KICS The Korean Institute of Communications and Information Sciences

Available online at www.sciencedirect.com

ScienceDirect

ICT Express 6 (2020) 220–228

www.elsevier.com/locate/icte

Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data Fusion and Machine Learning Approach

Nwamaka U. Okafor[a,*], Yahia Alghorani[b], Declan T. Delaney[a]

[a] *School of Electrical and Electronic Engineering, University College Dublin, Ireland*
[b] *Department of Electrical Engineering, Lakehead University Ontario, Canada*

Received 28 February 2020; received in revised form 9 June 2020; accepted 22 June 2020
Available online 3 July 2020

Abstract

Environmental monitoring has become an active research area due to the current rise in the global climate change crises. Current environmental monitoring solutions, however, are characterized by high cost of acquisition and complexity of installation; often requiring extensive resources, infrastructure and expertise. It is infeasible to achieve with these solutions, high density in-situ networks such as are required to build refined scale models to facilitate robust monitoring, thus, leaving large gaps within the collected dataset. Low-Cost Sensors (LCS) can offer high-resolution spatiotemporal measurements which could be used to supplement existing dataset from current environmental monitoring solutions. LCS however, require frequent calibration in order to provide accurate and reliable data as they are often affected by environmental conditions when deployed on the field. Calibrating LCS can help to improve their data quality and ensure they are collecting accurate data. Achieving effective calibration, however, requires identifying factors that affect sensor's data quality for a given measurement. This study evaluates the performance of three Feature Selection (FS) algorithms including Forward Feature Selection (FFS), Backward Elimination (BE) and Exhaustive Feature Selection (EFS) in identifying factors that affect data quality of low-cost IoT sensors in environmental monitoring networks. Applying the concept of data fusion, sensors data were merged with environmental factors and integrated into a single calibration equation to calibrate cairclipO_3/NO_2 and cairclipNO_2 sensors using Linear Regression (LR) and Artificial Neural Networks (ANN). The study showed the effectiveness of calibration in improving low-cost IoT sensor data quality and also demonstrated the convenience of feature selection and the ability of data fusion to provide more consistent, accurate and reliable information for calibration models. The analysis showed that the cairclipO_3/NO_2 sensor provided measurements that have good correlation with reference measurements whereas the cairclipNO_2 sensor showed no reasonable correlation with the reference data. Calibrating the cairclipO_3/NO_2 yielded good improvement in its measurement outputs when compared to reference measurements (R^2=0.83). However, calibrating the cairclipNO_2 sensor data yielded no significant improvement in its data quality.

Keywords: Internet of Things (ioT); Machine learning; Data fusion; Feature selection; Environmental monitoring

1. Introduction

Advances in sensor networks and Internet of Things (IoT) technologies have created new epoch in environmental monitoring [1,2], facilitating the collection of high-resolution spatiotemporal dataset and filling the gaps that existed within current dataset [3]. Although, IoT technology presents plausible tools to expand current capacity in environmental monitoring [4], the introduction of Low-Cost Sensors (LCS) is critical to have wider scope and adoption for this purpose [5]. The application of LCS in environmental monitoring, however, has raised several concerns especially pertaining to their accuracy, reliability, in-field applicability and performance. LCS are less precise and less sensitive to compound or variables of interest as their response is largely influenced by cross-sensitivities in the case of gas sensors, particle properties as with particulate matter sensors or environmental factors in both cases [6], amounting to trade-offs when being used to replace or to supplement existing monitoring solutions.

* Corresponding author.
E-mail addresses: nwamaka.okafor@ucdconnect.ie (N.U. Okafor), yahia.alghorani@ieee.org (Y. Alghorani), declan.delaney@ucd.ie (D.T. Delaney).

Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2020.06.004

Calibrating LCS can help to improve their data quality and field performance and ensure they are collecting accurate values [7]. Manufacturers of IoT sensors often calibrate the sensors before releasing them to the market, the calibration, however, is usually done under controlled laboratory conditions which does not represent the exact conditions faced by the sensors when deployed on the field. Hence, sensors may continue to report inaccurate values on the field even after being calibrated in the laboratory. Automatic/In-field sensor calibration has been suggested as a means to reduce these challenges [8]. Machine Learning (ML) techniques can be used to calibrate LCS on the field based on influencing environmental factors [9–11]. To effectively accomplish this, it is essential to

- determine the factors that affect data quality for the given sensor measurements
- Model the effects of these factors on sensor's response and
- Apply the model to correct sensor output errors

In this paper, we examine the performance of three feature selection algorithms for the selection of best subset of features suitable for building accurate calibration models for LCS in an environmental monitoring network . We apply statistical data fusion technique to merge and integrate selected features into calibration models using Linear Regression and Neural Network techniques. The developed models were then employed to calibrate cairclipO_3/NO_2 and cairclipNO_2 sensors and the performance of each of the model for this purpose was compared against the rest. Although, the models were designed for calibrating LCS in environmental monitoring networks, the tools and techniques, can be applied to calibrate sensors used in other domains. The contributions of this paper include to:

(i) compare the performance of Forward Feature Selection (FFS), Backward Elimination (BE) and Exhaustive Feature Selection (EFS) methods in determining the factors that affect IoT sensor data quality in environmental monitoring networks.
(ii) develop a process based on error modelling and sensor data fusion for in-field calibration of low-cost IoT sensors.
(iii) present reliable technique for error correction and data quality improvement of LCS in environmental monitoring systems.

The remaining parts of this paper are structured as follows, Section 2 details the current state of the art in IoT-based environmental monitoring. Section 3 presents the motivation for this work and Section 4 describes the dataset used in this study. Section 5 deals with error modelling for sensor calibration, outlining the techniques for feature selection, data fusion and data quality improvement in LCS. Results and performance evaluation are presented in Section 6 and Section 7 has the conclusion.

2. State of the art

Several studies have been conducted in the area of IoT-based environmental monitoring systems and showed significant benefits. Syafrudin et al. and Manes et al. utilized LCS to examine the environmental conditions of real sites [12,13]. The results from their studies demonstrated the practicability and effectiveness of the sensors in providing real-time environmental data. Although IoT sensors can provide high-resolution spatio-temporal dataset, this, however, may be subject to a short time period as low-cost sensors have exhibited limited applications in long term monitoring [14]. Current studies have also shown that IoT sensors are sensitive to environmental conditions [15]. Calibrating LCS can help to address these issues, improve the data quality of the sensors and ensure that they are collecting accurate values. Simple Linear Regression (SLR) method has been suggested for use in calibrating IoT sensors. Badura et al. suggested this method for calibrating PMS7003 Particulate Matter (PM2.5) sensor [16]. Although this approach was sufficient in relating the raw sensor output to reference measurements, it was not able to capture the temporal variability of sensor measurements with respect to environmental conditions. In a bid to overcome this shortcoming, they applied Multi Linear Regression (MLR) in fitting of the sensor's data to the reference measurement with the inclusion of environmental parameters; Temperature (T) and Relative Humidity (RH) in the calibration equation. Their work established the impact of high RH on the sensors output and showed T to have moderate correlation with RH.

Multi-parameter regression models and supervised machine learning methods have been applied for calibrating sensor devices. Munir et al. developed MLR models and Generalized Additive Models (GAM) in [17] for the calibration of the Environ watch E-MOTES for capturing concentrations of Nitrous Oxide (NO) and Nitrogen Dioxide (NO_2) for a one-year collocation period with a reference monitor. Their additive model was applicable to both normal and non-normal data distribution and does not assume a linear relationship between dependent (reference data) variables and explanatory variables. Although, various goodness of fit indicators upon which the models were evaluated, showed good agreement of the models results with reference measurements, the absence of previous work using similar sensors, however, does not allow for qualitative evaluation of this model especially in comparison to results from previous studies. Yamamoto et al. proposed a machine learning-based method for calibrating temperature sensor using Artificial Neural Network (ANN) [18]. Their investigation was based on a one year dataset collected from three locations in Japan using the SHT-71 sensors which measures air temperature and relative humidity, these sensors were collocated with a reference monitor (Automated Meteorological Data Acquisition System-AMeDAS), developed by the Japan Meteorological Agency. Abrupt changes in environmental conditions caused several calibration errors during their study. Their work showed the effect of using data from widely separated locations for sensor calibration, highlighting the impact of environmental differences between the sites. Zimmerman et al. investigated three calibration approaches including laboratory calibration, MLR and Random forest calibration techniques for the RAMP sensor package which measures CO, NO_2, O_3 and CO_2 in [19]. Their work investigated the accuracy of the models across different concentration ranges and also examined the importance of model variables.

Multiple variables included in a multi-parameter calibration model can help improve the performance of the model and produce more accurate results [20,21]. Integrating data from different sensor nodes into a calibration model through the process of data fusion can also produce more consistent, accurate and useful information for the model in predicting the target variable far beyond what an individual sensor can provide [22]. Including many irrelevant parameters, however, can make the calibration equation more complex and difficult to interpret [23]. Excluding redundant variables from a calibration models can improve the model's accuracy and simplify the signal processing and data acquisition processes. Stepwise regression has been suggested as a viable strategy for feature selection when building multi-parameter models [23]. The Best Subset Regression (BSR) technique was proposed in [24] for selecting the best subset of features/predictors for calibration models.

While there have been significant research in the individual fields of IoT, Machine Learning (ML), data fusion and sensor calibration, there is currently no study on the aggregation of these technologies into a complete solution in environmental monitoring, where significant improvement in data quality of LCS can be achieved by the integration of data fusion and machine learning techniques for error modelling and IoT sensor calibration.

3. Motivation

Achieving high spatial density ground-based coverage in environmental monitoring networks is essential to provide additional dataset that could be used for validating remote and satellite-based monitoring. LCS play significant role in achieving high density in-situ monitoring networks. The major challenge of using LCS for this purpose, however, is the high chance of recruiting erroneous data. Identifying and eliminating these errors are important for the adoption of LCS for environmental monitoring purposes. The processes involved in calibrating sensors in order to eliminate sensor response errors and ensure collection of quality data are costly, cumbersome and time consuming. Several environmental factors may affect sensor outputs, it is therefore necessary to identify and account for the effects of these variables on the sensor's output. Furthermore, effective feature selection is essential to make the learning task of calibration model more efficient and accurate while also facilitating the modelling of sensor output errors. The task of selecting variables with high influence on a model's predictive power is quite challenging particularly in small sized samples [25]. The approach adopted in this study involves identifying environmental features which affect LCS outputs using FS, BE and EFS and modelling sensor errors based on identified features. This method facilitates sensor calibration and makes the process of data quality improvement more efficient.

4. Dataset description and processing

The dataset used in this study was presented by Duvall et al. in [26], it consists of measurement concentration of Ozone (O_3) and Nitrogen Dioxide (NO_2) which were collected using CairclipO_3/NO_2 and CairclipNO_2 sensors in Houston Texas. The cairclipO_3/NO_2 sensor provides the sum concentration of O_3 and NO_2 while the cairclipNO_2 sensor measures only NO_2. The sensors were collocated with Federal Reference Monitors (FRM) within the period of 4–27 September 2013. O_3 measurement was obtained using ethylene-chemiluminescence FRM with a Bendix Model 8002 analyzer. NO_2 was measured by a gas-phase chemiluminescence FRM using a Teledyne Model T200U analyzer (Teledyne API; San Diego, CA, USA). Further details regarding the sensors and FRM can be obtained in [26]. Values of O_3 were obtained from the cairclipO_3/NO_2 sensor by subtracting the NO_2 values from the cairclipNO_2 sensor closest to the cairclipO_3/NO_2 sensor following the procedure described by the manufacturer in [27]. Although the cairclipO_3/NO_2 sensor was designed to measure the sum concentration of O_3 and NO_2, previous studies have identified the sensor to exhibit less sensitivity to NO_2 [28]. The separate O_3 data obtained from the cairclipO_3/NO_2 sensor showed good agreement with the data from the O_3 reference instrument (Pearson correlation coefficient r=0.82). The NO_2 data from the cairclipNO_2 sensor however showed low agreement with reference NO_2 data(r=0.08).

All implementation in this study was completed in python 3 on a Jupyter notebook available with Anaconda distribution. Additional library packages were installed including mlxtend which promotes convenience in model implementation, Keras [29] and TensorFlow [30] used for Neural Network implementation. This study was conducted using data from six CairclipO_3/NO_2 sensor (S_1, S_2, S_3, S_4, S_5, S_6), four CairclipNO_2 sensors (N_1,N_2, N_3, N_4), data from O_3 and NO_2 reference instruments as well as data from Temperature (T) and Relative Humidity (RH) sensors. Sensors S_1 and N_1 along with T and RH were collocated on the same location with the FRMs and the sensors were placed on the roof of the sampling trailer near the inlet of the FRM analyzers. The other sensors were operated by citizen scientists in schools within the vicinity. A correlation analysis performed on the collected data showed good correlation between most of the O_3 sensors and the O_3 reference data, with no correlation observed between data from the different O_3 sensor nodes. All the NO_2 sensors showed no correlation with each other as well as with the reference data. For all the sensors and reference monitors, data was collected every minute and was averaged into hourly measurements. Initial evaluation of the dataset showed few missing values, missing data were handled using mean imputation which allowed for the replacement of each missing data point with the mean of the observed values. A total of 576 data points were obtained and used in this study. 70% of the dataset was used for training the model and 30% was reserved for testing the model. The testing data being the most recent part of the dataset. For all the models, their performances were evaluated on the testing dataset which was not used in training the model.

Table 1
Performance comparison of FFS, BE and EFS feature selection models.

		FFS	BE	EFS
O_3	Total no of features	8	8	8
	No of features selected	6	6	6
	R^2	0.806	0.808	0.977
	Selected features	S1, S3, S4, S5, S6, T	S1, S2, S4, S6, T, RH	S1, S2, S4, S5, S6, RH
NO_2	Total no of features	6	6	6
	No of features selected	5	5	5
	R^2	0.513	0.513	0.936
	Selected features	N1, N3, N4, T, H	N1, N3, N4, T, RH	N1, N3, N4, T, RH

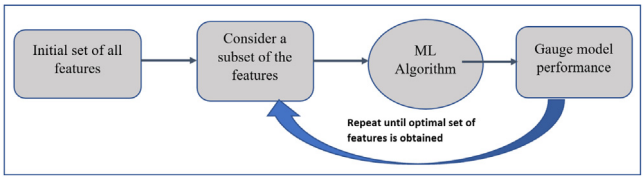


Fig. 1. General feature selection procedure.

5. Error modelling for sensor calibration

Previous studies have identified the weakness of SLR models in capturing the temporal variability of sensor response in relation to environmental conditions [31]. To achieve effective calibration in a network of low-cost IoT sensors, it is essential to model the error rate of the sensors by comparing sensors observations to observations from reference instrument. This process would be more effective if the effects of environmental factors such as temperature and relative humidity on the sensor outputs can adequately be accounted for. In this study, factors that impact concentration estimation for O_3 and NO_2 are identified through a feature selection method. The process of data fusion is then employed to merge identified influencing factors into a calibration equation to correct errors in LCS to improve the data quality of the sensors.

The inclusion of insignificant features in calibration models can lead to complexity, overfitting and low performance of the model. It is therefore necessary to ensure that only features with strong correlation to the target variable are included in the calibration equation. A general method for feature selection is illustrated in Fig. 1.

5.1. Feature selection

The performance of three methods of feature selection including Forward Feature Selection (FFS), Backward Elimination(BE) and Exhaustive Feature Selection (EFS) were evaluated to determine their ability to select the best subset of features necessary for the quantification of O_3 and NO_2 concentrations using cairclipO_3/NO_2 and cairclipNO_2 sensors respectively. The system is able to determine the impacts of environmental variables T, RH on the estimation of the target gasses. For each of the method, a random forest regressor was used to find the optimal features by evaluating all possible combination of the features(sensor nodes and environmental factors) and selecting the combination that produces the best results for estimating the true values of the target gases. In all the three methods, the number of trees (n_estimators) were set to 100 and n_jobs=-1. A 4-fold cross validation was performed on both FFS and BE, however, no cross validation was performed on EFS. A grid search cross validation (GridSearchCV) performed showed these parameters to be the best for the models, hence, the reason for choosing them for model implementation.

R^2 score was used as the scoring metric for evaluating the performance of the feature selection models. R^2 is the coefficient of determination and it measures the strength of the linear relationship between the predicted and observed values. Table 1 presents a comparison of all three feature selection methods. For O_3 estimation, the total number of available features is 8 and each of FFS, BE and EFS algorithm was set to select 6 optimal performing subset of features from the whole feature set. The list of feature selected by each method is shown in Table 1. From the analysis, EFS outperformed the rest of the methods with R^2 score of 0.977 against 0.808 and 0.806 for BE and FFS respectively. Also, for NO_2 estimation, the EFS performed best with R^2 of 0.936 outperforming both BE and FFS which both had R^2 of 0.513. The feature selection process is important for the task aimed to be completed in this work as it facilitates efficient selection of nodes and environmental factors with significant contributions to the target gasses estimation. It eliminates redundancy by excluding from the model, variables that are highly correlated to other variables, helping the calibration model to train faster, reducing computational cost and complexity, thereby promoting convenience in the overall calibration processes. Subsequently, we apply data fusion technique to merge the features selected by EFS (the best performing feature selection method in this study) to build linear regression and artificial neural network calibration models for cairclipO_3/NO_2 and cairclipNO_2 sensors.

5.2. Data fusion

To ensure provision of more consistent and accurate information for the calibration model, data fusion technique was applied to enable the merging of data from different sensor nodes into the calibration equation used for model implementation. Here, we present an objective means of combining observations from different nodes through a linear estimation method to provide more useful information for the calibration model. This approach would not only help to close any gap

existing on the dataset of individual sensor but would facilitate the provision of detailed spatial pattern of observations far beyond what a single sensor can provide. Using data fusion, we first combine the data from all available sensors including data from the temperature and relative humidity sensors to estimate the values for O_3 and NO_2 using observations from the FRMs as reference. Finally, the best subset of features needed for the estimation of the gasses as identified by EFS feature selection model were applied in a separate equation to model and correct sensor errors. The estimated concentration ($\hat{Y}$) at a particular time (t_0) is computed using Eq. (1)

$$\hat{Y}(t_0) = \alpha + \beta_1 x_1(t_0) + \beta_2 x_2(t_0) + \cdots \beta_n x_n(t_0) + \epsilon(t_0) \quad (1)$$

*w*here α is the intercept, β_1 to β_n are the regression coefficients, x_1 to x_n are the values of the predictor variables and ϵ is a constant error term.

5.3. Sensor calibration

The cairclipO_3/NO_2 sensors used in this study for O_3 measurements and the cairclipNO_2 sensors for NO_2 measurements were calibrated using SLR, MLR and ANN models. The SLR model used only one independent variable/feature (i.e. measurements from only one sensor), using this method, each sensor was calibrated separately to correct the bias in their outputs. With the MLR and ANN models, the sensors were calibrated using:

(i) All available features.
(ii) subset of feature selected by EFS.

5.4. Simple Linear Regression (SLR)

Out of the six O_3 sensors examined, three of them showed sufficient correlation (r>0.6) with the reference measurements however, all the NO_2 sensors showed poor correlation(r<0.1) with the reference data. SLR model was used to calibrate each individual sensors to correct their measurement errors. During calibration the sensor measurements were regressed against reference measurements using Eqs. (2) and (3) for O_3 and NO_2 sensors respectively. This process was able to reduce the average RMSE of the O_3 sensors from 28.72 ppb to 20.65 ppb.

$$O_3\_Reference = \alpha + \beta_1 S_1 + \epsilon \quad (2)$$

$$NO_2\_Reference = \alpha + \beta_1 N_1 + \epsilon \quad (3)$$

O_3_Reference and NO_2_Reference are concentrations from reference monitors, S_1 and N_1 are values of O_3 and NO_2 concentrations from one of the cairclipO_3/NO_2 and cairclipNO_2 sensors respectively.

5.5. Multiple Linear Regression (MLR)

To account for the effect of multiple variables on the sensors measurements, MLR was used to calibrate the sensors. The process of data fusion was applied to merge data from different sensor nodes, including environmental data as described in Section 5.2 into a single calibration equation, providing more useful information for the model in estimating the target gasses. The relationship between the variables is described in Eqs. (4) and (5).

$$O_3\_Reference = \alpha + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 S_4 + \beta_5 S_5 + \beta_6 S_6 + \beta_7(Temp) + \beta_8(RH) + \epsilon \quad (4)$$

$$NO_2\_Reference = \alpha + \beta_1 N_1 + \beta_2 N_2 + \beta_3 N_3 + \ + \beta_4 N_4 + \beta_5(Temp) + \beta_6(RH) + \epsilon \quad (5)$$

The measurement unit of all sensors and FRMs are in part per billion(ppb).

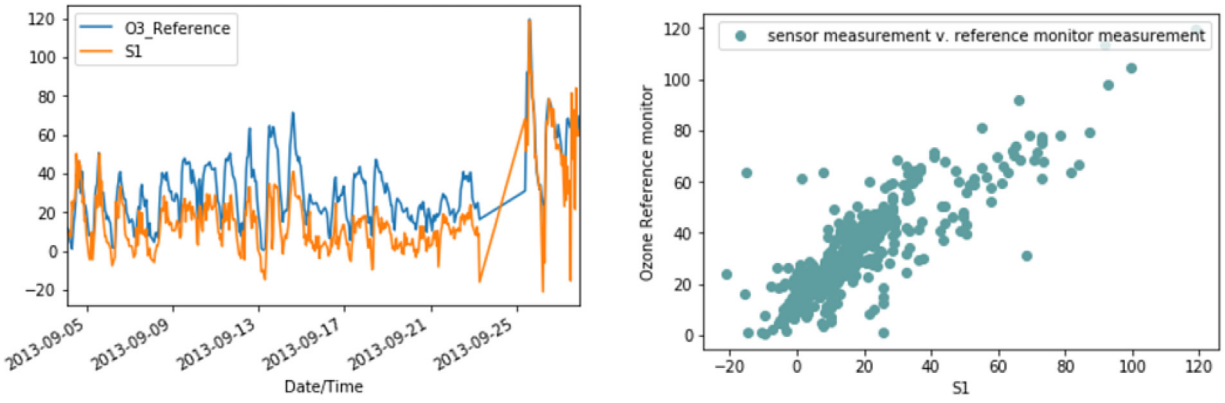
The sensors were first calibrated using all available features, then subset of features selected by the EFS feature selection algorithm was also used. The sensors data, T and RH data were used as the predictor variables and the FRM data were used as reference. Calibrating the sensors using this approach yielded better agreement between the sensor and reference data, reducing the Mean Absolute Error (MAE) between the reference and sensor data by 7.15 ppb.

5.6. Artificial Neural Network (ANN)

A three-layer back propagation ANN model was implemented and used for the calibration of the sensors. The model was built using all available features as well as features selected by the EFS model as input variables. For O_3 measurements, the calibration model had 8 inputs (S_1...S_6, temperature, RH data), six hidden layers (each with 13 neurons) and 1 output layer (FRM measurements). This architecture was chosen after evaluation of different architectures mostly involving n hidden layers (n=3, 6, 9, 12) and different training rounds. The network architecture with n=6 and trained for 250 epochs achieved the best performance and was therefore choosing for this study. Adam was used as the optimizer, the activation function used was ReLu, mean squared error was used as the loss function and the learning rate was set to 0.003. Calibrating the cairclipO_3/NO_2 sensor using this method yielded a good improvement in the data quality of the sensor, however, calibrating the caiclipNO_2 sensor yielded no significant improvement in the result.

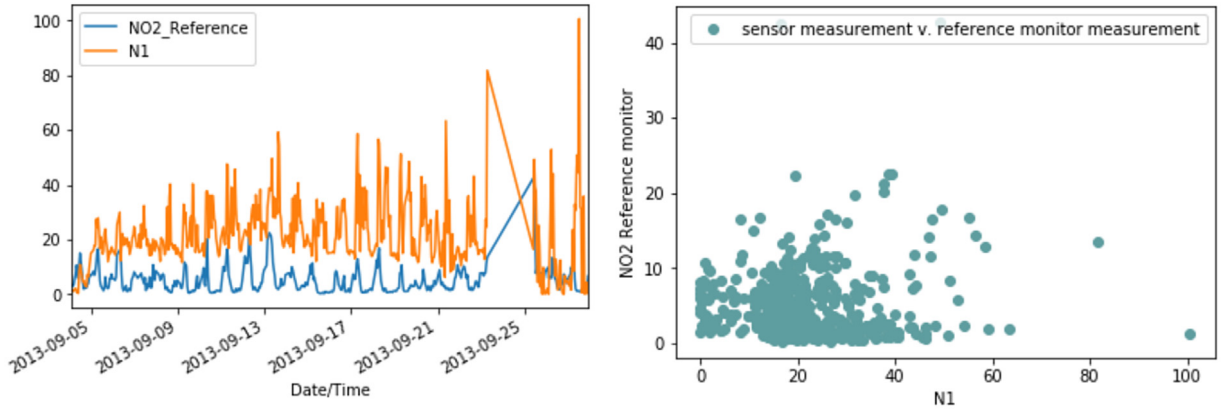
5.7. Evaluation metrics

To evaluate the performance of the calibration models, the sensors measurements before and after calibration were compared to reference measurements using three goodness of fit indicators; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and coefficient of determination (R^2). MAE measures the average magnitude of the errors in a set of prediction without considering their direction, in this study, we take MAE as the average over the test dataset of the absolute differences between the predicted and observed values. RMSE provides a measure of the model error by calculating the distance between predicted values and the actual values. A



(a) (b)

Fig. 2. O_3 Reference and sensor (a) Time series measurement (b) Correlation plot.



(a) (b)

Fig. 3. NO_2 reference and sensor (a) Time series measurement (b) Correlation plot.

lower value of both MAE and RMSE indicates good model performance. While MAE is a high performing metric for evaluation, we have also used RMSE in this study as it turns out to be useful in this case where large errors are undesirable in the calibration models. R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable. It is the square of the correlation coefficient (r), where r is the strength of the linear relationship between the observed and predicted values. Usually in the range of 0 to 1, R^2 closer to 1 reflects good agreement with reference measurements whereas small R^2 indicates poor fitting quality.

Table 2
Comparison of cairclipO_3/NO_2 sensor outputs with reference measurements when the sensor is uncalibrated (S_1Raw), calibrated using simple linear regression method S_1(SLR), calibrated using multiple linear regression (S_n(MLR) and Neural Networks (S_n(ANN) with all available features(All) and features selected by EFS.

	S_1(Raw)	S_1(SLR)	S_n(MLR)		S_n(ANN)	
			ALL	EFS	ALL	EFS
Intercept		18.17	120.5	31.6		
MAE	16.18	9.03	10.65	7.81	9.01	7.19
RMSE	18.57	12.58	12.93	10.24	12.26	9.69
R^2	0.03	0.71	0.69	0.81	0.71	0.83

6. Results and performance evaluation

Figs. 2 and 3 present the time series data and correlation plot between the sensors and reference data for O_3 and NO_2 measurements respectively. While the O_3 measurements from the sensor have shown good correlation with reference data, NO_2 sensor measurements showed poor correlation to reference data.

To test the performance of the calibration models, the models were applied to the testing dataset that were not used for model fitting. The concentrations of O_3 and NO_2 from the sensors were compared to data from their respective reference monitors. A comparison between the performance of the different models when used to calibrate the O_3 data is presented in Table 2. The result showed a good improvement in the correlation between the raw sensor data and reference data after calibrating the sensor data using SLR method. The process reduced the MAE existing between the sensor and reference data by more than 40% and increased R^2 from 0.027 to 0.706. The MLR calibration model yielded an improvement to this result, however, the ANN model outperformed both the MLR and SLR models. The result also showed that the MLR and ANN models which were built using the features selected

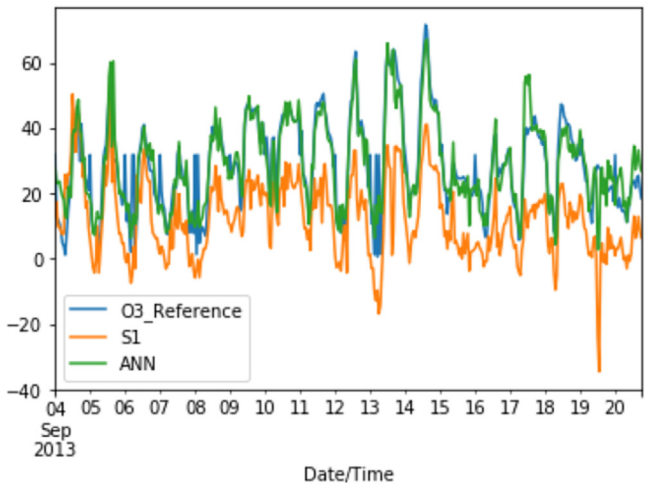


Fig. 4. Time series of hourly average data of O_3 sensor (S1) and reference measurements (O_3_Reference) before calibration and after calibration with ANN.

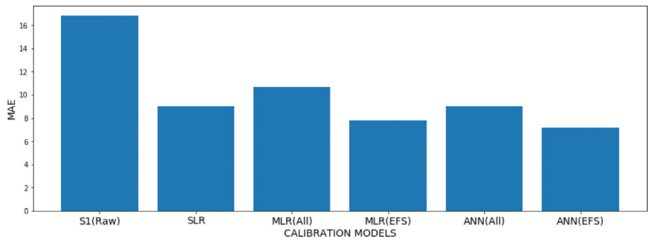


Fig. 5. MAE of O_3 sensor and reference measurements when the sensor is uncalibrated (S1Raw), calibrated using SLR method (SLR), calibrated using MLR with all available features (MLR(All)) and features selected by EFS (MLR(EFS), calibrated using ANN(All) and ANN(EFS).

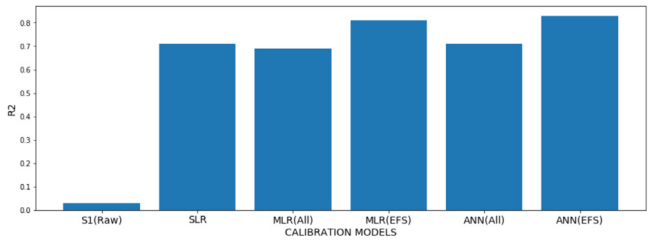


Fig. 6. R^2 of O_3 sensor and reference measurements when the sensor is uncalibrated (S1Raw), calibrated using SLR method (SLR), calibrated using MLR with all available features (MLR(All)) and features selected by EFS (MLR(EFS), calibrated using ANN(All) and ANN(EFS).

by the EFS feature selection algorithm yielded better results than models developed using all available features. The result presented by this study not only showed the effectiveness of sensor calibration in improving the data quality of low-cost IoT sensors and the high performance of artificial neural network for this purpose but it also showed the importance and convenience of feature selection in in-field calibration of sensors in environmental monitoring networks.

In Fig. 4, we present the time series of the O_3 reference measurement with the raw and ANN calibrated sensor outputs. The plot shows how the calibration process was able to close the gap in the error existing between the sensor and reference measurements. The bar plots of MAE and R^2 between the reference and sensor raw data, SLR, MLR and ANN calibrated data are presented in Figs. 5 and 6 respectively. The figures show the effect of the various calibration methods in improving the data quality of the sensors against the reference data.

The analysis shows that the MAE between the raw (uncalibrated) cairclipO_3/NO_2 sensor output and FRM output is significantly high, showing the limitation of the sensor in accurately capturing O_3 concentration, this error was halved when a SLR model was used to calibrate the sensor. An improvement to the SLR model was observed when MLR calibration was applied to calibrate the sensor outputs. The ANN model built using the features selected by EFS as predictors, however, gave the best result. Observe also in Fig. 6, the low R^2 value existing between the uncalibrated sensor data and reference data and the improvement in this value when the various calibration models were applied, signifying the improvement in agreement between the sensor and reference data after the calibration process.

7. Conclusion

Low-cost IoT sensors have the capacity to contribute to real time environmental monitoring, providing high resolution spatiotemporal dataset. Low-cost sensors, however, require frequent calibration when deployed on the field to ensure collection of accurate data. Machine learning methods such as linear regression and neural networks can be employed for calibrating low-cost sensors; adjusting the sensors measurements to compare to concentrations from reference monitors. Low-cost sensors can be affected by environmental factors such as temperature and humidity. It is therefore necessary to include these factors in the calibration model to account for their effects on the sensors response in order to ensure effective calibration. Including several insignificant variables in a calibration equation, however, will increase its computational complexity and reduce the accuracy of the calibration model. Therefore, it is important to identify the environmental factors that affect the data quality of sensor for a given measurement, and then apply these factors in developing the calibration model. In this study, three methods of feature selection including forward selection method, backward elimination and exhaustive feature selection method were applied in determining factors that affect the cairclipO_3/NO_2 sensors for measuring O_3 concentration and cairclipNO_2 sensor for NO_2 measurements. The performance of all three methods were compared against each other and the exhaustive feature selection method gave the best result. The result from the EFS method was then applied to linear regression and artificial neural network models to calibrate the sensors, using the process of data fusion to merge and integrate data from different nodes into the calibration model. The results from this study showed the importance of feature selection in building accurate multi-parameter calibration models as well as the effectiveness of SLR, MLR and ANN in the calibration of low-cost sensors in environmental monitoring network.

CRediT authorship contribution statement

Nwamaka U. Okafor: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization, Investigation, Validation, Writing - review & editing. **Yahia Alghorani:** Resources, Writing-review & editing. **Declan T. Delaney:** Supervision, Investigation, Validation, Writing - review & editing.

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

This work is supported by Schlumberger Foundation, Netherlands through the Faculty for the Future programme, Tertiary Education Trust Fund (TETFUND-Nigeria) and the Smart-BOG project under EPA Research Programme, Ireland 2014–2020 42617/03

References

[1] J. Talavera, et al., Review of IoT applications in agro-industrial and environmental fields, Comput. Electron. Agric. 142 (2017) 283–297, Available: http://dx.doi.org/10.1016/j.compag.2017.09.015.

[2] F.M. Bublitz, et al., Disruptive technologies for environment and health research: An overview of artificial intelligence, blockchain, and Internet of Things, Int. J. Environ. Res. Public Health 16 (20) (2019) 3847, Available: http://dx.doi.org/10.3390/ijerph16203847.

[3] M. Gao, J. Cao, E. Seto, A distributed network of low-cost continuous reading sensors to measure spatiotemporal variations of PM2.5 in Xi'an, China, Environ. Pollut. 199 (2015) 56–65, Available: http://dx.doi.org/10.1016/j.envpol.2015.01.013.

[4] N. Okafor, D. Delaney, Considerations for system design in IoT-based autonomous ecological sensing, Procedia Comput. Sci. 155 (2019) 258–267, Available: http://dx.doi.org/10.1016/j.procs.2019.08.037.

[5] F. Mao, K. Khamis, S. Krause, J. Clark, D. Hannah, Low-cost environmental sensor networks: Recent advances and future directions, Front. Earth Sci. 7 (2019) Available: http://dx.doi.org/10.3389/feart.2019.00221.

[6] A. Rai, et al., End-user perspective of low-cost sensors for outdoor air pollution monitoring, Sci. Total Environ. 607–608 (2017) 691–705, Available: http://dx.doi.org/10.1016/j.scitotenv.2017.06.266.

[7] R. Williams, et al., Air Sensor Guidebook, Science Inventory, US EPA, 2019, [Online]. Available: https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=277996. (Accessed 02 November 2019).

[8] C. Borrego, et al., Assessment of air quality microsensors versus reference methods:The EuroNetAir joint exercise, Atmos. Environ. 147 (9) (2016) 246–263.

[9] D. Hagan, et al., Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments, Atmos. Meas. Tech. 11 (1) (2018) 315–328, Available: http://dx.doi.org/10.5194/amt-11-315-2018.

[10] L. Sun, D. Westerdahl, Z. Ning, Development and evaluation of a novel and cost-effective approach for low-cost NO_2 sensor drift correction, Sensors 17 (8) (2017) 1916, Available: http://dx.doi.org/10.3390/s17081916.

[11] F. Delaine, B. Lebental, H. Rivano, In situ Calibration algorithms for environmental sensor networks: A review, IEEE Sens. J. 19 (15) (2019) 5968–5978, Available: http://dx.doi.org/10.1109/jsen.2019.2910317.

[12] M. Syafrudin, G. Alfian, N. Fitriyani, J. Rhee, Performance analysis of IoT-based sensor, big data processing, and machine learning model for real-time monitoring system in automotive manufacturing, Sensors 18 (9) (2018) 2946, Available: http://dx.doi.org/10.3390/s18092946.

[13] G. Manes, et al., Realtime gas emission monitoring at hazardous sites using a distributed point-source sensing infrastructure, Sensors 16 (1) (2016) 121, Available: http://dx.doi.org/10.3390/s16010121.

[14] S. Pandey, K. Kim, The relative performance of NDIR-based sensors in the near real-time analysis of CO_2 in air, Sensors 7 (9) (2007) 1683–1696, Available: http://dx.doi.org/10.3390/s7091683.

[15] Jiao, et al., Community Air Sensor Network (CAIRSENSE) project: evaluation of low-costsensor performance in a suburban environment in the southeastern unitedstates, Atmos. Meas. Tech. 9 (11) (2016) 5281–5292, Available: http://dx.doi.org/10.5194/amt-9-5281-2016.

[16] M. Badura, P. Batog, A. Drzeniecka-Osiadacz, P. Modzel, Regression methods in the calibration of low-cost sensors for ambient particulate matter measurements, SN Appl. Sci. 1 (6) (2019) Available: http://dx.doi.org/10.1007/s42452-019-0630-1.

[17] S. Munir, M. Mayfield, D. Coca, S. Jubb, O. Osammor, Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in Sheffield, Environ. Monit. Assess. 191 (2) (2019) Available: http://dx.doi.org/10.1007/s10661-019-7231-8.

[18] K. Yamamoto, T. Togami, N. Yamaguchi, S. Ninomiya, Machine learning-based Calibration of low-cost air temperature sensors using environmental data, Sensors 17 (6) (2017) 1290, Available: http://dx.doi.org/10.3390/s17061290.

[19] N. Zimmerman, et al., A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmos. Meas. Tech. 11 (1) (2018) 291–313, Available: http://dx.doi.org/10.5194/amt-11-291-2018.

[20] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre, F. Bonavitacola, Calibration of a cluster of low-cost sensors for the measurement of air pollution in ambient air, Sensors (2014) 21–24, IEEE.

[21] B. Maag, O. Saukh, D. Hasenfratz, L. Thiele, Pre-Deployment Testing, Augmentation and Calibration of Cross-Sensitive Sensors, ACM, 2016, pp. 169–180.

[22] P. Schneider, N. Castell, M. Vogt, F. Dauge, W. Lahoz, A. Bartonova, Mapping urban air quality in near real-time using observations from low-cost sensors and model information, Environ. Int. 106 (2017) 234–247, Available: http://dx.doi.org/10.1016/j.envint.2017.05.005.

[23] X. Fang, I. Bate, Using multi-parameters for calibration of low-cost sensors in urban environment, in: Proceedings of the International Conference on Embedded Wireless Systems and Networks, 2017, pp. 1–11.

[24] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a cluster of low-cost available: sensors for air quality monitoring Part A: Ozone and nitrogen dioxide, Sensors Actuators B 215 (2015) 249–257.

[25] Powerpoint presentation, 2020, [Online]. Available: https://webcache.googleusercontent.com. (Accessed 21 Febraury 2020).

[26] R. Duvall, R. Long, M. Beaver, K. Kronmiller, M. Wheeler, J. Szykman, Performance evaluation and community application of low-cost sensors for ozone and nitrogen dioxide, Sensors 16 (10) (2016) 1698, Available: http://dx.doi.org/10.3390/s16101698.

[27] Envea Cairclip O_3/NO_2, VAQUUMS project website, 2020, [Online]. Available: https://vaquums.eu/sensor-db/sensors/envea-cairclip-o3-no2. (Accessed 18 May 2020).

[28] R. Williams, R. Long, M. Beaver, A. Kaufman, F. Zeiger, M. Heimbinder, I. Hang, R. Yap, B. Acharya, B. Ginwald, K. Kupcho, S. Robinson, O. Zaouak, B. Aubert, M. Hannigan, R. Piedrahita, N. Masson, B. Moran, M. Rook, P. Heppner, C. Cogar, N. Nikzad, W. Griswold, Sensor Evaluation Report, U.S. Environmental Protection Agency, Washington, DC, 2014, EPA/600/R-14/143 (NTIS PB2015-100611).

[29] Keras Documentation, Keras.io, 2020, [Online]. Available: https://keras.io/. (Accessed 08 February 2020).

[30] T. Hope, Y. Resheff, I. Lieder, Learning TensorFlow, Vol. 1, O'Reilly, 2017.

[31] L. Spinelle, M. Gerboles, M. Aleixandre, Report of Laboratory and in-Situ Validation of Micro-Sensor for Monitoring Ambient Air-Ozone Micro-Sensors, Asense, Model:B4 O_3 Sensors, Publications Office of the European Union, 2013, 26681.