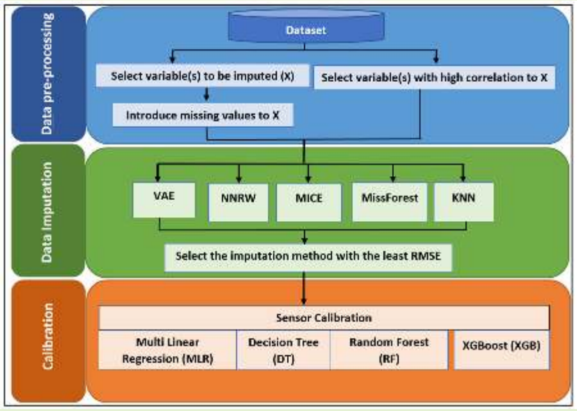
Missing Data Imputation on IoT Sensor Networks: Implications for on-site Sensor Calibration

Nwamaka U. Okafor *Graduate Student Member, IEEE*, Declan T. Delaney

***Abstract*—IoT sensors are becoming increasingly important supplement to traditional monitoring systems, particularly for in-situ based monitoring. Data collected using IoT sensors are often plagued with missing values occurring as a result of sensor faults, network failures, drifts and other operational issues. Missing data can have substantial impact on in-field sensor calibration methods. The goal of this research is to achieve effective calibration of sensors in the context of such missing data. To this end, two objectives are presented in this paper. 1) Identify and examine effective imputation strategy for missing data in IoT sensors. 2) Determine sensor calibration performance using calibration techniques on data set with imputed values. Specifically, this paper examines the performance of Variational Autoencoder (VAE), Neural Network with Random Weights (NNRW), Multiple Imputation by Chain Equations (MICE), Random Forest-based Imputation (missForest) and K-Nearest Neighbour (KNN) for imputation of missing values on IoT sensors. Furthermore, the performance of sensor calibration via different supervised algorithms trained on the imputed dataset were evaluated. The analysis showed VAE technique to outperform the other methods in imputing the missing values at different proportions of missingness on two real-world datasets. Experimental results also showed improved calibration performance with imputed dataset.**


Data pre-processing
Dataset
Select variable(s) to be imputed (X)
Select variable(s) with high correlation to X
Introduce missing values to X
Data Imputation
VAE
NNRW
MICE
MissForest
KNN
Select the imputation method with the least RMSE
Calibration
Sensor Calibration
Multi Linear Regression (MLR)
Decision Tree (DT)
Random Forest (RF)
XGBoost (XGB)

***Index Terms*—Calibration, Imputation, Internet of Things (IoT), Missing Data, Neural Network, Regression, Sensors, Variational Autoencoder, XGBoost.**

I. Introduction

Expanding the measurement networks for Green House Gases (GHG) is vital for understanding GHG global emission trends and the effectiveness of emission mitigation policies, strategies and initiatives, making it possible to ascertain how far emission reduction targets are being met at the local, regional and global scales [1].

Low Cost Sensors (LCS) have the potentials to enhance the spatio-temporal resolution of data acquisition for key GHG variables. LCS, however, are prone to diverse issues including bias, drifts, precision degradation, and loss of considerable amount of data due to operational issues [2]. Missing data is a pervasive issue, affecting most real-world datasets including medical records [3], [4], geo-informatics [5], traffic flow [6] and industrial applications [7], [8].

The European Union Data Quality Directive (EU-DQD) defined the data quality objective (DQO) that a monitoring method needs to comply with to be used as indicative measurement for regulative purposes [9]. The EU-DQD is a measure of the acceptable uncertainty for indicative measurements. The directive also defined the degree of data completeness for such monitoring method. To meet these requirements and to present LCS as suitable for adoption for this purpose, data completeness is essential for the sensors. In addition to this, complete data points consisting of sensor outputs and labels are neccessary for building reliable calibration models to ensure the collection of accurate and robust data by LCS.

Several methods have been proposed for handling missing data in diverse application domains. Common among these methods are downsampling which is also known as *Complete Case Analysis (CCA)* and imputation. The main idea behind downsampling is discarding the incomplete observations i.e. dropping missing data records. Although downsampling is a very simple method for handling missing data, it comes at the cost of losing useful information on data which may be valuable even though the data is incomplete. Applying downsampling as a means of handling missing values in any statistical analysis is mostly useful when there is a large number of samples [10] but performs poorly when the rate of missing values is high [11]. Imputation can be classified as:

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

Nwamaka U. Okafor is with the School of Electrical and Electronic Engineering, University College Dublin, on study leave from Federal Polytechnic Nekede, Nigeria (e-mail: nwamaka.okafor@ucdconnect.ie).

Declan T. Delaney is with the School of Electrical and Electronic Engineering, University College Dublin (e-mail: declan.delaney@ucd.ie).

(i) single imputation (SI) and (ii) Multiple Imputation (MI). SI involves replacing all missing values on a variable with a single value i.e., zero or the mean of the observation. MI is an iterative model-based approach [12].

State of the art multiple imputation techniques can be classified into Discriminative and Generative methods. Discriminative methods include Multiple Imputation by Chain Equations (MICE) [13], Random Forest-based Imputation (Missforest) [14] and matrix completion [15]. Generative methods consist mostly of techniques based on Deep Learning (DL) e.g Variational Autoencoders (VAE) [16], [17], Neural Networks with Random Weights(NNRW) [18], Denoising Autoencoders (DAE) [19] and Generative Adversarial Networks(GAN) [20], [21].

In this study, we evaluate the performance of different algorithms including VAE, NNRW, MICE, MissForest and K-Nearest Neighbour (KNN) for handling missing values in LCS networks where values could be missing over consecutive periods or at random points in time. Specifically, two tasks were conducted: first, Ozone (O_3) and NO_2/O_3 concentration data collected using Aeroqual and Cairclip sensors respectively over a six months data collection period were corrupted by removing data intervals at different missing periods (p) where $p \in \{1day, 1week, 2weeks, 1month\}$ and also at random points on the dataset at varying proportion (r) where $r \in \{5\%, 10\%, 30\%, 50\%, 70\%\}$. The missing data were then filled using the different imputation strategies and their imputation accuracy calculated. Second, the performance of sensor calibration by different regression models including Multi Linear Regression (MLR), Decision Tree (DT), Random Forest (RF) and XGBoost (XGB) trained on the imputed datasets were evaluated.

The key contributions of this paper include to:

1) identify suitable imputation technique to handle missing values on LCS networks.
2) develop a strategy based on efficient data imputation to support on-site sensor calibration.
3) present reliable technique for improving data quality of LCS in environmental monitoring networks.

In section II, we present the motivation for this work while section III details the current state of the art with respect to imputation and sensor calibration. Section IV presents the dataset used in this study while a detailed description of the methodology is presented in section V. The sensor calibration process is described in section VI and in section VII we present the model evaluation. Section VIII contains the results of the analysis and section IX has the conclusion and recommendations for future work.

II. Motivation

Inspired by the numerous successes of modern machine learning processes, especially in the development of strategies which have been found useful for handling missing values in areas such as medical data [22], sentence generation [23], image concealment [24] and data compression [25]. We explore the ability of diverse techniques for imputing missing data in LCS in environmental monitoring networks. Due to the nature of LCS, they are often challenged by the problems of missing data and this hinders the application of advanced analysis on the data collected by these sensors. Most researchers resort to deleting the cases with missing values when applying the dataset for further analysis. This method is ineffective as it simply ignores the cases with missing data, and does not take into consideration the complex distribution in environmental data, thereby leading to imprecision and bias. Imputation is capable of learning missing data either as a single value or as multiple possible values to address uncertainties. In cases where data distribution is of interest, imputation can estimate the most probable distribution of the data rather than estimating the unobserved data [26]. In this study, we evaluate the potentials of imputation on LCS dataset and the implications of data imputation for on-site sensor calibration.

III. State of the art

Missing or inconsistent data have been a major issue in data analysis since the origin of data collection. Methods for handling missing data ranges from the naive deletion of instances with missing values to modern machine learning imputation techniques. The suitability of an imputation method can be influenced by the missingness mechanism. Three different missing data mechanisms exist and these mechanisms can affect the accuracy of an imputation method. Techniques for handling missing values are generally assessed based on the three missingness mechanisms: *Missing Completely at Random (MCAR)* where the missingness occur completely at random with no dependency on any of the variables i.e. the distribution of missingness is independent on either the observed values or the missing values. *Missing at Random (MAR)* where the missingness depends only on the observed values but not on the missing values. *Missing Not at Random (MNAR)* where the missingness depends both on the observed and missing values [27].

Hedge et al. compared the performance of Probabilistic Principal Component Analysis (PPCA) and MICE for the imputation of missing data in healthcare dataset [28]. Their analysis began with a complete baseline dataset which included medical and dental variables, simulating missing data and its imputation assuming that values were MCAR. Their work shows PPCA outperforming MICE for this purpose. Stekhoven et al. proposed an iterative imputation strategy based on random forest (missForest) by averaging over several unpruned classification or regression trees [14]. They performed their analysis on multiple datasets from a diverse selection of biological fields with artificially introduced missing values at different rates. Their work shows that missForest is able to handle missing data on dataset consisting of different data types including continuous and categorical data. Comparatively analysing missForest with other imputation methods such as KNN, they presented results showing missForest outperforming the other methods, particularly in data settings where complex interactions and non-linear relationships were suspected.

In [29], Gupta et al., applied Neaural Networks (NN) for imputing missing values in classification problems. They used backpropagation algorithm to reconstruct missing values on datasets and shows that the reconstruction of the dataset using NN was better than reconstruction using statistical method. Their analysis showed that classification accuracy increases with the inclusion of reconstructed data values. Ravi et al., proposed missing data imputation on different tasks including classification, regression, bankrupcy prediction and credit scoring datasets using auto associative NN [30]. Their work showed interesting imputation results by the NN on the different tasks.

Although traditional neural networks have shown interesting capabilities, achieving state of the art results in many real-world applications, they are still challenged by a number of issues such as difficulty in training the network especially as the size of the network increases, slow convergence and local minima problems. To solve these problems, Cao et al. proposed Neural Network with Random Weights (NNRW) [31]. NNRW is a non-iterative algorithm in which the hidden wieghts and biases are randomly selected from a given range of values and kept the same throughout the training process while the weights between the hidden layer and output layer are obtained analytically, this process helps the NN to train faster with acceptable accuracy [32]. In [33], Cao et al. employed a type of NNRW known as Random Vector Functional LinK Network (RLFV) for semi-supervised learning. Their work was based on fuzzy theory and shows improved generalization of the fuzziness based RVFL in classification problems.

In [34], Beaulieu-Jones et al. proposed a deeply learned Auto Encoders (AE) technique for imputing missing electronic health record data. They compared the performance of AE to other imputation strategies and noted that AE, though computationally intensive, outperformed competing imputation methods. However, they noted that with GPU resources, AE trains in similar time to KNN and Singular Value Decomposition (SVD) methods. In [2], Loy-Benitez et al. proposed Variational Autoencoders with Convolutional layers (VAE-CNN) for imputing missing indoor air quality data. They compared this method to other imputation methods and VAE-CNN demonstrated improved results over the other methods in the imputation task. McCoy et al. also used VAE for imputing missing heavily corrupted (90% of records) and lightly corrupted (20% of record) data in a simulated milling circuit [17]. Their analysis showed that for both the heavily and lightly corrupted datasets, the Root Mean Squared Error (RMSE) of the VAE method was lower than other traditional methods including mean replacement and PCA.

In the environmental monitoring field where LCS are usually employed to supplement existing traditional monitoring solutions, data quality assurance is essential for the sensors and can be achieved through frequent sensor calibration. However, most studies in the IoT sensor calibration domain have drawn conclusions based on the assumption of a complete dataset [35], [36]. Although, some researchers have explored missing data issues on IoT sensor calibration [37], [38], more research is still needed in this space. Missing data reduces the representativeness of samples, thereby reducing the statistical power of a study which in turn produces bias estimates and can lead to invalid conclusions [39]. This distortion can adversely affect sensor calibration processes.

Calibrating sensors on imputed dataset rather than on a dataset where incomplete records have been discarded can help to yield more accurate calibration result, improving on-site sensor performance and ensuring that the sensors are collecting accurate data.

In this study, we examine the effect of missing and imputed data on LCS calibration. We investigate VAE-a deep learning-based generative model, NNRW, MICE, MissForest, and KNN for imputing missing sensor data and subsequently used the imputed dataset for sensor calibration.

Previous studies have proposed different methods for sensor calibration. Spinnelle et al. applied SLR, MLR and ANN for the calibration of a cluster of low-cost O_3, NO_2, NO, CO and CO_2 sensors over a two week calibration period. Based on the measurement uncertainty estimated by orthogonal regressions of sensors and reference data, their work shows ANN to be a suitable calibration model for the sensor clusters while both simple and multiple linear regressions provided high level of measurement uncertainties [40], [41]. De Vito et al. proposed and evaluated the calibration of low cost gas multi-sensor devices in an urban air pollution monitoring mesh using NN and a two week on-site recorded data for benzene, CO, NO_2 and NO_x pollutants. Their work shows the feasibility of obtaining a neural calibration capable of allowing multi sensor devices to successfully operate on the field carrying out pollutant estimation with optimal result even for a limited number of training periods [42], [43]. Okafor et al. applied SLR, MLR and ANN for the calibration of low-cost O_3 and NO_2 sensor devices. Their work evaluated the performance of different feature selection techniques in identifying factors that affect on-site sensor measurements and applied data fusion to exploit the correlation existing between similar sensors. Experimental results from their work shows the calibration methods to minimize estimation errors from the sensors with respect to conventional station outputs [44], [45].

IV. Dataset Description

The dataset used in this study is presented by Feinberg et.al in [46], it is a publicly available dataset which is available at the EPA environmental dataset gateway [47]. The dataset consists of measurements from particles and gas sensors. The sensors were deployed in triplicates in a static network configuration in co-location with Federal Equivalent Method (FEM) Monitors at an urban regulatory site in Denver Colorado, United States of America, over a six months monitoring period. Similar static sensor network configuration for air quality monitoring can be found in [48]. Sensor network deployments can be static or mobile [49], [50], where the methods rely on the mobility of the sensor nodes to achieve calibration [51].

For the purpose of this study, we exclude the particle sensors and concentrate our analysis on the gas sensors. Two datasets were considered. (i) Ozone (O_3) and (ii) combined Nitrogen dioxide and Ozone (NO_2/O_3) datasets. For O_3, the dataset consists of measurements from three units of Aeroqual O_3 sensors, Temperature (T) and Relative Humidity (RH) measurements as well as measurements from O_3 FEM monitor. Likewise, for NO_2/O_3, the dataset consist of measurements from three units of CairclipNO_2/O_3 sensors, T and RH measurements as well as measurements from NO_2/O_3 FEM monitor. All sensors and FEM monitors logged data per minute in part per billion (ppb) from 8 September 2015 to 22 February 2016.

The sensors are commercially available, low cost and with relatively high market prevalence. More information about the sensors is presented in Table I.

Manufacturer	Sensor	Operating mechanism	Quantity
SM50/Aeroqual (New Zealand)	O_3	Electrochemical sensors	3
Cairclip/Cairpol (France)	NO2 and O3 combined (ppb)	Electrochemical sensors	3

TABLE I
SELECTED SENSORS

The sensors exhibited moderate to strong correlation with the FEM reference monitors. O_3 Pearson correlation of the sensors and reference measurement is >0.9 while that of NO_2/O_3 sensors to NO_2/O_3 reference ranges between 0.32 to 0.57. The heatmap in figure 1 shows the correlation between the variables. The O_3 sensors (S1, S2, S3) show high correlation among sensors and also exhibited varying positive correlation with T and negative correlation with RH. This relationships can be exploited by the imputation models to predict missing values on the variables.

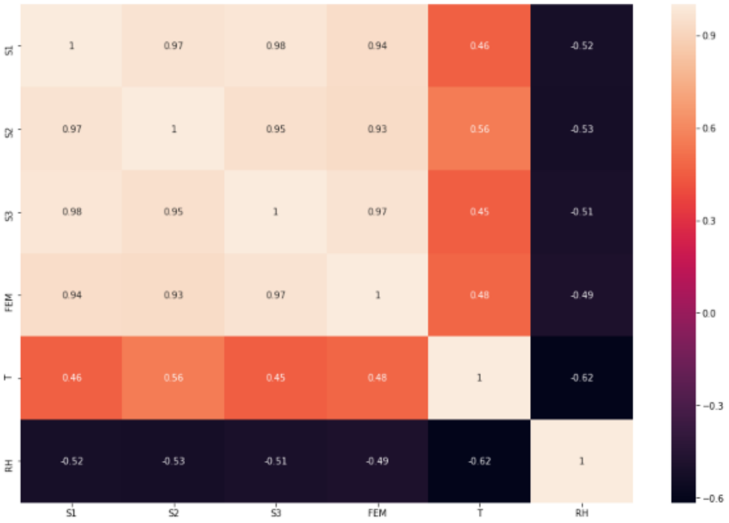


Fig. 1. Correlation coefficients of O_3 sensors and FEM measurements

The datasets originally consist of real missing values, 5.77% and 19.54% of values were missing from the O_3 and NO_2/O_3 dataset respectively. It is impossible to assess the performance of the imputation strategies when the real values are unknown. We therefore created simulated missing values on the datasets to assess the ability of the imputation algorithms. Further description about this is on the methodology section.

V. METHODOLOGY

We employed multiple imputation techniques to impute missing values on both the O_3 and NO_2/O_3 sensors datasets. The techniques considered in this study include VAE, NNRW, MICE, MissForest and KNN. To asses the performance of the imputation algorithms, we adopt a five step approach extending the method proposed in [52]. The first step involves an initial evaluation of the correlation between features; features exhibiting high correlation with the variable(s) to be imputed were selected and included in the imputation model. Step two involves artificially creating missing values at random points on the variable (as illustrated on table II) and at consecutive periods (as per table III). Step three involves applying the imputation techniques to impute the missing values. Step four involves comparing all imputation techniques using the RMSE performance indicator. RMSE is defined as the average squared difference between imputed and original values. In general, the best performing imputation technique will have the lowest RMSE. Finally, in step five, the imputed dataset from the best performing algorithm is used for sensor calibration and the performance of the imputed data on sensor calibration is compared to calibration on *Complete Case (CCA)* i.e, eliminating missing data records from the analysis.

t_1	t_2	t_3	t_4	t_5	t_6
x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x
x_{21}	x_{22}	x_{23}	x	x_{25}	x_{26}
x_{31}	x	x_{33}	x_{34}	x_{35}	x_{36}
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x
x_{51}	x_{52}	x_{53}	x_{54}	x	x_{56}

TABLE II
MISSING DATA PATTERN:MISSING VALUES OCCURRING AT RANDOM POINTS

t_1	t_2	t_3	t_4	t_5	t_6
x_{11}	x_{12}	x	x	x	x
x_{21}	x_{22}	x_{23}	x	x_{25}	x_{26}
x	x	x	x_{34}	x_{35}	x_{36}
x_{41}	x_{42}	x_{43}	x_{44}	x_{45}	x
x_{51}	x_{52}	x	x	x	x_{56}

TABLE III
MISSING DATA PATTERN: MISSING VALUES OCCURRING OVER CONSECUTIVE PERIODS

Specifically, missing values were introduced following the patterns listed below and we examined how the imputation techniques performed on imputation tasks under the different missingness scenarios:

1) missing values were simulated on a single variable with missingness spanning p period of time, where $p \in \{1day, 1wk, 2wks, 1month, ...4months\}$
2) missing values were introduced on multiple variables (3 sensor units) with missingness spanning p period of time, where $p \in \{1day, 1wk, 2wks, 1month\}$

3) missing values were introduced on multiple variables (3 sensor units) with missingness occurring at random points and at different proportion *(r), where* $r \in \{5\%, 10\%, 30\%, 50\%, 70\%\}$.

To the best of our knowledge, this is the first study that have considered multiple imputation on IoT sensors deployed for the quantification of GHGs, with missing values occurring not just at random points but over consecutive periods.

All data handling and processing in this study were performed using python 3 on a jupyter notebook which is included as part of the Anaconda distribution [53].

A. Multiple Imputation (MI)

As opposed to single imputation, multiple imputation supports analysis that makes use of all possible information on a dataset [54]. MI accounts for the statistical uncertainties in imputations and also yields more accurate results. They involve filling in the missing values multiple times, creating multiple complete datasets. MI techniques were developed to handle uncertainties in imputation beyond what SI can carter for. MI techniques are able to recover information which would otherwise be lost when observations with missing values are excluded in an analysis, thereby helping to minimize bias.

The validity of analysis relying on imputed data would however depend on the correct specification of the imputation model, a common question when dealing with missing values is what proportion of missingness is acceptable before inference with imputation becomes valid. Previous studies have identified various upper and lower limits, observing also that the availability of other auxiliary variables which can predict the missingness and/or are associated with the missing values may be an important consideration [55]. The inclusion of auxiliary variables in the imputation model can yield unbiased estimates even at high proportion of missing values. [55].

Correctly specifying the imputation model is necessary for obtaining accurate analysis with the imputed data. However, for practitioners, deciding which imputation method is most suitable for their particular problem still remains a challenge. To address this challenge, Meyer et.al proposed and launched a unified platform; R-miss-tastic [56]. A platform which provides an overview of standard missing values problems, with relevant implementations of methodologies (in R and python) that can be used to assess missing values in an analysis. Their work aims to create standard analysis workflow and to unify the community.

In the current study, the performance of different imputation methods for imputing missing values on IoT sensors deployed for estimating GHGs were assessed. Furthermore, the effects of imputation on sensor calibration models were also investigated. In the following subsections, we describe in more details, the imputation methods used.

B. Variational Autoencoders (VAE)

VAE is an autoencoder (AE) network with generative capability. AE and VAE however, differ in how they represent model input data. While AE learns a compressed representation of the input data, VAE learns a set of distribution parameters which describes the data, usually the mean and variance of a gaussian probability function. By sampling from these parameters, VAE can generate data instances that closely resembles the original data [57]. The VAE structure consist of two main components (i) encoder $q_\phi(Z|X)$ and (ii) decoder $p_\theta(X|Z)$ as shown in figure 2 .

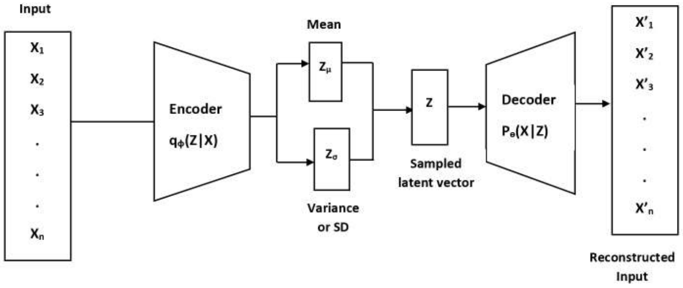


Fig. 2. Structure of a Variational Autoencoder

Both the encoder and decoder are multilayered NN with parameters ϕ and θ respectively [2]. VAE follows the assumption that the input data X is generated by some underlying distribution $p(X)$ which can be represented by the latent variable Z, where Z itself is generated by a distribution $p(Z)$. The joint distribution of $p(X, Z)$ can be represented as:

$$p(X, Z) = p_\theta(X|Z)p(Z) \tag{1}$$

This joint distribution can be generated by sampling from the distribution of Z, also known as the prior, $p(Z)$ and the distribution of X given Z, also called likelihood $p_\theta(X|Z)$. The likelihood function with the parameter θ is learnt from the data. $p(Z)$ is usually chosen to follow a normal distribution with zero mean and unit variance with no additional parameters to learn. The posterior is the distribution of the latent variable Z, given X. The posterior which is typically a NN is approximated by $q_\phi(Z|X)$. Maximizing the log of the marginal likelihood (log evidence) $lnp_\theta(X)$ will ensure that the best representation for the data X is obtained.

The log evidence, $lnp_\theta(X)$ can be expressed in terms of the Evidence Lower Bound (ELBO).

$$lnp_\theta(X) = ELBO + KL[q_\phi(Z|X)||p_\theta(X|Z)] \tag{2}$$

KL is the Kullback-Liebler divergence, it describes the agreement between two distributions such as the encoder network $q_\phi(Z|X)$; an approximation to the posterior distribution and the true posterior distribution $p_\theta(X|Z)$. KL divergence is zero when both $q_\phi(Z|X)$ and $p_\theta(X|Z)$ are identical and positive if they are not identical. The maximum value of $lnp_\theta(X)$ can be obtained by finding the parameters ϕ and θ which minimizes ELBO as per equation 3.

$$ELBO = E_{\mathrm{q}_\phi}(Z|X)[\ln p_\theta(X|Z)] - KL[q_\phi(Z|X)\|p(Z)] \tag{3}$$

The first term in equation 3 is the expectation of the log likelihood of the decoder network given the encoder network's output. The second term is the KL divergence between the posterior distribution (i.e the encoder network), $q_\phi(Z|X)$ and the prior distribution of the latent variable Z, p(Z). Minimizing this term brings the posterior $q_\phi(Z|X)$ closer to the prior p(Z).

In this study, the VAE network was trained and subsequently used to predict the missing values. We tested different network architectures to determine the network with optimal performance for the imputation task. The networks tested differs in the number of hidden layers and neurons for the encoder and decoder networks. Optimal result was achieved using two hidden layers in both the encoder and decoder networks with each layer having 25 neurons, the network was trained for 500 epochs using Adam as the optimizer and a learning rate of 0.001. Missing values were then imputed following the steps below:

1) First, the missing points in the dataset were replaced with zero
2) The dataset was then passed to the trained VAE network
3) Samples were drawn from the latent variable distribution i.e. output of the encoder network to generate Z, given X.
4) given Z, samples were drawn from the reconstructed data distribution. i.e. output of the decoder network to generate X'
5) the missing values were then replaced with the reconstructed values, leaving the observed values unchanged.
6) The imputation iteration limit was set to 25 as optimal result was achieved at this point during the training process, setting a higher limit overfits the model. Step 2 to 5 is repeated until this limit is reached.

C. Neural Network With Random Weights (NNRW)

In the current study, a feedforward NNRW was also applied to predict the missing values. The network topology used is a fully connected layered network. The network is divided into input layer, hidden layers and output layer. The number of neurons in the input and first hidden layer corresponds to the number of input variables present on the datasets and the output layer has neuron(s) corresponding to the target variable(s). The network diagram is shown in figure 3, and for clarity, depicts only one hidden layer on the diagram.

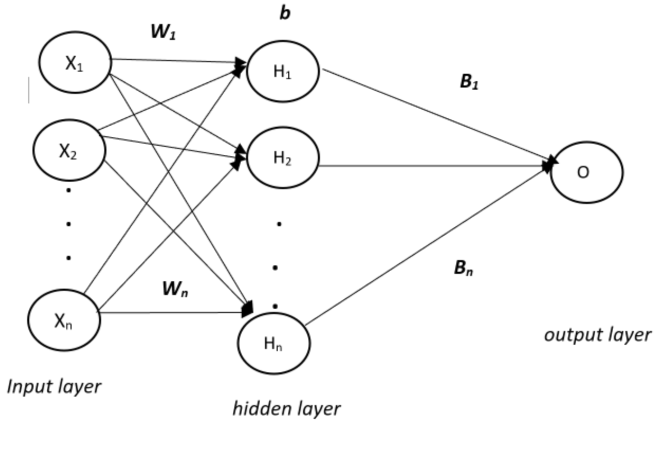


Fig. 3. Structure of a three layered NNRW

We tested different architectures differing only in the number of hidden layers H and hidden neurons (N) with H in $\{2,4,6\}$ and N in $\{3,5,10,15\}$. So far, there has not been much guidance regarding the number of hidden layers and hidden neurons, the choice of the same number of neurons in the input and first hidden layer in this study is to ensure that each input neuron has a corresponding neuron in the first hidden layer. Each neuron in the input layer is fully connected in the forward direction to all the neurons in the first hidden layer through a set of weights W. Similarly each of the neurons in the hidden layer is fully connected to all the neurons in the next hidden layer through the same set of weights. The output weights (weights of neurons between the hidden and output neuron) are another set of weights β. The weights W and the threshold of the hidden biases b were selected randomly from a range of values. The range of these values depends largely on the activation function used, although some authors recommend [-1:1] and some as small as[-0.1:0.1]. Empirically, a range of [-1:1] for W and [0:1] for b is found adequate for the current use case. These weights were kept fixed throughout the training process. The weights β between the hidden and output layer were determined following the method described in [33]. The activation function used was Rectified Linear Unit (ReLu) for the hidden layers and Linear activation function for the output and $Adam$ was the optimizer used.

To impute the missing values, we adopted an approach similar to that described in [58] following the steps below:

1) the variable to be imputed is set as the target and the other variables on the dataset are the predictors.
2) the dataset is splitted into training and test subsets.
3) the training subset contains non-missing values and is used to train the network.
4) in the test subset, missing values are introduced following the patterns described in section V, and was used to evaluate the network, comparing the predicted values to the true values.
5) the best trained network is chosen and applied to records with missing values.

D. Multiple Imputation by Chain Equation (MICE)

The MICE method imputes data on a variable by variable basis by specifying an imputation model per variable. We briefly outline below, the MICE procedure used in this study:

1) mean value imputation (a single imputation process) is first carried out for every missing value on the dataset (this mean imputation could be regarded as a place holder).
2) the place holder created in(1) for one of the variables (i.e. output from one of the sensors (S_j) is set back to missing.
3) the observed values from S_j in 2 are applied in a linear regression with the other variables. S_j being the dependent variable and the other variables being the independent variables on the regression model.
4) The missing values in S_j are then replaced by predictions from the regression model.

5) Step 2-4 is repeated for each variable with missing values.

A complete cycle for each variable constitutes one iteration. At the end of one cycle, all of the missing data were replaced with predicted values from the regression, with the predictions reflecting the relationships observed on the dataset.

The entire process of iterating through all the variables were repeated until convergence, at the end, the final imputations were retained, this final set of imputed values and the observed values resulted in one complete dataset.

It is essential to know the number of imputation rounds necessary for a good statistical inference. The authors in [59] suggested that small value of imputation rounds (n) on the order of 3 to 5 yields an excellent result. Schafer et.al stated that not more than 10 imputations rounds are usually required [60], however, Graham et.al after using a Monte Carlo simulation to test multiple imputation models across several scenarios in which the fraction of missing information for the parameter being estimated and n were varied, recommended that many more imputation rounds than previously considered sufficient should be performed [61] .

In the current study, 5, 10, 30 and 50 imputation rounds were tested and optimal performance was obtained with the 30 imputation rounds, hence we applied 30 imputation rounds to the MICE technique.

E. Missforest

missForest is an iterative imputation method that is based on Random Forest (RF). In previous literature, missForest was identified to exhibit attractive computational efficiency and capable of handling missing value imputation on high dimensional datasets [14]. In this study, we assessed the performance of missForest for data imputation. An iterative imputer scheme was used by first training an RF on the observed values, followed by predicting the missing values and then proceeding iteratively. The RF algorithm has an inbuilt function capable of handling missing data by weighing the frequency of the observed values on a variable with the RF proximities after being trained on the initially mean imputed dataset. We adopted an approach which involved training an RF on the observed data, similar to the method proposed in [14].

For any variable $\mathrm{S_j}$ containing missing values at points $\mathrm{i_{missing}}$, where $i_{missing} \in \{1...n\}$ the dataset was separated into 4 parts:

1) The observed values of $\mathrm{S_j}$ denoted by $\mathrm{y_{observed}}$
2) The missing values of $\mathrm{S_j}$ denoted by $\mathrm{y_{missing}}$
3) The other variables contained on the dataset (other than $\mathrm{S_j}$ with observations at point $\mathrm{i_{observed}}$=$\{1,...\mathrm{n}\}$ denoted by $\mathrm{X_{observed}}$
4) The other variables contained on the dataset (other than $\mathrm{S_j}$) with observations $\mathrm{i_{missing}}$ denoted by $\mathrm{X_{missing}}$

The iterative imputer algorithm began by making an initial guess for the missing values in S using mean imputation. The variables $\mathrm{S_j}$, j=1,...p are then sorted in accordance with the number of missing values beginning with the variable with the lowest amount of missing values. For each variable $\mathrm{S_j}$, the missing values are imputed by first fitting an RF with response $\mathrm{y_{observed}}$ and predictors $\mathrm{X_{observed}}$. $\mathrm{y_{missing}}$ is then predicted by applying the trained RF model to $\mathrm{X_{missing}}$.To avoid overfitting the model, we set an early stopping rounds criterion and repeated the imputation iteration until this criterion was reached.

F. K-Nearest Neighbor

Inspired by the work in [62], we also employed KNN to handle missing values on the IoT sensors datasets and compared the performance to the other imputation algorithms previously described. With KNN, for each variable on the dataset, the missing values were imputed by finding the k non-missing values on the sample which are closest to the missing data point. The average of these k closest values are then taken and used to fill in the missing point. We determined the closest neighbours to the missing point by using an euclidean distance metric given by equation 4 to calculate the distance between the missing point (x) and the non missing neighbour (y).

$$\sqrt{\sum_{i=1}^{k}(x_{\mathrm{i}}-y_{\mathrm{i}})^2} \tag{4}$$

Euclidean distance remains the most obvious way for representing the distance between two points. It measures the length of a segment connecting the two points together. We set k=3 and used the average of these 3 nearest neighbours to the missing point to fill the missing point. Due to the fact that KNN depends on the distance between samples, the scale of the predictor variables can significantly influence the distance among samples, for this reason, we centred and scaled all predictor variables before applying KNN to avoid this potential bias and to enable each predictor to contribute equally to the distance calculation.

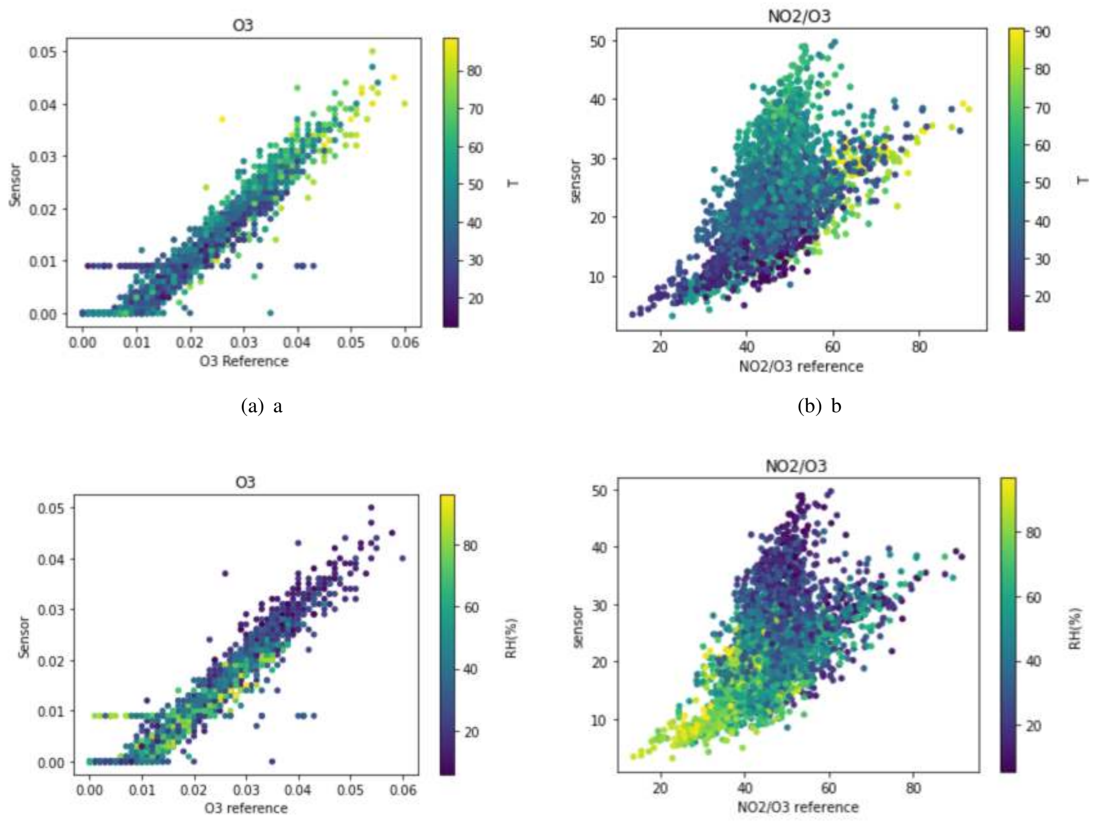
VI. Sensor calibration

After the data imputation process, we investigated the effect of using imputed dataset for sensor calibration purposes and evaluated the performance of calibration on imputed and CCA data.

To meet the maximum level of accuracy and to ensure that data quality of LCS is sufficient, frequent calibration and data validation are essential. Calibrating sensors however, requires the availability of complete data points for sensors and reference measurements, hence the need for efficient data imputation strategy.

While calibrating the sensors, we aim to correct the gain and offset errors by mapping raw sensor measurements to pollutant concentration provided by reference monitor [49]. Previous studies have identified data fusion technique to provide more consistent, reliable and accurate results when applied to sensor calibration [44]. We employ this technique in the current study as it allows for the use of multiple sensors outputs in the calibration model.

Furthermore, meteorological factors such as Temperature (T) and Relative Humidity (RH) can affect LCS outputs. Figure 4 shows the scatter plot of the O_3 and NO_2/O_3 sensors and their corresponding reference data color-coded



(a) a

(b) b

(c) c

(d) d

Fig. 4. Scatter plots of sensor and reference data, color-coded with meteorological factors: (a-b) temperature, (c-d) relative humidity

with meteorological factors, indicating the effects of both T and RH on the sensors measurements. At low T, the O_3 and NO_2/O_3 values measured by the sensors were lower than that of the reference monitors but as T increases, the sensors values tend to increase in proportion to the reference measurements, showing a positive interference of T with sensor measurements. Also, both O_3 and NO_2/O_3 sensor values were higher at low RH, but lower at high RH. Previous studies have also reported similar trends, showing the sensitivity of LCS to be influenced by changing environmental conditions. Pang et al. observed low sensitivity in electrochemical sensors with an increase in RH [63] while Rai et.al observed that LCS experienced a loss in sensitivity with changing ambient temperature [50].

It is important to account for the effects of meteorological factors on sensors output, thus, we incorporated changing environmental conditions (T and RH) into the calibration model to ensure improvement in the overall measurement accuracy of the LCS.

Before calibration, the variables were analysed for multicollinearity using Variance Inflation Factor (VIF) tool from the statsmodels package in python. Multicollinearity among independent variables will result in less reliable statistical inference, hence variables exhibiting high linear relationships with other variables as well as those exhibiting low level of significance were eliminated through a backward elimination process and were not included in the calibration model. We used different machine learning algorithms including Multi Linear Regression (MLR), Decision Tree (DT), Random Forest (RF) and XGBoost to build the calibration models. The models were tested on both O_3 and NO_2/O_3 datasets with 30% of missing values imputed using VAE imputation method. The VAE method was choosen for this particular task based on its significant imputation performance over the rest of the other algorithms. We describe in more details, the implementation of the calibration models in subsequent sections.

A. Multiple Linear Regression (MLR)

A MLR calibration model was used to fit the explanatory variables to the FEM monitor data using the formular in equation 5.

$$y_{\text{ref}} = \beta_0 + \beta_1 S_1 + \beta_2 S_2 + \beta_3 S_3 + \beta_4 T + \beta_5 RH \quad (5)$$

where y_{ref} is the target concentration from FEM monitor, S_1, S_2, S_3 are measurements from the three sensor units, T is temperature measurements and RH is relative humidity measurements

As per Cordero et.al, MLR is a suitable model for electrochemical sensors because their response to gas concentrations is linear and the cross-sensitivities are additive [64]. MLR has previously been applied in the calibration of low cost air quality sensors [64] [65].

B. Decision Tree (DT)

Decision Tree (DT) provides non-parametric method for partitioning dataset. It can be used to solve both regression and classification problems. The technique can aid the description, generalization and categorization of a given set of data by breaking the dataset into smaller subsets while incrementally developing an associated decision tree with decision nodes and leaf nodes. We used Grid search to get the best set of hyperparameters for the model. As per [66], grid search considers several hypaperameter combinations and chooses the one that returns a lower error score. It is most useful when there are only a few hyperparameters to optimize but would usually be outperformed by other weighted-random search algorithms when the model grows in complexity. We tested different values for the min sample split s=[5,10,15,20] and s=10 was found to be the best for the model with max depth of 3. A 10-fold cross validation was used to estimate the performance of the model.

C. Random Forest (RF)

RF works by constructing an ensemble of DTs through a bootstrap aggregation technique. This process involves training each DT on different data samples where sampling is done by replacement. The mean value from the ensemble is then used to predict the value of the new input data. By considering a random subset of the explanatory variables, the root node of the DT is split into sub nodes. The tree is split based on which of the explanatory variable in each random subset is the strongest predictor of the target [67]. The process of node splitting is repeated until a terminal node is reached.

In [68], Wang et.al applied RF method to calibrate a low-cost particle monitor (HK-B3) using measurements from MicroPEM monitor (RTI) as reference. Their work showed that RF was able to establish an accurate calibration function between the sensor and the reference device. In [67], Zimmerman et.al also applied RF algorithm in developing calibration model to calibrate LCS deployed for air quality monitoring.

In this study, we passed the explanatory variables and reference data to an RF regression model, a grid search cross validation was used to determine the optimal values of the hyperparameters of the model from a specified range of values. Here, we choosed two hyperparameters i.e max_depth and n_estimators to be optimized. We tested max_depths of 3,5,7 and n_estimators [50, 100, 150, 200]. A max_dept of 3 was found to be the best while n_estimator of 100 was found to be the best for the model. max_depth is the maximum depth of the tree and n_estimators is the number of trees in the forest [69]. we used a 10-fold cross validation method to determine the model performance.

D. XGBoost(XGB)

XGBoost is a decision tree-based ensemble algorithm which uses a gradient boosting framework [70]. It's a scalable and powerful algorithm especially where speed and acuracy are concerned. In fact, it has been a winning choice algorithm for participants in Kaggle-a data science and machine learning competition platform. Previous studies have identified XGBoost as an excellent algorithm for sensor calibration. In [71], the authors assessed the performance of XGB and other machine learning algorithms including MLR, and Feedforward NN for the calibration of low-cost $PM_{2.5}$ sensor and found that after calibrating with XGB, the variance of the $PM_{2.5}$ values were not statistically significantly different from the values measured by a highly accurate reference instrument with which the sensors were co-located, indicating a better agreement of the sensor values with the reference instrument after calibration. In the current study, we compared the performance of XGB with the earlier discussed algorithms for the calibration of LCS sensors.

While building an XGB model, it is important to consider different parameters and their values. XGB requires parameter tuning to improve and fully leverage its benefit. A random search method [72] was used in this study to tune the hyperparameters in the XGB model. We choosed three hyperparameters to tune. In table IV, we present the best hyperparameters choosen by the random search algorithm for the XGB model as well as the early stoping rounds used to avoid over-fitting the model.

Hyperparameters	Value
Number of trees to fit (n_estimators)	500
Number of parallel threads used (n_jobs)	10
Step size shrinkage (learning_rate)	0.03
Early stopping rounds	10

TABLE IV
HYPERPARAMETERS FOR THE XGB MODEL

VII. MODEL EVALUATION

The performance of the calibration models were evaluated by comparing the calibrated sensors responses to measurements from the FEM monitors using the error metrics: Root Mean Squared Error (RMSE) as given by equation 6, Mean Absolute Error (MAE) as per equation 7 and Coefficient of Determination (R^2). The lower the RMSE and MAE, the better the model and an R^2 value closer to 1 indicates good model performance.

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (Y_{\mathrm{i}} - y_{\mathrm{i}})^2} \tag{6}$$

$$MAE = 1/n \sum_{i=1}^{n} (|Y_{\mathrm{i}} - y_{\mathrm{i}}|) \tag{7}$$

(where n is the number of samples, Y is calibrated response and y is target response).

The models were built using imputed dataset and Complete Case Analysis (i.e eliminating missing observations from the datasets). The datasets were split into training and test subsets in proportions of 80% and 20% respectively with the test subset being the most recent part of the dataset. The training subset was used for the model training process and the test data was used to evaluate the model performance.

VIII. Results

In the following subsections, we present the results of the analysis undertaken in this study including the performances of the different imputation techniques as well as effect of imputation on sensor calibration.

A. Imputation

The imputation accuracy of the different imputation techniques were compared. The imputation accuracy is defined by the Root Mean Squared Error (RMSE) between the original values and imputed values.

At first, values were artificially removed from one sensor variable (S1) in a consecutive manner over specified period of time i.e., 1 day, 1 week, 1 month etc. At each period, the different imputation techniques (VAE, NNWR, MICE, missForest and KNN) were used to impute missing data and the imputation accuracy calculated. Table V shows the imputation accuracy (RMSE) of the different methods. For this case, all the imputation methods performed reasonably well with minimal errors (<0.1) even for a long period of missingness (up to 4 months), this could be explained by the availability of other auxiliary variables (i.e S2, S3, FEM, T and RH measurements) which were included in the imputation model and were able to predict the missingness on S1.

Result of the performance of the imputation methods when missing values were artificially introduced on multiple variables (i.e three sensor variables) over consecutive period of time is shown on figure 5.

Furthermore, results of the analysis of randomly introduced missing values are presented in figure 6. We ensured that all variables had at least one missing data point and that missing values were distributed across all the variables. For all the imputation tasks carried out in this study, VAE shows improved performance over the rest of the imputation methods for handling missing values.

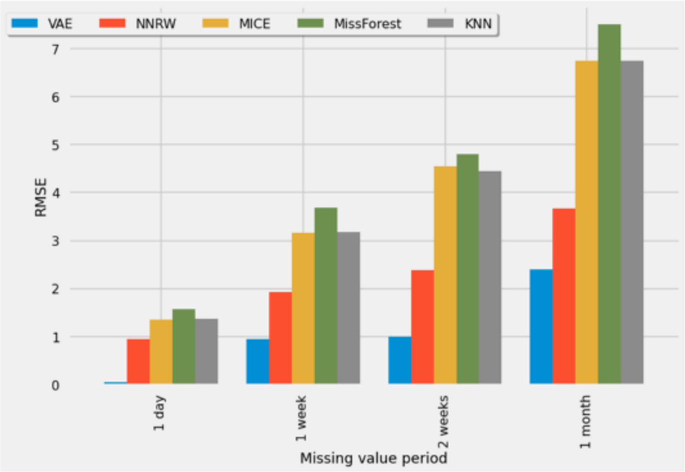


Fig. 5. *Comparison of imputation methods with missing values occurring on multiple variables over consecutive periods on the O_3 sensors dataset*

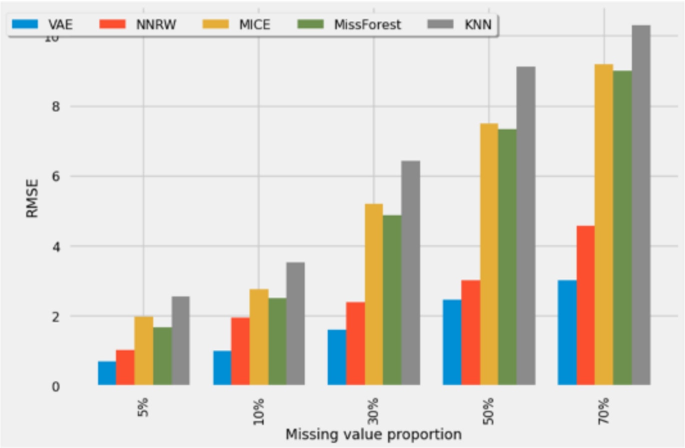


Fig. 6. *Comparison of imputation methods at different proportion of missingness on O_3 dataset*

B. Calibration

After the imputation processes, the sensors were calibrated using different supervised machine learning algorithms including MLR, DT, RF and XGB. To understand any effect imputed data may have on sensor calibration, the calibration was conducted using:

1) VAE-imputed dataset (with 30% missing values imputed)
2) Complete Case Analysis (CCA)- this involves deleting missing value records from the dataset

Calibrating the sensors using imputed dataset showed more promising performance when compared to *C*CA. In Tables VI and VII we present the calibration results from the different algorithms using imputed and CCA data for both O_3 and NO_2/O_3 sensors. For each of the calibration methods, the results showed calibration on imputed data to be more accurate than with CCA.

In the case of the O_3 dataset, all the calibration methods performed significantly well in correcting sensors errors with R^2 score > 0.9, XGB having the lowest error and the highest R^2 score however, outperformed the rest of the algorithms in the calibration task. The RMSE and MAE existing between uncalibrated/raw O_3 sensor and O_3 reference outputs is 0.0091 and 0.0082 respectively. After calibrating the sensor with XGB model trained on VAE-imputed dataset, the errors were significantly reduced (see Table VI). Also with calibration done on the CCA data, reduction in error between the sensor output and reference data was also observed, indicating the importance of sensor calibration in handling sensor output errors. Calibration on imputed dataset however showed better performance than calibration done using the CCA data, with XGB performing better than the other algorithms in the calibration task.

Furthermore, the R^2 score between the uncalibrated sensor data and the reference data was 0.50, however, after calibrating the sensors with VAE-imputed dataset, the R^2 score saw a significant improvement even with less sophisticated MLR model (R^2 = 0.9485) and the more sophisticated XGB model offering even better agreement between the calibrated sensor and reference data (R^2 = 0.9980).

Missing period	VAE	NNRW	MICE	missForest	KNN
1 day	$0.26\text{x } 10^{-5}$	$1.97\text{x } 10^{-5}$	$4.00 \text{ x } 10^{-5}$	$5.31\text{x}10^{-3}$	$1.11\text{x}10^{-2}$
1 week	$2.67\text{x } 10^{-5}$	$3.91\text{x } 10^{-5}$	$1.46 \text{ x } 10^{-3}$	$1.02\text{x}10^{-2}$	$1.74\text{x}10^{-2}$
1 month	$2.00\text{x } 10^{-4}$	$4.86\text{x } 10^{-4}$	$2.29 \text{ x } 10^{-3}$	$1.09\text{x}10^{-2}$	$2.00\text{x}10^{-2}$
2 months	$3.71\text{x } 10^{-4}$	$6.92\text{x } 10^{-4}$	$2.25 \text{ x } 10^{-3}$	$1.13\text{x}10^{-2}$	$2.06\text{x}10^{-2}$
3 months	$0.94\text{x } 10^{-3}$	$1.26\text{x } 10^{-3}$	$2.36\text{x } 10^{-3}$	$1.41 \text{ x } 10^{-2}$	$4.91\text{x}10^{-2}$
4 months	$1.67\text{x } 10^{-3}$	$3.52\text{x } 10^{-3}$	$6.64\text{x } 10^{-3}$	$1.49 \text{ x } 10^{-2}$	$6.00\text{x}10^{-2}$

TABLE V

Comparison of imputation methods with missing values occurring over consecutive periods on a single O_3 sensor

Similar trend was observed for the NO_2/O_3 data, with calibration on imputed data performing better than calibration done on CCA, even with less sophisticated algorithm such as MLR (see Table VII)

IX. Conclusion

This study explored imputation techniques for predicting missing values on the datasets of LCS deployed for the quantification of GHGs. Five different imputation techniques were investigated including VAE, NNRW, MICE, missForest and KNN. The analysis shows that at any measurement point, the concentrations of auxiliary variables such as T, RH, and other sensor variables with non-missing values exhibit important correlation that could be exploited by the imputation methods to predict missing values on a target variable. As it would be impossible to assess the performance of imputation strategies when the real values are unknown, we introduced missing values to the datasets following two distinct patterns to assess the ability of the imputation strategies. VAE method shows improved performance over the rest of the competing algorithms for imputation tasks conducted on two real-world datasets including O_3 dataset which consists of three aeroqual O_3 sensors, T, RH and FEM measurements and NO_2/O_3 dataset consisting of three cairclipNO_2/O_3 sensors, T, RH and NO_2/O_3 FEM measurements collected over a six months measurement campaign. Furthermore, the dataset imputed using the VAE method (30% of values imputed) was employed in sensor calibration to ascertain any impact imputed data may have on sensor calibration . The performance of different calibration models including MLR, DT, RF and XGB trained on the imputed datasets were evaluated. The analysis showed that applying imputation to handle missing values on LCS before calibration improved the performance of the sensors, reducing the RMSE existing between raw sensor outputs and FEM monitor outputs by more than 85%.

Due to time and resource constraint, this research has focused on dataset from a limited number of electrochemical gas sensors (3 units each of O_3 and NO_2/O_3 sensors), future research direction can focus on other type of gas sensors such as Non-Dispersive Infrared (NDIR) sensors. While this study has been able to show the effectiveness of data imputation on missing LCS values and the importance of imputation on sensor calibration, further research is required to ascertain the maximum period of time upon which multiple sensors can have missing values for imputation to be valid and applicable for use in sensor calibration tasks.

Acknowledgment

The authors wish to thank the reviewers for their invaluable feedback.

References

[1] MacFaul, L. ‘Monitoring greenhouse gases’, Verification Yearbook 2004, VERTIC, London.

[2] J. Loy-Benitez, S. Heo and C. Yoo, ”Imputing missing indoor air quality data via variational convolutional autoencoders: Implications for ventilation management of subway metro systems”, Building and Environment, vol. 182, p. 107135, 2020. Available: 10.1016/j.buildenv.2020.107135.

[3] B. Wells, A. Nowacki, K. Chagin and M. Kattan, ”Strategies for Handling Missing Data in Electronic Health Record Derived Data”, eGEMs (Generating Evidence and Methods to improve patient outcomes), vol. 1, no. 3, p. 7, 2013. Available: 10.13063/2327-9214.1035.

[4] K. Ha and K. Kwok, ”Dealing with Missing Values in Healthcare Data — BioSymetrics — Drug Discovery Accelerated”, BioSymetrics — Drug Discovery Accelerated, 2020. [Online]. Available: https://www.biosymetrics.com/missing-values-healthcare-data/. [Accessed: 22- Nov- 2020].

[5] K. Sanjar, O. Bekhzod, J. Kim, A. Paul and J. Kim, ”Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method”, ISPRS International Journal of Geo-Information, vol. 9, no. 4, p. 227, 2020. Available: 10.3390/ijgi9040227.

[6] Y. Chen, Y. Lv and F. Wang, ”Traffic Flow Imputation Using Parallel Data and Generative Adversarial Networks”, IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 4, pp. 1624-1630, 2020. Available: 10.1109/tits.2019.2910295.

[7] K. Lakshminarayan, S. Harp and T. Samad, ”Imputation of Missing Data in Industrial Databases”, Applied Intelligence, vol. 11, pp. 259–275, 2020. [Accessed 22 November 2020].

[8] L. Ehrlinger, T. Grubinger, B. Varga, M. Pichler, T. Natschläger and J. Zeindl, ”Treating Missing Data in Industrial Data Analytics,” 2018 Thirteenth International Conference on Digital Information Management (ICDIM), Berlin, Germany, 2018, pp. 148-155, doi: 10.1109/ICDIM.2018.8846984.

[9] ”Air Sensor Guidebook — Science Inventory — US EPA”, Cfpub.epa.gov, 2020. [Online]. Available: https://cfpub.epa.gov/si/si_public_record_report [Accessed: 16- Dec- 2020].

[10] M. S. Osman, A. M. Abu-Mahfouz and P. R. Page, ”A survey on data imputation techniques: water distribution system as a use case,” IEEE Access, vol. 6, pp. 63279-63291, Jun. 2017.

[11] John W Graham. Missing data analysis: Making it work in the real world. Annual review of psychology, 60:549–576, 2009.

[12] M. P. Gómez-Carracedo, J. M. Andrade, P. López-Mahía, S. Muniategui and D. Prada, ”A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets,” Chemom. Intell. Lab. Syst., vol. 134, pp. 23-33, 2014.

[13] M. Azur, E. Stuart, C. Frangakis and P. Leaf, ”Multiple imputation by chained equations: what is it and how does it work?”, International Journal of Methods in Psychiatric Research, vol. 20, no. 1, pp. 40-49, 2011. Available: 10.1002/mpr.329.

[14] D. Stekhoven and P. Buhlmann, ”MissForest–non-parametric missing value imputation for mixed-type data”, Bioinformatics, vol. 28, no. 1, pp. 112-118, 2011. Available: 10.1093/bioinformatics/btr597.

[15] X. Chen, Z. He and L. Sun, ”A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation”, Transportation Research Part C: Emerging Technologies, vol. 98, pp. 73-84, 2019. Available: 10.1016/j.trc.2018.11.003.

	Uncalibrated	MLR		DT		RF		XGB	
		Imputed	CCA	Imputed	CCA	Imputed	CCA	Imputed	CCA
RMSE	0.0091	**0.0029**	0.0036	**0.0014**	0.0036	**0.0013**	0.0019	**0.0011**	0.0020
MAE	0.0082	0.0023	0.0289	0.0013	0.0029	0.0012	0.0015	0.0010	0.0017
R^2	0.5004	**0.9485**	0.9343	**0.9878**	0.9117	**0.9927**	0.9449	**0.9980**	0.9494

TABLE VI

Comparison of imputation and complete case analysis on O_3 sensor calibration

	Uncalibrated	MLR		DT		RF		XGB	
		Imputed	CCA	Imputed	CCA	Imputed	CCA	Imputed	CCA
RMSE	31.7010	**0.1752**	1.0221	**0.9802**	10.2419	**0.7841**	10.4080	**0.7408**	5.9454
MAE	31.1802	0.0284	0.6147	0.2607	9.0950	0.1963	8.9858	0.3083	4.5106
R^2	0.0056	**0.9994**	0.9391	**0.9811**	0.4024	**0.9877**	0.4679	**0.9895**	0.7272

TABLE VII

Comparison of imputation and complete case analysis on NO_2/ O_3 sensor calibration

[16] R. Xie, N. Jan, K. Hao, L. Chen and B. Huang, ”Supervised Variational Autoencoders for Soft Sensor Modeling With Missing Data”, IEEE Transactions on Industrial Informatics, vol. 16, no. 4, pp. 2820-2828, 2020. Available: 10.1109/tii.2019.2951622.

[17] J. McCoy, S. Kroon and L. Auret, ”Variational Autoencoders for Missing Data Imputation with Application to a Simulated Milling Circuit”, IFAC-PapersOnLine, vol. 51, no. 21, pp. 141-146, 2018. Available: 10.1016/j.ifacol.2018.09.406.

[18] D. Mesquita, J. Gomes and L. Rodrigues, ”Artificial Neural Networks with Random Weights for Incomplete Datasets”, Neural Processing Letters, vol. 50, no. 3, pp. 2345-2372, 2019. Available: 10.1007/s11063-019-10012-0.

[19] B. Jiang, M. Siddiqi, R. Asadi and A. Regan, ”Imputation of Missing Traffic Flow Data Using Denoising Autoencoders”, Procedia Computer Science, vol. 184, pp. 84-91, 2021. Available: 10.1016/j.procs.2021.03.122.

[20] J. Yoon, J.Jordon,M.van der Schaar,”GAIN: Missing Data Imputation using Generative Adversarial Nets”, Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden,2018.

[21] D. Snow, ”MTSS-GAN: Multivariate Time Series Simulation Generative Adversarial Networks”, SSRN Electronic Journal, 2020. Available: 10.2139/ssrn.3616557

[22] A. Purwar and S. Singh, ”Hybrid prediction model with missing value imputation for medical data”, Expert Systems with Applications, vol. 42, no. 13, pp. 5621-5631, 2015. Available: 10.1016/j.eswa.2015.02.050.

[23] S.Bowman, L.Vilnis, O.Vinyals, A. Dai, R.Jozefowicz and S.Bengio,”Generating Sentences from a Continuous Space”, Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), pages 10–21,2016.

[24] J. Koloda, J. Ostergaard, S. Jensen, V. Sanchez and A. Peinado, ”Sequential Error Concealment for Video/Images by Sparse Linear Prediction”, IEEE Transactions on Multimedia, vol. 15, no. 4, pp. 957-969, 2013. Available: 10.1109/tmm.2013.2238524.

[25] M. Peralta, P. Jannin, C. Haegelen and J. Baxter, ”Data Imputation and Compression For Parkinson’s Disease Clinical Questionnaires”, Hal.archives-ouvertes.fr, 2021. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02570967v1. [Accessed: 06- Jan- 2021].

[26] A. Kazemi and H. Meidani, ”IGANI: Iterative Generative Adversarial Networks for Imputation Applied to Prediction of Traffic Data”, arXiv:2008.04847

[27] M.Brown, J.Kros, ”Data mining and the impact of missing data”, Emeralds: Industrial Management and data systems vol 103,pp.611-621, 2003

[28] H. Hegde, N. Shimpi, A. Panny, I. Glurich, P. Christie and A. Acharya, ”MICE vs PPCA: Missing data imputation in healthcare”, Informatics in Medicine Unlocked, vol. 17, p. 100275, 2019. Available: 10.1016/j.imu.2019.100275.

[29] A. Gupta and M. Lam, ”Estimating Missing Values Using Neural Networks”, The Journal of the Operational Research Society, vol. 47, no. 2, p. 229, 1996. Available: 10.2307/2584344.

[30] V. Ravi and M. Krishna, ”A new online data imputation method based on general regression auto associative neural network”, Neurocomputing, vol. 138, pp. 106-113, 2014. Available: 10.1016/j.neucom.2014.02.037.

[31] W. Cao, X.Wang, Z. Ming and J.Gao, ” A review on neural networks with random weights”, Neurocomputing, 2018, 275, pp278-287.

[32] W. Cao, Z. Xie, J. Li, Z. Xu, Z. Ming and X. Wang, ”Bidirectional stochastic configuration network for regression problems”, Neural Networks, vol. 140, pp. 237-246, 2021. Available: 10.1016/j.neunet.2021.03.016.

[33] W. Cao, J. Gao, Z. Ming, S. Cai and Z. Shan, ”Fuzziness based Random Vector Functional-link Network for Semi-Supervised Learning”, IEEE Xplore, 2021. Available: 10.1109/CSCI.2017.135 [Accessed 20 June 2021].

[34] B. Beaulieu-Jones and J Moore, ”Missing data imputation in the electronic health record using deeply learned autoencoders”,Pacific Symposium on Biocomputing 2017, pp 207-218

[35] S. Munir, M. Mayfield, D. Coca, S. Jubb and O. Osammor, ”Analysing the performance of low-cost air quality sensors, their drivers, relative benefits and calibration in cities—a case study in Sheffield”, Environmental Monitoring and Assessment, vol. 191, no. 2, 2019. Available: 10.1007/s10661-019-7231-8.

[36] K. Yamamoto, T. Togami, N. Yamaguchi and S. Ninomiya, ”Machine Learning-Based Calibration of Low-Cost Air Temperature Sensors Using Environmental Data”, Sensors, vol. 17, no. 6, p. 1290, 2017. Available: 10.3390/s17061290.

[37] C. Dorffer, M. Puigt, G. Delmaire and G. Roussel, ”Informed Nonnegative Matrix Factorization Methods for Mobile Sensor Network Calibration”, IEEE Transactions on Signal and Information Processing over Networks, vol. 4, no. 4, pp. 667-682, 2018. Available: 10.1109/tsipn.2018.2811962.

[38] D. Hasenfratz et al., ”Deriving high-resolution urban air pollution maps using mobile sensor nodes”, Pervasive and Mobile Computing, vol. 16, pp. 268-285, 2015. Available: 10.1016/j.pmcj.2014.11.008.

[39] H. Kang, ”The prevention and handling of the missing data”, Korean Journal of Anesthesiology, vol. 64, no. 5, p. 402, 2013. Available: 10.4097/kjae.2013.64.5.402.

[40] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre and F. Bonavitacola, ”Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide”, Sensors and Actuators B: Chemical, vol. 215, pp. 249-257, 2015. Available: 10.1016/j.snb.2015.03.031.

[41] L. Spinelle, M. Gerboles, M. Villani, M. Aleixandre and F. 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, vol. 238, pp. 706-715, 2017. Available: 10.1016/j.snb.2016.07.036.

[42] S. De Vito, E. Massera, M. Piga, L. Martinotto and G. Di Francia, ”On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario”, Sensors and Actuators B: Chemical, vol. 129, no. 2, pp. 750-757, 2008. Available: 10.1016/j.snb.2007.09.060.

[43] S. De Vito, M. Piga, L. Martinotto and G. Di Francia, ”CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization”, Sensors and Actuators B: Chemical, vol. 143, no. 1, pp. 182-191, 2009. Available: 10.1016/j.snb.2009.08.041.

[44] N. Okafor, Y. Alghorani and D. Delaney, ”Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data

Fusion and Machine Learning Approach", ICT Express, vol. 6, no. 3, pp. 220-228, 2020. Available: 10.1016/j.icte.2020.06.004.

[45] N. Okafor and D. Delaney, "Application of Machine Learning Techniques for the Calibration of Low-cost IoT Sensors in Environmental Monitoring Networks," 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), New Orleans, LA, USA, 2020, pp. 1-3, doi: 10.1109/WF-IoT48130.2020.9221246.

[46] S. Feinberg et al., "Long-term evaluation of air sensor technology under ambient conditions in Denver, Colorado", Atmospheric Measurement Techniques, vol. 11, no. 8, pp. 4605-4615, 2018. Available: 10.5194/amt-11-4605-2018.

[47] "EPA Environmental Dataset Gateway", Edg.epa.gov, 2021. [Online]. Available: https://edg.epa.gov. [Accessed: 06- Apr- 2021].

[48] W. Jiao et al., "Community Air Sensor Network (CAIRSENSE) project: evaluation of low-costsensor performance in a suburban environment in the southeastern UnitedStates", Atmospheric Measurement Techniques, vol. 9, no. 11, pp. 5281-5292, 2016. Available: 10.5194/amt-9-5281-2016.

[49] B. Maag, Z. Zhou and L. Thiele, "A Survey on Sensor Calibration in Air Pollution Monitoring Deployments", IEEE Internet of Things Journal, vol. 5, no. 6, pp. 4857-4870, 2018. Available: 10.1109/jiot.2018.2853660.

[50] A. Rai et al., "End-user perspective of low-cost sensors for outdoor air pollution monitoring", Science of The Total Environment, vol. 607-608, pp. 691-705, 2017. Available: 10.1016/j.scitotenv.2017.06.266.

[51] F. Delaine, B. Lebental and H. Rivano, "In Situ Calibration Algorithms for Environmental Sensor Networks: A Review", IEEE Sensors Journal, vol. 19, no. 15, pp. 5968-5978, 2019. Available: 10.1109/jsen.2019.2910317.

[52] T. Phan, É. Poisson Caillault, A. Lefebvre and A. Bigand, "Dynamic time warping-based imputation for univariate time series data", Pattern Recognition Letters, vol. 139, pp. 139-147, 2020. Available: 10.1016/j.patrec.2017.08.019.

[53] "Anaconda — The World's Most Popular Data Science Platform", Anaconda, 2020. [Online]. Available: https://www.anaconda.com/. [Accessed: 17- Dec- 2020].

[54] Bouhlila, D. and Sellaouti, F., 2013. Multiple imputation using chained equations for missing data in TIMSS: a case study. Large-scale Assessments in Education, 1(1)
.

[55] "A practical guide to multiple imputation of missing data in nephrology", Kidney International. 6-Apr.-2020. [Online]. Available: www.kidney-international.org/article/S0085-2538(20)30951-0/fulltext. [Accessed: 4-Dec.-2020].

[56] I. Mayer, J. Josse, N. Tierney and N. Vialaneix, "R-miss-tastic: a unified platform for missing values methods and workflows", arXiv.org, 2021. [Online]. Available: https://arxiv.org/abs/1908.04822. [Accessed: 23- Mar- 2021].

[57] D. Kingma and M.Welling, "Auto-Encoding Variatioanl Bayes".https://arxiv.org/abs/1312.6114

[58] "Euredit", Cs.york.ac.uk, 2021. [Online]. Available: https://www.cs.york.ac.uk/euredit/. [Accessed: 21- Jun- 2021].

[59] J.Schafer and M. Olsen, "Multiple imputation for multivariate missing data problems: a data analyst's perspective", Multivariate Behavioral Research 1998, 33: 545–571. 10.1207/s15327906mbr33045

[60] J.Schafer "NORM: Multiple imputation of incomplete multivari-ate data under a normal model [computer software]", University Park, PA:Department of Statistics, Pennsylvania State University.

[61] Graham, J., Olchowski, A. and Gilreath, T., "How Many Impu-tations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory", Prevention Science,2017, 8(3), pp.206-213.

[62] "Applied Predictive Modeling", Applied Predictive Modeling, 2020. [Online]. Available: http://appliedpredictivemodeling.com/. [Accessed: 16- Dec- 2020].

[63] X. Pang, M. Shaw, A. Lewis, L. Carpenter and T. Batchellier, "Electrochemical ozone sensors: A miniaturised alternative for ozone measurements in laboratory experiments and air-quality monitoring", Sensors and Actuators B: Chemical, vol. 240, pp. 829-837, 2017. Available: 10.1016/j.snb.2016.09.020.

[64] J. Cordero, R. Borge and A. Narros, "Using statistical methods to carry out in field calibrations of low cost air quality sensors", Sensors and Actuators B: Chemical, vol. 267, pp. 245-254, 2018. Available: 10.1016/j.snb.2018.04.021.

[65] M. Badura, P. Batog, A. Drzeniecka-Osiadacz and P. Modzel, "Regression methods in the calibration of low-cost sensors for ambient particulate matter measurements", SN Applied Sciences, vol. 1, no. 6, 2019. Available: 10.1007/s42452-019-0630-1.

[66] M. Cueto, "Grid Search in Python from scratch— Hyperparameter tuning", Medium, 2020. [Online]. Available: https://towardsdatascience.com/grid-search-in-python-from-scratch-hyperparameter-tuning-3cca8443727b. [Accessed: 19- Apr- 2021].

[67] N. Zimmerman et al., "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring", Atmospheric Measurement Techniques, vol. 11, no. 1, pp. 291-313, 2018. Available: 10.5194/amt-11-291-2018.

[68] Y. Wang, Y. Du, J. Wang and T. Li, "Calibration of a low-cost PM2.5 monitor using a random forest model", Environment International, vol. 133, p. 105161, 2019. Available: 10.1016/j.envint.2019.105161.

[69] "3.2.4.3.2.sklearn.ensemble.RandomForestRegressor-scikit-learn 0.23.2 documentation",Scikit-learn.org,2020.[Online].Available: https://scikit-learn.org/stable/modules/generated/
sklearn.ensemble.RandomForestRegressor.html.[Accessed: 17 Dec 2020].

[70] Y. Lin, W. Dong and Y. Chen, "Calibrating Low-Cost Sensors by a Two-Phase Learning Approach for Urban Air Quality Measurement", Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol. 2, no. 1, pp. 1-18, 2018. Available: 10.1145/3191750.

[71] M. Si, Y. Xiong, S. Du and K. Du, "Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods", Atmospheric Measurement Techniques, vol. 13, no. 4, pp. 1693-1707, 2020. Available: 10.5194/amt-13-1693-2020.

[72] J.Bergstra, Y.Bengio, "Random Search for Hyper-Parameter Optimization", Journal of Machine Learning Research no.13, pp. 281-305, 2012

Nwamaka U. Okafor is currently a PhD student in the School of Electrical and Electronic Engineering, University College Dublin. Her research is focused on the application of Internet of Things (IoTs) technologies, Machine Learning and AI in ecological sensing. Working with the team on the SmartBOG project on Irish peatlands, she develops and deploys cost-effective IoT and AI solutions to corroborate remote and satellite-based surveillance. Nwamaka received her MSc in Computer Forensics and Cyber Security (Distinction) from the University of Greenwich, London in 2016. Nwamaka is also a lecturer in the department of Computer Science, Federal Polytechnic Nekede, Nigeria.

Declan T. Delaney is currently an Assistant Professor at the School of Electrical and Electronic Engineering, UCD. Declan received his Ph.D. in network analysis and design for IoT at the School of Computer Science, UCD in 2015. Declan is an SFI Funded Investigator on the project CONSUS (www.consus.com), an SFI-industry funded collaboration focused on precision agriculture, and a Principal Investigator for the SmartBOG project (www.smartbog.com). Declan's research interests are in the areas of network data analytics for adaptable programmable networks and infrastructure and data assurance for IoT and sensor systems. Having previously worked at LMI Ericsson and collaborations with SMEs in H2020 funding proposals, Declan maintains strong links with industry partners.