estimation index. MAE is one of the most relevant, and often selected, performance indices for air quality multisensors performance assessment, (see [7]). Here, it has been computed by averaging, along all of each test sets samples, the absolute estimation error, i.e. the absolute difference among hourly SVR estimations y_i and hourly reference concentrations $\hat{y}_i$ as provided by the co-located regulatory grade analyser:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{8}$$

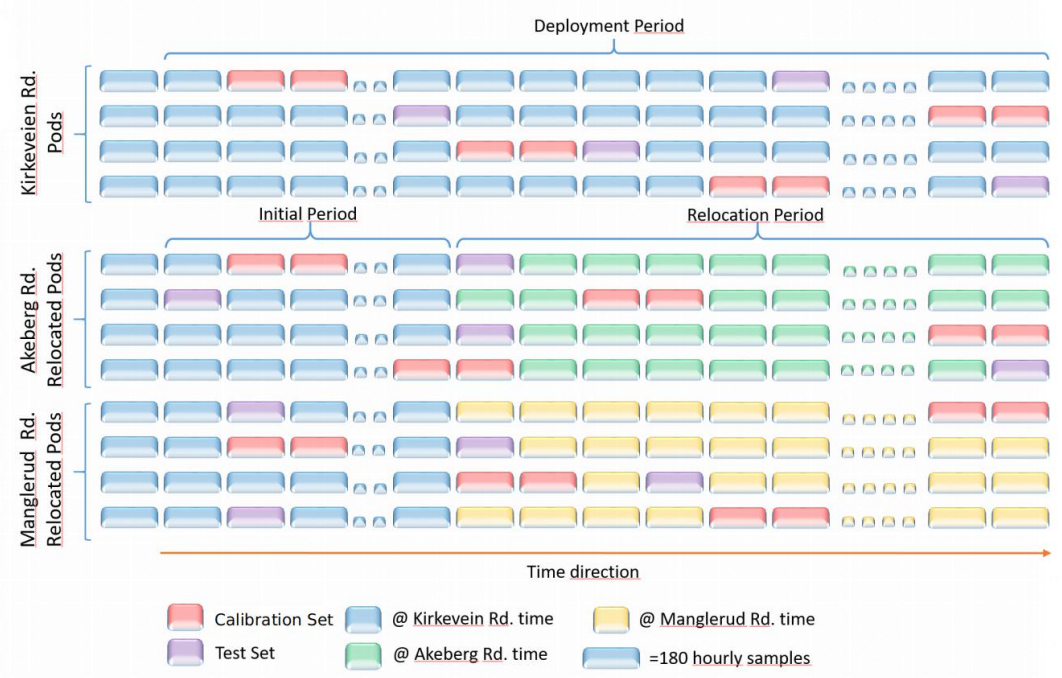


Figure 4: Graphical description of the dataset including pods locations and the calibration/test partitioning and combination procedure used for deriving SVR based calibration functions.

3.3 Distribution dissimilarity relevance testing procedure

After performance evaluation, a linear Pearson correlation index has been computed between the test MAE values associated to each *calibration-test* sets pair and the distribution dissimilarity indices evaluated on the same sets pair. The specific goal was, in fact, to show that the knowledge of the forcers distribution change between calibration and operative conditions is a good predictor of the operative performances of a field calibration procedure. More precisely, for each different calibration and test sets pairs, the MAE index obtained on test set samples, have been *normalized* by dividing it for the MAE value obtained during the respective calibration set, specifically, using its associated *validation subset*. This allowed a) to assess initial performance on a separate but ideally equally distributed set of data with respect to training data; b) to rule out, by normalization, the inherent performance diversity among sensor arrays as well as the one induced by outcomes of the training procedure. Hence, we are now assessing only the actual performance worsening (or improvement) when applying the learned calibration function to a different set of samples. Dissimilarity indexes were similarly obtained considering test set and the corresponding *validation* subset. In this way, the correlation coefficient among them, or more precisely its squared value R^2, has provided a tool to estimate how much of the performance variance could be linearly explained by the concept drifts effects accounted for by the chosen dissimilarity tool.

Among the introduced indices, the choice initially focused on evaluating the impact of the dissimilarity between the univariate statistical distribution of Temperature (T) interference occurring in calibration, and test sets. Slightly afterward we also checked for the sum of distances of T and the target gas (NO_2) concentration distributions

among the same sets. Finally, we computed the distance between the multivariate distributions (T, NO_2) occurring in validation and test sets.

4. Results and Discussion

4.1. Univariate dissimilarity impact assessment

Following the above mentioned experimental procedures, as a first step, we have fitted NO_2 empirical distributions computed for each of the designated training and test sets, with a Lognormal probability density function (*pdf*) [28]. The same procedure have been repeated for Temperature readings, this time using a Gaussian distribution model.

We have then numerically computed the first of the proposed *pdf* dissimilarity indices (i.e. Kullback-Leibler divergence) between each pair of validation and test sets fitted *pdf*s for both target gas and temperature for all pods. The dissimilarity index term obtained for NO_2, specifically $D_{KL}(NO_{2.Validation}||NO_{2.Test})$, have been added to the dissimilarity index obtained for temperature $D_{KL}(T_{Validation}||T_{Test})$ obtaining a composed term hereafter named SUM-D. Afterward, for each of the above mentioned *calibration-test* pairs, we have trained and optimized SVR machine using training set and, respectively, validation samples; MAE on NO_2 concentration estimations was then computed on the relevant test sets. Finally, the *normalized* MAE and dissimilarity indices, as computed on all the *calibration-test* pairs for each pod, have been correlated. Table 3, shows the Pearson correlation coefficients r values as obtained for each pod of all locations for the case of *KL* divergence. The obtained r values range from a minimum of 0.42 to a maximum of 0.84, averaging at 0.65, with respect to Temperature changes as captured by *KL* divergence while they ranges from 0.44 to 0.84, averaging at 0.66, when considering the sum of the *KL* obtained for both forcers. All of the computed r values have been tested for statistical significance at $\alpha = 0.01$ level. For all the sets pair, test results indicate that there is sufficient evidence of a statistically significant correlation among the variables under investigation.

Kirkeveien Rd. Stable Pods			Manglerud Rd. Relocated Pods			Akebergveien Rd. Relocated Pods		
#Pod	$r_{D(T)}$	$r_{SUM\text{-}D}$	*#Pod*	$r_{D(T)}$	$r_{SUM\text{-}D}$	*#Pod*	$r_{D(T)}$	$r_{SUM\text{-}D}$
715150	0.76	0.75	**718150**	0.80	0.80	**712150**	0.82	**0.84**
764150	0.55	0.53	**737150**	0.66	0.65	**743150**	0.73	0.84
785150	0.45	0.48	**751150**	0.67	0.65	**828150**	<u>0.42</u>	0.44
849150	0.63	0.62	**856150**	0.58	0.58	**850150**	0.72	0.72

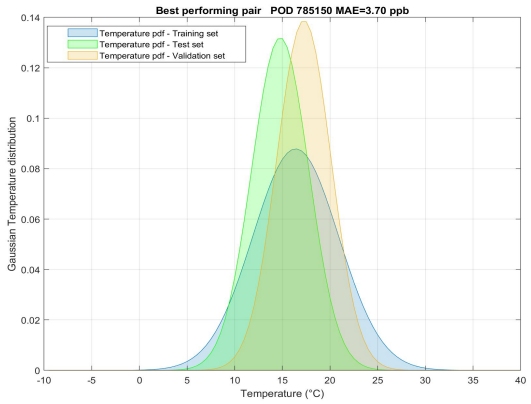
Table 3: Correlation indexes obtained among KL diversity indexes and normalized MAEs obtained for all Pods during the entire experimental campaign. In particular, $r_{D(T)}$ = Correlation between MAE and KL dissimilarity for temperatures; $r_{SUM\text{-}D}$ = correlation between MAE and the sum of KL dissimilarity for temperatures and target gas. The bold/underlined values highlights the maximum and minimum computed values, respectively.

Furthermore, the results show that except for two pods, the correlation among concept drift and performance worsening always exceeds 0.50. Hence, it can be observed that in all cases, the dissimilarity between the Temperature distributions and/or the sum of dissimilarities of T and NO_2, is a good predictor of the performances obtained in the calibration test set. However, the increase in correlation obtained by adding the target concentrations' distribution dissimilarity term was very small (if any). Actually in 5 pods, the correlation index slightly decreased; 3 pods showed equal results while the remaining 4 pods showed a limited increase. This is probably due to the relative stability of the target pollutant concentration, with respect to temperatures behaviour, during all the considered time interval.

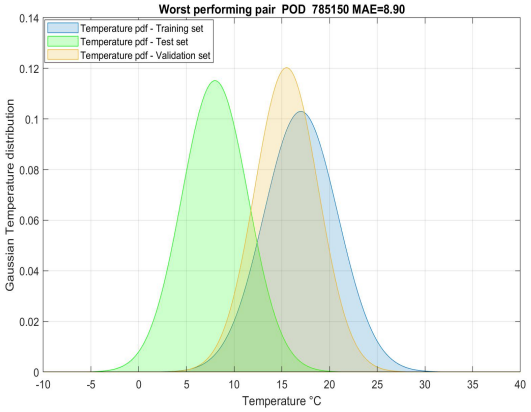
R^2 reached an average of 0.43 and 0.44 respectively, highlighting how the sum of univariate D_{KL} captured fractions of the considered concept drift is capable to explain more than 40% of the variance of the considered performance index.

Summarizing, by relying on the proposed methods, knowledge of the single forcers distributions during the calibration could be exploited to obtain quantitative insights about the calibration algorithm (normalized) performances in future operational conditions.

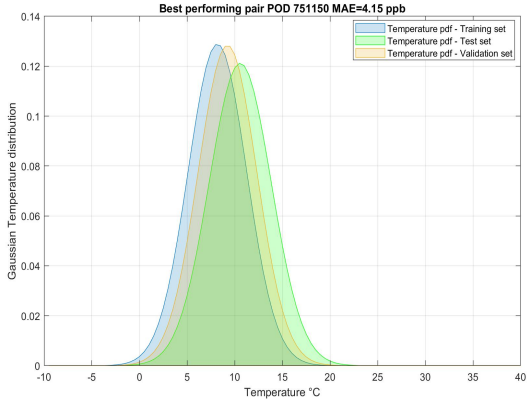
a



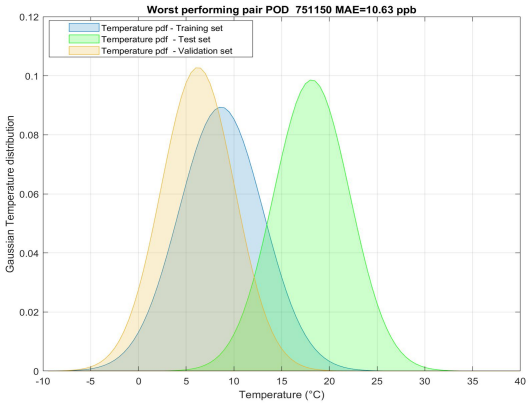
b



c



d



e

f

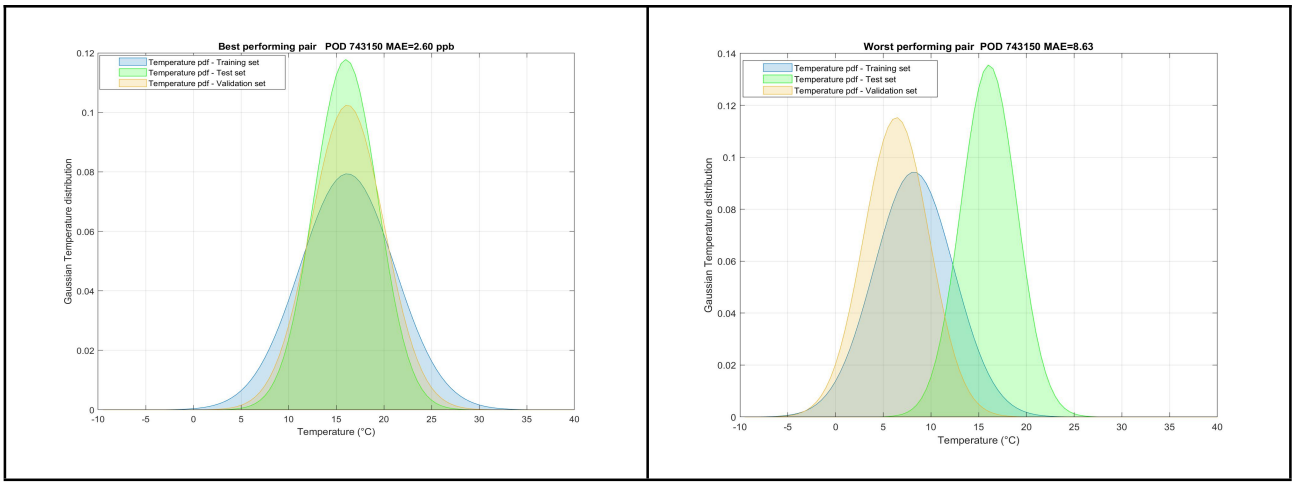


Figure 5: Normal (Gaussian) estimates of temperature probability distribution functions during training (blue), validation (orange) and test (green) as computed on the best performing calibration/test set pair (left column, lowest normalized MAE) and the worst performing pair (highest normalized MAE, right column), respectively recorded for the Kierkeveien rd. stable (a,b) pods, the Manglerud rd relocated pods (c,d) and the Akeberg relocated pods (e,f). Note how best performance are consistently obtained when temperature pdf as estimated on test set data are very similar to the one estimated on calibration set data (both training and validation). On the contrary, large performance degradation have been found to be consistently associated with significant temperature distribution dissimilarities.

Figure 5 helps to visualize how the dissimilarity in temperature distribution between calibration and test (operative) conditions is positively associated with performance degradation in single pods. In particular, pods have been partitioned by the three final locations. For each of the 4 pods belonging to each of the 3 partitions, calibration and test pairs were scanned searching for best and worst *normalized* MAE performance. Among the pods of each partition, the ones expressing the widest difference between best and worst case were selected. For these pods, the temperature *pdf* representative of calibration and test conditions and belonging to the best and worst performing pair, were depicted respectively on the left and right column of fig. 5. It is evident how different environmental conditions shown by relative position (μ) and shape (σ) of the temperature normal distributions during train and test are reflected by the worsening of performance indices. Worst cases actually show a very different distribution of temperatures between calibration and operative phase that make the learnt models to work outside the calibration manifold causing inaccurate estimations.

4.2. Multivariate dissimilarity impact assessment

In an attempt to take into account as much as possible of the properties of the forcers' sub-space in which the pods are calibrated and operated, we repeated the process computing dissimilarities of joint multivariate distributions. In particular, we computed the bivariate empirical joint distribution of temperature and NO_2 concentrations, $p(T, NO_2)$, for each *calibration/test* pair, more specifically for the validation subset and for the test set. Afterwise, the *Euclidean*, *KL* and *Hellinger*

dissimilarity indices between them, have been computed. Table 4 shows the correlation coefficient between normalized MAE and corresponding multivariate dissimilarity measures, for each pod considering all relative conditions of time and locations. All of the computed values have been positively tested for statistical significance under α=0.01 conditions.

Kirkeveien Rd. Stable Pods				Manglerud Rd. Relocated Pods				Akebergveien Rd. Relocated Pods			
#Pod	r_{Eu}	r_{He}	r_{KL}	#Pod	r_{Eu}	r_{He}	r_{KL}	#Pod	r_{Eu}	r_{He}	r_{KL}
715150	0.80	0.79	0.72	718150	0.81	**0.84**	080	712150	0.79	0.74	0.55
764150	0.57	0.56	0.50	737150	0.64	0.70	0.68	743150	0.83	0.76	0.72
785150	0.57	0.62	0.55	751150	0.66	0.71	0.68	828150	0.57	0.46	0.40
849150	0.67	0.68	0.62	856150	0.57	0.64	0.64	850150	0.75	0.76	0.70

Table 4: Correlation Coefficient (r) of dissimilarity indexes between normalized MAE and joint empirical probability distributions p(T, NO_2). In particular, rEu = Correlation between normalized MAE and Euclidean distance; rHe = correlation between normalized MAE and Hellinger distance; rKL = correlation between normalized MAE and Kullback-Leibler dissimilarity. The bold/underlined values highlights the maximum and minimum computed values, respectively.

In particular, the correlation coefficient ***r*** averaged 0.69 for both Hellinger and Euclidean distance while averaging 0.65 for KL divergence. R^2, averaged 0.48 for Euclidean distance, 0.47 for Hellinger distance and 0.43 for the KL one. That means that according to the first two indices the forcers subset distribution change is capable to linearly explain almost 50% of the chosen performance index variance. This improves the results reached by either univariate or multivariate assessments based on D_{KL}. Empirically, Euclidean and Hellinger dissimilarity measures seams capable to slightly better explain the observed performance variance with respect to KL divergence.

A qualitative evaluation of the relationship among the performance and forcers distribution may improve our understanding beyond simple ***r*** assessments. Since Hellinger and Euclidean distances show best performances and an extremely similar average correlation with a slight advantage for Euclidean distance, we choose to focus on the latter. Actually, Figure 6 (a,b,c) shows a scatter plot based depiction of the relationship among the value of *normalized* MAE and Euclidean distance between joint (bivariate) empirical probability distributions *p(T, NO_2)* in calibration and operative phase, for all the considered nodes. All the *s*catter plots highlight a significant to good correlation (0.57<r<0.83). It is also noticeable how unknown performance drivers, ostensibly not captured by our model, hampers a full linear prediction capability. Specifically, the shape of the depicted cloud points consistently shows that, when low dissimilarity values are concerned, the recorded dissimilarity is not able to satisfactorily explain most of the observed performance variance. In these conditions, performance may be dominated by irreducible uncertainty of the node or unconsidered forcers. Correlation values significantly improve when growing dissimilarity values are reflected in a consistent increase in calibration error, as captured by normalized MAE. This confirms the existence of a sufficient condition

whereby, if the dissimilarity between the training and operative forcers distribution is sufficiently high then this condition will be invariably associated with significant performance degradation. Furthermore, the higher the dissimilarity the higher the average estimation error will be. This relationship show to be capturable by a linear regression. Comparing the results obtained by different nodes we could see that linear prediction bias ranged from 0.75 to 0.95 while angular coefficient ranges from 0.075 to 2.5, in half of the cases, though, they are confined in the [0.13, 0.16] range.
Finally, Figure 7 shows a summary of the obtained correlation indices, using the different distance measures, for the different pods.

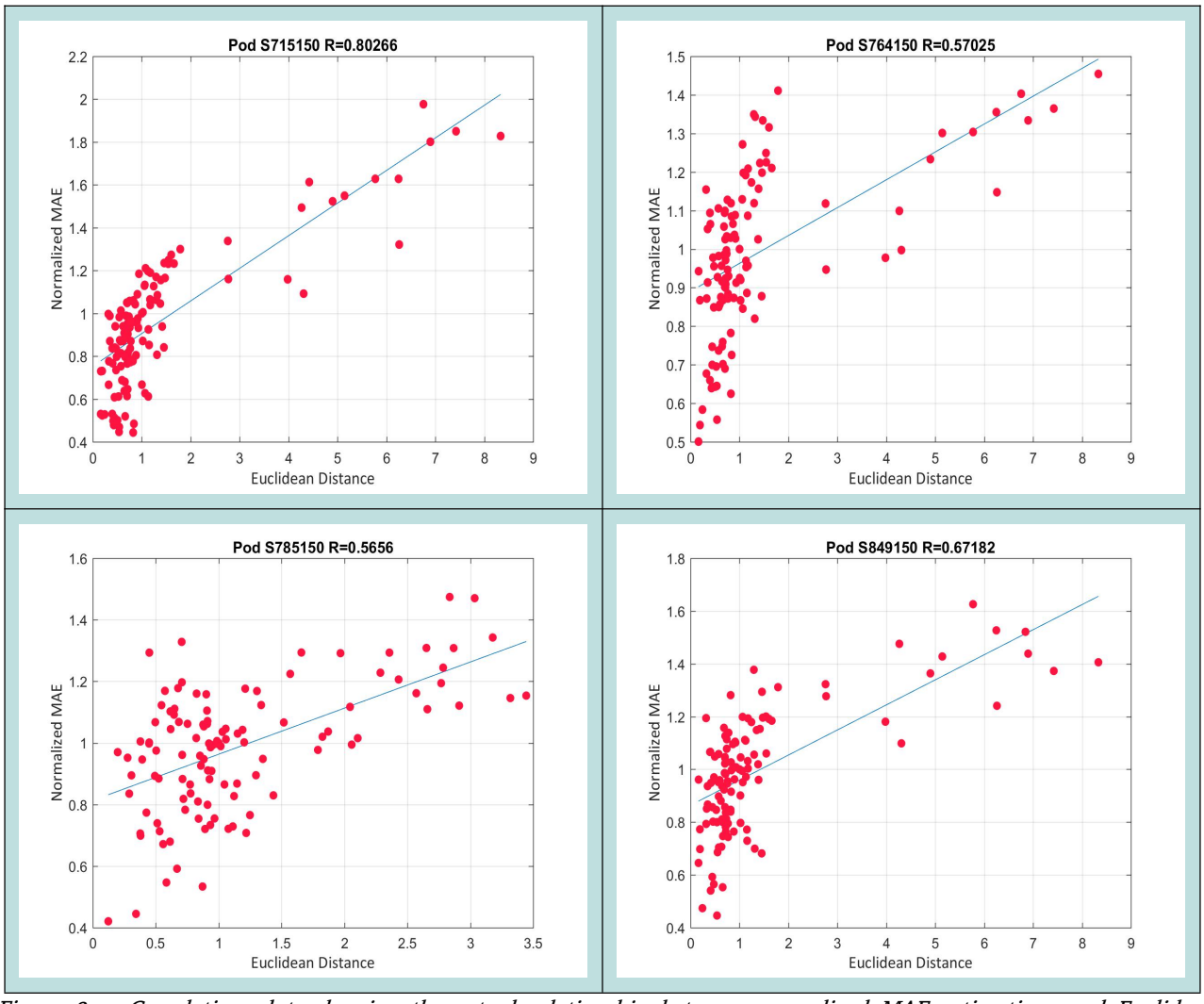


Figure 6a : Correlation plots showing the actual relationship between normalized MAE estimations and Euclidean distance applied to joint empirical distribution p(T, NO2) for the 4 stable pods.

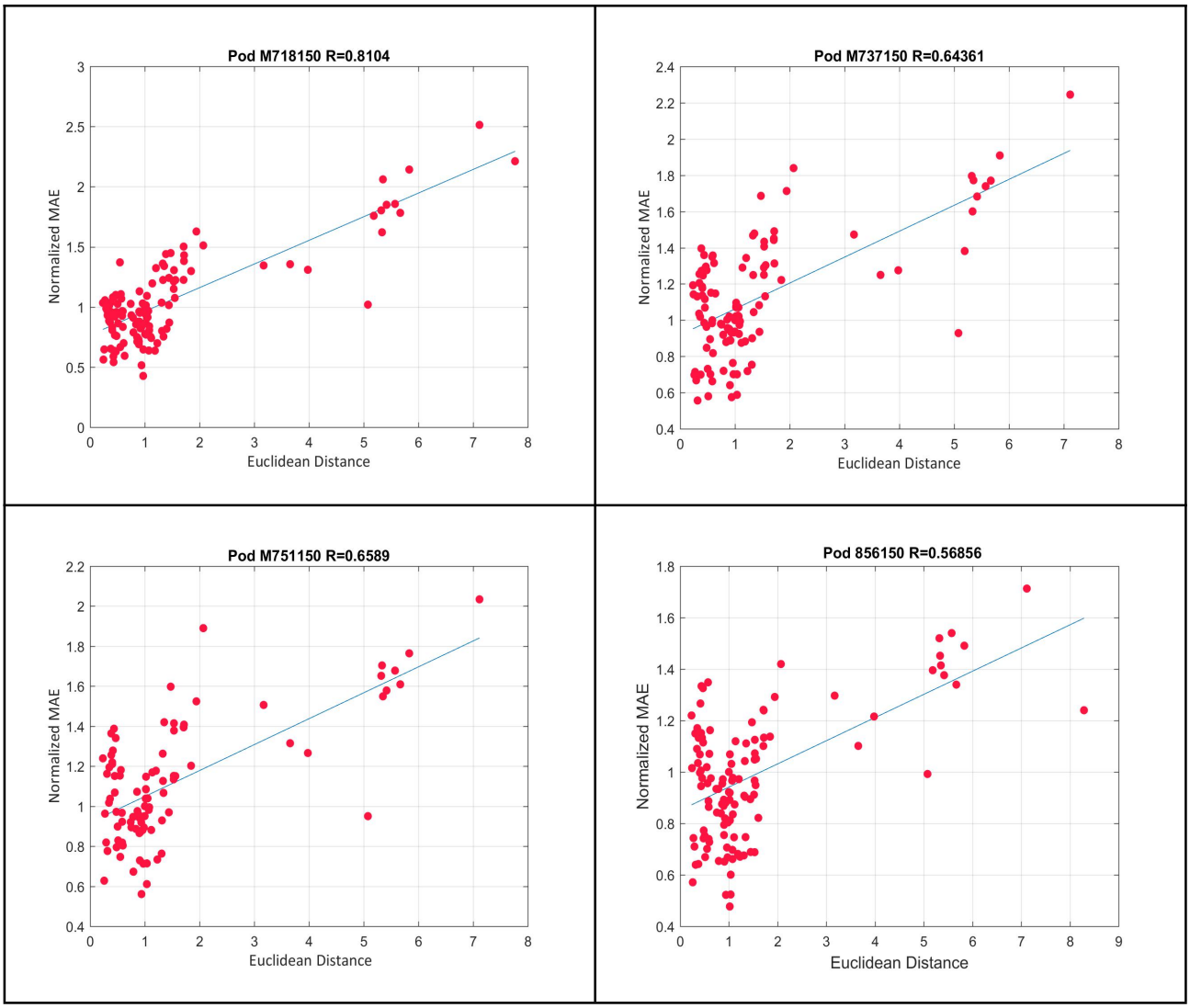


Figure 6b : Correlation plots showing the actual relationship between normalized MAE estimation nad Euclidean distance applied to joint empirical distribution p(T, NO2) for the 4 pods relocated in Manglerud rd.

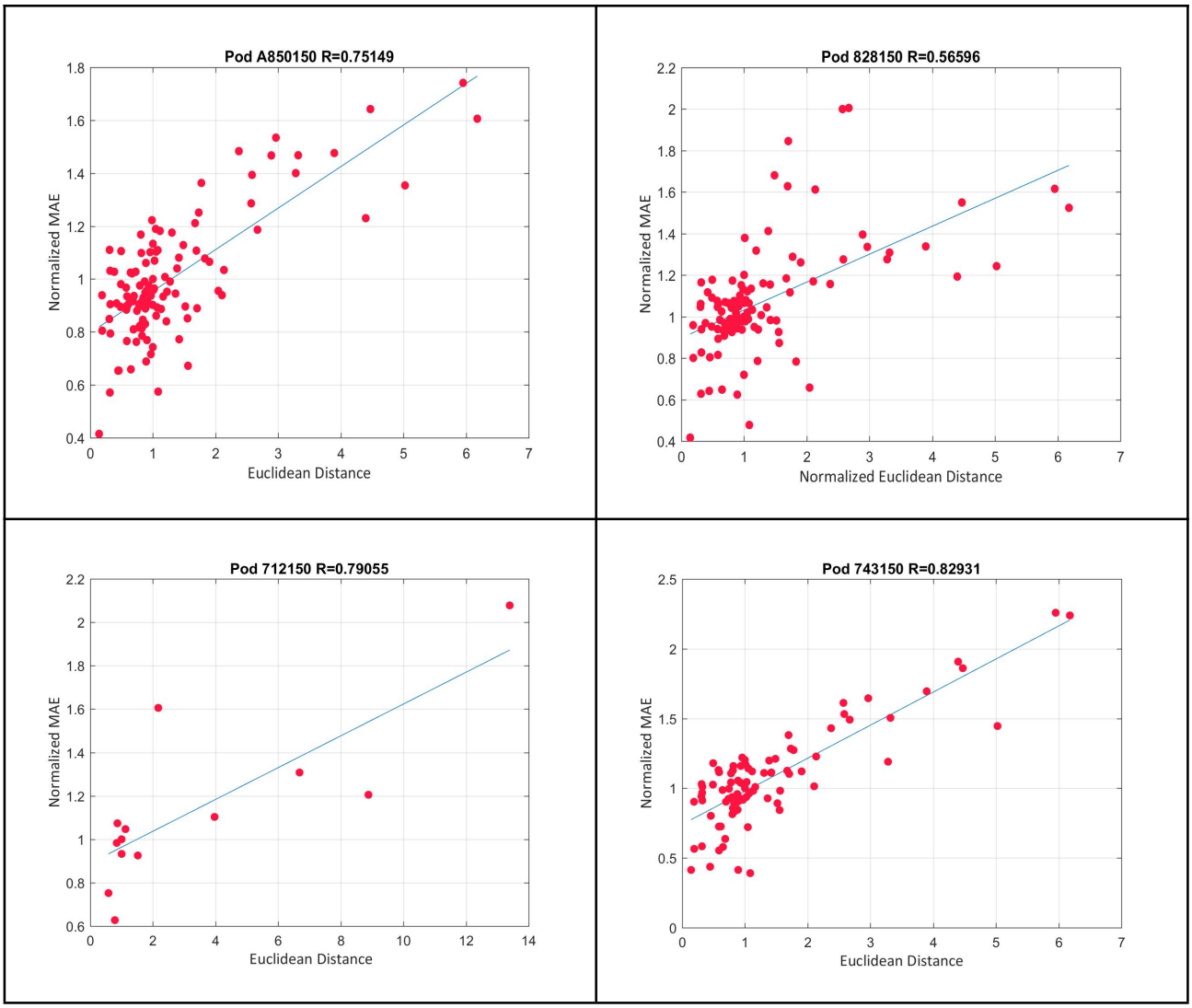


Figure 6c : Correlation plots showing the actual relationship between normalized MAE estimation and Euclidean distance applied to joint empirical distribution p(T, NO2) for the 4 pods relocated in Akebergveien rd.

Figure 7: Summary plot of the correlation results obtained using different distribution dissimilarity indexes - Multivariate distributions in greenish colors (rKL - KL Dissimilarity, rHe-Hellinger Distance, rEu-Euclidean Distance), Univariate distribution results in bluish colors (rSUM-D - sum of KL Dissimilarity for Temperature and Target gas concentration, rD(T) - KL Dissimilarity for Temperature) .

In particular, the greenish bars show the correlation coefficient r of *KL* divergence, Hellinger distance and Euclidean distance computed between the bivariate distribution and the normalized calibration error. Bluish bars indicate computed correlation coefficient r for the univariate case. As it can be seen, all the considered distances show a very similar behaviour, except in the case of POD no. 828150 where Euclidean bivariate value significantly differs from the value obtained by the other indicators. In general, a slightly better performance is expressed by Hellinger and Euclidean distance in the bivariate case. We can summarize these results as confirming that the knowledge of the manifold in which our sensors have been calibrated, including the distribution of at least the target (NO_2) and temperature interferent, allows to generally forecast the field calibration performances for the concerned gas multi-sensors devices in operative conditions. Furthermore, distribution dissimilarity indices seem to be capable to linearly explain significant fractions of the performance variance. These results have been obtained in very different conditions characterized by different locations and different weeks of the year, with multi-sensor platforms that were relocated or operated in the same locations in which they have been trained. Generalizing these findings, forcers' distribution changes including targets and non-targets/environmental interferents could be used to predict performance issues for field calibrated multi-sensors. In some cases, and, in

particular for two of the twelve analysed multi-sensors, results clearly indicate the presence of unknown performance drivers that, being not taken into account, generates outliers results.

5. Conclusions and Remarks

In this work, we analyse the subject of *in field calibration* robustness by pursuing a way to objectively identify and quantify the main drivers of performance degradation for field calibrated smart air quality monitors. In particular, we proposed the use of *pdf* dissimilarities indices to predict performance losses occurring when they are forced to operate in different places and, more exactly, in different environmental, non-target and target gases concentrations conditions, with respect to those encountered during calibration function learning. Our results are obtained with measurements of twelve multisensory units, deployed over several months and in different locations. They show that widely adopted performance indicators like MAE are significantly correlated with changes occurring in forcers' probability distribution between calibration and operation phase. Actually, distributions dissimilarity indices seems capable to numerically capture and so quantify these changes. In particular, when high values of distributions dissimilarity were recorded, significant performance worsening can be expected. These findings highlight the significance of the role of concept drift in determining the performance of field calibrated environmental multi-sensors platforms. Until recently, this role was often underestimated with respect to the role universally recognized to sensors drift. Given the enhanced role of concept drifts, these findings may also stimulate further research tackling semi-supervised learning strategies for adaptive drift correction.

In general, to which fraction the concept drift may explain the observed lack of robustness of field calibration, may depend on multiple factors including size of the calibration dataset, the current amount of sensors ageing and generalization properties of the calibration algorithm itself. These effects may negatively affect linear correlation, however normalization procedures may help to reduce their impact.

These results can be exploited in several ways. Just as an example, practitioners may use the performance prediction capability to identify when a new calibration is needed. If concept drift is consistently detected in available forcers, or sensors, distributions for a significant period of time then a new calibration can be autonomously requested by the pod. Furthermore, the results can be used to invalidate data when needed. This will avoid using bad quality data in the subsequent steps of the data processing path such as for exposome monitoring or for high resolution pollutant concentrations mapping. Most importantly, these insights can be furthermore explored to define criteria for the selection of field calibration sites and timing aiming to match operative and calibration conditions. Ultimately, we should aim to sites and time of the year in which we could, preferably in short times, obtain a forcers measurement dataset spanning a volume wide enough to

include a sufficient range of final operative conditions and dense enough to effectively describe the nonlinearity regions of the calibration function support. Complete datasets, in this sense, will reduce or rule out the occurrence of dissimilarity and their relative impacts. As a consequence, the recorded performance will be entirely explained by sensors precision.
As a final remark, further work is needed to better identify the causes of the fraction of performance changes that appear not to be captured by the proposed methodology. Enlarging the number of known interferents (e.g. in our case, Ozone) that are taken into account may improve the explanatory capabilities of our model though unknown interferents will still represent a threat to them.

Author Contributions

N. Castell, P. Schneider and A. Bartonova devised and supervised the dataset acquisition including multisensors location, deployment, operation. P. Schneider and E. Esposito pre-processed and prepared the acquired dataset for the experimental work. S. De Vito and E. Esposito devised the theoretical framework and the concept drift assessment experimental design. E. Esposito and S. De Vito designed and implemented the machine learning components for which results have been collected. S. De Vito and A. Bartonova supervised the entire work.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

S. De Vito and E. Esposito have received funding by European Union through Flag-ERA JTC 2016 call Convergence project and by EU-European Regional Development Fund through UIA (Urban Innovation Actions) 3rd Call Air-Heritage project. A. Bartonova, P. Schneider and N. Castell have received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 308524. Finally, the authors wish to dedicate this work to the beloved memory of Vienna D' Auria.

6. References

[1] Castell N, Liu H-Y, Schneider P, Cole-Hunter T, Lahoz W, Bartonova A. Towards a personalized environmental health information service using low-cost sensors and crowdsourcing. 2015. EGU General Assembly.

[2] Steinle, S.; Reis, S.; Sabel, C.E. Quantifying human exposure to air pollution—Moving from static monitoring to spatio-temporally resolved personal exposure assessment. Sci. Total Environ. 2013, 443, 184–193.

[3] Jiang XQ, Mei XD, Feng D. Air pollution and chronic airway diseases: what should people know and do?. *J Thorac Dis*. 2016;8(1): E31-40.

[4] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?", Environment International, 99, (2017), 293-302.

[5] Philipp Schneider, Nuria Castell, Matthias Vogt, Franck R. Dauge, William A. Lahoz, Alena Bartonova, Mapping urban air quality in near real-time using observations from low-cost sensors and model information, Environment International, Volume 106, 2017, Pages 234-247, ISSN 0160-4120.

[6] A. C. Lewis, P. Edwards, Validate personal air-pollution sensors, *Nature*, 535, (2016), 29-31.

[7] C. Borrego, J. Ginja, M. Coutinho, C. Ribeiro, K. Karatzas, Th Sioumis, N. Katsifarakis, K. Konstantinidis, S. De Vito, E. Esposito, M. Salvato, P. Smith, N. Andr, P. Grard, L.A. Francis, N. Castell, P. Schneider, M. Viana, M.C. Minguilln, W. Reimringer, R.P. Otjes, O. von Sicard, R. Pohle, B. Elen, D. Suriano, V. Pfister, M. Prato, S. Dipinto, M. Penza, Assessment of air quality microsensors versus reference methods: The EuNetAir Joint Exercise – Part II, Atmospheric Environment, Volume 193, 2018, Pages 127-142, ISSN 1352-2310, https://doi.org/10.1016/j.atmosenv.2018.08.028.

[8] S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia, On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical, Volume 129, Issue 2, 2008, Pages 750-757.

[9] 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, Sensors and Actuators B: Chemical, Volume 231, 2016, Pages 701-713.

[10] Zimmerman, N., Presto, A. A., Kumar, S. P. N., Gu, J., Hauryliuk, A., Robinson, E. S., Robinson, A. L., and R. Subramanian: A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmos. Meas. Tech., 11, (2018), 291-313.

[11] Cross, E. S., Williams, L. R., Lewis, D. K., Magoon, G. R., Onasch, T. B., Kaminsky, M. L., Worsnop, D. R., and Jayne, J. T.: Use of electrochemical sensors for measurement of air pollution: correcting interference response and validating measurements, Atmos. Meas. Tech., 10, (2017), 3575-3588.

[12] Laurent Spinelle, Michel Gerboles, Maria Gabriella Villani, Manuel Aleixandre, Fausto Bonavitacola, Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. Part B: NO, CO and CO2, Sensors and Actuators B: Chemical, Volume 238, 2017, Pages 706-715.

[13] E. Esposito *et al.*, "Is on field calibration strategy robust to relocation?," *2017 ISOCS/IEEE International Symposium on Olfaction and Electronic Nose (ISOEN)*, Montreal, QC, 2017, pp. 1-3.do i: 10.1109/ISOEN.2017.7968904.

[14] Hagan, D. H., Isaacman-VanWertz, G., Franklin, J. P., Wallace, L. M. M., Kocar, B. D., Heald, C. L., and Kroll, J. H.: Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments, Atmos. Meas. Tech., 11, (2018), 315-328.

[15] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, G. Di Francia, Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches, Sensors and Actuators B: Chemical, Volume 255, Part 2, 2018, Pages 1191-1210.

[16] Vapnik, V. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[17] J. G. Casey, M. P. Hannigan, Testing the performance of field calibration techniques for low-cost gas sensors in new deployment locations: across a county line and across Colorado,Atmos. Meas. Tech., 11, 6351-6378, 2018.

[18] Nicola Masey, Jonathan Gillespie, Eliani Ezani, Chun Lin, Hao Wu, Neil S. Ferguson, Scott Hamilton, Mathew R. Heal, Iain J. Beverland, Temporal changes in field calibration relationships for Aeroqual S500 O3 and NO2 sensor-based monitors, Sensors and Actuators B: Chemical, Volume 273, 2018, Pages 1800-1806.

[19] McCloskey, M. & Cohen, N. (1989), Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (ed.) *The Psychology of Learning and Motivation*,*24*, 109-164.

[20] Ditzler, Gregory & Roveri, Manuel & Alippi, Cesare & Polikar, Robi. (2015). Learning in Nonstationary Environments: A Survey. Computational Intelligence Magazine, IEEE. 10. 12-25. 10.1109/MCI.2015.2471196.

[21] Alphasense Website - www.alphasense.com - Last accessed in May 2019.

[22] José María Cordero, Rafael Borge, Adolfo Narros, Using statistical methods to carry out in field calibrations of low cost air quality sensors, Sensors and Actuators B: Chemical, Volume 267, 2018, Pages 245-254, ISSN 0925-4005, https://doi.org/10.1016/j.snb.2018.04.021.

[23] AQMesh website - www.aqmesh.com - Last accessed in May 2019.

[24] Landrigan et al., The lancet commission on pollutant and health, The Lancet, 2018; 39:462-512.

[25] WHO. Review of Evidence on Health Aspects of Air Pollution—REVIHAAP Project; Technical Report; World Health Organisation: Copenhagen, Denmark, 2013..

[26] De Vito, S.; Esposito, E.; Formisano, F.; Massera, E.; Fiore, S.; Fattoruso, G.; Salvato, M.; Buonanno, A.; Veneri, P.D.; Francia, G.D. Enabling Citizen Science with A Crowdfunded and Field Validated Smart Air Quality Monitor. Proceedings 2018, 2, 932.

[27] Pang, X., Shaw, M. D., Lewis, A. C., Carpenter, L. J., and Batchel-lier, T.: Electrochemical ozone sensors: a miniaturised alter-native for ozone measurements in laboratory experiments andair-quality monitoring, Sensor. Actuat. B-Chem., 240, 829–837,https://doi.org/10.1016/j.snb.2016.09.020, 2017.

[28] K. E. Bencala, John H. Seinfeld, On frequency distributions of air pollutant concentrations, Atmospheric Environment (1967), Volume 10, Issue 11, 1976, Pages 941-950, ISSN 0004-6981, https://doi.org/10.1016/0004-6981(76)90200-6.

[29] B. Scholkopf and A.J. Smola, Learning with Kernels, The MIT Press; 1st edition (December 15, 2001)

[30] Kullback, S.; Leibler, R. A. On Information and Sufficiency. Ann. Math. Statist. 22 (1951), no. 1, 79--86. doi:10.1214/aoms/1177729694. https://projecteuclid.org/euclid.aoms/1177729694

[31] Liese, F.; Vajda, I. (2006). "On divergences and informations in statistics and information theory". *IEEE Transactions on Information Theory*. **52** (10): 4394–4412. doi:10.1109/TIT.2006.881731

[32] Pollard, David E. (2002). *A user's guide to measure theoretic probability*. Cambridge, UK: Cambridge University Press. ISBN 0-521-00289-3.

[33] G. Ditzler and R. Polikar, "Hellinger distance based drift detection for nonstationary environments," *2011 IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments (CIDUE)*, Paris, 2011, pp. 41-48. doi: 10.1109/CIDUE.2011.5948491

[34] Cha, Sung-Hyuk (2007) *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions.* International Journal of Mathematical models and Methods in Applied Sciences, 1 (4). pp. 300-307.