sensors

MDPI

Article

A Novel LSTM for Multivariate Time Series with Massive Missingness

Nazanin Fouladgar [1,*] and Kary Främling [1,2]

[1] Department of Computing Science, Umeå University, 901 87 Umeå, Sweden; kary.framling@umu.se
[2] School of Science and Technology, Aalto University, P.O. Box 15500, 00076 Aalto, Finland
* Correspondence: nazanin@cs.umu.se

Received: 15 March 2020; Accepted: 12 May 2020; Published: 16 May 2020

Abstract: Multivariate time series with missing data is ubiquitous when the streaming data is collected by sensors or any other recording instruments. For instance, the outdoor sensors gathering different meteorological variables may encounter low material sensitivity to specific situations, leading to incomplete information gathering. This is problematic in time series prediction with massive missingness and different missing rate of variables. Contribution addressing this problem on the regression task of meteorological datasets by employing Long Short-Term Memory (LSTM), capable of controlling the information flow with its memory unit, is still missing. In this paper, we propose a novel model called forward and backward variable-sensitive LSTM (FBVS-LSTM) consisting of two decay mechanisms and some informative data. The model inputs are mainly the missing indicator, time intervals of missingness in both *forward* and *backward* direction and *missing rate* of each variable. We employ this information to address the so-called missing not at random (MNAR) mechanism. Separately learning the features of each parameter, the model becomes adapted to deal with massive missingness. We conduct our experiment on three real-world datasets for the air pollution forecasting. The results demonstrate that our model performed well along with other LSTM-derivation models in terms of prediction accuracy.

Keywords: multivariate time series; regression; massive missingness; LSTM

1. Introduction

A great number of time series problems are stuck in Multivariate Time Series (MTS) on which multiple variables follow the interdependency between/within variables. Predicting future values of these variables by modeling previous observed sequences of values has been investigated widely among researchers for decades to make more accurate decisions [1–5]. For instance, meteorological data collected with different sensors is among multivariate time series problems involving different variables that change over time and accordingly predict future situations for target purposes. However, some difficulties in collecting such data range from faulty sensors to costly efforts of establishing them. This gives rise to the problem of missing data [6,7].

Multivariate time series with missing data is a challenge in different tasks, specifically in prediction. Since missingness causes bias in results, modeling any approach requires investigation of different types of missing data [8]. In general, one could categorize these data into three classes: *missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR). In the first class, the missing data lies independent of both the observed and unobserved variables, while in the second class, missingness falls dependent of the observed variables. In the case that there are some patterns of missingness but the observed variables cannot explain these patterns, the last class arises [9].

Various approaches have the potentiality of dealing with aforementioned missingness in MTS prediction. A straightforward policy is to discard the incomplete information and feed the complete information to traditional models like autoregressive moving average (ARMA) [10] or its generalized model, autoregressive integrated moving average (ARIMA) [11]. This policy could diminish the accuracy of prediction, specifically when it comes to MNAR missingness, due to the loss of rich information in the missing points of variables. Some researchers have applied other policies to impute the missing values with either simple statistics like mean or median [12] or more sophisticated statistics like polynomial interpolation [13], matrix factorization [14] and expectation maximization (EM) [15]. Furthermore, machine-learning algorithms like ANN [16], kNN [17] and decision tree [18] have been employed for the purpose of approximation and imputation of missing values. However, some of these approaches fall short in capturing the complex dynamisms of temporal dependencies within variables and some cannot deal with missingness when large amounts of data are lost or not collected.

Taking into account recurrent methods, the variants of Recurrent Neural Network (RNN) like Gated Recurrent Unit (GRU) [19–21] and Long Short-Term Memory (LSTM) [22–24] have demonstrated promising results in extracting temporal features and some could deal with huge missingness as well. It is notable that the approaches in [19–22] are among few works on addressing massive MNAR missingness in MTS (more than 70%). However, the main focus of these works have been on the classification task of medical domains. Exploring massive MNAR missingness in the regression task of other domains is still an open research area. More importantly, since not all variables follow the same missing rate in many applications, it is worth investigating the missing rate of each variable along with other missing information in the recurrent methods [21]. To the best of our knowledge, this has not yet been explored in an LSTM-based structure, having the power of controlling the information flow with its memory unit, jointly with other missing information.

This paper proposes a novel LSTM-based model called FBVS-LSTM to contribute to:

- dealing with massive MNAR missingness over three real-world meteorological and air quality MTS datasets [25–27], under the flag of regression task;
- exploring a new LSTM-based architecture, integrating jointly two decay mechanisms with the missing rate of each variable, to learn the missing pattern informatively;
- concluding that not all missing patterns provide informative data in the meteorological settings.

The rest of the paper is organized as follows. Section 2 considers related work in the domain of MTS with missingness. Section 3 gives details of the proposed method. We conduct our experiments on air pollution forecasting in Section 4. Finally, the conclusion and future works will be discussed in Section 5.

2. Related Works

Multivariate time series with missing values has been the main challenge in a great amount of literature for a long time [28]. Some works [29–31] put their attention on dealing with the MCAR missingness category mentioned in Section 1. In [29] a combination of three statistical models, namely vector autoregressive (VAR) [32], expectation and minimization (EM) and prediction error minimization (PEM) were applied to impute missing values. Normally, in an autoregressive model (AR) each sample is a linear combination of some previous observations with a stochastic term, while in a more generalized form, a vector autoregressive captures the linear interdependencies among multiple time series. The applied method identifies the series with different time lag, selects the best time lag and accordingly maximizes the likelihood of parameter estimation in the incomplete time series to impute missing values. Since the parameter estimation is computationally expensive in the EM process, Liu et al. [30] proposed sampling from Gaussian and gamma distributions as an alternative of sampling from conditional distribution in the EM process. In addition to [30], exploiting the precise model of MTS distribution for missingness imputation was discussed in [31]. The model exploits the power of generative adversarial networks to train the complete set and then apply the trained

model for the imputation of the incomplete set. The imputation is done by finding the "closest" latent encoding of the missing value and then applying the generated samples of generator. Despite the effectiveness of these models [29–31], they did not prove the same performance in case of consecutive missingness. Addressing the consecutive gaps, Zhang et al. [33] proposed a deep-learning model consisting of encoder and decoder components. The model used the past and future information of latent factors in a bi-directional LSTM structure of encoder component. After applying a general attention mechanism to let the model focus on the data range of interest, uni-directional LSTM units accompanied by a fully connected layer were employed. Their main task was to reconstruct the learnt input data with missingness in a decoder component. Here, the assumption was that all variables follow the same missing rate, while in many applications the variables do not stand for the same characteristics and consequently the variables missing rate could differ.

Focusing on the MAR category of missingness in different applications [34,35], Feng et al. [34] manipulated the RNN structure to impute missingness in biomedical wearable recordings. Similar to our work, the model explored the latent features of missing values, using the forward and backward missing information. To impute missingness, while our model focuses on learning the past and future time lags in a jointly uni-directional manner, the current model puts attention on learning the history estimation as well as the feature estimation in a bi-directional manner. Shifting to the wind power prediction, Liu et al. [35] applied an EM-based estimation (like [29,30]) to impute missing values by estimating the mixture components of data distribution. Here, multiple imputation was performed to generate new samples and then let a Gaussian process regression signify how likely is a prediction, given the actual data.

Unlike the previous missingness categories, most of the literature on MNAR was mainly oriented towards health-related applications. Lipton et al. [36] worked on patient ICU records and modeled directly missing data by contributing a binary variable as well as other missing indicators. The variable basically indicated whether the data was measured or missed. Other missing indicators consisted of the mean and standard deviation of missing value, the relative times of the first and last recordings and also the frequency with which a variable switches from measured to missing or vice versa across adjacent sequence steps. All this information was attached to the input values and learnt in an LSTM-based structure. Similar to [36], Singh et al. [24] augmented the network input with some of the mentioned missing information, but in a bi-directional LSTM-based model. Moreover, the model was extended with some additional layers providing meaningful representations of missing data and more attention to the important spans of time series. Focusing again on an LSTM-based model, Kim et al. [22] presented a bio-inspired approach in terms of belief gate for the purpose of imputing missing data with either the last observation of each feature or its average. The work mainly focused on the impact of missingness forward and backward time intervals, interpreting as the extent of temporal belief one can trust on only the last observation at the current time. While this work considered the missing imputation within the stream, the work in [37] focused on the imputation and interpolation both across and within the stream, respectively. In general, all these models [22,24,36,37] exploited a fixed impact of time lags for the missingness imputations. Later, in [19,20], the authors addressed a dynamic impact of time lag under the flag of decay mechanism. The mechanism was imposed on both the input and latent factors in a GRU-based structure to explore more meaningful representations of missingness. Although in the latter works the missing rate of variables were huge (more than 70%) and different from each other, the model behaved the same for all variables. In [21], the missing rate of each variable was included in the inputs of GRU to reduce the harmful impact of variables with high missing rate on variables with low missing rate. Considering the fact that the mentioned literature of MNAR category was all analyzed over medical datasets, the work in [38] extended the analysis over the computer vision domain as well. The main idea was to reconstruct the missing values by not only the information of correlated observed features at the current time point but also other time points of the series. It is worth noting that all these models were articulated for the task of either classification or data reconstruction in MTS.

In this paper, we focus on the regression task of MTS in a different domain, mainly meteorological and air quality domain, with massive MNAR missingness. Moreover, different from [21], we investigate the potential capability of LSTM, by introducing a new structure. The structure mainly learns the joint information of two time lags under two decay mechanisms as well as the missing rate of each variable and its mask indicator.

3. Methods

3.1. LSTM

As a model of learning sequential data and capturing long term temporal dependencies, Long Short-Term Memory (LSTM) was first proposed by Hochreiter [39] in 1997. This model allows for constant error flow through self-connected units to impede from the gradient decay. In this regard, a memory cell along with three major gates construct the architecture of LSTM cell unit. The memory cell is mainly devised to keep or release the information by the contribution of three aforementioned gates. The gates are namely input gate i_t, forget gate f_t and output gate o_t, by each the extent of information for passing forward is controlled. The following equations give more insight into the gates and process:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{1}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{2}$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \tag{3}$$

$$c_t = f_t \circ c_{t-1} + (i_t \circ \phi(W_c x_t + U_c h_{t-1} + b_c)) \tag{4}$$

$$h_t = o_t \circ \phi(c_t) \tag{5}$$

In the equations above, σ stands for the sigmoid function, ϕ indicates the tanh function and $\circ$ stands for the element-wise multiplication. Furthermore, x_t and h_{t-1} are the inputs and hidden state at time t and $t-1$ respectively. The model parameters are also identified by W, U as weights and b as biases.

LSTM does not by itself contain any information of missingness to cope with multivariate time series with massive missingness. To make the model convenient for this purpose, we discuss two phases of modification. In the first phase, we propose an LSTM with two decay mechanisms, namely forward and backward LSTM (FB-LSTM). This model contains the *mask indicator* and *time intervals* as two important information of missing pattern. In the second phase, we extend FB-LSTM to a variable-sensitive version, namely FBVS-LSTM. Here, we incorporate the missing rate of each variable to deal with massive missingness problem. In the following, the aforementioned models are articulated in Sections 3.2 and 3.3 respectively.

3.2. FB-LSTM

To deal with MNAR category of missing data, we modify LSTM with two decay mechanisms, namely forward and backward LSTM (FB-LSTM). The decay mechanisms are devised to reinforce the imputation of missing values more accurately. Similar to [19], the missing data is formulated with a missing indicator $M = \{m_1, m_2, \ldots, m_T\}^T \in \mathbb{R}^{D \times T}$ for each observation x_t^d in the time series $X = \{x_1, x_2, \ldots, x_T\}^T \in \mathbb{R}^{D \times T}$ where $d \in \{1, \ldots, D\}$ and $t \in \{1, \ldots, T\}$ denote the variable and time of observation, respectively. In this regard, the observation, x_t^d is interpreted as the t-th observation of the variable d. Following the formulations, the missing indicator m at time stamp t of the variable d, m_t^d, is regarded as a binary mask as below:

$$m_t^d = \begin{cases} 1, & \text{if } x_t^d \text{ is observed} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

In addition to such information, suppose that in the process of meteorological data collection, the sensors of capturing air pollution have low sensitivity to the high pollution. Therefore, the data gathering would drop in such case of pollution. The process will be back on the track of data collection when the level of pollution decreases. In this context, considering the duration of missingness is critical to be explored as an informative missing pattern. To address this, in each time stamp, the last observation of each variable from current missing data, as well as the first observation after that, are calculated. The former follows the same formulation as [19], incorporating our first decay mechanism and the latter is formulated in this work, incorporating our second decay mechanism. While the first mechanism contributes in the decreasing impact of backward time interval, the second mechanism does the same in the forward time interval. In the following, similar to the definition in [19], the time interval between the last observation and current missing data for each variable d at time stamp t is defined as δ_t^{1d} in a set of $\delta^1 = \{{\delta_1}^1, {\delta_2}^1, \ldots, {\delta_T}^1\}^T \in \mathbb{R}^{D\times T}$:

$$\delta_t^{1d} = \begin{cases} s_t - s_{t-1} + \delta_{t-1}^{1d}, & t > 1, m_{t-1}^d = 0 \\ s_t - s_{t-1}, & t > 1, m_{t-1}^d = 1 \\ 0, & t = 1 \end{cases} \tag{7}$$

In the equation above, s_t is the time stamps when the t-th observation is recorded.

In this paper, we define $\delta^2 = \{{\delta_1}^2, {\delta_2}^2, \ldots, {\delta_T}^2\}^T \in \mathbb{R}^{D\times T}$ as the set of time intervals between the current value and first observation after that. Therefore, δ_t^{2d} indicates the forward time interval for each variable d at time t:

$$\delta_t^{2d} = \begin{cases} s_{t+1} - s_t + \delta_{t+1}^{2d}, & t > 1, m_{t+1}^d = 0 \\ s_{t+1} - s_t, & t > 1, m_{t+1}^d = 1 \\ 0, & t = 1 \end{cases} \tag{8}$$

Although the two defined sets of time intervals could reflect the useful information influencing the missing data, their impact on missingness decrease as the time intervals increase. Therefore, two decay rates are defined, implying for the first and second decay mechanisms. The rates mainly control the impacts of time intervals over time and contribute in the process of learning with other model parameters in LSTM. The first decay rate, γ^1, is formulated as [19] and the second one, γ^2, is introduced in this work:

$$\gamma^1 = exp\{-max(0, {W_\gamma}^1 \delta^1 + {b_\gamma}^1)\} \tag{9}$$

$$\gamma^2 = exp\{-max(0, {W_\gamma}^2 \delta^2 + {b_\gamma}^2)\} \tag{10}$$

where ${W_\gamma}^1$, ${W_\gamma}^2$, ${b_\gamma}^1$ and ${b_\gamma}^2$ are the model parameters. It should be mentioned that the decay rates range between 0 and 1.

To learn the parameters, the decay rates are imposed jointly to the input and hidden features of LSTM to capture the missing pattern informatively. This process constructs the main structure of LSTM with two decay mechanisms (FB-LSTM).

In FB-LSTM, the missing data is imputed with the values either close to the mean of the variable or close to the last/first observation of the variable. This basically implies the interpretation that the smaller the time intervals are, the closer the missing data is to the last/first observation of the variable. Moreover, this data is closer to the mean of the variable if the time intervals are larger. To formulate this, we denote $\tilde{x}^d$ as the mean of variable d. The last and first observations of variable d are also indicated by $x_{t'}^d$ and $x_{t''}^d$, respectively. The Equation (11) shows the decay process over the input mathematically:

$$\hat{x}_t^d = m_t^d x_t^d + (1 - m_t^d)\Big[(\gamma_{\mathbf{x}t}^{1d} x_{t'}^d + \gamma_{\mathbf{x}t}^{2d} x_{t''}^d) + ((1 - \gamma_{\mathbf{x}t}^{1d})(1 - \gamma_{\mathbf{x}t}^{2d})\tilde{x}^d)\Big] \tag{11}$$

where $\gamma_{\mathbf{x}t}^{1d}$ and $\gamma_{\mathbf{x}t}^{2d}$ represent the input decay rates of the first and second decay mechanisms, respectively.

In addition to imposing the two decay mechanisms on the input, we apply the same mechanisms on the hidden state to facilitate exploring the rich information of missing data in the latent space. More clearly, the two decay rates make their influence simultaneously on the previous hidden state as Equation (12):

$$\hat{h}_{t-1} = \gamma_{\mathbf{h}t}^{1} \gamma_{\mathbf{h}t}^{2} \odot h_{t-1} \tag{12}$$

in which $\gamma_{\mathbf{h}t}^{1}$ and $\gamma_{\mathbf{h}t}^{2}$ represent the hidden state decay rates.

Then, the obtained input and hidden state from Equations (11) and (12) directly incorporate in LSTM gates, described in Section 3.1, to construct the structure of FB-LSTM. Furthermore, the masking indicator is added to all the three gates to let the model learn from the missingness directly. The following equations illustrate the FB-LSTM functionality over the cell memory and all the gates:

$$f_t = \sigma(W_f \hat{x}_t + U_f \hat{h}_{t-1} + V_f m_t + b_f) \tag{13}$$

$$i_t = \sigma(W_i \hat{x}_t + U_i \hat{h}_{t-1} + V_i m_t + b_i) \tag{14}$$

$$o_t = \sigma(W_o \hat{x}_t + U_o \hat{h}_{t-1} + V_o m_t + b_o) \tag{15}$$

$$c_t = f_t \circ c_{t-1} + (i_t \circ \phi(W_c \hat{x}_t + U_c \hat{h}_{t-1} + V_c m_t + b_c)) \tag{16}$$

$$h_t = o_t \circ \phi(c_t) \tag{17}$$

in which V_f, V_i, V_o and V_c are the added parameters in FB-LSTM.

3.3. FBVS-LSTM

In many MTS, each variable follows its own characteristics, implying different frequency in the case of missingness. This characteristic is critical to explore, specifically when the variables with high missing frequency negatively influence those with low missing frequency. Since FB-LSTM only considers the missing indicator and the time intervals of missingness for each variable, we extend this model to a variable-sensitive version, namely forward and backward variable-sensitive LSTM (FBVS-LSTM). In the following, we explain how this model works.

First, the missing rate of each variable d, μ^d, over all time steps t is calculated by the mask indicator m_t^d. We formulate this rate similar to [21]:

$$\mu^d = 1 - \frac{1}{T} \sum_{t=1}^{T} m_t^d \tag{18}$$

where μ^d ranges between 0 and 1.

Then, to make the model adapted to the missing rate of each variable independently, there is a possibility of contributing μ^d in the learning process of previously articulated model (FB-LSTM) and accordingly constructing FBVS-LSTM model. The advantage of this contribution is to make the model sensitive to the variables with low missingness. However, subsuming μ^d directly in FB-LSTM, the model cannot discern the particular missingness feature of variables for those with close missing frequency. This is notable, specifically when μ^d is close to 1 in case of massive missingness. Therefore, the missing rate of each variable is decayed within a negative exponential function to construct a missing factor β. Later, this factor directly takes part in the learning process in FBVS-LSTM. The formulation of β is similar to [21] as below:

$$\beta = exp\{-max(0, W_\beta \mu + b_\beta)\} \tag{19}$$

where W_β is a vector as size as the transpose of missing rates vector μ, and b_β is also a vector as size as μ. W_β stands for the weights and b_β indicates the bias, integrating with μ.

Incorporating the missing factor β in FB-LSTM, the gates are rectified as the equations below and construct our final model as FBVS-LSTM:

$$f_t = \sigma(W_f \hat{x}_t + U_f \hat{h}_{t-1} + V_f m_t + P_f \beta + b_f) \tag{20}$$

$$i_t = \sigma(W_i \hat{x}_t + U_i \hat{h}_{t-1} + V_i m_t + P_i \beta + b_i) \tag{21}$$

$$o_t = \sigma(W_o \hat{x}_t + U_o \hat{h}_{t-1} + V_o m_t + P_o \beta + b_o) \tag{22}$$

$$c_t = f_t \circ c_{t-1} + (i_t \circ \phi(W_c \hat{x}_t + U_c \hat{h}_{t-1} + V_c m_t + P_c \beta + b_c)) \tag{23}$$

$$h_t = o_t \circ \phi(c_t) \tag{24}$$

in which W, U, and V are the parameters of model. These parameters are regarded as vectors instead of matrixes. This could accelerate the learning process in FBVS-LSTM with less parameter computation than FB-LSTM. Another parameter of FBVS-LSTM compared with FB-LSTM is the vector P, responsible for learning the missing factor. Considering massive reduction of computation in FBVS-LSTM, this vector does not overload more than the entire computations in FB-LSTM. The cell structure of FBVS-LSTM is depicted in Figure 1.

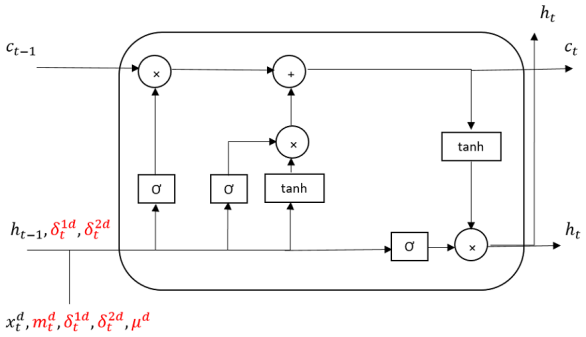


Figure 1. FBVS-LSTM unit.

4. Experiments

4.1. Dataset Description and Preprocessing

The proposed method is performed on three real datasets, namely Beijing PM2.5 [25], Italy Air Quality [26] and Beijing Multi-Site Air Quality [27], collections of hourly meteorological and air quality data. The first dataset consists of 8 main attributes, namely *PM2.5 concentration*, *dew point*, *temperature*, *pressure*, *wind direction*, *cumulated wind speed*, *cumulated hours of snow* and *cumulated hours of rain*. The values have been gathered for the period of 1 January 2010 to 31 December 2014. This period contains 43,824 hourly instances quite big data. To reduce the time complexity, we select only one year of data over the period of 1 January 2010 till 31 December 2010, consisting of 8760 samples. The second dataset encompasses the average responses of 5 metal oxide chemical sensors in one of the polluted areas in an Italian city. The attributes are mainly PT08.S1 (tin oxide), PT08.S2 (titania), PT08.S3 (tungsten oxide), PT08.S4 (tungsten oxide) and PT08.S5 (indium oxide), nominally stand for CO, NMHC, NOx, NO2 and O3. Considering 8760 out of 9358 samples, we choose again only one year of data, from 11 March 2004 to 11 March 2005. Moreover, employing the data collected from multiple sites in Beijing, the third dataset stands for 6 features in Guanyuan site, namely PM2.5, PM10, SO2, NO2, CO and O3 concentrations. Similar to two previous datasets, we consider a subset of data to reduce time complexity. Here, 7346 out of 35,065 instances from the period of 1 March 2013 till the third data of 1 January 2014 are opted.

We reshape the samples and generate a multivariate time series of 24 h within 8, 5 and 6 variables, standing for each of our three datasets, respectively. These samples are required to feed into our models, discussed further in Section 4.3, for the purpose of short-term (next-hour) prediction. In order to explore the imputation impact of variables with massive missingness on the prediction task, first we forecast a variable with a high missing rate in the first dataset and later we predict variables with low missing rate in the second and third datasets, respectively. In the first dataset, a one-step-ahead prediction of PM2.5 concentration over the last 24 h is performed. In the cases of the second and third datasets, we focus on the next-hour prediction of PT08.S1 (tin oxide), saying PT08.S1 (CO), and O3, respectively over the last 24 h. It should be mentioned that we drop the first 24 h in the first dataset, making our data uniform when generating the MNAR missingness.

To synthetically generate MNAR missingness, we make usage of each feature median [40] in our datasets. Since we aim to achieve high rate of missingness, we consider different formulations, yet the same logic, in each dataset. In Beijin PM2.5, we subtract $0.6 \times median$ of each feature from its median. In fact, two groups of values are defined given this subtraction. One group has the higher values than the subtraction value and the other group has the lower values. The first group is represented as missingness while the second group keeps the observed values. The following formulation indicates this process for all features:

$$x_t^d = \begin{cases} observed, & \text{if } x_t^d < median(x^d) - 0.6 * median(x^d) \\ missing, & \text{otherwise.} \end{cases} \tag{25}$$

However, by applying Equation (25) on the *PM2.5 concentration* feature, all values of this feature are lost. Therefore, we apply only the median of this variable as the decision point of missingness. It should be pointed out that we refer to this feature as PM2.5 for simplicity. Equation (26) shows this formulation:

$$x_t^{PM2.5} = \begin{cases} observed, & \text{if } x_t^{PM2.5} < median(x^{PM2.5}) \\ missing, & \text{otherwise.} \end{cases} \tag{26}$$

With the same policy, we formulate missingness in each feature of the second and third datasets. Here, we discriminate PT08.S3 (tungsten oxide), saying PT08.S3(NOx) and SO2 in these two datasets, applying different formulations to provide massive missingness. Equations (27) and (28) indicate the formulations of MNAR missingness generation over all features of the second and third datasets respectively, excluding PT08.S3(NOx) and SO2 features.

$$x_t^d = \begin{cases} observed, & \text{if } x_t^d < median(x^d) + 0.1 * median(x^d) \\ missing, & \text{otherwise.} \end{cases} \tag{27}$$

$$x_t^d = \begin{cases} observed, & \text{if } x_t^d < median(x^d) + 0.5 * median(x^d) \\ missing, & \text{otherwise.} \end{cases} \tag{28}$$

Additionally, we generate missing data over PT08.S3(NOx), and SO2 attributes, employing the following equations:

$$x_t^{NOx} = \begin{cases} observed, & \text{if } x_t^{NOx} < median(x^{NOx}) - 0.3 * median(x^{NOx}) \\ missing, & \text{otherwise.} \end{cases} \tag{29}$$

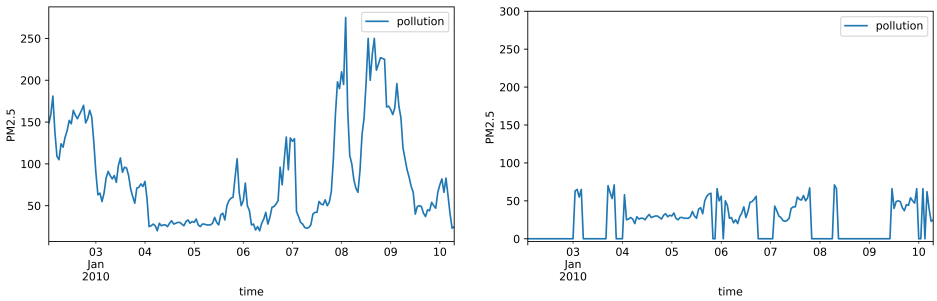
$$x_t^{SO2} = \begin{cases} observed, & \text{if } x_t^{SO2} < median(x^{SO2}) - 0.6 * median(x^{SO2}) \\ missing, & \text{otherwise.} \end{cases} \tag{30}$$

It should be mentioned that we used "NOx" as an abbreviation for "PT08.S3(NOx)" variable in Equation (29).

After applying the formulations above, the approximate missing rate of all datasets features are calculated and presented in Table 1. In addition, to visually compare the actual values of features and their generated missingness, Figures 2–4 are depicted over 200 samples standing for PM2.5, the attribute of first dataset, as well as NOx and SO2, the attributes of second and third datasets, respectively.

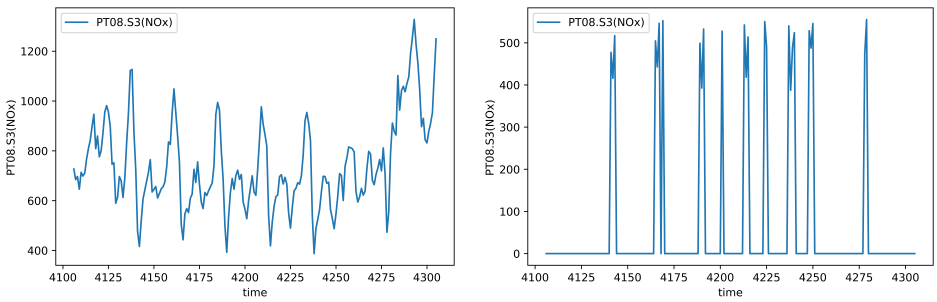
Table 1. Missing rate of datasets features.

Dataset	Features	Missing Rate
Beijing PM2.5	PM2.5	75%
	dew	52%
	temperature	66%
	pressure	49%
	wind direction	89%
	wind speed	67%
	snow	1%
	rain	5%
Italy Air Quality	PT08.S1(CO)	34%
	PT08.S2(NMHC)	38%
	PT08.S3(NOx)	88%
	PT08.S4(NO2)	32%
	PT08.S5(O3)	45%
Beijing Multi-Site Air-Quality	PM2.5	34%
	PM10	28%
	SO2	81%
	NO2	22%
	CO	36%
	O3	38%



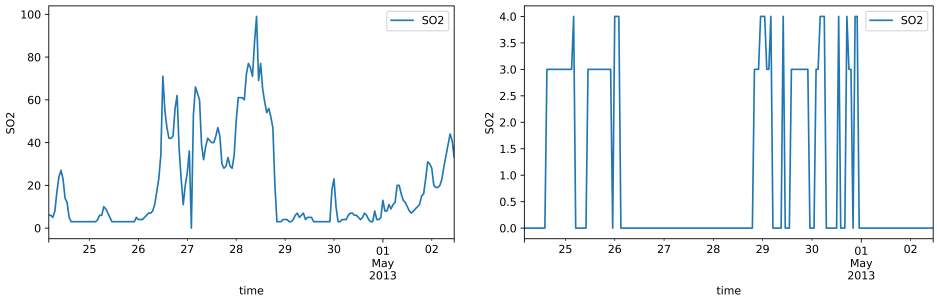
(**a**) (**b**)

Figure 2. PM2.5 concentration over 200 samples in Beijin PM2.5. (**a**) Actual data; (**b**) Generated missing data.



(**a**) (**b**)

Figure 3. NOx over 200 samples in Italy Air Quality. (**a**) Actual data; (**b**) Generated missing data.



(**a**) (**b**)

Figure 4. SO2 over 200 samples in Beijing Multi-Site Air-Quality. (**a**) Actual data; (**b**) Generated missing data.

4.2. Metric

To measure the performance of models in the regression task of all datasets, we consider the mean squared error (MSE), mathematically represented as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{31}$$

4.3. Evaluation and Results

In this section, we evaluate the results of proposed method along with five variations of LSTM as assessment models over all datasets. The assessment models mainly consist of LSTM with zero imputation, LSTM with mean imputation, LSTM with first decay mechanism, LSTM with second decay mechanism and variable-sensitive LSTM equipped with the first decay mechanism. We refer these models as LSTM-0, LSTM-mean, B-LSTM, F-LSTM and BVS-LSTM, respectively. The first model imputes missing data with zero and the second model imputes with the mean value of each variable in all time stamps. While the third model imputes missingness considering the forward time interval, δ^1, the fourth model follows the same procedure considering the backward time interval, δ^2. Finally, the last model employs both δ^1 and μ, missing rate of each variable, for imputation. Applying any of these models, the mask indicator is also fed into the model to contribute in identifying the missing pattern directly.

It is worth pointing out that time series data of each dataset are normalized before feeding into the network to scale all features in the same range. This process is accomplished by max-min normalization and ranges data between 0 and 1 as follows:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{32}$$

The normalized time series are then appraised with 5-fold cross-validation on the assessment models as well as the proposed model. In each fold, the models of each dataset are separately trained within 24 LSTM units implying for 24 h training data. More clearly, each unit stands for one hour of training data. Figure 5 illustrates the whole architecture of the proposed model with the 24 units.

Moreover, the training process of each fold accomplishes a 30-epoch run for each model. As mentioned before, the output is one-step-ahead prediction of appropriate variable in each dataset, indicating the next-hour prediction from the current time. Therefore, we tune the number of outputs to 1 for all datasets. Since there are 8, 5 and 6 features in the first, second and third datasets respectively, the input size of each dataset is adjusted correspondingly. Furthermore, the learning rate is initiated with 0.01 and the optimizer to train each model is considered to be stochastic gradient descent. The parameter settings of all datasets are shown in Table 2.

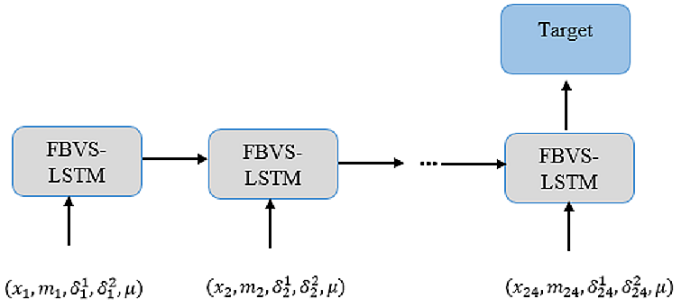


Figure 5. The general architecture of FBVS-LSTM units.

Table 2. Parameter settings.

Datasets	Parameters				
	Epoch Number	**Learning Rate**	**Hidden Layers**	**Features (Input Size)**	**Output Size**
Beijing PM2.5	30	0.01	24	8	1
Italy Air Quality	30	0.01	24	5	1
Beijing Multi-Site Air-Quality	30	0.01	24	6	1

Table 3 shows the results of comparison between each assessment model and the proposed method in three datasets. The results are represented in terms of training and test sets MSE errors as well as their standard deviation. It is worth pointing out that the errors imply for the average loss errors of 5-fold cross-validation in each model of each dataset. Considering the results of first dataset, we could verify that FBVS-LSTM performs rather similar to BVS-LSTM during the training process with a slight error difference. This reveals the fact that the forward time interval does not provide much more effective information than the integration of backward time interval and missing rate provides. This claim is also asserted by comparing the test errors of these two models. In this case, FBVS-LSTM performs again tightly in line with BVS-LSTM. Here as well as the training set, F-LSTM has the highest amount of errors among other models and this could strengthen the truth of our previous argument that the latent factors of future are not strong representations of missing pattern in this dataset. This is mainly due to the rather non-discriminative future pattern in a local window. Comparing other assessment models, LSTM-0 and LSTM-mean show quite similar performance to each other during the training and testing process, while B-LSTM outperforms these two models. However, statistically all these three models as well as BVS-LSTM perform quite similar to FBVS-LSTM. In general, closing the errors to zero proves the robustness of all these LSTM-based models and indicates that the variables with high rate of missingness influence those with low rate less. This is because the models with the power of their gates and cell memory could regulate the extent of missingness imputation in a local window within the variables. This is true even if there is no information of missingness. In case of such information, the models also generate an estimation of each feature for regulation. Moreover, by separately learning the information of missing rate, the variables with high missingness could less influence those with low missingness.

The results extracted from the second dataset in terms of the training and test errors indicate that FBVS-LSTM outperforms other models with the minimum error difference around 0.01 over BVS-LSTM and the maximum error difference around 0.8 over LSTM-0. This verifies that all four information of missingness employed in FBVS-LSTM influenced the imputation process more accurate than other models, specifically LSTM-mean and LSTM-0. These two models performed more weakly relative to other models. Considering the fact that this dataset encompasses high variations (see Figure 3), LSTM-mean and LSTM-0 encounter with the lack of additional information to extract latent factors. Yet, relying on the forget gate in LSTM-0 as well as the mean of variables in LSTM-mean provide promising results. Among other models, here F-LSTM also shows promising results and this is mainly due to equipping with additional information when it comes to a dataset with high variation.

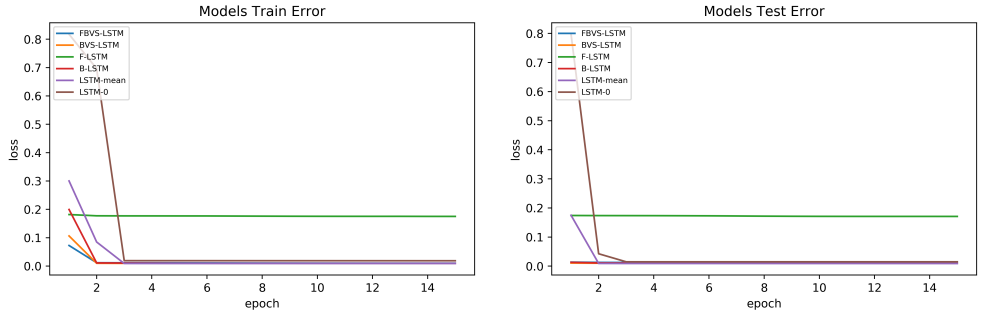
The same argument stands for B-LSTM as well. In fact, F-LSTM and B-LSTM models accompanied with BVS-LSTM show quite similar performance to our proposed model in this dataset.

Table 3. All models performance in each dataset.

Dataset	Model	MSE ± STD	
		Train Error	Test Error
Beijin PM2.5	LSTM-0	0.021 ± 0.020	0.016 ± 0.009
	LSTM-mean	0.021 ± 0.016	0.015 ± 0.011
	B-LSTM	0.016 ± 0.008	0.010 ± 0.004
	F-LSTM	0.180 ± 0.324	0.172 ± 0.323
	BVS-LSTM	0.013 ± 0.006	0.010 ± 0.004
	FBVS-LSTM	0.012 ± 0.004	0.011 ± 0.005
Italy Air Quality	LSTM-0	0.122 ± 0.123	0.130 ± 0.142
	LSTM-mean	0.063 ± 0.078	0.066 ± 0.082
	B-LSTM	0.027 ± 0.003	0.027 ± 0.009
	F-LSTM	0.030 ± 0.007	0.029 ± 0.015
	BVS-LSTM	0.031 ± 0.011	0.031 ± 0.013
	FBVS-LSTM	0.023 ± 0.002	0.024 ± 0.006
Beijing Multi-Site Air-Quality	LSTM-0	0.049 ± 0.029	0.06± 0.032
	LSTM-mean	0.038 ± 0.015	0.03 ± 0.023
	B-LSTM	0.031 ± 0.011	0.03 ± 0.016
	F-LSTM	0.179 ± 0.3	0.148 ± 0.25
	BVS-LSTM	0.034 ± 0.008	0.040 ± 0.024
	FBVS-LSTM	0.026 ± 0.019	0.031 ± 0.002

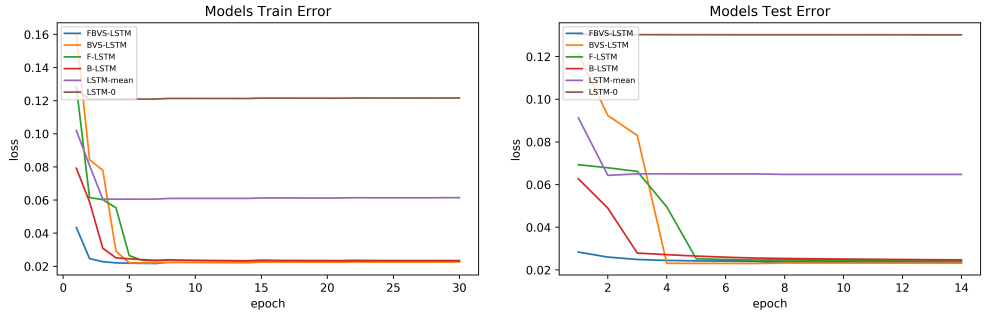
In the third dataset, the presented model performs tightly in line with B-LSTM in both training and test set errors. Comparing FBVS-LSTM with BVS-LSTM, B-LSTM and F-LSTM, it proves that forward time intervals in missing data does not provide informative representation of missingness. The claim is similar to the results obtained from the first dataset. Focusing on LSTM-mean and LSTM-0, the former presents more accurate imputation due to the input inclination toward the mean of variable. In future works, we could test and analyze whether the same situation exists for other statistical attributes of variable. In general, all models except F-LSTM shows promising results.

Visually exploring how the models of each dataset perform in the training and test sets, Figures 6–8 are depicted in 15 epochs. Each epoch illustrates the average errors of the same epoch within 5-fold cross-validation. As shown in Figure 6, in the first dataset, all models except F-LSTM follow a decreasing pattern until they reach constant error levels. As mentioned before, F-LSTM could not provide informative representation of missingness in this dataset. Therefore, this method pursues a constant trend within all epochs. The same analysis is valid in the third dataset as well (Figure 8), except the decreasing trend of F-LSTM within the primitive epochs in the training set. Among other models in the first and third datasets, although FBVS-LSTM learns the missing pattern with slightly lower loss error than LSTM-0, LSTM-mean and BVS-LSTM from the primitive stages, these models also perform well with loss errors close to zero. This argument stands for both the training and test sets in both datasets. In case of B-LSTM, the model produces higher errors than FBVS-LSTM in the training sets of first and third datasets (Figures 6a and 8a). However, it shows similar performance to FBVS-LSTM in the test sets of first dataset (Figure 6b) and a lower loss start in the test set of third dataset (Figure 8b). Considering model performance in the second dataset, LSTM-0 and LSTM-mean encounter higher errors in the training and test sets from the primitive epochs as expected (Figure 7a,b). Other models as well as the proposed model follow a decreasing pattern in both training and test sets, yet lower loss in the beginning refers to FBVS-LSTM. In general, we could conclude that most of LSTM-based models perform well dealing with massive MNAR missingness in the regression task of meteorological settings. To realize the accuracy of proposed model visually, the prediction and ground truth of PM2.5, PT08.S1(CO) and O3 variables over 70 test data samples are depicted for their corresponding dataset in Figure 9.



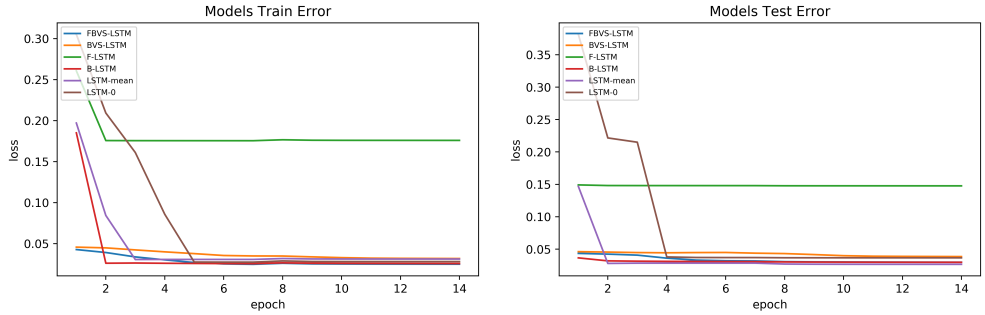
(**a**) (**b**)

Figure 6. Models performance in Beijing PM2.5 dataset. (**a**) Training set errors; (**b**) Test set errors.



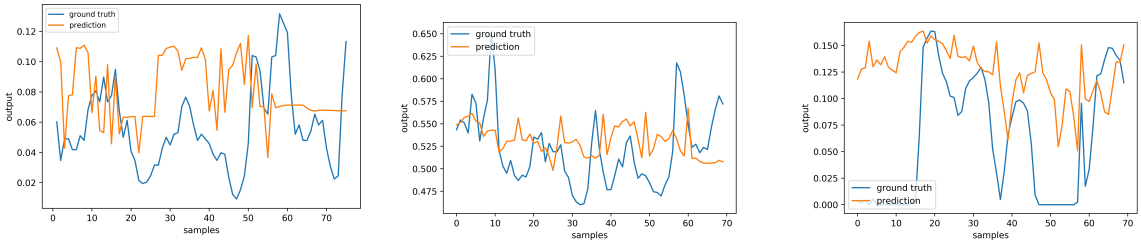
(**a**) (**b**)

Figure 7. Models performance in Italy Air Quality dataset. (**a**)Training set errors; (**b**) Test set errors.



(**a**) (**b**)

Figure 8. Models performance in Beijing Multi-Site Air-Quality dataset. (**a**) Training set errors; (**b**) Test set errors.



(**a**) (**b**) (**c**)

Figure 9. Prediction and ground truth outputs over 70 test samples. (**a**) PM2.5 in Beijing PM2.5 dataset; (**b**) PT08.S1(CO) in Italy Air Quality dataset; (**c**) O3 in Beijing Multi-Site Air-Quality.

4.4. Statistical Analysis

To acknowledge the achieved results of Section 4.3, a statistical analysis has been employed to compare the proposed algorithm (FBVS-LSTM) with the introduced assessment models in each of the three datasets. In this paper, we conduct the t-test as one of the most commonly applied analysis. Two hypotheses, H0 and H1, are defined as below:

- **H0**: *The proposed method performed similarly w.r.t. other assessment models.*
- **H1**: *The proposed method performed differently w.r.t. other assessment models.*

Here, we consider 30 tests. In addition, the degrees of freedom and alpha are adjusted to 58 and 0.05, respectively. Therefore, there is 95% confidence that the conclusion of test is valid. Table 4 shows the result of *t*-test in terms of *t*-value and *p*-value in each of the applied datasets and their models. It can be seen that in the first dataset, F-LSTM has the *p*-value equals to 0.0001, much less than 0.05, indicating the acceptance of rejecting the null hypothesis. With the same logic, LSTM-0 and LSTM-mean reject the null hypothesis in the second dataset with p-values equal to 0.0002 and 0.0005, respectively. In the third dataset, B-LSTM model accepts the null hypothesis with the *p*-value equals to 0.71, much higher than 0.05. All other models of this dataset reject H0 and accept H1.

Table 4. Statistical analysis with *t*-test.

Dataset	Model	FBVS-LSTM	
		***t*-Value**	***p*-Value**
Beijin PM2.5	LSTM-0	−1.62	0.11
	LSTM-mean	−0.87	0.39
	B-LSTM	−0.5	0.61
	F-LSTM	−78.23	0.0001
	BVS-LSTM	−0.02	0.98
Italy Air Quality	LSTM-0	−135.68	0.0002
	LSTM-mean	−24.17	0.0005
	B-LSTM	−1.58	0.12
	F-LSTM	−1.87	0.06
	BVS-LSTM	−1.51	0.13
Beijing Multi-Site Air-Quality	LSTM-0	−15.07	0.0001
	LSTM-mean	−868.89	0.0004
	B-LSTM	−0.37	0.71
	F-LSTM	−52.92	0.0008
	BVS-LSTM	−6.00	0.0001

5. Conclusions and Future Works

This paper was conducted to address the massive MNAR missingness on the regression task of meteorological multivariate time series. We proposed a novel LSTM-based model, FBVS-LSTM, consisting of four effective pieces of information as the augmentation of model input. The information included the missing indicator, two time intervals of missingness in forward and backward direction and missing rate of each variable. Experiments were conducted on Beijing PM2.5, Italy Air Quality and Beijing Multi-Site Air Quality datasets, filtering out around one year of data of all datasets for short-term prediction. The results proved promising performance of the proposed model along with some LSTM-based derivative methods. More importantly, we concluded that not all missing patterns provide meaningful representation of missingness. Future study aims to replicate the experiments in a health-related domain, mainly a daily stress-monitoring dataset collected from different sensors. Additionally, we will explore how the deeper layers of LSTM could affect the performance of model. Finally, further evaluations with other deep-learning models like GRU will be performed to investigate which gates have the most influential role to deal with missingness.

Author Contributions: Methodology, N.F.; writing—original draft, N.F.; writing—review and editing, N.F.; supervision, K.F.; funding acquisition, K.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Umeå University.

Acknowledgments: We are immensely grateful to Frank Drewes from Umeå University, Sweden for his valuable comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lewis, R.; Reinsel, G.C. Prediction of multivariate time series by autoregressive model fitting. *J. Multivar. Anal.* **1985**, *16*, 393–411. [CrossRef]
2. Han, M.; Fan, M.; Xi, J. Study of Nonlinear Multivariate Time Series Prediction Based on Neural Networks. In *Advances in Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2005; Volume 3497, pp. 618–623.
3. Cai, Y.; Wang, H.; Ye, X.; An, L. Multivariate Time Series Prediction Based on Multi-Output Support Vector Regression. In *Knowledge Engineering and Management*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 385–395.
4. Jin, X.; Yu, X.; Wang, X.; Bai, Y.; Su, T.; Kong, J. Prediction for Time Series with CNN and LSTM. In Proceedings of the 11th International Conference on Modelling, Identification and Control (ICMIC2019), Tianjin, China, 13–15 July 2019; Springer: Singapore, 2019; pp. 631–641.
5. Du, S.; Li, T.; Yang, Y.; Horng, S.J. Multivariate time series forecasting via attention-based encoder–decoder framework. *Neurocomputing* **2020**, *388*, 269–279. [CrossRef]
6. Pratama, I.; Permanasari, A.E.; Ardiyanto, I.; Indrayani, R. A review of missing values handling methods on time-series data. In Proceedings of the IEEE International Conference on Information Technology Systems and Innovation (ICITSI), Bandung-Bali, Indonesia, 24–27 October 2016; pp. 1–6.
7. Kotsiantis, S.; Kostoulas, A.; Lykoudis, S.; Argiriou, A.; Menagias, K. Filling missing temperature values in weather data banks. In Proceedings of the 2nd IET International Conference on Intelligent Environments, IE 06, IET, Athens, Greece, 5–6 July 2006; Volume 1, pp. 327–334.
8. Howell, D.C. The treatment of missing data. In *The SAGE Handbook of Social Science Methodology*; Outhwaite, W., Turner, S.P., Eds.; SAGE Publications Ltd.: Thousand Oaks, CA, USA, 2007; pp. 212–226. Available online: https://study.sagepub.com/sites/default/files/Howell.pdf (accessed on 15 May 2020).
9. Ghorbani, A.; Zou, J.Y. Embedding for informative missingness: Deep learning with incomplete data. In Proceedings of the 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–5 October 2018; pp. 437–445.
10. Gómez, V.; Maravall, A.; Peña, D. Missing observations in ARIMA models: Skipping approach versus additive outlier approach. *J. Econ.* **1999**, *88*, 341–363. [CrossRef]
11. Damsleth, E. Interpolating missing values in a time series. *Scand. Stat. Theory Appl.* **1980**, *7*, 33–39.
12. García-Laencina, P.J.; Sancho-Gómez, J.L.; Figueiras-Vidal, A.R. Pattern classification with missing data: A review. *Neural Comput. Appl.* **2010**, *19*, 263–282. [CrossRef]
13. Gnauck, A. Interpolation and approximation of water quality time series and process identification. *Anal. Bioanal. Chem.* **2004**, *380*, 484–492. [CrossRef]
14. Yu, H.F.; Rao, N.; Dhillon, I.S. Temporal regularized matrix factorization for high-dimensional time series prediction. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 847–855.
15. Schneider, T. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *J. Clim.* **2001**, *14*, 853–871. [CrossRef]
16. Rivero, C.R.; Pucheta, J.; Laboret, S.; Patiño, D.; Sauchelli, V. Forecasting short time series with missing data by means of energy associated to series. *Appl. Math.* **2015**, *6*, 1611–1619. [CrossRef]
17. Oehmcke, S.; Zielinski, O.; Kramer, O. KNN ensembles with penalized DTW for multivariate time series imputation. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2774–2781.

18. Amato, A.; Calabrese, M.; Di Lecce, V. Decision trees in time series reconstruction problems. In Proceedings of the 25th IEEE Instrumentation and Measurement Technology Conference, Victoria, BC, Canada, 12–15 May 2008, pp. 895–899.
19. Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; Liu, Y. Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.* **2018**, *8*, 6085. [CrossRef]
20. Strauman, A.S.; Bianchi, F.M.; Mikalsen, K.Ø.; Kampffmeyer, M.; Soguero-Ruíz, C.; Jenssen, R. Classification of postoperative surgical site infections from blood measurements with missing data using recurrent neural networks. In Proceedings of the IEEE International Conference on Biomedical & Health Informatics (BHI), Las Vegas, NV, USA, 4–7 March 2018; pp. 307–310.
21. Li, Q.; Xu, Y. VS-GRU: A variable sensitive gated recurrent neural network for multivariate time series with massive missing values. *Appl. Sci.* **2019**, *9*, 3041. [CrossRef]
22. Kim, Y.J.; Chi, M. Temporal belief memory: Imputing missing data during RNN training. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18, Stockholm, Sweden, 13–19 July 2018, pp. 2326–2332.
23. LEE, M.; AN, J.; LEE, Y. Missing-value imputation of continuous missing based on deep imputation network using correlations among multiple IoT data streams in a smart space. *IEICE Trans. Inf. Syst.* **2019**, *E102.D*, 289–298. [CrossRef]
24. Singh, B.P.; Deznabi, I.; Narasimhan, B.; Kucharski, B.; Uppaal, R.; Josyula, A.; Fiterau, M. Multi-resolution networks for flexible irregular time series modeling (Multi-FIT). *arXiv* **2019**, arXiv:1905.00125.
25. Liang, X.; Zou, T.; Guo, B.; Li, S.; Zhang, H.; Zhang, S.; Huang, H.; Chen, S.X. Assessing Beijing's PM2.5 pollution: Severity, weather impact, APEC and winter heating. *Proc. R. Soc. A* **2015**, *471*, 20150257. [CrossRef]
26. Vito, S.D.; Massera, E.; Piga, M.; Martinotto, L.; Francia, G.D. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sens. Actuators B Chem.* **2008**, *129*, 750–757. [CrossRef]
27. Zhang, S.; Guo, B.; Dong, A.; He, J.; Xu, Z.; Chen, S.X. Cautionary tales on air-quality improvement in Beijing. *Proc. R. Soc. A* **2017**, *473*, 20170457. [CrossRef]
28. Cai, X.; Zhang, N.; Venayagamoorthy, G.K.; Wunsch, D.C. Time series prediction with recurrent neural networks trained by a hybrid PSO–EA algorithm. *Neurocomputing* **2007**, *70*, 2342–2353. [CrossRef]
29. Bashir, F.; Wei, H.L. Handling missing data in multivariate time series using a vector autoregressive model-imputation (VAR-IM) algorithm. *Neurocomputing* **2018**, *276*, 23–30. [CrossRef]
30. Liu, J.; Kumar, S.; Palomar, D.P. Parameter Estimation of Heavy-Tailed AR Model With Missing Data Via Stochastic EM. *IEEE Trans. Signal. Process.* **2019**, *67*, 2159–2172. [CrossRef]
31. Guo, Z.; Wan, Y.; Ye, H. A data imputation method for multivariate time series based on generative adversarial network. *Neurocomputing* **2019**, *360*, 185–197. [CrossRef]
32. Holden, K.; Broomhead, A. An examination of vector autoregressive forecasts for the U.K. economy. *Int. J. Forecast.* **1990**, *6*, 11–23. [CrossRef]
33. Zhang, Y.; Thorburn, P.J.; Xiang, W.; Fitch, P. SSIM—A Deep Learning Approach for Recovering Missing Time Series Sensor Data. *IEEE Internet Things J.* **2019**, *6*, 6618–6628. [CrossRef]
34. Feng, T.; Narayanan, S.S. Imputing Missing Data In Large-Scale Multivariate Biomedical Wearable Recordings Using Bidirectional Recurrent Neural Networks with Temporal Activation Regularization. In Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Berlin, Germany, 23–27 July 2019; pp. 2529–2534.
35. Liu, T.; Wei, H.; Zhang, K. Wind power prediction with missing data using Gaussian process regression and multiple imputation. *Appl. Soft. Comput.* **2018**, *71*, 905–916. [CrossRef]
36. Lipton, Z.C.; Kale, D.; Wetzel, R. Directly modeling missing data in sequences with RNNs: Improved classification of clinical time series. In Proceedings of the 1st Machine Learning for Healthcare Conference, Los Angeles, CA, USA, 19–20 August 2016; pp. 253–270.
37. Yoon, J.; Zame, W.R.; van der Schaar, M. Estimating missing data in temporal data streams using multi-directional recurrent neural networks. *IEEE Trans. Biomed. Eng.* **2017**, *66*, 1477–1490. [CrossRef] [PubMed]
38. Fortuin, V.; Baranchuk, D.; Rätsch, G.; Mandt, S. GP-VAE: Deep Probabilistic Time Series Imputation. *arXiv* **2020**, arXiv:1907.04155.

39. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
40. Santos, M.S.; Pereira, R.C.; Costa, A.F.; Soares, J.P.; Santos, J.; Abreu, P.H. Generating synthetic missing data: A review by missing mechanism. *IEEE Access* **2019**, *7*, 11651–11667. [CrossRef]