Building and Environment 182 (2020) 107135

Contents lists available at ScienceDirect

Building and Environment

journal homepage: http://www.elsevier.com/locate/buildenv

Imputing missing indoor air quality data via variational convolutional autoencoders: Implications for ventilation management of subway metro systems

Jorge Loy-Benitez, SungKu Heo, ChangKyoo Yoo *

Department of Environmental Science Engineering, College of Engineering, Kyung Hee University, Yongin, 446-701, South Korea

ARTICLE INFO

Keywords:
Convolutional neural networks
Monitoring sensor reliability
Indoor air quality
Subway ventilation management
Variational autoencoders

ABSTRACT

Missing data represents a common problem in environmental and building-related processes, especially in the indoor air quality (IAQ) system of subway stations, where the collected information leads to actions in ventilation management. For these reasons, imputation approaches have been used to avoid information loss due to downsampling or sensor malfunction. This paper introduces an imputation approach for IAQ data via variational autoencoders (VAE) coupled with convolutional layers (VAE-CNN). Two scenarios were introduced: first, the IAQ dataset was corrupted by removing data intervals at different missing rates (i.e., 20%, 50%, and 80%), and second, a point-to-point removal of three sensors was conducted. The performance of the proposed method was compared with different techniques, showing that the VAE-CNN was superior to other methods even for massive amounts of missing data. Finally, the effects of missing and imputed data on the IAQ system in the D-subway station were evaluated in terms of ventilation energy demand, CO_2 emissions, and IAQ level. The IAQ management with the imputed data could decrease by approximately 20% of CO_2 emissions by reducing the energy demand, while the IAQ level increased by 3% in another scenario.

1. Introduction

Along with worldwide rapid urbanization, the concept of smart cities (SCs) has become a strong trend. Accordingly, many technologies have emerged to meet the requirements of SCs, including the internet of things (IoT), machine learning, and big data applications [1]. The research community has employed these technologies paying particular attention to the need for SCs to monitor air quality in the context of environmental sustainability [2,3]. Moreover, smart monitoring and control of indoor air quality (IAQ) have become necessary to deal with public health threats in highly contaminated indoor environments [4,5].

In the subway metro environment, several air pollutants are caused by emissions from the piston effect, rail-wheel-break interaction, catenaries, and infiltration from the outdoors [6]. These microenvironments are affected by the accumulation of volatile organic compounds (VOCs), CO_2, and particulate matter (PM) rich in iron. Additionally, long-term exposure to these pollutants is closely related to severe illnesses, including pulmonary and cardiovascular diseases, premature births, cancer, and even death [5,7–10]. These facts have triggered alerts, especially in developed countries, where the metro is a significant transportation system. For instance, in Seoul, South Korea, approximately seven million people commute in underground spaces daily, spending a considerable amount of time exposed to these conditions [11]. Therefore, the Korean ministry of environment (MOE) acknowledged the subway space as a potential threat to public health, proposing several countermeasures such as the installation of telemonitoring systems (TMS), and ventilation systems to regulate the accumulation of indoor pollutants [12,13].

The TMS is generally installed in the subway platform and canopy; it collects hourly measurements of indoor pollutants, including CO, CO_2, NO, NO_2, NO_X, and PM with aerodynamic diameters of less than 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$). Moreover, meteorological variables are collected, including temperature and humidity [14]. These monitoring sensors provide crucial information for the supervisory management of the ventilation systems. However, hardware sensors are prone to suffer various failures, including bias, precision degradation, drifting, and the loss of a considerable volume of data due to a hostile underground environment or operability issues, making the sensors report unrealistic IAQ measurements [15,16].

Missing data is an issue that occurs in several real-world datasets

* Corresponding author.
E-mail address: ckyoo@khu.ac.kr (C. Yoo).

https://doi.org/10.1016/j.buildenv.2020.107135
Received 29 April 2020; Received in revised form 26 June 2020; Accepted 13 July 2020
Available online 27 July 2020

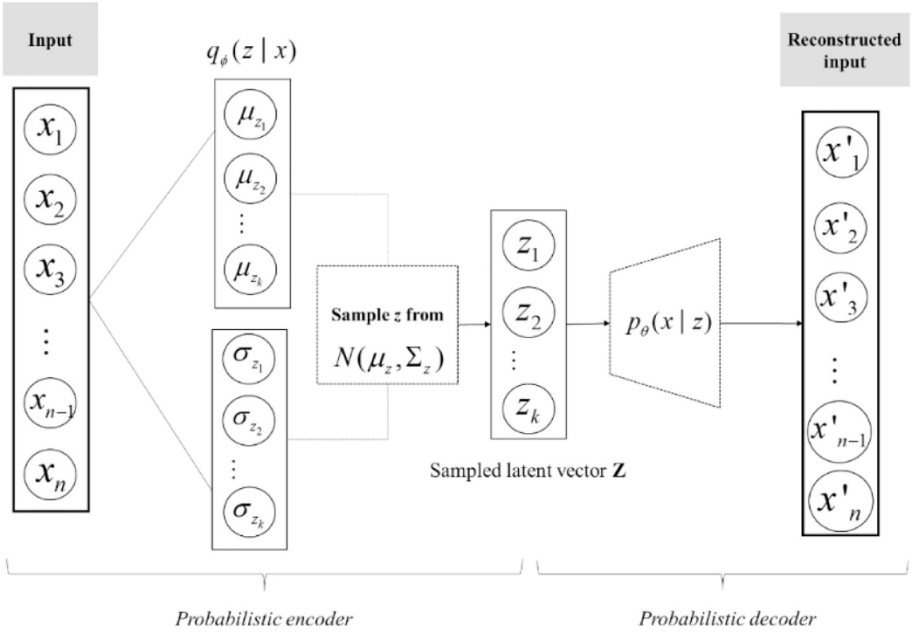


Fig. 1. Illustration of the VAE architecture.

regarding industrial operations, traffic information, medical records, and mechanical systems [17–19]. Various issues emerge when the data is subsequently used for tasks like supervisory control and tuning controllers resulting in slow control responses when downsampling data to match the measurement with the lowest frequency or even take erroneous control actions when discarding a data point. Additionally, on the development of monitoring techniques, missing data decrease the statistical power by reducing the representativeness of samples and evidencing a biased parameter estimation when modeling anomaly detection or data reconciliation approaches [15,18]. Considering that this matter has become a practical challenge for data mining researchers, many techniques have been developed to impute (estimate) missing data. These imputation strategies aim to replace missing values with plausible values to reduce the loss of information [20].

In the case of the metro environment, the ventilation system is controlled based on the data collected from the TMS. Therefore, missing values can cause incorrect interpretations when managing the IAQ, resulting in consequences such as energy waste, releasing unnecessary CO_2 emissions to the atmosphere, or sudden contamination of the indoor environment [15,16]. For these reasons, introducing a robust imputation method for the subway IAQ monitoring could result in more uptime for ventilation management, refining the sustainable balance between the competing objectives in subway stations: ventilation energy demand and the IAQ level.

Several investigations have been devoted to identifying effective imputation methods. These strategies include statistical-based methods [17,21,22], model-based methods [23], and machine learning methods [18,22,24]. Active research in machine learning methods has been widely conducted in the branch of neural networks for missing data imputation approaches [25,26]. Particularly, autoencoders (AE) have shown great performance in the field of missing data imputation when handling to learn hidden representations of real-world datasets with non-linear dependencies [17]. For instance, Duan et al. (2016) [27] developed a denoising stacked autoencoder (DSAE) for traffic data imputation, outperforming popular approaches including the autoregressive integrated moving average (ARIMA) and backpropagation neural network (BPNN) models. Bianchi et al. (2019) [28] proposed the temporal kernelized autoencoder (TKAE) model to learn compressed representations of multivariate time series. Then, two frameworks were suggested for imputing missing data and for one-class classification tasks.

Further studies acknowledged the advantages of utilizing variational autoencoders (VAE) over typical AE structures. As both of these structures can handle non-linear data, and aim to reconstruct the input data while learning hidden representations; unlike AE, the latent space of the VAE is continuous, the encoder comes from an inference model, and the decoder is a generative model [29,30]. VAE models are novel architectures that have been employed to impute image-based data and industrial processes, including milling circuits [18]. Contrary to model-based and standard AE models, VAE models are known as probabilistic AE since they come from Bayesian inference, modeling the potential probability distribution of the data, and generating new samples from this distribution as highly desired in imputation tasks [29,31].

In general, AE models use different types of layers within their architectures to map meaningful representations of the data for different purposes, including feature learning, fault detection, multivariate monitoring, dimensionality reduction, or denoising reconstructions [16, 32–34]. The tendency of utilizing neural methods for feature learning lays in the flexibility emerging for non-linear data provided by available non-linear activation functions, i.e., Sigmoid or Tanh, contrary to state-of-art methods as principal component analysis (PCA) or independent component analysis (ICA) that consider only linear transformations [34]. Within the AE structure, fully connected dense layers are the most common evidenced in previous studies [34,35]. However, new approaches have determined that dense layers are not an ideal alternative when dealing with dynamic information in a time series.

Therefore, recurrent neural networks (RNN), and memory gated structures as gated recurrent units (GRU) and long short term memory (LSTM) are introduced to the research community to cope with *seq2seq* problems [36,37]. Most of these architectures have been employed in the fields of forecasting [38–41], video representation [42], and fault diagnosis [36,43]. Although these architectures have exhibited

state-of-the-art performances in time series modeling, many recent approaches indicate that convolutional neural networks (CNN) outperform the RNN in certain tasks such as machine translation and audio synthesis [44]. For instance, Bai et al. (2018) [45] evaluated the performance of CNN via a temporal convolutional network (TCN) and RNN utilizing several tasks of sequence modeling. Here, simple CNN structures outperformed the canonical LSTM structures in several datasets. They concluded that the CNN models should be first considered as the starting point for *seq2seq* problems.

The present study aims to introduce a deep learning-based imputation approach for the missing IAQ data in a subway station by combining the VAE structures with 1D-CNN layers (VAE-CNN). Joining these structures aim to exploit the capability of modeling the data distribution to generate new samples for proper imputation considering the multivariate IAQ non-linear dependencies and dynamic characteristics. First, the VAE-CNN model is trained using a complete record of the IAQ measurements. Second, different missing data rates were synthetically applied to a validation dataset yielding two imputation cases: 1) imputation of interval missing data and 2) point-to-point imputation. Then, the proposed method was evaluated and compared with other imputation techniques. Finally, experiments were conducted in the ventilation system of the D-subway station, Seoul metro, in which the effects of missing and imputed data were assessed in terms of ventilation energy demand and quantitative health risk. The remainder of this paper is as follows. Section 2 introduces the background of CNN and VAE. Section 3 presents the IAQ data collection and ventilation management in the D-subway station alongside the model training scheme and validation cases. Section 4 presents the procedure for the assessment of the effects of the proposed imputation model on ventilation management in the D-subway station. Section 5 discusses the performance of the VAE-CNN against other imputation approaches; moreover, the implications of the data imputation in the ventilation system of the D-subway station are analyzed. Finally, the conclusions are presented in Section 6.

2. Background

2.1. Variational convolutional autoencoders

This study leverages the advantages of incorporating convolutional layers into a VAE structure to account for time dynamic representation as stacked layers of deep AE for IAQ data imputation. In the following subsections, background information related to the VAE models and the CNN architectures is explained.

2.1.1. Variational autoencoders

Contrary to regular AE structures, VAEs are well known as generative models that combine Bayesian inference with deep neural networks (DNN). While AE learns an underlying representation of the data, VAE learns the distribution of the latent representation space. As shown in Fig. 1, VAE structure consists of two main parts: 1) the encoder $q_\phi(Z|X)$ and 2) the decoder $p_\theta(X|Z)$. Both of them are multilayered neural networks with parameters ϕ and θ, respectively [18,46].

VAE in the context of a generative model follows the assumption that the data X is generated by an intrinsic distribution $p(X)$ that can be represented by the latent variable Z. At the same time, Z is generated by a distribution $p(Z)$. Therefore, the joint distribution $p(X, Z)$ is represented by $p(X,Z) = p_\theta(X|Z)p(Z)$. Here, the distribution is generated by sampling the prior distribution of Z (i.e., $p(Z)$), and the distribution of X given Z, known as the likelihood $p_\theta(X|Z)$, is the probabilistic decoder. On the other hand, the prior $p(Z)$ follows a normal distribution with zero mean and unit variance with no additional parameters to be learned.

The optimal parameter θ is obtained by maximizing the marginal likelihood $p_\theta(X) = \int p(Z)p_\theta(X|Z)dZ$. The maximization of the marginal likelihood represents a severe problem since it is not possible to exhaust the latent variable, making this process unmanageable [31]. To avoid this problem, the log-likelihood $\log p_\theta(X)$ is represented in Eq. (1).

$$\log p_\theta(X) = E_{q_\phi(Z|X)}[\log p_\theta(X|Z)] - D_{KL}(q_\phi(Z|X)||p(Z)) + D_{KL}(q_\phi(Z|X)||p_\theta(Z|X)) \tag{1}$$

Here, the calculation includes both the probabilistic encoder (i.e., $q_\phi(Z|X)$) and decoder (i.e., $p_\theta(X|Z)$). Moreover, $(D_{KL}(S||T)$ is the Kullback-Leibler divergence describing the agreement of two continuous distributions S and T [47]. The form of the divergence is explained in Eq. (2).

$$D_{KL}(S||T) = \int_{-\infty}^{\infty} s(x)\log\frac{s(x)}{t(x)}dx, \tag{2}$$

where s and t are the probability densities of S and T.

Since the Kullback-Leibler divergence tosses non-negative expressions, the third term of Eq. (1) can be neglected when maximizing the log-likelihood of the marginal distribution, resulting in a variational lower bound (*ELBO*), represented as a loss function in Eq. (3) [31].

$$L(\phi,\theta;X) = E_{q_\phi(Z|X)}[\log p_\theta(X|Z)] - D_{KL}(q_\phi(Z|X)||p(Z)) \tag{3}$$

When the *ELBO* is maximized, optimal parameters ϕ and θ are obtained for the encoder and the decoder, respectively. These parameters can be obtained by stochastic gradient descent (SGD) or other approaches suitable for the training of neural networks. The optimal decoder is then utilized as a generator to fill the voids of missing values according to the data distribution.

2.1.2. Convolutional neural networks

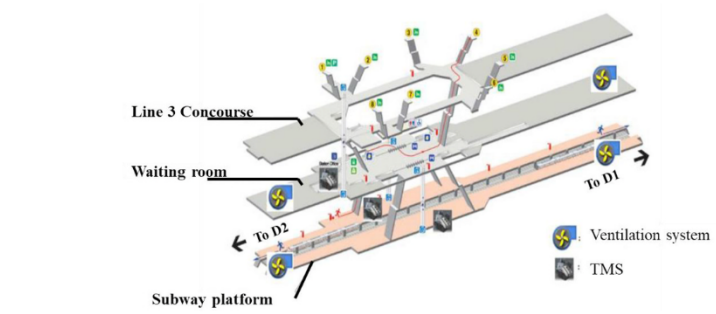
Originally proposed by LeCun et al. (1989), CNN has been employed for sequence modeling over the past decades in such tasks as computer vision and natural language processing [45,48]. Contrary to the conventional neural networks, the layers of a CNN model use a convolutional operation instead of general matrix multiplication followed by pooling operation [44]. Time series datasets are considered to be 1-D grids taking samples at regular time intervals. The input for the one-dimensional convolutional layer (Conv-1D) is represented by X. Then, the network aims to learn a set of parameters Φ to generate a reconstruction X', which is ideally identical to the input. This process is conducted by a hierarchical feature extraction as follows:

$$X' = F(X|\Phi) = f_R(\ldots f_2(f_1(X|\Phi_1)|\Phi_2)|\Phi_R) \tag{4}$$

where R represents the number of hidden layers. Then, the operation for the rth layer is described by Eq. (5).

$$X'_r = f_r(X_r|\Phi_r) = \sigma(W \otimes X_r + b), \quad \Phi_r = [W, b] \tag{5}$$

Here, X_r is a two-dimensional input matrix counting with N feature maps, W is a set of N one dimensional kernels, b is a bias vector, $\otimes$ is the convolutional operation, and $\sigma(\cdot)$ is the activation function. In addition to the convolutional layers, pooling layers are applied for increasing the area covered by the next fields. Then, the output from the convolutional layer is flattened so it can be used as an input of fully connected layers.



(a)

(b)

(c)

Fig. 2. (a) Schematic of the D-subway station with underground levels. TMS installed in the (b) waiting room and in (c) the subway platform.

Table 1
Features of the ventilation system in the D-subway station.

Feature	Unit	Value
Ventilation units	n	4
Ventilation capacity	m^3/h	60,000
Frequency inverter interval	Hz	20–60
Filter efficiency	–	0.8
Static pressure	mmHg	105
Airflow rate per frequency inverter unit	m^3/Hz	1000

3. Materials and methods

3.1. Subway telemonitoring system and ventilation control

The validation of this study is based on experiments conducted in the D-subway station as part of line 3 in the Seoul metro, South Korea. This section aims to explain the conventional regulation of the ventilation system and monitoring techniques as well as their relationship.

The D-subway station has a commuting capacity of approximately 2000 people per hour through its two underground levels. The shallow

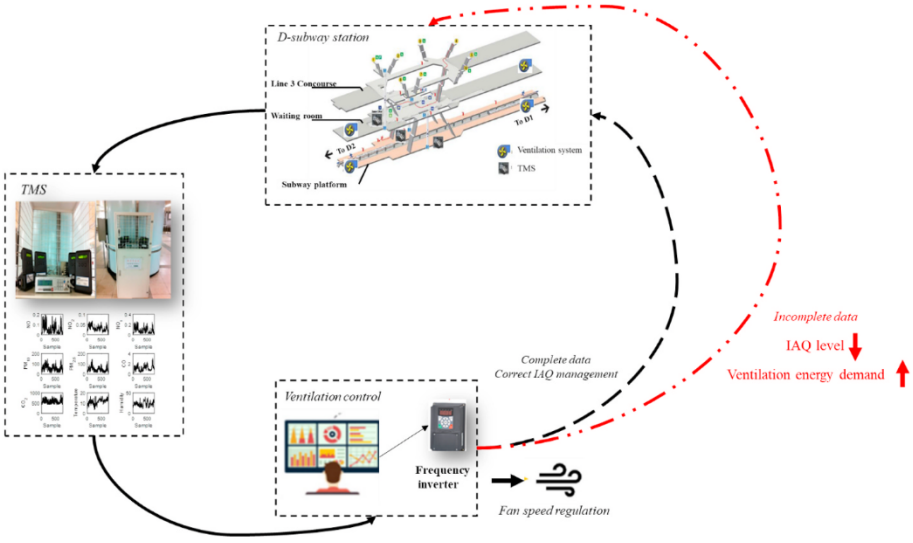


Fig. 3. Illustration of the conventional ventilation control system governed by the TMS in the D-subway station.

Table 2
Specifications of the TMS sampling devices.

Device	Accuracy	Detection limit
CO_2 analyzer (NDIR gas analyzer)	±1% of full span	0.1 ppm (0–5000 ppm)
NO_2 analyzer (NA-623)	±1% of full span	0.5 ppm (0–1 ppm)
PM_{10} analyzer (SPM-613D)	±0.5% of full span	less than ± 1 $\mu g/m^3$
$PM_{2.5}$ analyzer (SPM-613)	±2% of full span	less than ± 1 $\mu g/m^3$

level consists of the waiting room, located around 10 m below ground, and the subway platform located approximately 24 m underground. The subway platform is the "facing" type since the people who want to transit to one direction face the others who want to go on the opposite side separated by the railways (see Fig. 2(a)). The trains operate from 5:00 a.m. to 11:00 p.m. with a waiting interval of 10 min, except in rush hours, in which the waiting time tends to decrease to 2–3 min. Each level has two ventilation units with similar characteristics, as detailed in Table 1.

TMS is installed in both the waiting room and the subway platform (Fig. 2(b and c)). Moreover, it plays a crucial role in the subway ventilation control since the IAQ level monitored in this stage allows the air delivery from the ventilation system to increase/decrease accordingly. The IAQ management of the D-subway station is governed by the regulation of a mechanical ventilation system. Frequency inverter (FI) apparatuses generate control signals for the manipulation of the revolution speed of the ventilation fans by altering the frequency to power at the desired speed and torque [16,49]. Fig. 3 depicts the conventional monitoring system, which consists of capturing the IAQ measurements from the TMS and sending these signals to the FI. This way, air delivery is regulated by adjusting the ventilation speed; therefore, a void in the IAQ information may lead to improper corrective actions, resulting in poor regulation of the ventilation system.

3.2. Data collection and IAQ measurement

For this study, the collected IAQ data consists of 9 variables, including NO, NO_2, NO_X, PM_{10}, $PM_{2.5}$, CO, CO_2, temperature, and humidity. Table 2 shows some specifications related to the sampling apparatuses used by the TMS. The dataset consists of 716 samples collected for a month at an hourly resolution, as depicted in Fig. 4(a). Then, it is separated into two parts: 1) the training set, and 2) the test set, in a proportion of 60% and 40%, respectively. The training set is utilized for modeling the VAE-CNN and obtaining the optimal parameters for the encoder and decoder. On the other hand, the test set is used for

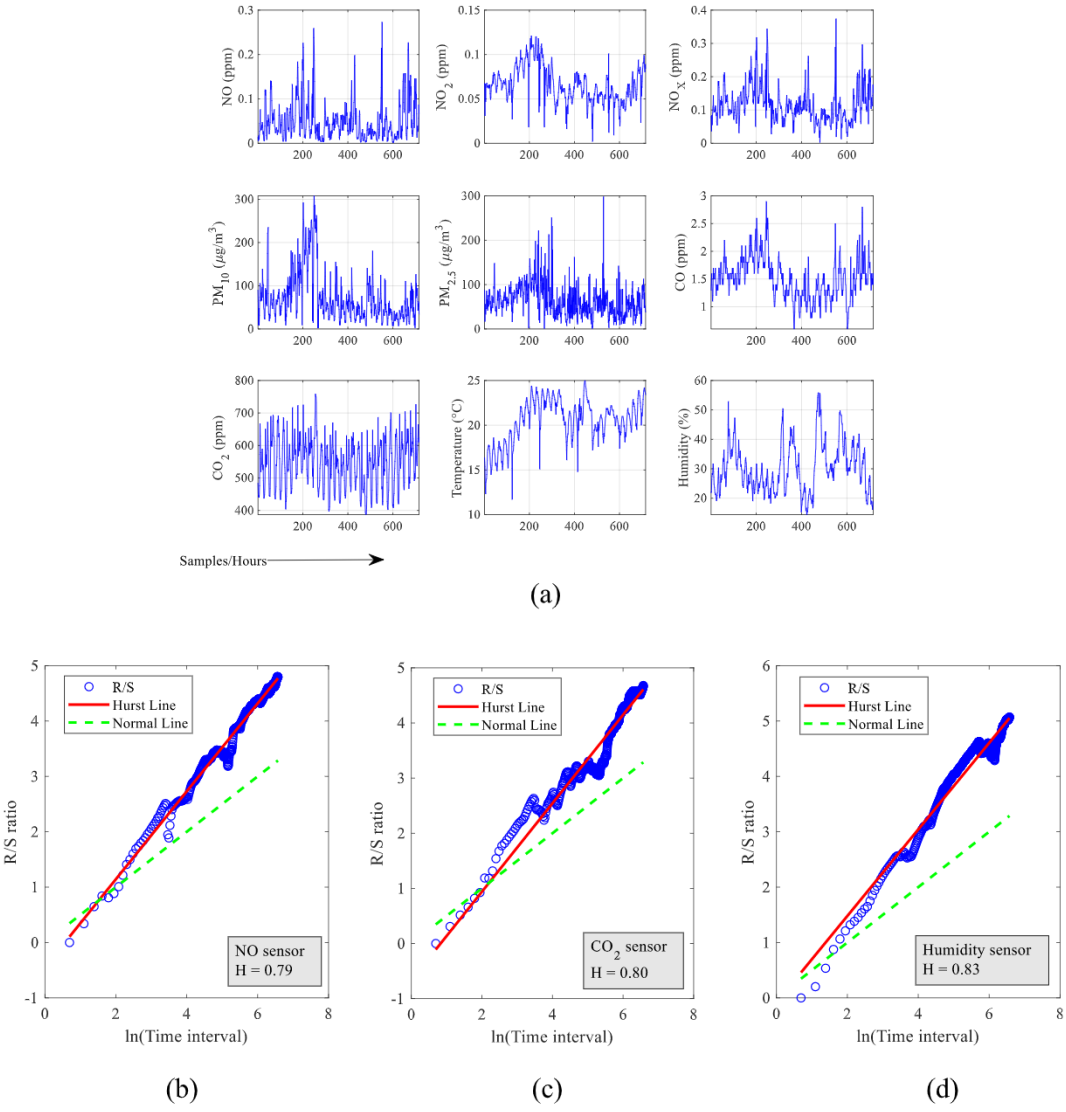


Fig. 4. (a) Variations of the TMS measurements for the IAQ in the D-subway station. R/S analysis for Hurst exponent determination for the (b) NO sensor, (c) the CO_2 sensor, and (d) the humidity sensor.

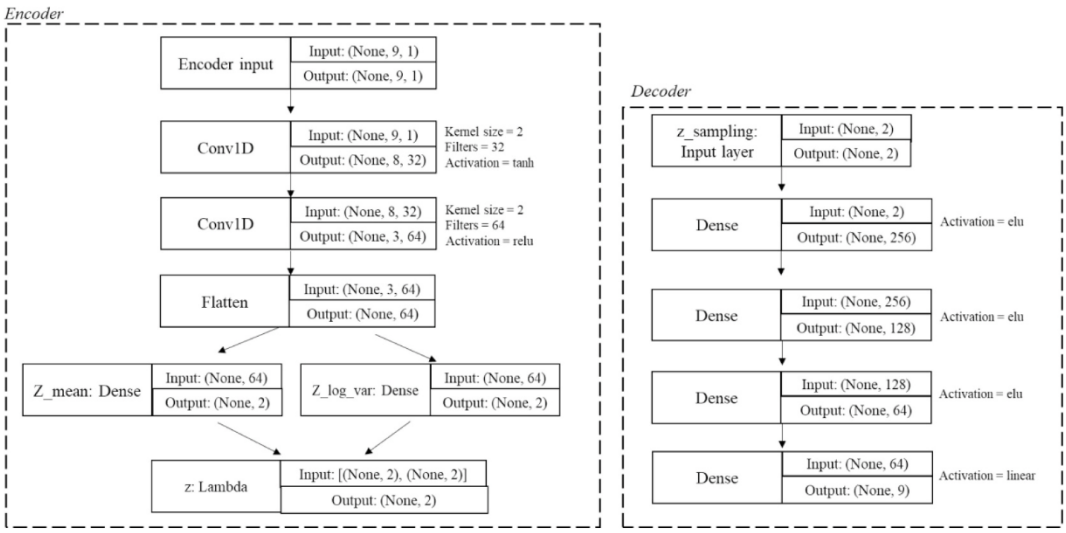


Fig. 5. Illustration of the implemented VAE-CNN model for training the missing IAQ data imputation.

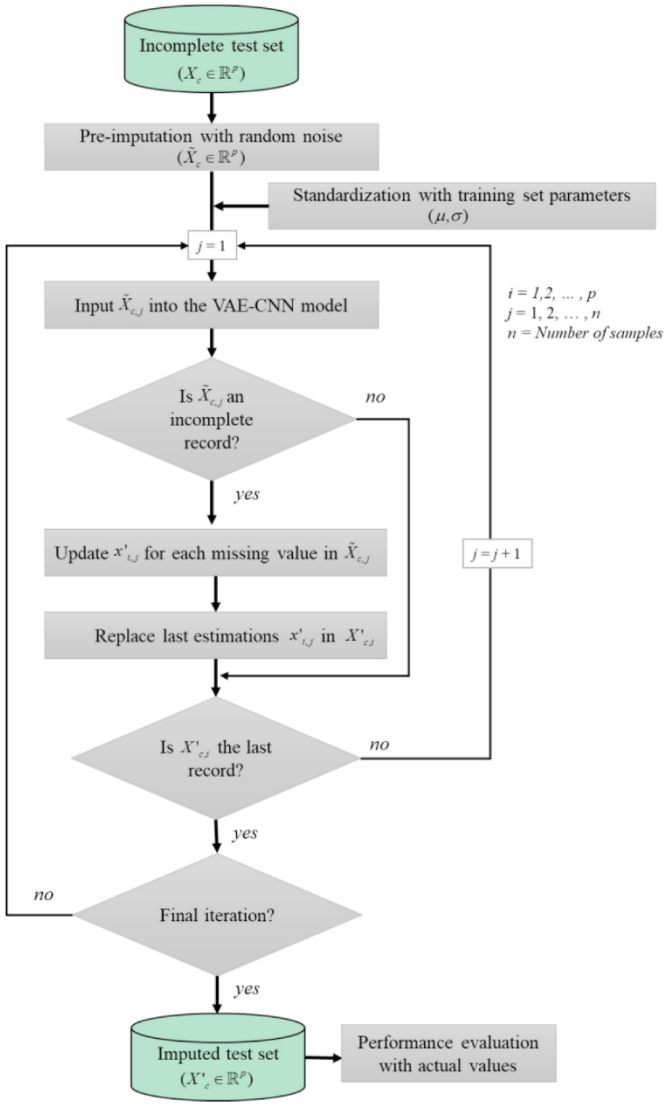


Fig. 6. Flowchart of the validation procedure for the proposed VAE-CNN imputation method.

validating the model on two main scenarios described in subsequent sections.

Previous investigations have employed static and dynamic approaches to monitor the subway IAQ. For instance, PCA [50] and ICA [51] have been applied for sensor fault validation. However, two principal assumptions are considered in this study. First, data follow a multivariate Gaussian distribution, and second, the data are independent of past values. To address these issues, dynamic approaches, such as the dynamic independent component analysis (DICA), account for the non-Gaussian and dependency characteristics of the subway IAQ data [15].

Preliminary examinations were conducted to illustrate these features of the dataset. The Hurst exponent (H) is a parameter between 0 and 1 that indicates whether the observations are independent or not. If H is equal to 0.5, the time-series data have Brownian motion properties (independent). In contrast, if H values are lower or higher than 0.5, there is dependence on past values, for which the data follows an anti-persistent ($0 < H < 0.5$) or persistent ($0.5 < H \leq 1$) behavior [52]. For instance, Fig. 4(b-d) depicts the rescaled range (R/S) plot for the NO, CO_2, and humidity sensors with values of $H = 0.7936$, $H = 0.8024$, and $H = 0.8301$, respectively. It should be noted that these sensors show a high H value, leading us to conclude that the time-series is persistent. In other words, the correlation with past values is positive. The Hurst line indicates the actual tendency of the analyzed sensor measurements, while the green dashed line (Normal line) is a reference line that traces a *Gaussian* distribution tendency. Therefore, if the Hurst line and the normal line coincide, it can be concluded that the dataset is *Gaussian* distributed. Otherwise, as shown in Fig. 4(b–d), it is evident that the distribution of the dataset is far from normality for the analyzed sensors, being these datasets *non-Gaussian* distributed. For these reasons, static and linear approaches such as PCA cannot be used for feature-learning of subway IAQ due to its *non-Gaussian* distribution and dynamism. Therefore, this study aims to leverage a neural approach that learns the distribution existing among the IAQ sensors to generate reliable estimations of missing values.

3.3. Development of the imputation approach via VAE-CNN

The proposed method is presented in this section. First, the training set corresponding to 60% of the total dataset was used to train the *VAE-CNN* network. Here, the structure of the model is explained. Second, two validation scenarios are suggested using the remaining 40% of the dataset (i.e., test set). Selecting the proportions for splitting the dataset

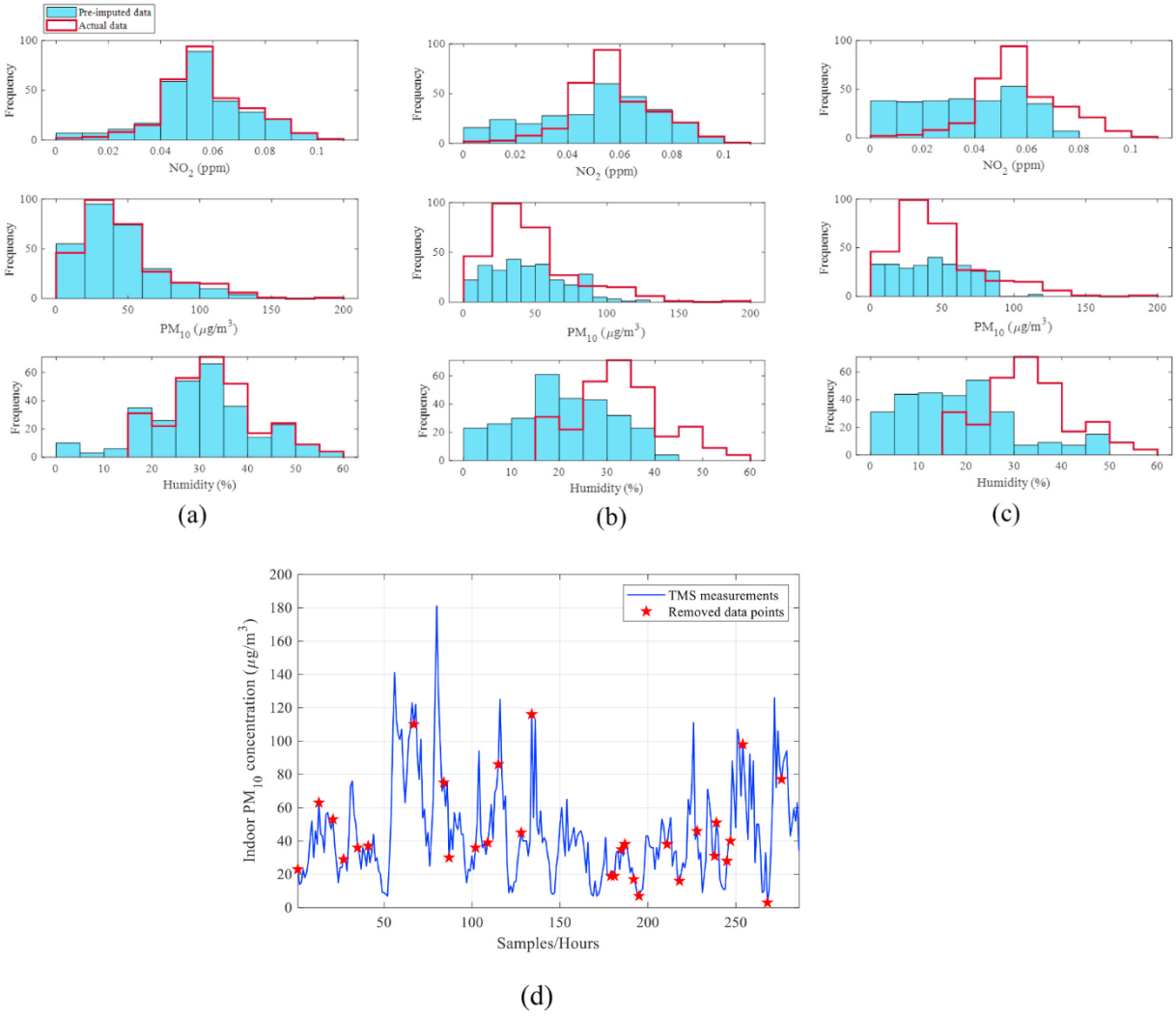


Fig. 7. Comparison of pre-imputed and actual IAQ data based on their distributions with missing data rates of (a) 20%, (b) 50%, and (c) 80%, and (d) point-to-point removed data points for the PM_{10} sensor.

affects the performance of any machine learning task due to high variability in the parameter estimation [53]. The dataset split in this study was based on a thumb rule as considered in past investigations [38,54,55]. Finally, a performance analysis was conducted between the proposed method and other imputation approaches.

3.3.1. *Training the VAE-CNN model*

In general, the objective of the *VAE* model as a deep generative model is to generate new data based on distribution encoding/decoding [31]. For model training, the input and output vectors are the same ($X \rightarrow X$). These vectors consist of 9 neurons, representing the number of TMS sensors. The model encodes the input as a distribution over the latent space; then, a point from the latent space is sampled from this distribution to finally be decoded into a set (X'), that is ideally identical to X. The training procedure ensures that the model had learned the necessary parameters to generate a reliable estimation/imputation when introducing missing data into the model. The concrete description for the training procedure is explained in the following steps:

Step 1: Collect the IAQ training dataset (X) containing complete measurements in normal conditions.
Step 2: Normalize the training set with zero-mean and unit variance, saving the parameters for the validation procedure.
Step 3: Design the architecture of the VAE-CNN and tune associated hyperparameters.
Step 4: Train the VAE-CNN model, setting the input and the output vectors of the model as the same X.

Let $X \subseteq \mathbb{R}^p$ represent the training set, where p is the number of variables in the TMS measurements. Then, this set is standardized with zero-mean and unit variance; these parameters are saved to normalize the test set in the validation scenarios. The ultimate goal of the proposed model is to generate the data imputation $\{\mathbf{x}_j^{'}\}_{j=1}^{N}$, from a latent variable $\mathbf{z}$. Here, N represents the number of samples. As explained in section 2.1.1, two main parts are conducted for training the VAE-CNN. First, $\mathbf{z}^j$ is generated from the prior distribution $p_\theta(\mathbf{z})$. Second, the imputed data $\left(\mathbf{x}_\mathbf{j}^{'}\right)$ is generated from some conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$. θ represents the parameters of the model, obtained by maximizing the ELBO. The Adam optimization method is selected given its capability to be invariant to the constant rescaling of the objective. Fig. 5 illustrates the core of the VAE-CNN model with its corresponding dimensions and main parameters. The encoder comprises two stacked Conv-1D layers having in common 2 kernels, the number of filters and activation functions are 32 and tanh, and 64 and relu, for the first and second layer, respectively. Then, one dense layer is added to generate the mean and log-variance to simplify the calculation of the loss. The output from the encoder is the sampling

of z (latent representation). The decoder, on the other hand, corresponds to three stacked dense layers with 256, 128, and 64 neurons, respectively. Moreover, the activation function for these layers is elu. The output layer is comprised of nine neurons with a linear activation. Here, the latent representation is converted to an imputation approximately similar to the inputs.

Computationally, the model was trained with Keras, a Python deep learning library, with 250 epochs and a batch size of 72.

3.3.2. Validation scenarios and performance analysis

Once the model is trained, two validation scenarios are introduced to assess the performance of the *VAE-CNN* as an imputation approach. The validation procedure for these scenarios is similar, as described by the flowchart depicted in Fig. 6. This framework consists of three parts. First, the data preparation, which aims to provide an incomplete set ($X_c \in \mathbb{R}^p$) to the VAE-CNN model. Validation scenarios consist of intentionally removing data in intervals or point-to-point of a given sensor to provide an incomplete set. Then, the generated incomplete sets are normalized with the parameters obtained from the training set (i.e., μ and σ). Then, depending on the validation scenario, a naïve pre-imputation technique is conducted to fulfill the empty spots in the time-series. For instance, scenario 1 considers intervals of removed data; then, the pre-imputation is given through the introduction of random noise in (0,1] multiplied by the mean of the sensor in the training set. For scenario 2, which consists of a point-to-point removed dataset, the pre-imputation method consists of the mean substitution (MS) that replaces a missing value (x_i) with the average of the previous (x_{i-1}) and posterior value (x_{i+1}). The pre-imputed data for each scenario replaces the empty spaces, conforming to a pre-imputed vector ($\tilde{X}_c \in \mathbb{R}^p$). Second, the pre-imputed data is propagated through the VAE-CNN model to obtain a reliable approximation of the actual values (namely, imputed set $X_c^{'} \in \mathbb{R}^p$), replacing the pre-imputed values representing missing data. Finally, the proposed method is compared with other imputation methods such as MS, PCA, a standard AE, a standard AE using Conv-1D layers in its structure (AE-CNN), and a VAE with multi-layer perceptron (VAE-MLP) using several performance metrics.

Fig. 7 illustrates the validation scenarios for the imputation of the IAQ data. The first scenario consists of interval-based data removal that was conducted for several sensors in the TMS. In contrast, the second scenario consists of a point-to-point discard in one of the sensors. These scenarios are explained as follows:

- *Scenario 1:* The first validation scenario involves deleting two missing data intervals (i.e., interval A and B) for each time-series measurement of the NO_2, PM_{10}, and humidity sensors to achieve different missing data rates (20%, 50%, and 80%). As the test set consists of 286 samples, Table 3 describes the missing data intervals with their respective occurrence periods, indicating the necessary data points to be removed from the sensor measurements to reach the missing data rate. To fulfill the empty spaces in these intervals, a naïve pre-imputation method is conducted, consisting of replacing the missing value by random noise in (0,1] multiplied by the mean value of each sensor in the training set. Fig. 7(a–c) depicts the distributions of the selected IAQ sensors' actual and pre-imputed measurements for the given missing rates It should be noted that the mismatching between the distributions of pre-imputed and actual data increases as the missing data rate is increased.
- *Scenario 2:* The second scenario consists of the removal of point-to-point data for the PM_{10} sensor. A total of 30 points are randomly eliminated from the measurements (see Fig. 7(d)). For this scenario, the pre-imputation method is the MS technique by replacing a missing value xi, with the average of the previous (x_{i-1}) and posterior (x_{i+1}) data points. This pre-imputation technique is included in the performance evaluation since it is considered as a common method for data imputation [25].

Two metrics are used to assess the performance of the proposed method in contrast with the other approaches, including the root mean squared error (RMSE) (Eq. (6)) and mean absolute error (MAE) (Eq. (7)) [14].

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - x'_i)^2} \tag{6}$$

$$MAE = \frac{\sum_{i=1}^{N}|x_i - x'_i|}{N} \tag{7}$$

where x_i is the actual value of the IAQ measurements, x'_i represents the reconstructed/imputed value of the IAQ data and N is the number of samples.

In addition to the fitness of the imputation to the real IAQ values, the computational cost of the introduced models needs to be assessed to determine whether an increase in the complexity of these structures represents an improvement on the imputation task. The computational complexity (CC) is selected as the metric to assess the computational cost, defined in Eq. (8) [56].

$$\text{Computational complexity (CC)} = \frac{CT_M}{CT_N} \tag{8}$$

CC is defined as the mean computational time (milliseconds) required by the model to impute or estimate a value in a time-series (CT_M), divided by the corresponding time needed by a naïve method to conduct the same task (CT_N). In this regard, the PCA is considered as the naïve method, and the CC for each model results from averaging the computational time to impute the missing data in scenario 1. The estimation of the computational time was conducted with the following features: Intel ® Core™ I3-6100 CPU @ 3.70 GHz, 16.0 GB RAM, ×64 based processor.

4. Assessment of implications for the ventilation control of the D-subway station

The superior imputation approach from the previous analyses is then utilized when considering the implications of missing data in subway ventilation control. For this purpose, we conducted experiments on the ventilation system of the D-subway station. The control of the ventilation system is important in two main respects: 1) ventilation energy demand, and 2) public health risk [12].

In general, when the reliability of the sensors is compromised, the control system misinterpretation leads to a poor selection of the air delivery flow. For instance, faulty sensors may overestimate the air quality in the subway platform; then, the air delivery is set to its minimum capacity (20 Hz) since there is no need to provide the space with fresh outdoor air. In this case, when the actual pollutant concentration is higher than that recorded by the sensors, the air management may result in a deterioration in the subway air quality. In contrast, when the air quality is underestimated, the efforts of the control system focus on alleviating the IAQ by setting the fan speed at its maximum capacity (60 Hz). Following this kind of failure, the IAQ management may waste energy since the ventilation system is providing extra outdoor air when it is not needed [15,16,37]. However, when the sensors stop recording the measurements, there is no specific action that the ventilation system can perform. Therefore, in this study, two assumptions are made: 1) the ventilation control system sets the fan speed to its maximum (60 Hz) and 2) to its minimum capacity (20 Hz). The effects of these actions are evaluated and contrasted with the response of the ventilation system under the imputed IAQ data using the VAE-CNN method.

Loy-Benitez et al. (2018) [7] discussed a mathematical representation of the relationship between the IAQ and the ventilation fan speed (i.e., IAQ system). After the imputation of the PM_{10} missing data, the

Table 3
Missing data intervals of IAQ data for validation scenario 1.

Missing data rate (%)	Missing data interval	
	A	B
20%	20–35	60–75
50%	20–92	120–192
80%	15–130	171–286

reconstructed values from the proposed method go through a control loop deciding proper management from the ventilation system. The transfer function describing the IAQ system in the D-subway station is as follows:

$$G_p = \frac{-1.2127\exp(-0.063s)}{0.0168s+1} \tag{9}$$

On assessing the ventilation control system responses for all the cases, two crucial aspects are considered. First, given the occupancy and considerable commuting time in the subway stations, monitoring and control of public health receive significant attention [57]. The comprehensive indoor air quality index (CIAI) suggested by the Environmental Protection Agency of the USA (EPA) [12] has been frequently utilized to quantify the IAQ level, as explained in Eq. (10).

$$CIAI = \frac{I_{HI}-I_{LO}}{BP_{HI}-BP_{LO}}\left(C_p - BP_{LO}\right) + I_{LO} \tag{10}$$

Here, BP_{HI} and BP_{LO} represent the concentration breakpoints of each pollutant, I_{HI} and I_{LO} are the index breakpoints for each level of concern for human health, and C_p is the concentration of the indoor pollutant. The specifications of the CIAI can be found in Table A1.

On the other hand, evaluating the energy ventilation demand is critical since it occupies approximately 50% of the total power consumption of the building sector [7]. Exploiting the potential of energy-savings in buildings is a viable way of addressing climate change to accomplish the objectives established by the Intergovernmental Panel on Climate Change (IPCC) [58]. Liu et al. [59] proposed a third-order polynomial equation for quantifying the energy demand for the ventilation system as in Eq. (11).

$$E_{Vent}(\text{kWh}) = 0.0007 \times Hz^3 - 0.046 \times Hz^2 + 2.01 \times Hz + 8.8, \tag{11}$$

where Hz is a unit of ventilation FI.

Experiments at the D-subway station were conducted to demonstrate the benefits of the proposed imputation approach by comparing ventilation system performance with missing and imputed data given the above quantitative aspects. In this case, the PM_{10} sensor was considered, given that it is a general tracer of the IAQ in the D-subway station [7]. Some intervals of the PM_{10} sensor were removed to corrupt 50% of the set, as in scenario 1 in Section 3.2.2 (see Table 3). Within these corrupted intervals, three cases regarding the actions of the ventilation control system were selected: 1) the ventilation fan speed was set to 60 Hz or 2) to 20 Hz, and 3) the fan was set according to the imputed data from the proposed method. The experiments at the D-subway station are illustrated in Fig. 8.

5. Results and discussion

5.1. Imputation of the IAQ measurements in the D-subway station

As described in the previous sections, two scenarios were introduced for evaluating the performance of the proposed method against other approaches. As there is no fixed rule for the selection of the number of principal components (PCs), the setting for the PCA approach consisted of the retention of PCs able to explain 85% of the total data variance, resulting in a total of 4 PCs [60]., The structures are different for the neural techniques. For the standard AE, the structure consisted of three fully connected layers (i.e., input, hidden, and output layers), the hidden layer had 128 neurons. Additionally, for the AE-CNN the dense layer was replaced with a Conv-1D layer with 2 kernels and 64 filters. The activation function for these structures is the tanh, the optimizer is Adam, the loss function is the mean squared error (mse), and the number of epochs is 250. Additionally, the VAE-MLP approach follows the structure of the VAE-CNN described in Fig. 5 with the difference that the encoder consisted of dense layers instead of Conv-1D layers.

For scenario 1, two intervals (i.e., A and B) of three IAQ sensors were intentionally removed to meet the missing rate. As shown in Fig. 7, the distribution of the IAQ data was progressively corrupted as the missing data rate increased, and pre-imputation was utilized to fill these voids. The idea behind the imputation approach is to recover the missing data (from minimal to high corruption) accurately. Therefore, we expected that the corrupted distribution would be forced to follow the real distribution of each IAQ sensor after applying these techniques. Table 4 shows the performance metrics of the different imputation methods with their respective missing data rate and the interval for which it is imputed. Here, the most accurate imputation is distinguished in bold. It should be noted that the proposed method successfully imputed the majority of the given scenarios, most notably for those with long removed intervals (i.e., missing rate = 80%). These cases are the most critical since the estimation of missing values becomes erratic and intractable for other methods given the high corruption levels.

To further assess the effectiveness of the proposed method, the relationship between the missing data rate and the imputation performance for the removed data interval A and B were studied and are illustrated in Figs. 9 and 10, respectively. The curves represent the performance of the different methods. For the NO_2 sensor, the performance of the variational neural methods remains relatively similar despite the increasing missing data level for interval A in both metrics (Fig. 9(a)). Moreover, the AE-CNN shows similar performance with the VAE-CNN for the lowest corruption level, deteriorating the estimation as the missing data rate gets higher. The PCA overperformed AE, AE-CNN, and VAE-CNN when the corruption level was 50%. For the PM_{10} sensor (Fig. 9(b)), the performance of the proposed method outperformed the other methods for the missing data rates of 20% and 80%. However, for 50% of corruption, the VAE-MLP showed superiority over the proposed

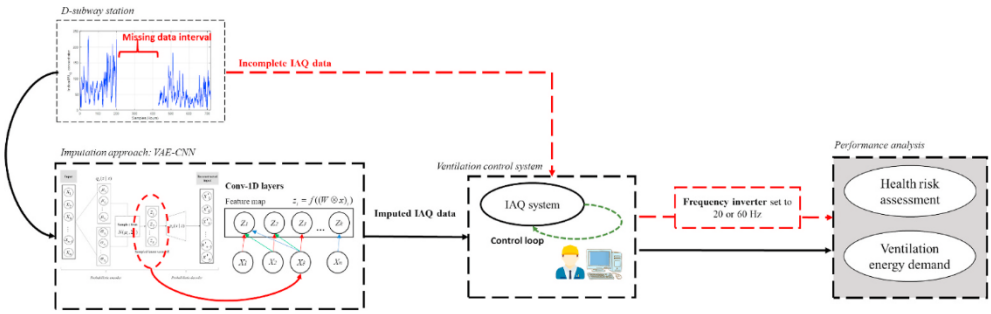


Fig. 8. Analysis of the D-subway station to compare the response of the ventilation system given incomplete and imputed IAQ data.

Table 4
Imputation performance and computational complexity for three IAQ sensors following the conditions of interval-based removal scenario.

Imputation approach	*Computational complexity (CC)*	*Sensor*	NO_2				PM_{10}				Humidity			
			A		B		A		B		A		B	
		Missing data rate	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
PCA	1	20%	0.021	0.019	0.0131	0.0124	40.8199	33.5639	**25.869**	**19.6788**	15.1485	12.505	18.0259	16.8568
		50%	**0.01218**	**0.01**	0.0188	0.0169	**37.8412**	30.0947	44.1337	36.8286	14.5807	11.9461	16.1877	12.7462
		80%	0.0173	0.015	0.0272	0.0223	49.5093	38.1036	84.88	72.5576	13.5019	10.8737	12.4856	10.1576
AE	6.02	20%	0.0243	0.019	0.0324	0.0268	32.4532	24.5292	47.505	40.3293	18.3638	15.1331	20.0529	18.5957
		50%	0.0231	0.019	0.025	0.0208	45.3525	34.866	36.3874	29.495	26.9781	24.1957	24.8804	22.7267
		80%	0.0254	0.021	0.0314	0.0265	40.6344	30.5243	39.2921	31.6236	24.1628	21.1797	17.5169	14.9181
AE-CNN	7.96	20%	0.02088	0.0125	0.024	0.012	24.2529	19.9672	62.8993	57.149	10.8587	8.2591	9.0142	7.8472
		50%	0.0143	0.0117	0.0108	0.016	43.9374	34.8443	34.904	27.614	14.0376	11.467	11.0376	9.4166
		80%	0.01535	0.012	**0.0194**	0.016	38.826	27.2799	29.6434	23.6005	12.2474	9.2698	8.2499	7.0077
VAE-MLP	9.86	20%	0.0183	0.014	**0.0057**	**0.005**	34.7521	25.9402	33.2657	28.8216	**9.5807**	8.2196	11.0529	10.4191
		50%	0.0154	0.013	0.0212	0.0191	34.9956	**28.7674**	81.2749	43.7554	14.735	12.1541	12.8376	10.9207
		80%	0.0176	0.014	0.0204	0.018	64.115	36.8325	88.3929	54.0443	12.7693	9.7491	**7.3523**	**6.3501**
VAE-CNN	16.92	20%	**0.0182**	**0.01**	0.0167	0.0127	**18.513**	**15.5867**	52.2057	47.2168	10.0733	**7.5121**	**7.924**	**5.8496**
		50%	0.0139	0.011	**0.0094**	**0.0122**	39.2097	29.8037	**27.9429**	**22.6681**	**13.4699**	**10.999**	**9.7183**	**8.2201**
		80%	**0.0146**	**0.011**	0.0202	**0.0159**	**35.3227**	**25.4804**	**27.4968**	**22.3332**	**11.8765**	**8.8903**	8.0166	6.8111

and additional methods. On the humidity sensor (Fig. 9 (c)), the VAE-CNN accounted for the highest performance over the other techniques, except for the 20% of data corruption, for which the VAE-MLP showed better performance, based on the RMSE, as compared to the other methods.

For interval B in the NO_2 sensor (Fig. 10(a)), the metrics indicate that the VAE-CNN method was more effective for the higher missing rates, followed by the AE-CNN that showed greater performance than the standard AE and PCA, while the VAE-MLP technique outperformed the other methods in the lowest missing rate. For the PM_{10} sensor (Fig. 10 (b)), the same performance was observed for the VAE-CNN method, showing the best performance on both metrics for the higher missing data rate. In this case, the PCA approach outperformed the other methods for the lowest missing data rate. Finally, for the humidity sensor (Fig. 10(c)), both variational and the AE-CNN methods showed similar performance trends for all the missing rates. However, the VAE-CNN accounted for the best imputation approach since it outperformed all methods for 20% and 50% missing data rates. Therefore, the proposed VAE-CNN showed superiority for the imputation of the majority removed intervals at different missing rates, giving a more accurate estimation of the actual measurements.

Fig. 11 illustrates the CC of each model versus their respective averaged imputation performance metric. The mean computational times for imputing the missing IAQ data in scenario 1 were 1.097, 6.648, 8.732, 10.816, and 18.556 ms for the PCA, AE, AE-CNN, VAE-MLP, and VAE-CNN, respectively. The CC results from dividing each method's computational time by the time that takes the naïve method (PCA) to achieve the same task. Then, the CC, as reported in Table 4, was 1, 6.06, 7.96, 9.86, and 16.92 for the PCA, AE, AE-CNN, VAE-MLP, and VAE-CNN, respectively. These values represent a relative metric indicating the additional, proportional time required for the missing data imputation task from more complex techniques.

Considering long-term goals for online monitoring in subway stations, the computational time is a critical aspect when massive data need to be processed. Compared to the naïve method, the neural methods exhibited greater CC, being the proposed model, the most computational demanding. For instance, the AE model showed poor performance for the NO_2 (Fig. 11(a)) and humidity (Fig. 11(c)) sensors. This issue may emerge given the training procedure of this model, consisting of copying the input to the output, this way the parameters are set based on this replication mechanism without any generation capability. Moreover, as the AE-CNN shows a higher CC, the performance of the imputation seems to improve given the capability of the convolutional layers to model the sequences in the IAQ time series for all cases, even exhibiting better imputation performance than the VAE-MLP. On the other hand, the VAE-MLP showed poor performance in the PM_{10} sensor (Fig. 11(b)), this may occur given the characteristics of the intrinsic architecture of the network since this model is built with fully connected dense layers, which have evidenced drawbacks when modeling and processing sequential data.

It should be noted that despite the highest CC was exhibited by the proposed model, the performance of the imputation is maintained high in comparison to the naïve and neural methods. In this case, the VAE-CNN model is capable of modeling sequential data considering non-linear dependencies, as well as introducing variational inference models as a framework to learn the distribution from the data and generate meaningful outputs. In conclusion, the PCA is computationally cheap, showing poor imputation performance, while the proposed model exhibits the most accurate imputation with its complex structure. The trade-off between CC and imputation performance plays an important factor in decision-making for the implementation of novel systems. Then, as the IAQ management in the D-subway station depends on the sensors' reliability, having a robust system that accurately imputes missing data implies a safer environment for commuters and avoids energy-waste. Therefore, the accuracy of the imputation method should be first considered over the CC; however, high CC may be tackled

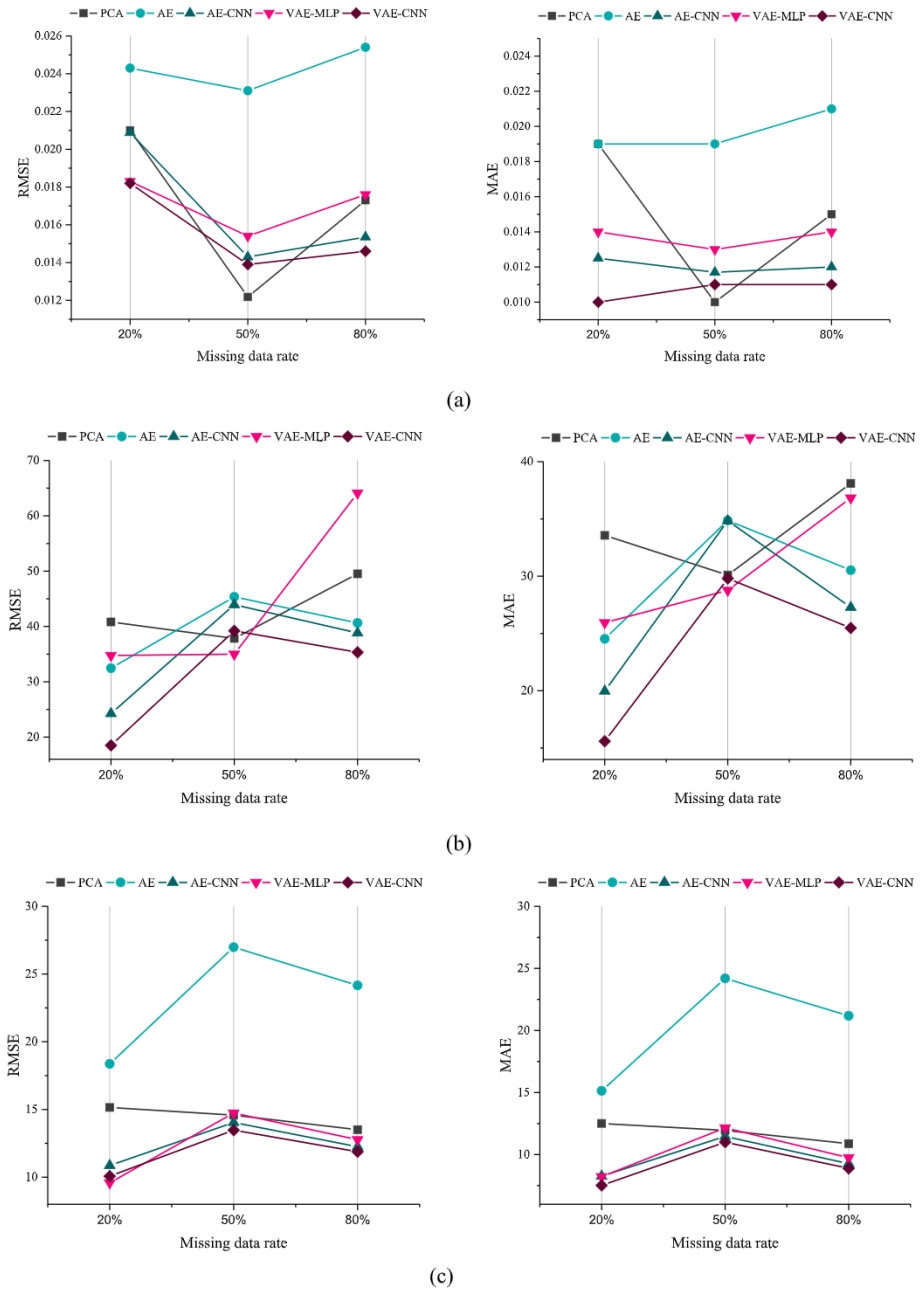


Fig. 9. Imputation performance from four imputation approaches given different missing rates on interval A for the (a) NO_2, (b) PM_{10}, and (c) humidity sensors.

by utilizing more modern systems than the used in this study.

A point-to-point removal was conducted for scenario 2 by randomly eliminating a total of 30 points in the PM_{10} sensor. In contrast to missing intervals, the MS method can be conducted when the missing information does not represent a large number of losses. Contrary to scenario 1, applying a fixed amount during the entire intervals represents a complete failure of the sensor; therefore, this method is only used for this scenario. The performance of imputation methods is depicted in Fig. 12. The MS method showed superior performance compared with the PCA, AE, and AE-CNN by 57%, 47%, and 6% for the RMSE, respectively. Additionally, the MS method for the MAE was 54%, 33%, and 9% better than the PCA, AE, and AE-CNN, respectively. When considering the variational techniques (i.e., VAE-MLP and VAE-CNN), the estimations are more accurate than the PCA, AE, and AE-CNN.

Nevertheless, the metrics of the neural variational approaches are closer to the MS method. The VAE-MLP technique showed similar performance to the MS method based on both parameters, with a minor difference of 9.3% and 3.3% for the RMSE and the MAE, respectively. On the other hand, the VAE-CNN outperformed both approaches for the computed values, giving an accurate estimation of the missing points for the PM_{10} sensor.

These results show that the proposed method surpassed the performance of the typical PCA and the neural approaches for both introduced scenarios. Therefore, the next experiment at the D-subway station was conducted based on the imputation given by the VAE-CNN method.

5.2. *Assessing the influence of the proposed imputation method on the ventilation control system*

As explained in previous sections, the IAQ management may be affected by the missing data points from the TMS. In this section, three cases of IAQ data conditions were evaluated to determine the response of the ventilation control system and its influence on energy consumption and public health aspects. The missing data consist of two intervals

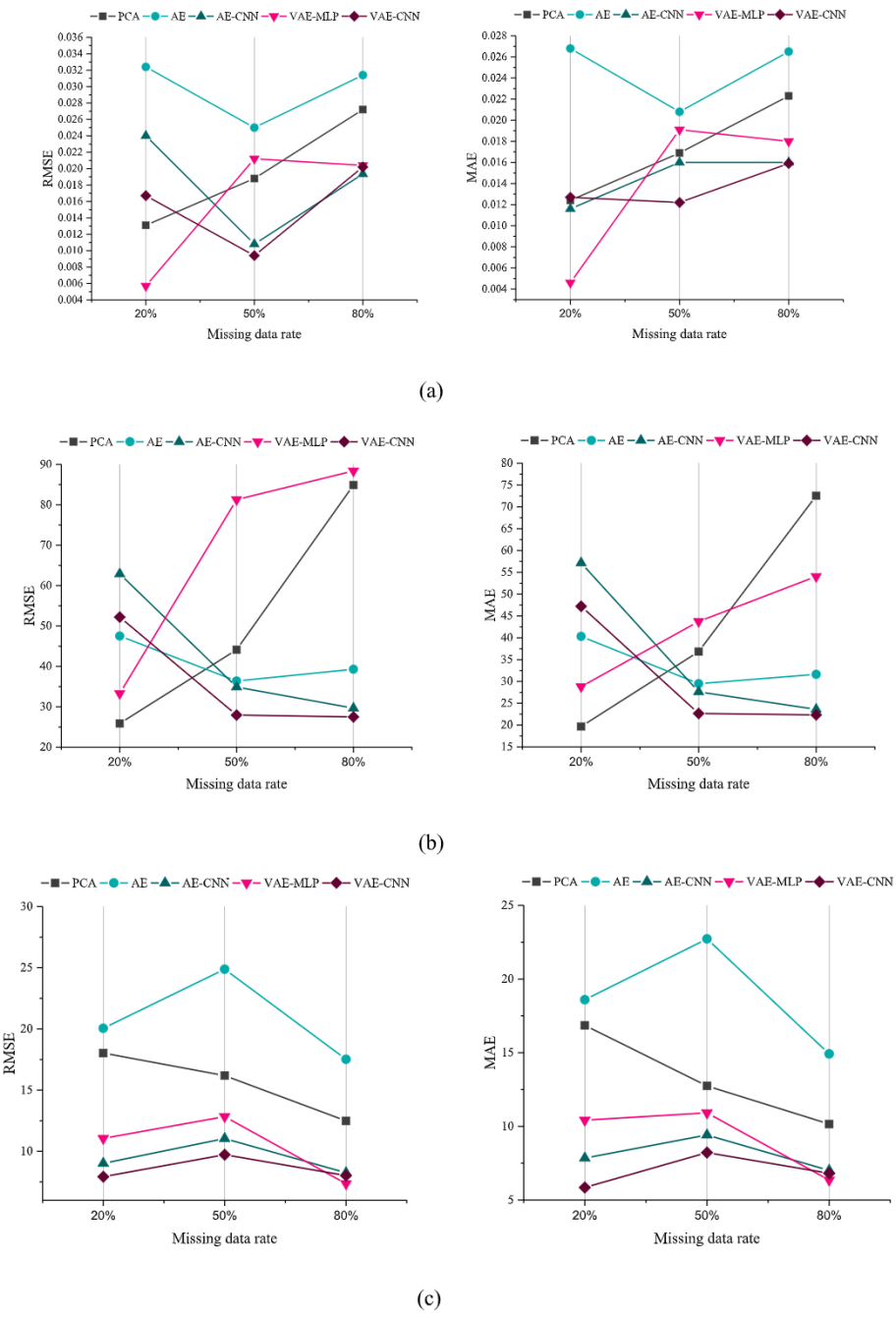
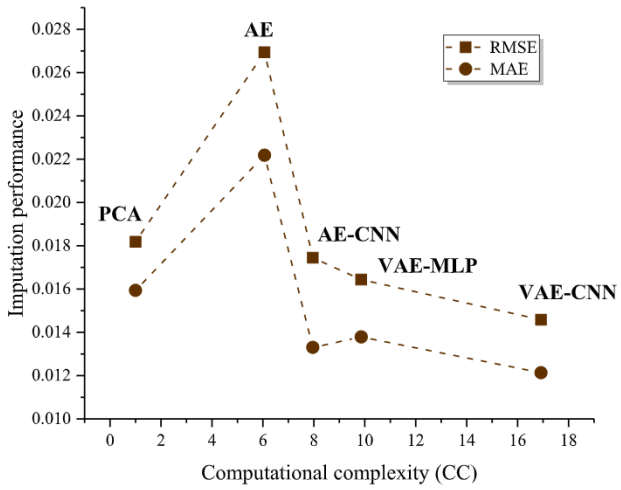


Fig. 10. Imputation performance from four imputation approaches given different missing rates on interval B for the (a) NO_2, (b) PM_{10}, and (c) humidity sensors.

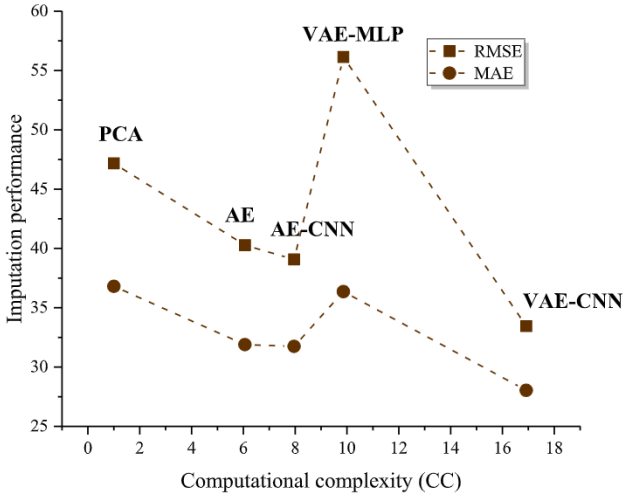
accounting for 50% of the total in the PM_{10} sensor, as in Table 3. For the first and second cases, the ventilation control system assumed a high and low concentration of PM_{10}, respectively, when a void in the PM_{10} measurements appeared. Therefore, for the first case, the efforts of the IAQ management focused on the rapid improvement of the subway platform air by setting the fan speed at its maximum capacity (i.e., 60 Hz). On the other hand, the second case provides low concentrations of PM_{10}. Consequently, the indoor ambient is assumed to be in the healthy range of conditions, and the fan speed is set to its minimum capacity (i.e., 20 Hz). The third case consists of the imputed PM_{10} data via the VAE-CNN method to obtain a reliable approximation of the real IAQ; therefore, proper IAQ management is expected to occur by the manipulation of the fan speed.

The control system includes the relationship between the indoor PM_{10} concentration and the frequency inverter as explained in Eq. (8). Fig. 12 illustrates the influence of the PM_{10} data conditions on the ventilation control system response for a total of 12 days (including the fan speed in terms of the IF and health risk assessment via the CIAI), while Table 5 shows quantitative values regarding the ventilation energy demand and health risk conditions.

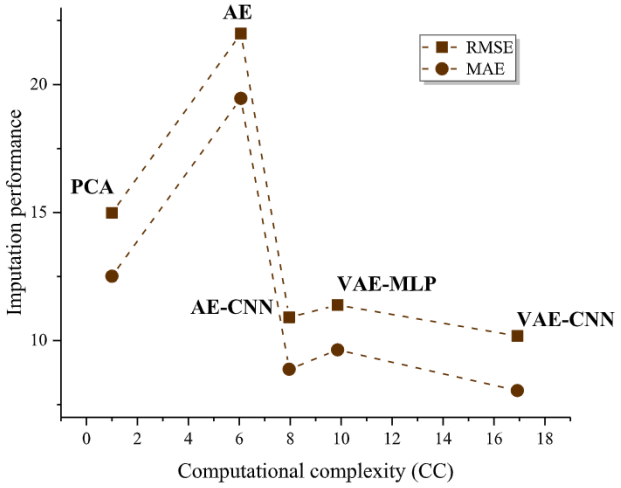
Fig. 13(a) shows the response of the ventilation system when assuming an overestimation of the IAQ deterioration. For these intervals, the FI is set to its maximum capacity of 60 Hz. As was expected for this case, the ventilation energy demand may result in wasted energy, representing 1246.8 kWh/day. In the case of the health risk assessment, the red dashed line is the breakpoint limit marking good and moderate indoor air conditions (CIAI = 50). It should be noted that the CIAI is generally maintained below the breakpoint limit for the given intervals. The breakpoint limit is surpassed for approximately 54% of the analyzed range (155 h), ensuring that, for a significant part of the



(a)



(b)



(c)

Fig. 11. Imputation performance versus computational complexity for (a) NO_2, (b) PM_{10}, and (c) humidity sensors.

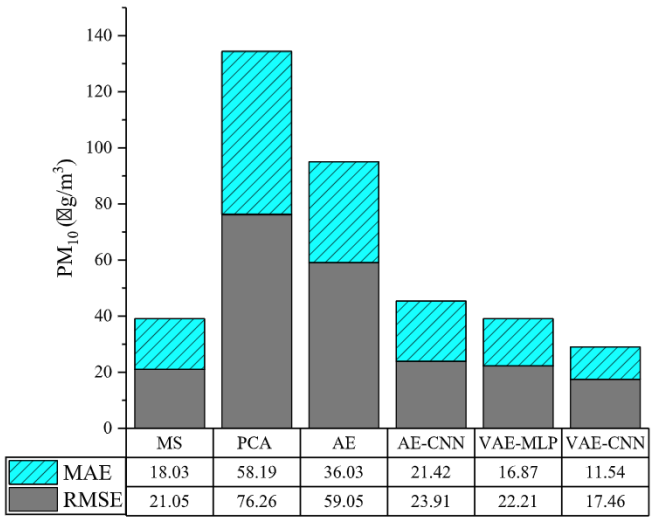


Fig. 12. Performance of the point-to-point imputation using six different techniques.

Table 5
Evaluation of the ventilation system response under the assumed cases for missing data and the VAE-CNN based imputed data.

	Ventilation energy demand (kWh/day)	Average PM_{10} concentration ($\mu g/m^3$)[a]	Points surpassing CIAI moderate level breakpoint
Case 1: Overestimation of IAQ deterioration	1246.8	50.31	155
Case 2: Underestimation of IAQ deterioration	650.6	72.62	287
Case 3: VAE-CNN based imputed IAQ data	824.1	68.31	258

[a] The results represent the average of 12 days.

time, the indoor air is in good condition for human health with an average value of 50.31 μg/m^3.

Fig. 13(b) depicts the response for the assumption in which the air delivery is set to its minimum capacity of 20 Hz, with an energy demand for 650.69 kWh/day. Contrary to the first case, the control system interprets this case as having good indoor air conditions; therefore, an energy-saving mechanism is activated by delivering a low air rate. This action deteriorates the IAQ level as it can be seen that a significant portion, approximately 99%, of the analyzed interval (287 h), falls over the breakpoint limit, representing an average of 72.62 μg/m^3.

Finally, Fig. 13 (c) shows the response of the ventilation system with the imputed IAQ data for the given intervals. It can be noted that the ventilation system takes some action to control the air delivery rate instead of using a fixed value as in the previous cases. The ventilation energy demand is reported as 824.16 kWh/day. On the other hand, when assessing the CIAI variation for these intervals, ventilation allows the indoor air to be in good condition for approximately 10% of the total measurements since the level is maintained over the moderate breakpoint for 258 h, showing an average of 68.31 μg/m^3.

The discussion of this study is based on the conflict between the IAQ level and the ventilation energy demand. Accordingly, the IAQ level deteriorates when the fan speed is low and becomes excellent when the fan speed is high. Therefore, sustainable balance is desired for these systems. Considering case 1, the IAQ was maintained in a good region with an expensive ventilation air rate. These expenses affect the system from an environmental point of view. For instance, the rate of CO_2

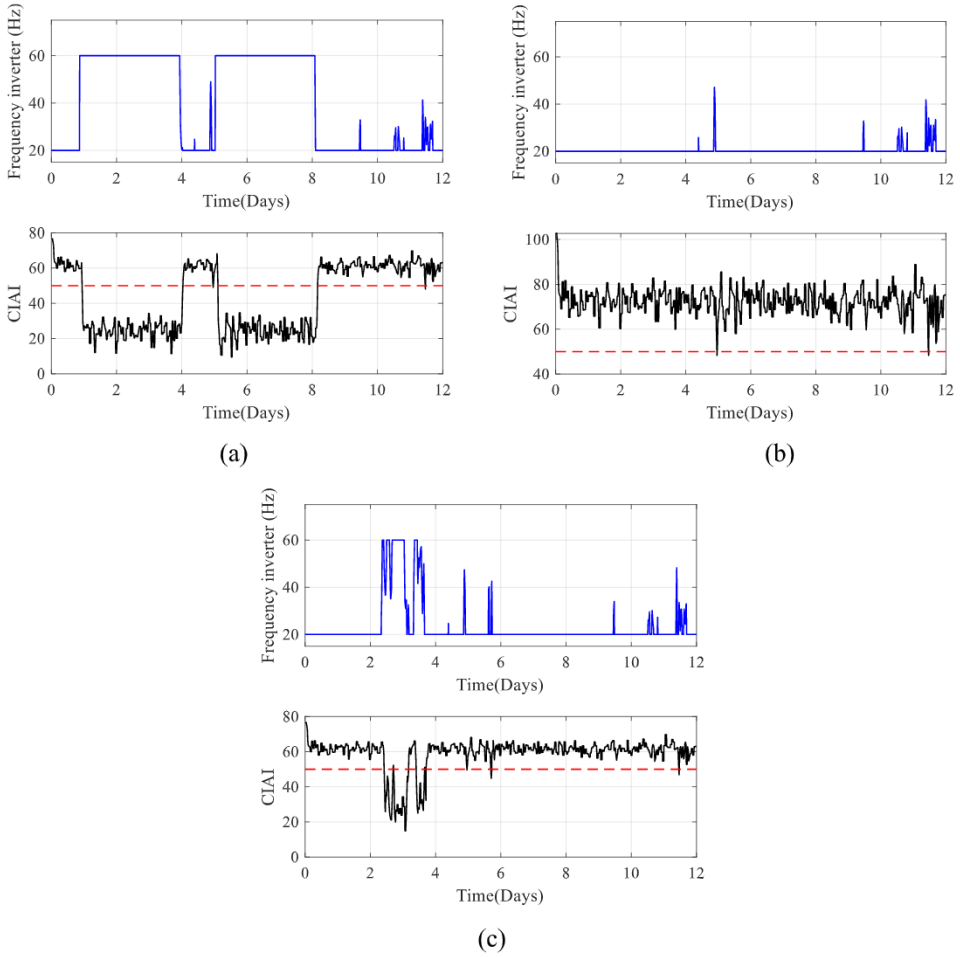


Fig. 13. Frequency inverter and CIAI variations in the subway platform for (a) case 1: assuming high PM_{10} concentrations, (b) case 2: assuming low PM_{10} concentration, and (c) case 3: response with imputed PM_{10} data using the proposed method.

emitted per unit of energy is 1.7 kg CO_2/kWh [7]. Therefore, for case 1, the emissions of CO_2 given the ventilation demand are 2119.6 kg CO_2/day, while the ventilation system in case 3 with the imputed data lowers the energy demand, thus decreasing the emission of CO_2 by approximately 20% to 1401.07 kg CO_2/day. A counter scenario can be noted for case 2, in which the IAQ is exhausted by the scarce air delivered. Based on energy demand, this system required the least energy among the presented cases. However, it is notable that for the analyzed interval, most of the CIAI is above the control limit, representing a decrease in the IAQ conditions. In comparison to the ventilation system with the imputed data, the IAQ level was improved by approximately 3%, while the ventilation energy demand was increased by 12%. Therefore, the proposed method ensures reliable imputation of the IAQ measurements, allowing proper management of the subway ventilation system to achieve a balance between the energy demand and the IAQ level.

6. Conclusions

An imputation approach for estimating missing data in multivariate IAQ data was developed by utilizing VAE-CNN models. Given that the IAQ data are *non-Gaussian* and show dynamism, static methods like PCA showed poor performance, while methods as AE are limited when acting as generators. The training procedure consisted of propagating a complete IAQ set through the probabilistic encoder and decoder. Then, the necessary parameters were learned by backpropagation via ELBO maximization to generate a reliable estimation from the data distribution. The validation was conducted by corrupting an IAQ dataset, eliminating intervals of different sensors, and eliminating data points on the PM_{10} sensor. Then, the imputation approach was compared with other methods, demonstrating the superiority of the VAE-CNN.

Nonetheless, lacking open source data of the subway TMS resulted in a limitation for the development of this study, considering that the proposed neural method is ideally selected for the processing of massive data. However, this study aimed to demonstrate the proposed method's feasibility of imputing missing data in the IAQ subway environment to enhance the sensors' reliability along with its implications on energy and public health management. Furthermore, for real-world implementation of the presented imputation technique in the subway environment, complete access to continuous IAQ information at higher time resolution must be granted.

Additionally, an experiment on the IAQ system in the D-subway station was conducted, in which the ventilation system response with missing and imputed data were assessed. The IAQ management with the imputed dataset verified the importance of having accurate estimations of the missing data since it affects the entire system in terms of CO_2 emissions, IAQ levels, and ventilation energy demand. Concluding that the proposed VAE–CNN–based imputation approach provided an accurate estimation of data voids for ensuring sustainable management of the IAQ in subway stations for the development of smart environmental systems within the building sector.

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.

Acknowledgments

This research was supported by a grant from the National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (No. 2017R1E1A1A03070713), and from the Subway Fine Dust Reduction Technology Development Project of the Ministry of Land Infrastructure and Transport from the Republic of Korea (20QPPW-B152306-02).

Appendix A. The comprehensive indoor air quality index

Table A1
Specifications of the CIAI suggested by the U.S. Environmental Protection Agency [61].

Index level	Good		Moderate		Unhealthy for sensitive groups		Unhealthy		Very unhealthy		Hazardous	
I_{LO}	0		51		101		151		251		351	
I_{HI}	50		100		150		250		350		500	
Conc. Level	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}	BP_{LO}	BP_{HI}
NO_2 (ppm)	0	0.03	0.031	0.05	0.051	0.15	0.151	0.25	0.251	0.5	0.501	2
CO (ppm)	0	5	5.01	10	10.01	20	20.01	30	30.01	40	40.01	50
CO_2 (ppm)	0	500	501	1000	1001	1500	1501	2000	2001	3000	3001	5000
PM_{10} (μg/㎥)	0	50	51	150	151	250	251	350	351	450	451	600
$PM_{2.5}$ (μg/㎥)	0	15	16	40	41	140	141	250	251	350	351	500

References

[1] S. Metia, Q.P. Ha, H.N. Duc, Y. Scorgie, Urban air pollution estimation using unscented Kalman filtered inverse modeling with scaled monitoring data, Sustain. Cities Soc. 54 (2020) 101970, https://doi.org/10.1016/J.SCS.2019.101970.
[2] S.E. Bibri, The IoT for smart sustainable cities of the future: an analytical framework for sensor-based big data applications for environmental sustainability, Sustain. Cities Soc. 38 (2018) 230–253, https://doi.org/10.1016/J.SCS.2017.12.034.
[3] Y. Shi, X. Xie, J.C.-H. Fung, E. Ng, Identifying critical building morphological design factors of street-level air pollution dispersion in high-density built environment using mobile monitoring, Build. Environ. 128 (2018) 248–259, https://doi.org/10.1016/J.BUILDENV.2017.11.043.
[4] S. Heo, K. Nam, J. Loy-Benitez, Q. Li, S. Lee, C. Yoo, A deep reinforcement learning-based autonomous ventilation control system for smart indoor air quality management in a subway station, Energy Build. 202 (2019), 109440, https://doi.org/10.1016/J.ENBUILD.2019.109440.
[5] K. Huang, J. Song, G. Feng, Q. Chang, B. Jiang, J. Wang, W. Sun, H. Li, J. Wang, X. Fang, Indoor air quality analysis of residential buildings in northeast China based on field measurements and longtime monitoring, Build. Environ. 144 (2018) 171–183, https://doi.org/10.1016/J.BUILDENV.2018.08.022.
[6] B. Xu, J. Hao, Air quality inside subway metro indoor environment worldwide: a review, Environ. Int. 107 (2017) 33–46, https://doi.org/10.1016/j.envint.2017.06.016.
[7] J. Loy-Benitez, Q. Li, P. Ifaei, K. Nam, S. Heo, C. Yoo, A dynamic gain-scheduled ventilation control system for a subway station based on outdoor air quality conditions, Build. Environ. 144 (2018) 159–170, https://doi.org/10.1016/J.BUILDENV.2018.08.016.
[8] M. Junaid, J.H. Syed, N.A. Abbasi, M.Z. Hashmi, R.N. Malik, D.-S. Pei, Status of indoor air pollution (IAP) through particulate matter (PM) emissions and associated health concerns in South Asia, Chemosphere 191 (2018) 651–663, https://doi.org/10.1016/J.CHEMOSPHERE.2017.10.097.
[9] J.J. Figueroa-Lara, J.M. Murcia-González, R. García-Martínez, M. Romero-Romo, M. Torres Rodríguez, V. Mugica-Álvarez, Effect of platform subway depth on the presence of Airborne PM2.5, metals, and toxic organic species, J. Hazard Mater. 377 (2019) 427–436, https://doi.org/10.1016/J.JHAZMAT.2019.05.091.
[10] J. Loy-Benitez, P. Vilela, Q. Li, C. Yoo, Sequential prediction of quantitative health risk assessment for the fine particulate matter in an underground facility using deep recurrent neural networks, Ecotoxicol. Environ. Saf. 169 (2019) 316–324, https://doi.org/10.1016/J.ECOENV.2018.11.024.
[11] Y.-S. Son, J.-H. Jeong, H.J. Lee, J.-C. Kim, A novel control system for nitrogen dioxide removal and energy saving from an underground subway stations, J. Clean. Prod. 133 (2016) 212–219, https://doi.org/10.1016/J.JCLEPRO.2016.05.116.
[12] S. Lee, S. Hwangbo, J.T. Kim, C.K. Yoo, Gain scheduling based ventilation control with varying periodic indoor air quality (IAQ) dynamics for healthy IAQ and energy savings, Energy Build. 153 (2017) 275–286, https://doi.org/10.1016/j.enbuild.2017.08.021.
[13] Q. Li, J. Loy-Benitez, S. Heo, S. Lee, H. Liu, C. Yoo, Flexible real-time ventilation design in a subway station accommodating the various outdoor PM 10 air quality from climate change variation, Build. Environ. 153 (2019) 77–90, https://doi.org/10.1016/j.buildenv.2019.02.029.
[14] H. Liu, C. Yang, M. Huang, D. Wang, C. Yoo, Modeling of subway indoor air quality using Gaussian process regression, J. Hazard Mater. 359 (2018) 266–273, https://doi.org/10.1016/J.JHAZMAT.2018.07.034.
[15] M. Kim, H. Liu, J.T. Kim, C. Yoo, Evaluation of passenger health risk assessment of sustainable indoor air quality monitoring in metro systems based on a non-Gaussian dynamic sensor validation method, J. Hazard Mater. 278 (2014) 124–133, https://doi.org/10.1016/J.JHAZMAT.2014.05.098.
[16] J. Loy-Benitez, Q. Li, K. Nam, C. Yoo, Sustainable subway indoor air quality monitoring and fault-tolerant ventilation control using a sparse autoencoder-driven sensor self-validation, Sustain. Cities Soc. 52 (2020) 101847, https://doi.org/10.1016/J.SCS.2019.101847.
[17] X. Lai, X. Wu, L. Zhang, W. Lu, C. Zhong, Imputations of missing values using a tracking-removed autoencoder trained with incomplete data, Neurocomputing 366 (2019) 54–65, https://doi.org/10.1016/J.NEUCOM.2019.07.066.
[18] J.T. McCoy, S. Kroon, L. Auret, Variational autoencoders for missing data imputation with application to a simulated milling circuit, IFAC-PapersOnLine. 51 (2018) 141–146, https://doi.org/10.1016/J.IFACOL.2018.09.406.
[19] H. Kang, The prevention and handling of the missing data, Korean J. Anesthesiol. 64 (2013) 402–406, https://doi.org/10.4097/kjae.2013.64.5.402.
[20] S. Sedghi, A. Sadeghian, B. Huang, Mixture semisupervised probabilistic principal component regression model with missing inputs, Comput. Chem. Eng. 103 (2017) 176–187, https://doi.org/10.1016/J.COMPCHEMENG.2017.03.015.
[21] O. Kalaycioglu, A. Copas, M. King, R.Z. Omar, A comparison of multiple-imputation methods for handling missing data in repeated measurements observational studies, J. R. Stat. Soc. Ser. A Stat. Soc. 179 (2016) 683–706, https://doi.org/10.1111/rssa.12140.
[22] J.M. Jerez, I. Molina, P.J. García-Laencina, E. Alba, N. Ribelles, M. Martín, L. Franco, Missing data imputation using statistical and machine learning methods in a real breast cancer problem, Artif. Intell. Med. 50 (2010) 105–115, https://doi.org/10.1016/j.artmed.2010.05.002.
[23] M.G. Rahman, M.Z. Islam, FIMUS: a framework for imputing missing values using co-appearance, correlation and similarity analysis, Knowl. Base Syst. 56 (2014) 311–327, https://doi.org/10.1016/J.KNOSYS.2013.12.005.
[24] N. Abiri, B. Linse, P. Edén, M. Ohlsson, Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems, Neurocomputing 365 (2019) 137–146, https://doi.org/10.1016/J.NEUCOM.2019.07.065.
[25] I.A. Gheyas, L.S. Smith, A neural network-based framework for the reconstruction of incomplete data sets, Neurocomputing 73 (2010) 3039–3065, https://doi.org/10.1016/J.NEUCOM.2010.06.021.
[26] S.J. Choudhury, N.R. Pal, Imputation of missing data with neural networks for classification, Knowl. Base Syst. 182 (2019), 104838, https://doi.org/10.1016/J.KNOSYS.2019.07.009.
[27] Y. Duan, Y. Lv, Y.-L. Liu, F.-Y. Wang, An efficient realization of deep learning for traffic data imputation, Transport. Res. C Emerg. Technol. 72 (2016) 168–181, https://doi.org/10.1016/J.TRC.2016.09.015.
[28] F.M. Bianchi, L. Livi, K.Ø. Mikalsen, M. Kampffmeyer, R. Jenssen, Learning representations of multivariate time series with missing data, Pattern Recogn. 96 (2019), 106973, https://doi.org/10.1016/J.PATCOG.2019.106973.
[29] S. Lee, M. Kwak, K.-L. Tsui, S.B. Kim, Process monitoring using variational autoencoder for high-dimensional nonlinear processes, Eng. Appl. Artif. Intell. 83 (2019) 13–27, https://doi.org/10.1016/J.ENGAPPAI.2019.04.013.
[30] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd Int. Conf. Learn. Represent. ICLR 2014 - Conf. Track Proc, 2014, pp. 1–14.

[31] Z. Zhang, T. Jiang, C. Zhan, Y. Yang, Gaussian feature learning based on variational autoencoder for improving nonlinear process monitoring, J. Process Contr. 75 (2019) 136–155, https://doi.org/10.1016/J.JPROCONT.2019.01.008.
[32] L. Gondara, K. Wang, MIDA: multiple imputation using denoising autoencoders, in: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 10939 LNAI, 2018, pp. 260–272, https://doi.org/10.1007/978-3-319-93040-4_21.
[33] H. Liu, J. Zhou, Y. Xu, Y. Zheng, X. Peng, W. Jiang, Unsupervised fault diagnosis of rolling bearings using a deep neural network based on generative adversarial networks, Neurocomputing (2018), https://doi.org/10.1016/J.NEUCOM.2018.07.034.
[34] W. Yan, P. Guo, L. gong, Z. Li, Nonlinear and robust statistical process monitoring based on variant autoencoders, Chemometr. Intell. Lab. Syst. 158 (2016) 31–40, https://doi.org/10.1016/J.CHEMOLAB.2016.08.007.
[35] C. Lu, Z.-Y. Wang, W.-L. Qin, J. Ma, Fault diagnosis of rotary machinery components using a stacked denoising autoencoder-based health state identification, Signal Process. 130 (2017) 377–388, https://doi.org/10.1016/J.SIGPRO.2016.07.028.
[36] H. Zhao, S. Sun, B. Jin, Sequential fault diagnosis based on LSTM neural network, IEEE Access 6 (2018) 12929–12939, https://doi.org/10.1109/ACCESS.2018.2794765.
[37] J. Loy-Benitez, S. Heo, C. Yoo, Control Engineering Practice Soft sensor validation for monitoring and resilient control of sequential subway indoor air quality through memory-gated recurrent neural networks-based autoencoders, Contr. Eng. Pract. 97 (2020), 104330, https://doi.org/10.1016/j.conengprac.2020.104330.
[38] Q. Li, J. Loy-Benitez, K. Nam, S. Hwangbo, J. Rashidi, C. Yoo, Sustainable and reliable design of reverse osmosis desalination with hybrid renewable energy systems through supply chain forecasting using recurrent neural networks, Energy 178 (2019) 277–292, https://doi.org/10.1016/J.ENERGY.2019.04.114.
[39] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, T. Chi, Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation, Environ. Pollut. (2017), https://doi.org/10.1016/j.envpol.2017.08.114.
[40] X. Qing, Y. Niu, Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM, Energy 148 (2018) 461–468, https://doi.org/10.1016/j.energy.2018.01.177.
[41] Y. Su, C.-C.J. Kuo, On extended long short-term memory and dependent bidirectional recurrent neural network, Neurocomputing 356 (2019) 151–161, https://doi.org/10.1016/J.NEUCOM.2019.04.044.
[42] N. Srivastava, E. Mansimov, R. Salakhutdinov, Unsupervised learning of video representations using LSTMs, in: 32nd Int. Conf. Mach. Learn. ICML 2015, vol. 1, 2015, pp. 843–852.
[43] D. Processes, An Intelligent Fault Diagnosis Method Using GRU Neural Network towards Sequential Data in, 2019, https://doi.org/10.3390/pr7030152.
[44] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: 52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf, vol. 1, 2014, pp. 655–665, https://doi.org/10.3115/v1/p14-1062.
[45] S. Bai, J.Z. Kolter, V. Koltun, An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, 2018. http://arxiv.org/abs/1803.01271.
[46] S. Lee, M. Kwak, K.-L. Tsui, S.B. Kim, Process monitoring using variational autoencoder for high-dimensional nonlinear processes, Eng. Appl. Artif. Intell. 83 (2019) 13–27, https://doi.org/10.1016/J.ENGAPPAI.2019.04.013.
[47] F. Guo, R. Xie, B. Huang, A deep learning just-in-time modeling approach for soft sensor based on variational autoencoder, Chemometr. Intell. Lab. Syst. 197 (2020), 103922, https://doi.org/10.1016/J.CHEMOLAB.2019.103922.
[48] Z. Chen, Handwritten digits recognition, in: Proc. 2009 Int. Conf. Image Process. Comput. Vision, Pattern Recognition, IPCV 2009, vol. 2, 2009, pp. 690–694, https://doi.org/10.31142/ijtsrd8384.
[49] M. Kim, R.D. Braatz, J.T. Kim, C. Yoo, Indoor air quality control for improving passenger health in subway platforms using an outdoor air quality dependent ventilation system, Build. Environ. 92 (2015) 407–417, https://doi.org/10.1016/j.buildenv.2015.05.010.
[50] H. Liu, M. Kim, O. Kang, B. Sankararao, J. Kim, J.C. Kim, C.K. Yoo, Sensor validation for monitoring indoor air quality in a subway station, Indoor Built Environ. 21 (2012) 205–211, https://doi.org/10.1177/1420326X11419342.
[51] M. Kim, H. Liu, J.T. Kim, C. Yoo, Sensor fault identification and reconstruction of indoor air quality (IAQ) data using a multivariate non-Gaussian model in underground building space, Energy Build. 66 (2013) 384–394, https://doi.org/10.1016/j.enbuild.2013.07.002.
[52] H. Lotfalinezhad, A. Maleki, TTA, a new approach to estimate Hurst exponent with less estimation error and computational time, Phys. A Stat. Mech. Its Appl. (2020) 124093, https://doi.org/10.1016/J.PHYSA.2019.124093.
[53] G. Afendras, M. Markatou, Optimality of Training/Test Size and Resampling Effectiveness of Cross-Validation Estimators of the Generalization Error, vol. 1, 2015. http://arxiv.org/abs/1511.02980.
[54] S. Hwangbo, K. Nam, S. Heo, C. Yoo, Hydrogen-based self-sustaining integrated renewable electricity network (HySIREN) using a supply-demand forecasting model and deep-learning algorithms, Energy Convers. Manag. 185 (2019) 353–367, https://doi.org/10.1016/J.ENCONMAN.2019.02.017.
[55] H. Liu, C. Yang, M. Huang, D. Wang, C.K. Yoo, Modeling of subway indoor air quality using Gaussian process regression, J. Hazard Mater. 359 (2018) 266–273, https://doi.org/10.1016/j.jhazmat.2018.07.034.
[56] S. Makridakis, E. Spiliotis, V. Assimakopoulos, Statistical and Machine Learning forecasting methods: concerns and ways forward, PloS One 13 (2018) 1–26, https://doi.org/10.1371/journal.pone.0194889.
[57] T. Moreno, N. Pérez, C. Reche, V. Martins, E. de Miguel, M. Capdevila, S. Centelles, M.C. Minguillón, F. Amato, A. Alastuey, X. Querol, W. Gibbons, Subway platform air quality.Assessing the influences of tunnel ventilation, train piston effect and station design, Atmos. Environ. (2014), https://doi.org/10.1016/j.atmosenv.2014.04.043.
[58] M.J. Kim, R.D. Braatz, J.T. Kim, C.K. Yoo, Economical control of indoor air quality in underground metro station using an iterative dynamic programming-based ventilation system, Indoor Built Environ. 25 (2016) 949–961, https://doi.org/10.1177/1420326X15591640.
[59] H. Liu, S. Lee, M. Kim, H. Shi, J.T. Kim, K.L. Wasewar, C. Yoo, Multi-objective optimization of indoor air quality control and energy consumption minimization in a subway ventilation system, Energy Build. 66 (2013) 553–561, https://doi.org/10.1016/j.enbuild.2013.07.066.
[60] P. Ifaei, A. Karbassi, S. Lee, C. Yoo, A renewable energies-assisted sustainable development plan for Iran using techno-econo-socio-environmental multivariate analysis and big data, Energy Convers. Manag. 153 (2017) 257–277, https://doi.org/10.1016/j.enconman.2017.10.014.
[61] S. Tariq, J. Loy-Benitez, K. Nam, S. Heo, C. Yoo, Energy-efficient time-delay compensated ventilation control system for sustainable subway air quality management under various outdoor conditions, Build. Environ. 174 (2020), 106775, https://doi.org/10.1016/j.buildenv.2020.106775.