Comparing the Performance of Recurrent Neural Network and Some Well-Known Statistical Methods in the Case of Missing Multivariate Time Series Data

Samira Zahmatkesh[1,2] and Philipp Zech[1,3]

[1] Department of Computer Science, University of Innsbruck, Austria
[2] Samira.zahmat-kesh@uibk.ac.at
[3] Philipp.zech@uibk.ac.at

Abstract. Missing values in multivariate time series data, often caused by network disruptions, device or power outages, and bad weather, can pose challenges for future analysis. Common statistical methods like multiple imputation and expectation-maximization are often used to impute missing time series data. However, these methods assume that the data is missing at random and may struggle with more complex missing data mechanisms and higher missing ratios. In these cases, advanced techniques like neural networks may offer improved imputation. This study assess the effectiveness of two recurrent neural network methods LSTM and GRU, enhanced with a time decay function, named LSTM-D and GRU-D, for analyzing missing multivariate time series. Their performance is compared with three well-known statistical methods: Bootstrapped-EM, EM-ARIMA, and MICE, across different missing data scenarios. Results indicate that LSTM-D and GRU-D perform better than traditional statistical methods for two different datasets, particularly when the missing data is not random.

Keywords: Time series · Missing data · Imputation · Neural networks · Statistical methods.

1 Introduction

Time series data, characterized by its temporal nature with each data point linked to a specific timestamp, enables the analysis of changes over time. Multivariate Time Series (MVTS) includes two or more variables observed at the same time steps. This data is used in various domains, such as financial analysis, healthcare, manufacturing, environmental monitoring, and transportation.

The presence of missing data poses a challenge in the analysis and utilization of time series data due to factors such as sensor failures, network issues, human errors, non-response in surveys, system upgrades, or data storage limitations [34]. Understanding the reasons for missing values is crucial due to their significant impact on future analysis. According to Rubin [41] Data with missing values can be categorized into three main groups, known as the missingness mechanisms. Missing data mechanisms include Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). In MCAR, missingness is unrelated to any data. In MAR, it depends on observed data. In MNAR, it depends on unobserved data, making it the most challenging

and complex mechanism, potentially leading to biases if not properly addressed. Handling missing data in time series analysis requires careful consideration of these mechanisms.

Common techniques for handling missing time series data include using complete case and imputation. Complete case or deletion is carried out either list-wise or pairwise [35], where mainly samples or variables are removed that are only partially observed. This method leaves gaps in the data set, possibly resulting in erroneous parameter estimations [22]. Imputation is predicting missing values based on available data. The choice of method depends on the nature of the missing data and aims. Many methods have been proposed based on statistical and Machine Learning models for imputing missing time series data.

Statistical imputation approaches have been mentioned as single imputation and Multiple Imputation (MI) methods. Single imputation is the process of filling distinctly one value for each missingness. Methods like mean/median averages [4], forward and backward imputation [45], or linear regression [53] are the most convenient single imputation methods that have been used, but there was significant bias and loss of precision. Also, the Auto-Regressive Moving Average Models (ARMA) [9] which is a combination of the basic linear processes, auto-regressive and moving average model, perform well in forecasting missing time series data. To address non-stationarity in time series data, Auto-Regressive Integrated Moving Average Models (ARIMA) are effective for imputing missing values. It is the integrated form of the ARMA that captures temporal patterns and trends for accurate data reconstruction [29, 6, 38, 2]. The ARIMA model is used when the dataset exhibit temporal pattern and is of substantial volumes. However, alteration in observation and model specification leads the model to be unstable. Multiple imputation proposed by Rubin [33] which replaces missing values by $m \geq 2$ possible values, each with a unique estimate reflecting the uncertainty attached. The m estimates are combined to yield a single estimate. A wide variety of approaches based on MI have been proposed such as Expectation-Maximization (EM) [20], Probabilistic Matrix Factorization (PMF) [36], MI by Chained Equation(MICE) [49, 54], Bayesian computational algorithm known as Markov Chain Monte Carlo (MCMC) [49]. Multiple imputation requires certain conditions to be met: data should be MAR, an appropriate imputation model should be used, and an adequate number of imputations should be performed to ensure convergence and address uncertainty.

ML methods like Bayesian network [47], support vector regression [57], and k-nearest neighbors [1], decision tree [21] have also been applied to the MVTS imputation problem. These methods are limited in covering complex temporal dependencies between observations [43]. Recently, various deep learning (DL) based methods have been introduced that are not only computationally feasible but also capable of addressing the complex missing data patterns in MVTS. The Recurrent Neural Networks (RNN) like Gated Recurrent Unit (GRU) [16, 14] and Long Short Term Memory (LSTM) [26, 31] have the ability to represent temporal dependencies in sequential data and have been widely considered in analyzing MVTS. Moreover, they can capture long-range dependencies in

time series data in an effective manner. Che et al., [14] proposed GRU-Decay (GRU-D) to handle missing values in medical data for classification purposes. A bidirectional RNN structure, which considers the input sequence in both forward and backward directions, based on LSTM [23] was presented by Cao et al. instead of the GRU-D to improve training accuracy [12]. Besides bi-directional RNN, Yoon et al., introduced the multi-directional RNN [58], which performs imputation across data streams. These models primarily address missingness in classification rather than prediction tasks. In multivariate time series, each variable may have different missingness ratios. Continuous and prolonged gaps, caused by sensor or equipment failures, pose challenges, leading to substantial information loss in consecutively correlated data. Therefore, a method capable of accurately reconstructing and analyzing such data is necessary. This study endeavors to analyze carefully the efficacy of two RNN-based models for imputation and prediction purposes. Specifically, we explore the performance of the GRU-D model proposed by Che et al. [14], and its extension to LSTM as LSTM-D. To implement these RNNs, we have used Python[4]. We have extended the code for the LSTM-D. Our objective is to compare their reconstruction capabilities for MVTS against three established statistical methods: the Bootstrapped-EM algorithm, the EM-ARIMA model, and multiple imputation using MICE, where we used reference implementations in R to conduct our experiments [59, 50, 56]. To better investigate our goal, we artificially create MAR, MNAR, and Long Gaps in the data. In the two latter cases, the missing data rate is deliberately high. In this paper, our primary objectives are outlined as follows:

- Assessing the performance the GRU-D and its extension to LSTM, known as LSTM-D in reconstructing MVTS under various rates and mechanisms of missingness.
- Assessing of three widely recognized statistical methods: Bootstrapped-EM, EM-ARIMA model, and MICE in imputation missing MVTS under various missing data scenarios.
- Conducting a comprehensive comparison between the performance of the RNNs and the selected statistical methods.

The rest of the paper is organized as follows. Section 2 gives an overview of RNN methods and details about LSTM-D as an extension of GRU-D. In Section 3 we present a brief introduction of the considered statistical methods. In Section 4 We implement the methods on air quality data and wind-turbin data. Finally, the conclusion and future works will be discussed in Section 5.

2 Recurrent Neural Networks for Multivariate Time Series Imputation

Recurrent neural networks, particularly LSTM [23] and GRU [16, 17] architectures, are instrumental in multivariate time series prediction. LSTM and GRU

[4] The contributed source code link is "https://github.com/samirazahmat-kesh/GRU-D.git"

are ideal for capturing dependencies in time-ordered data due to their memory cells and gating units, which address the vanishing gradient problem in traditional RNNs. These mechanisms enable selective retention and forgetting of information, crucial for long-term dependencies in time series analysis. They handle variable-length sequences and multivariate inputs effectively, making them versatile for various time series data. The choice between LSTM and GRU depends on model complexity and dataset characteristics. In essence, both excel in modeling sequential dependencies and predicting multivariate time series data.

When using RNNs for prediction and dealing with missing values, a straightforward approach involves replacing missing observations with the variable mean or assuming missing values are the same as their last measurement (forward imputation). However, these methods can result in loss of variability in the time series, assuming missing values have the same variability as observed ones, which may not be true, particularly with abrupt changes or fluctuations. Moreover, sensitivity to outliers, assuming constant time intervals, and artificially increasing variable correlation can lead to biased estimates. GRU-D, a variant of the GRU with a decay mechanism, was introduced by Che et al. [14] for clinical applications with a primary focus on classification tasks. The innovative GRU-D model improves sequential data analysis in medical contexts by integrating decay mechanisms into the standard GRU architecture. It demonstrates proficiency in handling missing data in clinical applications. This integration enhances sequence modeling and the model's robustness in managing incomplete or missing information in real-world medical datasets. An extension of GRU-D to LSTM, termed LSTM-D, is presented in this paper.

LSTM unit is made of the memory cell, which stores information over periods. The cell is controlled by three gates: the input gate ($\mathbf{i}$), the forget gate ($\mathbf{f}$), and the output gate ($\mathbf{o}$). These gates are responsible for regulating the flow of information into, out of, and within the memory cell. The input gate determines how much of the new information should be stored in the cell, the forget gate controls the extent to which the existing information is retained, and the output gate manages the information to be outputted to the next time step. By carefully adjusting these gates, LSTM can capture and preserve relevant information over long sequences, making it a powerful tool for tasks such as natural language processing, speech recognition, and time-series analysis. An LSTM unit also has a hidden state represented by $\mathbf{h}_{t-1}$ and $\mathbf{h}_t$ for the hidden state of the previous time stamp and the current timestamp, respectively. And has a cell state represented by $\mathbf{c}_{t-1}$ and $\mathbf{c}_t$ for the previous and current timestamps, respectively. Hidden state is known as short term memory, and the cell state is known as Long term memory.

We denote a multivariate time series with D variables observed at T times $X = (\mathbf{x}_1, ..., \mathbf{x}_T)^T \in R^{T \times D}$, thus x_t^d is the sample for d-th variable at time t, and $\mathbf{x}_t = (x_t^1, ..., x_t^D)$ is the measurement vector at time t for all the variables, where $t = 1, \ldots, T$ and $d = 1, \ldots, D$. The equations for LSTM gates are

$$
\begin{aligned}
\mathbf{f}_t &= \sigma(W_f \mathbf{h}_{t-1} + R_f \mathbf{x}_t + b_f) \\
\mathbf{i}_t &= \sigma(W_i \mathbf{h}_{t-1} + R_i \mathbf{x}_t + b_i) \\
\mathbf{o}_t &= \sigma(W_o \mathbf{h}_{t-1} + R_o \mathbf{x}_t + b_o) \\
\tilde{\mathbf{c}}_t &= \phi(W_c \mathbf{h}_{t-1} + R_c \mathbf{x}_t + b) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot \phi(\mathbf{c}_t)
\end{aligned}
$$

where σ and ϕ stand for activation functions which often are considered sigmid and tanh, respectively. W_f, W_f, W_f, R_f, R_i and R_o are the relevant weight matrices, b_f, b_i, b_o and b are the bias vectors, and all of them are as the model parameters. The cell architecture for LSTM is shown in Fig 1(a). To effectively address missing values in MVTS, LSTM-D employs a dynamic decay mechanism, adept at capturing temporal dependencies and mitigating the impact of incomplete data. To denote which measurement is missing or observed, we define a masking matrix M with the same dimension of X with the elements m_t^d as:

$$
m_t^d = \begin{cases} 1 & \mathrm{x}_t^d \text{ is observed} \\ 0 & \text{otherwise} \end{cases} \tag{1}
$$

To track missing values for each variable in X, the last time interval is maintained in a matrix $\Delta \in R^{T \times D}$ with elements $\delta_t^d \in R$ as

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

where s_t are the time stamps relative to each measurement. A vector of decay rates, γ is defined as

$$
\gamma_t = \exp\{-\max(0, W_\gamma \delta_t + b_\gamma)\}, \tag{3}
$$

W_γ and b_γ are also the model parameters. This decay rate will be applied to the input as γ_x and to the hidden state as γ_h. Thus the input valus are updated as

$$
x_t^d \leftarrow m_t^d x_t^d + (1 - m_t^d)\gamma_{xt}^d x_{t'}^d + (1 - m_t^d)(1 - \gamma_{xt}^d)\widetilde{x}^d \tag{4}
$$

where $x_{t'}^d$ is the last observed value and the $\widetilde{x}^d$ is the empirical mean of d-th variable. Also, to capture better the information of missingness, the previous hidden state $\mathbf{h}_{t-1}$ is decayed before computing the new hidden state h_t as

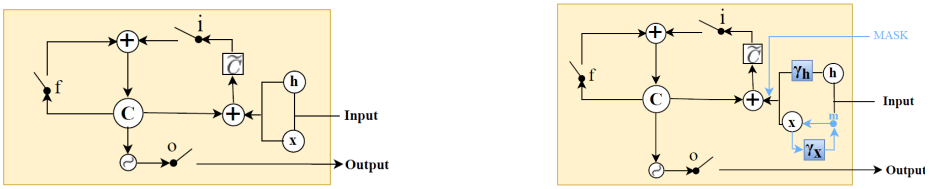
$$
\mathbf{h}_{t-1} \leftarrow \mathbf{h}_{t-1} \odot \gamma_{ht}. \tag{5}
$$

Moreover the masking vectors ($\mathbf{m}_t$) are fed directly into the model. Thus, the modification of LSTM with decay mechanism as LSTM-D will be with the fol-

lowing equations over the cell memory and all the gates:

$$
\begin{aligned}
\mathbf{f}_t &= \sigma(W_f \mathbf{h}_{t-1} + R_f \mathbf{x}_t + V_f \mathbf{m}_t + b_f) \\
\mathbf{i}_t &= \sigma(W_i \mathbf{h}_{t-1} + R_i \mathbf{x}_t + V_i \mathbf{m}_t + b_i) \\
\mathbf{o}_t &= \sigma(W_o \mathbf{h}_{t-1} + R_o \mathbf{x}_t + V_o \mathbf{m}_t + b_o) \\
\tilde{\mathbf{c}}_t &= \phi(W_c \mathbf{h}_{t-1} + R_c \mathbf{x}_t + V \mathbf{m}_t + b) \\
\mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot \phi(\mathbf{c}_t)
\end{aligned}
$$

where V_f, V_i and V_o are the new added parameters. Fig 1(b).



(a) (b)

Fig. 1: The architecture of LSTM and LSTM-D cell

3 Well-known Statistical Methods for Multivariate Time Series Imputation

Various methods for multivariate time series imputation have been introduced, including regression-based, auto-regressive, and Bayesian approaches. These methods provide robust solutions for handling missing data in complex temporal datasets by leveraging statistical relationships and dependencies among variables. In our study, we focus on three effective methods: Bootstrap-based EM, EM-ARIMA, and MICE, which we briefly introduce here.

3.1 Bootstrap-Based EM

In general, the Bootstrap-based EM algorithm draws m (the number of imputation datasets) samples of size n (the size of the original dataset) from original datasets, and point estimates of mean and variance are performed in each sample using the EM method. Let $X = (\mathbf{x}_1, ..., \mathbf{x}_D)^T \in R^{n \times D}$ be the data matrix with observed part X^{obs} and unobserved part X^{mis} and has a multivariate normal distribution as $X \sim \mathrm{N}_D(\mu, \Sigma)$. Also, It is assumed data are MAR. Similar to Section 2, Let M be the missingness mask matrix. Under MAR assumption

$p(M|X) = P(M|X^{obs})$, then the joint distribution of Observed data and mask can be broken up as

$$p(X^{obs}, M|\theta) = p(M|X^{obs})p(X^{obs}|\theta)$$

As the inferences are based on the complete data parameters, the likelihood can be written as

$$L(\theta|X^{obs}) \propto p(X^{obs}|\theta)$$

which can be rewritten using the law of iterated expectations [42] as

$$L(\theta|X^{obs}) \propto p(X^{obs}|\theta) = \int p(X|\theta)dX^{mis}.$$

The main computational difficulty in the analysis of incomplete data is taking draws from this posterior. The EM algorithm [20] is a simple computational approach to finding the mode of the posterior. Bootstrap-based EM algorithm combines the classic EM algorithm with a bootstrap approach to take draws from this posterior. For each draw, the data are bootstrapped to simulate estimation uncertainty and then the EM algorithm is run to find the mode of the posterior for the bootstrapped data, which gives fundamental uncertainty too [24]. After drawing the posterior of the complete-data parameters, imputations take place by drawing values of X^{mis} from its distribution conditional on X^{obs} and the draws of θ. Remember there are m sets of estimates. Then each set of estimates is used to impute the missing observations from original dataset. The result is m sets of imputed data that can be used for subsequent analyses.

3.2 EM-ARIMA

As the previous section, let $\mathbf{x}_t$ be the tth realization of a D-variates time series. If l components of $\mathbf{x}_t$ are unobserved, then it can be rearranged and divided in two missing and observed parts, which are denoted by $\mathbf{x}_t = (\mathbf{x}_{tm}, \mathbf{x}_{to})$. Also It is assumed X is of multivariate normal distribution with mean μ and covariance matrix Σ. Then they also can be partitioned as following:

$$\tilde{\mu}_t = \begin{bmatrix} \tilde{\mu}_{tm} & \tilde{\mu}_{to} \end{bmatrix} \qquad \text{and} \qquad \tilde{\Sigma}_t = \begin{bmatrix} \tilde{\Sigma}_{t(mm)} & \tilde{\Sigma}_{t(mo)} \\ \tilde{\Sigma}_{t(om)} & \tilde{\Sigma}_{t(oo)} \end{bmatrix}.$$

The method is proposed by Junger and Leon [25], and is a modification of EM algorithm [20]. In the imputation algoritm, first the missing values are replaced by some initial estimates and the the parameters μ and Σ are estimated. Then, the level for each of the uni-variate time series is estimated by ARIMA model, and finally re-estimate the missing values using updated estimates of the parameters and the level of the time series. These steps are iterated until some convergence criterion is reached. At the $(k+1)$-th iteration of the E (estimation)

step of the EM algorithm, the missing values are imputed with conditional means given the observed values and the previous estimates of the parameters given by

$$\begin{aligned}\tilde{\mathbf{x}}_{tm}^{(k+1)} &= \mathrm{E}\left[\mathbf{x}_{tm}|\ \mathbf{x}_{to}, \tilde{\mu}_t^{(k)}, \tilde{\Sigma}_t^{(k)}\right] \\ &= \tilde{\mu}_{tm}^{(k)} + \tilde{\Sigma}_{t(mo)}^{(k)} \tilde{\Sigma}_{t(oo)}^{(k)} \left(\mathbf{x}_{to} + \tilde{\mu}_{to}^{(k)}\right).\end{aligned}$$

In the M (maximization) step, the revised maximum likelihood estimates of μ_t and Σ_t are computed. As we mentioned, The temporal contribution to the level of each time series μ_t is independently estimated using ARIMA model [7].

3.3 MICE

A popular approach for implementing multiple imputation is sequential regression modeling, also called multiple imputation by chained equations (MICE) [39, 11]. The basic idea in MICE is imputing missing values in one variable from a regression of the observed elements of it condition on the other variables. Let $X = (\mathbf{x}_1, ..., \mathbf{x}_D)$ be the data matrix consists of D variables. Each variable may be partially observed so we divide the vector of each variable into two parts: observed (obs) and missing (mis) as $\mathbf{x}_d = (\mathbf{x}_d^{obs}, \mathbf{x}_d^{mis})$. The imputation problem is to draw the unconditional multivariate distribution of X from $P(X)$. Let n denote an iteration counter, one may repeat the following sequence of Gibbs sampler iterations:

$$\begin{aligned}&\text{For } \mathbf{x}_1: \quad \text{draw } \mathbf{x}_1^{n+1} \text{ from } P(\mathbf{x}_1|\mathbf{x}_2^n, \mathbf{x}_3^n, ..., \mathbf{x}_D^n) \\ &\text{For } \mathbf{x}_2: \quad \text{draw } \mathbf{x}_2^{n+1} \text{ from } P(\mathbf{x}_2|\mathbf{x}_1^{n+1}, \mathbf{x}_3^n, ..., \mathbf{x}_D^n) \\ &\vdots \\ &\text{For } \mathbf{x}_D: \quad \text{draw } \mathbf{x}_D^{n+1} \text{ from } P(\mathbf{x}_D|\mathbf{x}_1^{n+1}, \mathbf{x}_2^{n+1}, ..., \mathbf{x}_{D-1}^{n+1})\end{aligned}$$

where means condition each time on the most recently drawn values of all other variables the candidate variable is imputed. This can be performed for l time until convergence. All the procedure will be repeated m times, yielding m imputed sets which then the best imputed set can be selected based on an appropriate criterion. Using $l = 10$ typically yields satisfactory results. It is standard to use generalized linear models as the basis of the predictive draws, but other kind of models also have been applied. the details about specification of the imputation model or monitoring convergence can be found in references [11, 10].

4 Results and Discussion

In this section, we first introduce the data used in the experiment and explain how to implement different scenarios of missingness in them. Then the obtained results in imputation missing data with two RNNs and three statistical methods are presented. Finally, we have a comprehensive discussion about the results and the future works.

4.1 Experiments

In this study, we employed two distinct datasets: air quality data and wind-turbin data. Air quality data [52] is used as the basis for our analyses, while the wind-turbin data (from kaggle.com), was utilized to independently confirm and ensure the robustness of our findings, enhancing the reliability and generalizability of our research.

The air quality data set contains 8784 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multi-sensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city. Data were recorded from March 2004 to March 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor responses. The dataset consists of the hourly averaged of ten air quality variables. In this study, we have chosen five of them for our prediction and imputation purpose, which are: PT08.S1 (tin oxide, which is nominally CO targeted) PT08.S2 (titania, which is nominally NMHC[5] targeted), PT08.S3 (tungsten oxide, which is nominally NOx targeted), PT08.S4 (tungsten oxide, which is nominally NO2 targeted) and PT08.S5 (indium oxide, which is nominally O3 targeted). As the amount of missing values are limited, we have imposed artificially two missing data mechanisms MAR and MNAR, and also, we have created long gaps during the data so that every variable is missing for consecutive days of a month or more. In order to create MAR in each variable with specific missing rate p, we generated a random binary matrix M as Eq (1) with probability of being zero equal to p and probability of being one equal to 1-p for each m_t^d. Then we replace x_t^d with NA wherever $m_t^d = 0$. To creat MNAR in the data with different missing rates, we considered the median of each variables in X and if x_t^d be the observation of dth variable at time t then:

$$x_t^d = \begin{cases} Missing & \text{if } x_t^d > a + c * median \\ Observed & \text{otherwise} \end{cases} \tag{6}$$

We considered (a,c) equal to (-0.1,1), (0,0.9), (-0.1,1), (0,1) and (-0.1,1) for variables CO, NMHC, NO_x, NO_2 and O_3 respectively. Furthermore, to insert long gaps in the data, we chose randomly some periods for each variable and replaced the data with NAs. In the event of Long Gaps, some points are also MAR inherently. In Tbl 1 the rates of missing data in each variable under the specific type of missingness are reported.

The second dataset, wind-turbin data, consists of measurements of 19 variables from Jan 2018 to Mar 2020 of a wind turbine, at a 10 minute frequency. Here, we use data from Jan 2018 to Dec 2018, with a total 52704 observation. We use four variables "Active Power","Ambient temperature","Wind direction" and "Wind Speed" as input features. There are missing variables in all variables, but in order to fulfill the assumption of MNAR with higher missing rates,

[5] Non-methane Hydrocarbons

we manipulated the data similar to air quality data and created MNAR with missing rates 57, 58,79 and 67 percent respectively in "Active Power","Ambient temperature","Wind direction" and "Wind Speed" variables.

Variable	Missing Rates(%)	Missing type
CO	40	
NMHC	57.7	
NO_x	36.5	MAR
NO_2	14	
O_3	28	
CO	70.5	
NMHC	63.2	
NO_x	64	MNAR
NO_2	53.2	
O_3	60.2	
CO	49.6	
NMHC	64.5	
NO_x	42.5	Long Gaps
NO_2	34	
O_3	41.3	

Table 1: Missing rates of air-quality dataset variables for different kind of missingness.

It is worth to note that we normalize the time series data in each dataset before using it in each method. We do this by using max-min normalization, which scales all the variables to a range between 0 and 1. The parameters for LSTM-D and GRU-D model are set as: learning rate=0.01, batch size=64, optimizer=Adam, Epoch=1000 and time step=10. The EM-bootstrapping algorithm has been implemented via a multiple imputation with m=2000 iterations and ultimately the imputation with least error has been selected. We have considered EM-ARIMA model with several Parameter values for p,d and q, eventually ARIMA(1,1,1) was selected and reported here. The performance of each method is reported in aspect of the averaged mean absolute error (MAE) and averaged mean square error (MSE) in Tbl 2, where each one is calculated as follow:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|,$$

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

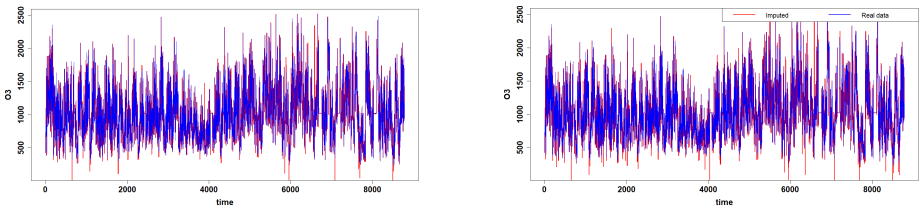
4.2 Results

In MAR case the accuracies of all the methods are very close to each other and near zero, so that they cannot be distinguished and all of them are doing well. In MNAR case, the results show the better performance of LSTM-D and GRU-D, with a significant difference in the MAE and MSE values compared to statistical methods. In Long Gaps Case, the two RNN methods performs again better. Of course, it cannot be concluded that this superiority is very high compared to statistical methods. In this case, among statistical methods, EM-ARIMA model

has better performance with lower MSE and MAE, and is doing as well as LSTM-D and GRU-D. Therefore, according to what we checked, considered RNNs are more capable of handling non-random missingness and long gaps than the well-known statistical methods, although this issue requires further investigation in other applications and in the case of other patterns of missingness. Since, the results for all the variables are equivalent, we have chosen "O3" and plot the real data and imputed data together in Fig 2 and 3. Obviously, the imputed data is almost identical with real data for LSTM-D and GRU-D methods, but for the three statistical methods especially with EM-Bootstrap and MICE, in some periods compliance has not been achieved.

Method	Train MSE	Test MSE	Test MAE	Missing Type
GRU-D	0.029	0.030	0.139	
LSTM-D	0.048	0.048	0.187	
EM-Bootstrap	-	0.161	0.184	MAR
EM-ARIMA	-	0.030	0.074	
MICE	-	0.163	0.186	
GRU-D	0.071	0.067	0.176	
LSTM-D	0.058	0.060	0.152	
EM-Bootstrap	-	1.605	1.033	MNAR
EM-ARIMA	-	1.642	1.005	
MICE	-	1.690	1.047	
GRU-D	0.033	0.033	0.137	
LSTM-D	0.024	0.025	0.118	
EM-Bootstrap	-	0.256	0.267	Long Gaps
EM-ARIMA	-	0.081	0.184	
MICE	-	0.279	0.282	

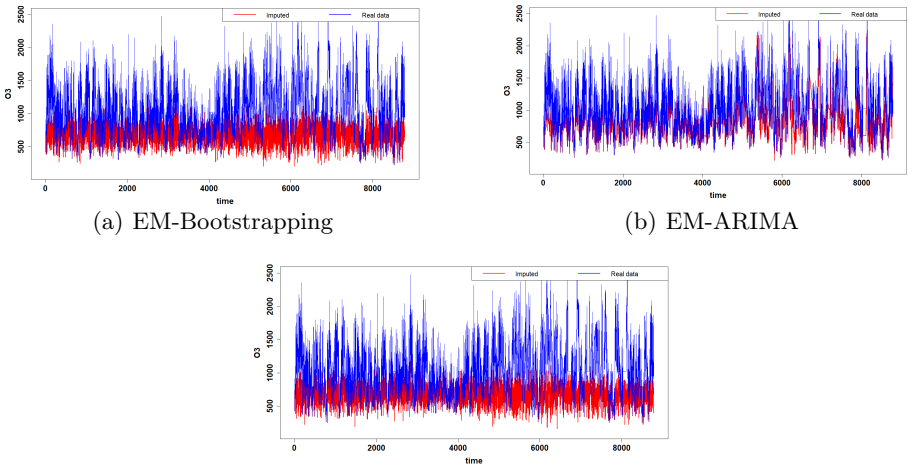
Table 2: Result of different methods in reconstruction air-quality data-set variables, under three type of missingness assumption



(a) LSTM-D (b) GRU-D

Fig. 2: Reconstruction (prediction) of the variable O3 in MNAR case with LSTM-D and GRU-D.

Furthermore, the MAE and MSE of different methods for reconstruction Wind-Turbin data variables are reported in Tbl 3. The two RNNs performs almost similarly, and among statistical methods EM-ARIMA performs better. It can be seen that the error of the two RNNs is much less than the statistical methods.



(a) EM-Bootstrapping

(b) EM-ARIMA

(c) MICE

Fig. 3: Reconstruction (prediction) of the variable O3 in MNAR case with statistical methods: EM-ARIMA, EM-Bootstrap and MICE methods.

Method	Train MSE	Test MSE	Test MAE
GRU-D	0.028	0.029	0.077
LSTM-D	0.029	0.030	0.080
EM-Bootstrap	-	2.174	1.179
EM-ARIMA	-	1.403	0.939
MICE	-	2.334	1.229

Table 3: Result of different methods for reconstruction Wind-Turbin data variables

4.3 Discussion

In this paper we compared the performance of two modified RNNs (LSTM-D and GRU-D) and three statistical methods for imputing missing multivariate time series (MVTS) data. The study focused on different missingness scenarios: Missing At Random (MAR), Missing Not At Random (MNAR), and Long Gaps, which were artificially applied to the data. In tests with air quality data, the RNNs generally performed better than statistical methods, particularly as the missing ratio increased, with MNAR showing significant improvement. Then, to ensure the stability of the result, we experimented on the wind turbine data and the results confirmed the superior performance of the RNNs as well.

Many time series data exhibit both temporal and spatial dependencies, known as spatio-temporal data. Traditional methods like KNN-based methods [15, 48], EM [20], singular value decomposition [46], random forest [8] often treat spatial and temporal dimensions separately. Oversimplified methods can lead to suboptimal results, especially with complex, dynamic scenarios and significant missing values in spatio-temporal data. Various statistical and ML/DL approaches

have been proposed to address these challenges by leveraging spatial and temporal correlations to improve missing data for classification or prediction [5, 51, 28, 44, 18, 32, 19, 30].

DL methods can generate accurate predictions by leveraging complex missing data mechanisms and dependencies but have flaws [3]. Despite their power, DL methods have limitations in statistical modeling for spatial and spatio-temporal data due to substantial uncertainties caused by inherent variability such as measurement errors or natural fluctuations, data gaps, mismatched prediction supports, and sampling. They also do not directly estimate prediction or classification errors and struggle to incorporate known mechanistic relationships in spatio-temporal data.

That is problematic when one is attempting to use the output from these models to make decisions, it is challenging as it is not obvious how much reliable the output is. Also, a more troubling issue is that most DL methods are "black boxes" and it is difficult to know why they are producing the prediction or classification that they do, meaning that their internal workings are not easily interpretable or explainable by humans [40].

In opposite, in traditional statistical models, such as linear regression, it is relatively easy to interpret the impact of each variable on the model's output. One can understand how changes in input variables relate to changes in the output. However, the main issue with these methods is that there may be cases where they are not sufficient to capture the broad nature of spatial non-linearity and complex features inside data. In some cases, data may exhibit complex multivariate features, non-Gaussianity, non-stationarity, and MNAR or any complex missingness patterns. Considering these cases, the spatio-temporal estimation of continuous variables could be a challenging task. DL models allow one to identify the patterns in the complex datasets and to make estimations/predictions based on them. The key feature of the DL models is that they learn from the data, and they do not require rigid statistical assumptions (such as stationarity and linearity) [55].

In order to mitigate the shortcomings of each of the DL-based and statistical-based methods, there has been an increasing number of works in recent years, that take a hybrid approach for analyzing different kinds of data where spatio-temporal data were not exempted, for example: [27, 37, 13]. These hybrid models borrow some of the effective ideas from each DL methods and statistical models in order to facilitate modeling the data, produce uncertainties for model outputs and enhance the interpretability of DL methods. Thus, they strike a balance between predictive accuracy and the ability to understand and trust the model's decisions. Despite promising efforts, our investigation reveals a significant gap in the availability of a hybrid framework that ensures high prediction/imputation accuracy, model-based uncertainty quantification, and explainability for missing spatio-temporal data across various applications. As a hypothesis, the fusion of DL methods with classical statistical models may offer an attractive way forward. Since this issue has not been addressed in the literature, working on developing a novel hybrid method in the case of complex missingness patterns

and mechanisms in spatio-temporal data is in our future work plan. We intend to present a comprehensive and hybrid model that in one hand uses the advantages of statistical methods to model the missing process and the measurement process together so that it is possible to explain the relationships between variables and missingness processes altogether in a spatio-temporal model framework with the ability of uncertainty quantification, and on the other hand, by using a suitable neural network, the complexities in the spatio-temporal data can be modeled with the aim of prediction/imputation with higher accuracy.

Conclusion

This study demonstrates that GRU-D and LSTM-D, modified RNN methods with a decay mechanism, outperform traditional statistical methods (Bootstrapped-EM, EM-ARIMA, and MICE) in reconstructing missing multivariate time series data, especially under high rates of missingness with MNAR or long gaps. Statistical methods have been used over the years and can produce interpretable results and quantify uncertainties, while deep learning methods do not have this ability due to being "black boxes" [3]. Our study paves the way for our future research, aiming to develop a hybrid framework combining deep learning and statistical methods. This approach will seek to enhance the neural networks-based method's effectiveness in handling missing data with complex correlations, including spatio-temporal data.

References

1. Acuna, E., Rodriguez, C.: The treatment of missing values and its effect on classifier accuracy. In: Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, 15–18 July 2004. pp. 639–647. Springer (2004)
2. Afrifa-Yamoah, E., Mueller, U.A., Taylor, S., Fisher, A.: Missing data imputation of high-resolution temporal climate time series data. Meteorological Applications **27**(1), e1873 (2020)
3. Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M., Farhan, L.: Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. Journal of big Data **8**, 1–74 (2021)
4. Aydilek, I.B., Arslan, A.: A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences **233**, 25–35 (2013)
5. Beckers, J.M., Rixen, M.: Eof calculations and data filling from incomplete oceanographic datasets. Journal of Atmospheric and oceanic technology **20**(12), 1839–1856 (2003)
6. Box, G.E., Jenkins, G.M., Reinsel, G.C.: Time series analysis: forecasting and control, vol. 734. John Wiley & Sons (2011)
7. Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: Time series analysis: forecasting and control. John Wiley & Sons (2015)
8. Breiman, L.: Random forests. Machine learning **45**, 5–32 (2001)
9. Broersen, P.M., Bos, R.: Time-series analysis if data are randomly missing. IEEE transactions on instrumentation and measurement **55**(1), 79–84 (2006)
10. Burgette, L.F., Reiter, J.P.: Multiple imputation for missing data via sequential regression trees. American journal of epidemiology **172**(9), 1070–1076 (2010)

11. Buuren, S.V., Oudshoorn, K.: Flexible multivariate imputation by mice (1999)
12. Cao, W., Wang, D., Li, J., Zhou, H., Li, L., Li, Y.: Brits: Bidirectional recurrent imputation for time series. Advances in neural information processing systems **31** (2018)
13. Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., Kashinath, K.: Towards physics-inspired data-driven weather forecasting: integrating data assimilation with a deep spatial-transformer-based u-net in a case study with era5. Geoscientific Model Development **15**(5), 2221–2237 (2022)
14. Che, Z., Purushotham, S., Cho, K., Sontag, D., Liu, Y.: Recurrent neural networks for multivariate time series with missing values. Scientific reports **8**(1), 6085 (2018)
15. Chen, J., Shao, J.: Nearest neighbor imputation for survey data. Journal of official statistics **16**(2), 113 (2000)
16. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
17. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
18. Ciampi, A., Appice, A., Guccione, P., Malerba, D.: Integrating trend clusters for spatio-temporal interpolation of missing sensor data. In: Web and Wireless Geographical Information Systems: 11th International Symposium, W2GIS 2012, Naples, Italy, April 12-13, 2012. Proceedings 11. pp. 203–220. Springer (2012)
19. Cui, Z., Lin, L., Pu, Z., Wang, Y.: Graph markov network for traffic forecasting with missing data. Transportation Research Part C: Emerging Technologies **117**, 102671 (2020)
20. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)
21. Fouladgar, N., Främling, K.: A novel lstm for multivariate time series with massive missingness. Sensors **20**(10), 2832 (2020)
22. Graham, J.W.: Missing data analysis: Making it work in the real world. Annual review of psychology **60**, 549–576 (2009)
23. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
24. Honaker, J., King, G.: What to do about missing values in time-series cross-section data. American journal of political science **54**(2), 561–581 (2010)
25. Junger, W., De Leon, A.P.: Imputation of missing data in time series for air pollutants. Atmospheric Environment **102**, 96–104 (2015)
26. Kim, Y.J., Chi, M.: Temporal belief memory: Imputing missing data during rnn training. In: In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-2018) (2018)
27. Kirkwood, C., Economou, T., Pugeault, N.: Bayesian deep learning for mapping via auxiliary information: a new era for geostatistics? arXiv preprint arXiv:2008.07320 (2020)
28. Kondrashov, D., Ghil, M.: Spatio-temporal filling of missing points in geophysical data sets. Nonlinear Processes in Geophysics **13**(2), 151–159 (2006)
29. Layanun, V., Suksamosorn, S., Songsiri, J.: Missing-data imputation for solar irradiance forecasting in thailand. In: 2017 56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE). pp. 1234–1239. IEEE (2017)
30. Le, T.T., Le Nguyen, P., Binh, H.T.T., Akerkar, R., Ji, Y., et al.: Gcrint: network traffic imputation using graph convolutional recurrent neural network. In: ICC 2021-IEEE International Conference on Communications. pp. 1–6. IEEE (2021)
31. 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 TRANSACTIONS on Information and Systems **102**(2), 289–298 (2019)
32. Li, Y., Parker, L.E.: Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks. Information Fusion **15**, 64–79 (2014)
33. Little, R., Rubin, D.: Multiple imputation for nonresponse in surveys. John Wiley & Sons, Inc.. doi **10**, 9780470316696 (1987)
34. Little, R.J., Rubin, D.B.: Statistical analysis with missing data, vol. 793. John Wiley & Sons (2019)
35. McKnight, P.E., McKnight, K.M., Sidani, S., Figueredo, A.J.: Missing data: A gentle introduction. Guilford Press (2007)

36. Mnih, A., Salakhutdinov, R.R.: Probabilistic matrix factorization. Advances in neural information processing systems **20** (2007)
37. Mohan, A.T., Lubbers, N., Livescu, D., Chertkov, M.: Embedding hard physical constraints in convolutional neural networks for 3d turbulence. In: ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations. vol. 520 (2020)
38. Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., Stork, J.: Comparison of different methods for univariate time series imputation in r. arXiv preprint arXiv:1510.03924 (2015)
39. Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P., et al.: A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey methodology **27**(1), 85–96 (2001)
40. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, f.: Deep learning and process understanding for data-driven earth system science. Nature **566**(7743), 195–204 (2019)
41. Rubin, D.B.: Inference and missing data. Biometrika **63**(3), 581–592 (1976)
42. Schervish, M.J., DeGroot, M.H.: Probability and statistics, vol. 563. Pearson Education London, UK: (2014)
43. Siami-Namini, S., Tavakoli, N., Namin, A.S.: A comparative analysis of forecasting financial time series using arima, lstm, and bilstm. arXiv preprint arXiv:1911.09512 (2019)
44. Song, C., Yang, X., Shi, X., Bo, Y., Wang, J.: Estimating missing values in china's official socioeconomic statistics using progressive spatiotemporal bayesian hierarchical modeling. Scientific reports **8**(1), 10055 (2018)
45. Song, S., Li, C., Zhang, X.: Turn waste into wealth: On simultaneous clustering and cleaning over dirty data. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1115–1124 (2015)
46. Stewart, G.W.: On the early history of the singular value decomposition. SIAM review **35**(4), 551–566 (1993)
47. Susanti, S.P., Azizah, F.N.: Imputation of missing value using dynamic bayesian network for multivariate time series data. In: 2017 International Conference on Data and Software Engineering (ICoDSE). pp. 1–5. IEEE (2017)
48. Tak, S., Woo, S., Yeo, H.: Data-driven imputation method for traffic data in sectional units of road links. IEEE Transactions on Intelligent Transportation Systems **17**(6), 1762–1771 (2016)
49. Takahashi, M.: Statistical inference in missing data by mcmc and non-mcmc multiple imputation algorithms: Assessing the effects of between-imputation iterations. Data Science Journal **16**, 37–37 (2017)
50. Van Buuren, S., Groothuis-Oudshoorn, K.: mice: Multivariate imputation by chained equations in r. Journal of statistical software **45**, 1–67 (2011)
51. Venturi, D., Karniadakis, G.E.: Gappy data and reconstruction procedures for flow past a cylinder. Journal of Fluid Mechanics **519**, 315–336 (2004)
52. Vito, S.: Air Quality. UCI Machine Learning Repository (2016), DOI: https://doi.org/10.24432/C59K5F
53. Von Hippel, P.T.: 4. regression with missing ys: an improved strategy for analyzing multiply imputed data. Sociological Methodology **37**(1), 83–117 (2007)
54. White, I.R., Royston, P., Wood, A.M.: Multiple imputation using chained equations: issues and guidance for practice. Statistics in medicine **30**(4), 377–399 (2011)
55. Wikle, C.K., Zammit-Mangion, A.: Statistical deep learning for spatial and spatiotemporal data. arXiv preprint arXiv:2206.02218 (2022)
56. Wu, R., Hamshaw, S.D., Yang, L., Kincaid, D.W., Etheridge, R., Ghasemkhani, A.: Data imputation for multivariate time series sensor data with large gaps of missing data. IEEE Sensors Journal **22**(11), 10671–10683 (2022)
57. Wu, S.F., Chang, C.Y., Lee, S.J.: Time series forecasting with missing values. In: 2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom). pp. 151–156. IEEE (2015)
58. Yoon, J., Zame, W.R., van der Schaar, M.: Estimating missing data in temporal data streams using multi-directional recurrent neural networks. IEEE Transactions on Biomedical Engineering **66**(5), 1477–1490 (2018)
59. Zhang, Z.: Multiple imputation for time series data with amelia package. Annals of translational medicine **4**(3) (2016)