2022

Lagged-kNN Based Data Imputation Approach for Multi-Stream Building Systems Data

Ojas Pradhan

David Hälleberg

Zhelun Chen

Jin Wen

Teresa Wu

See next page for additional authors

Lagged-kNN based data imputation approach for multi-stream building systems data

Ojas PRADHAN[1]*, David HÄLLEBERG[2], Zhelun CHEN[3], Jin WEN[4], Teresa WU[5], K. Selcuk CANDAN[6], Zheng O'NEILL[7]

[1] Drexel University, Department of Civil, Architectural and Environmental Engineering, Philadelphia, PA, USA
omp28@drexel.edu

[2] KTH Royal Institute of Technology, Department of Civil and Architectural Engineering, Stockholm, Sweden
halleb@kth.se

[3]Drexel University, Department of Civil, Architectural and Environmental Engineering, Philadelphia, PA, USA
zc98@drexel.edu

[4]Drexel University, Department of Civil, Architectural and Environmental Engineering, Philadelphia, PA, USA
jw325@drexel.edu

[5]Arizona State University, School of Computing, Informatics, Decision Systems Engineering, Tempe, AZ, USA
teresa.wu@asu.edu

[6]Arizona State University, School of Computing, Informatics, Decision Systems Engineering, Tempe, AZ, USA
kasim.candan@asu.edu

[7] Texas A&M University, Department of Mechanical Engineering, College Station, TX, USA
zoneill@tamu.edu

* Corresponding Author

ABSTRACT

Increasing advancements in building digitization, smart sensing, and metering technologies have allowed large amount of data to be collected and saved for monitoring, analyzing, and controlling building systems. However, due to sensors or communications failure, the data collected are often incomplete and poor in quality. Data imputation approaches to replace the missing values, specifically based on both statistical and predictive models have been widely adopted for multivariate datasets. It is hence of interest to find an effective way to impute building system data by leveraging the mutual information from strongly correlated sensors. In this paper, we evaluate multiple data imputation approaches using data collected from a medium sized, mixed-use institution building situated in Stockholm, Sweden. Sensors with widely varying characteristics from the case study building were selected to test the imputation methods. Artificial test data with ground-truth was first created for validation by removing randomly selected portions of data. The imputation accuracy was computed for each method and the impact of the chosen method on a short-term building forecasting model was evaluated. Results demonstrate that incorporating time-lagged cross correlations within the k-nearest neighbor (kNN) model provide the most accurate imputations.

1. INTRODUCTION

Recent advances in monitoring systems, communication and information technology make it possible to collect and store large amounts of time series data. Such databases are utilized by data-driven methods and algorithms for solving complex problems in different disciplinary fields. One of the sectors that has greatly benefitted from the advances in big data is building sciences. Use of artificial intelligence, machine learning, and deep learning analysis have been proved to have remarkable impacts on various research domains such as building energy forecasting, pattern recognition, and fault detection and diagnosis (Ma et al., 2020).

An essential prerequisite for implementing these technologies is high-quality data set from the building automation system (BAS). However, the databases obtained from the BAS are usually incomplete due to equipment failures, communication/transmission issues or data corruption, thus, leading to loss of valuable information. The issue of missing data exists in almost all kinds of data sets, and the size of the missing data can significantly affect the research outcomes since most data analysis and statistical tools are not designed to handle incomplete data. Before exploring methods to reconstruct the missing values, it is important to first understand the missing data mechanism. Rubin (1976) first described and categorized the types of missing data based on the assumptions for the missing data. In the literature, missing data mechanisms are generally divided into three categories (Allison, 2003; Rubin, 1976):

- Missing Completely at Random (MCAR) – data are defined as MCAR when the probability of missingness is not related to either the specific values or the set of observed responses. If data are missing completely at random, then dropping cases with missing data does not bias the inferences.
- Missing at Random (MAR) – defined when the probability of missingness depends on the set of observed responses, but the mechanism of data missingness is traceable or predictable from other variables in the database.
- Not Missing at Random (NMAR) – defined when the probability of missingness depends on the missing variable itself. A model for the missing data must be created for the specific dataset to handle this type.

Many approaches have been developed for handling missing values that are available in the literature. The simplest approach is to simply ignore it and perform the analysis based on the available data (also known as listwise deletion). This procedure, however, can lead to loss of efficiency due to discarding the incomplete observations and may output biased and unreliable results due to the systematic differences between the available and missing data. Hence, an effective technique for estimating missing values is through data imputation which retains the original sample size and allows subsequent analysis to be carried on the entire dataset and produce much reliable results. The imputation approaches vary from simple methods such as mean imputation to some more robust methods that leverage the relationships among variables within the dataset. The following section presents some popular imputation methods for data precleaning (Allison, 2003; Rubin, 1976; Schafer & Graham, 2002):

- Mean or mode imputation: This method replaces the unknown value for a given variable by the mean (quantitative variable) or mode (qualitative variable) of all known values of that variable. A major drawback of this method is that replacing all missing records with a single value distorts the input data distribution.
- Hot deck imputation: In this method, the missing values are substituted from the variables that are closest to the variable consisting of the missing values. This method attempts to preserve the input data distribution since the substituted values are based on the similarity with the other variables in the dataset.
- Prediction models: These methods involve creating a predictive model to estimate values that will replace the missing data. The incomplete variables are used as target, and the remaining variables are used as inputs for the model. These methods leverage the relationships (correlations) among the different variables within the dataset. Those correlations can be used to create a predictive model for either classification or regression. If there are no strong correlations between the incomplete variables and the remaining variables, the imputed values will not be precise. Also, many prediction models must be designed for cases with many incomplete variables, making it computationally expensive.

This paper aims to explore potential data imputation method for building systems data and is structured as follows: Section 2 provides a review of existing literature on data imputation methods. Section 3 describes the proposed method and Section 4 present the case study building used to evaluate and compare different imputation methods. The results are discussed in Section 5 and Section 6 presents the current conclusions and future work.

2. LITERATURE REVIEW

Existing literature on data imputation methods for building data has primarily been focused on univariate time series data using statistical methods. For example, Jin et al. (2006) used a stochastic binning method to estimate the missing weather data (outdoor temperature) for the estimation of building energy. Ouyang et al. (2017) implemented linear interpolation to tackle the missing values when predicting wind power. Kasam et al. (2014) focused on a Gaussian-distributed auto-regressive model to interpolate the missing meteorological data for use in building simulations. Although statistical methods are easy to implement and perform well for low rates of missing data, their performance is limited when the datasets are more complicated since these methods are based on linear assumptions and do not account for the nonlinearities of the dataset.

In order to address this issue, nonlinear machine learning and deep learning methods are utilized to fill in the missing values. For example, Garnier et al. (2012) implemented Artificial Neural Network (ANN) to estimate the missing values for energy resources management in buildings. Ma et al. (2020) have proposed a bi-directional missing data imputation scheme based on deep learning and transfer learning for building energy data. Rahman et al. (2018) and Yang et al. (2019) have used deep Recurrent Neural Networks (RNN) models to first perform imputation on the missing data and then forecast the building energy usage.

Since the single imputation method may not work for various types of sensors in a building system, ensemble methods that include both statistical and learning methods that utilize multiple imputation methods to make an improved prediction of the missing values are used. Inman et al. (2015) utilized two data imputation methods prior to performing clustering analysis on building energy consumption data. More recently, Zhang (2020) developed a pattern recognition-based ensemble framework to first create validation data that have similar characteristics with the missing data and then, use the optimal imputation method for each sensor. Research on ensembles methods have shown good performance compared to single imputation methods for building system data.

However, there remains a few key problems that the existing literature fails to address. First, most of the existing studies only explores univariate time series data, hence, these methods do not leverage the mutual information from other variables when a multivariate dataset is available. Since many building system sensors are strongly correlated with each other, there is potential in utilizing information from other variables. Secondly, almost all the studies and developed methods only consider short time periods of missing data and the more challenging case of continuous missing data for long time periods is not explored by these methods.

Given the limitations in existing literature for handling multivariate datasets in BAS data, studies from other fields such as biomedical, traffic flow, and radio transmission are investigated. A probabilistic principal component analysis (PPCA) based method is developed for traffic flow dataset by Qu et al. (2009). In PPCA-based imputation, the PCA is used to separate the significant and dominant parts of the traffic flow, whereas the maximum likelihood estimation (MLE) is applied to estimate the missing values based on the obtained significant parts determined by the PCA. A drawback of this method is the requirement of a large amount of historical data to train the model.

Another widely adopted method was the k-Nearest Neighbor (kNN) imputation method in which k nearest neighbors for a missing instance are identified using the observed instances of other variables. The missing instance is then estimated by combining the k estimates using approaches such as the weighted average or a kernel function (Rahman et al., 2015). These kNN-based methods are most appropriate for a multivariate dataset, and when the missing variables are correlated with the other observed variables. One drawback of kNN is that, because it relies on the values of other variables, it cannot impute a value when all variables are missing in an instance and may be less accurate as more variables are absent.

To overcome this issue, Rahman et al. (2015) developed a combined method using the Fourier transforms which uses past values of each variable and an extension of kNN called the lagged-kNN which also considers the time lagged cross-correlation between each of the variables. The Fourier lagged-kNN method (FL-kNN) overcomes the limitation of the nearest neighbors methods which require observed data to be presented at each instance and improves accuracy by handling both MAR and NMAR missing data. In their research, the imputation accuracy for multiple biomedical data using this method was reported to be the highest when compared with other imputation methods for up to 50% of the dataset missing.

Motivated by the limitations in existing literature for imputing continuous missing data for long time periods and the potential of the FL-kNN method, in this paper, we applied five different imputation methods to a multi-stream building dataset. Artificial test data with ground-truth was first created for validation by removing randomly selected portions of data. The imputation accuracy was computed for each method. Considering that the sensor measurements are typically used for further data analytics, it is of interest to also examine how data repaired by different imputation schemes might affect the quality of data analytics. Hence the impact of the chosen method on a short-term building forecasting model was also evaluated. The imputation schemes chosen for the study are as follows: (i) linear interpolation (from *pandas* library), (ii) kNN imputer (from *sklearn* library), (iii) lagged-kNN, (iv) Fourier method and (v) combined Fourier lagged-kNN imputation method (FL-kNN).

3. COMBINED FOURIER LAGGED-KNN METHOD

The combined Fourier lagged-kNN method is a combination of two imputation methods presented by Rahman et al., (2015). This imputation method consists of an extension of kNN imputation with lagged correlations, and the Fourier transformation. For each missing data point, the final estimated value is calculated by averaging the estimates from the lagged-kNN imputation and the Fourier method.

3.1 Lagged-kNN Method

First, a kNN with a time lag parameter (p) is developed to incorporate the time dependent correlations that may persist between the variables. In this step, cross-correlation is used to identify which variables are correlated and at which time lags. The cross-correlation r_{xy} between variables x and y, for time delay d is defined as:

$$r_{xy}(d) = \frac{C_{xy}(d)}{\sqrt{(C_{xx}(0)C_{yy}(0)}} \tag{1}$$

$$c_{xy}(d) = \begin{cases} \dfrac{1}{T-d}\displaystyle\sum_{t=1}^{T-d}(x_t - \bar{x})(y_{t+d} - \overline{y}),\ if\ d \geq 0 \\ \dfrac{1}{T+d}\displaystyle\sum_{t=1-d}^{T-d}(x_t - \bar{x})(y_{t+d} - \overline{y}), otherwise \end{cases} \tag{2}$$

where T is the length of the series, $\bar{x}$ and $\overline{y}$ are the mean of x and y respectively, d varies from – $(D-1)$ to $(D-1)$ and D is the maximum time delay. Based on the strength of the correlation, matrices are constructed for each of the p lags, with the correlations ordered from $1 \ldots p$ by decreasing strength. For each pair of variables, (L_1) contains the lag, d, with the strongest correlation ($max\ |r_{xy}|$) and L_p the lag with the weakest. Each L is an NxN matrix, where elements represent the time lags for each correlation between the N variables.

In contrast to kNN, the multiple lags that differ across variable pairs must be accounted for while forming the training and testing vectors with candidate values. In the lagged-kNN method, for a variable x that has missing value at time t and has a time lagged relationship with variables y and z, with lags l_{xy} and l_{xz} respectively, the test vector is formed using the values of y and z at $t + l_{xy}$ and $t + l_{xz}$. Training vectors are formed in a similar way and the values of x, which are the candidate values for imputation, are stored separately. Training vectors are generated from the existing values of x and the time instances resulting after adding the lags must be within 1 to T (length of data).

A weighted modification of the Euclidean distance is used as a proximity measure while finding the nearest neighbors for each missing instance. Since the strength of the correlation between variables and across the p lags may differ substantially, a weight is incorporated into the distance measure. This ensures that neighbors based on highly correlated variables with their associated lags are given more weight than weakly correlated variables. The distance between instance x and y is calculated through:

$$d(x,y) = \frac{\sqrt{\sum_{i=1}^{N}[(x_i \wedge y_i)(x_i - y_i)^2 \times w_i]}}{\sum_{i=1}^{N}(x_i \wedge y_i)} \quad (3)$$

where N is the number of variables, and w_i is the weight, which is the normalized correlation coefficients between missing variables and i^{th} variable. The result is p sets of k nearest neighbors (one set of neighbors for each L matrix). We then average the values for the k neighbors with the lowest weighted distance (out of the set of p x k neighbors).

3.2 Fourier Method

The Fourier imputation method uses past values of each variables to impute each missing value. First, a data segment is formed with the data from the beginning of the signal up to the last non-missing data point. For example, if values v_1 through v_{p-1} are present (or imputed), and $v_p \ldots v_q$ are missing, the Fourier descriptors are obtained with:

$$F_k = \sum_{j=1}^{p-1} v_j \times e^{-(2i\pi/p-1)(j-1)(k-1)} \quad (4)$$

where F_k is the k^{th} Fourier descriptor with $1 \leq k \leq (p-1)$ and $i = \sqrt{-1}$ Then, the imputed value for time m, where $p \leq m \leq q$, calculated from the Fourier descriptors with:

$$v_m = \frac{1}{p-1} \sum_{k=1}^{p-1} F_k \times e^{(2i\pi/p-1)(j-1)(m-1)(k-1)} \quad (5)$$

The proposed method estimates each missing value based on the observed data, hence, it is found that if the given data does not capture the high frequency components (i.e. sampling frequency is less than 2 * Nyquist frequency), the imputed value will not be accurate.

4. REAL BUILDING CASE STUDY

4.1 Multi-Stream Building Data

The study utilized data from a 13,434 sq. meter, mixed-use institution building situated in Stockholm, Sweden. It is acclimatized by an intelligent demand-control ventilation (DCV) system where the airflow can independently vary within and between different rooms, carbon- and temperature sensors controlling the air volume within in conference rooms, modern building management system (BMS) and energy meters throughout the building. Datapoints collected from the building include the heating, ventilating and air conditioning (HVAC) system, occupancy data for individual offices, and the energy-meter.

The RealEstateCore (REC) ontology is utilized for data integration in the building. RealEstateCore, which is described in more detail by Hammar et al. (2019), is a universal language that facilitates data-driven building control and the development of new services. REC implementation in the case study building enabled two-way communication which additionally allows operators to send control signals back to the BMS (called SAIA) and the DCV system (called Lindinvent) in REC format (Halleberg & Martinac, 2020). Both the BMS and the DCV system communicate by Modbus Transmission Control Protocol (TCP) and are translated using Internet-of-Things (IoT)-Edge Modules. The modules are written in C# and can be configured to poll the source system at different intervals.

A Microsoft Azure digital platform was also developed for the case study building by combining the potential of the cloud with security using an IoT-Edge solution (Halleberg & Martinac, 2020). Data from BMS and DCV system are extracted (according to the configuration specified in the two REC tag lists) using an IoT-Edge module running on an Edge Device physically located in the building. Additionally, data from the tenant's system for managing space booking, forecasted weather data from Swedish Meteorological and Hydrological Institute (SMHI) and energy management system (called Energiportalen) are imported and processed via application programming interfaces

(APIs) within an Azure Data Factory that triggers every hour. This digital platform solution enables storing, processing big data, and hosting various automated data-driven strategies for the building in the future.

To study the data quality and compare multi-stream data imputation strategies, thirteen (13) different sensors from the case study building were selected to test the imputation methods. A systematic feature selection procedure developed by Zhang & Wen (2019) was used to identify the air handling units (AHUs) and zones with the largest impact on the building electricity and heating energy. Additionally, typical sensors found in weather station, BMS and DCV system that cover different sensor types were included in study. The sensor list is presented in Table 1. Data collected from March 1st to 7th, 2021, with a 5-minute sample interval was used.

Table 1: Sensors used to evaluate data imputation accuracy

System	Sensor Name	Sensor Type	Description
BMS	LB100/GP11/MV	Pressure sensor	AHU supply air static pressure
BMS	LB101/GF11/MV	Flow sensor	AHU supply air flowrate
BMS	LB101/GT11/MV	Temperature sensor	AHU supply air temperature
BMS	LB101/TF11/R	Fan speed sensor	AHU supply fan speed
BMS	LB101/SV21/R	Valve position	AHU cooling coil valve position
BMS	VS130/GT11/MV	Temperature sensor	Radiator hot water temperature
Weather Station	R1B17/GM51/MV	Hygrometer	Outdoor humidity
Weather Station	VS110/GT31/MV	Flow sensor	Outdoor temperature
DCV	3B13/TD10:61/GF	Flow sensor	Actual supply-air flow
DCV	3B13/TD10:61/GT	Temperature sensor	Actual supply-air temperature
DCV	3B13/TD10:61/GP	Pressure sensor	Actual supply-air pressure
DCV	3B13/TD10:61/GT	Temperature sensor	Actual room temperature
DCV	3B13/TD10:56/CO	Carbon-dioxide sensor	Actual carbon dioxide concentration

4.2 Creating Validation Dataset

Artificial validation dataset was created by generating a non-uniform set of random indices through the Python 'pd.DataFrame.sample' function for each sensor. To ensure that the same index is removed for each imputation method, 'rng default' function was used to select the random seed.

'Randomized missing' validation sets for up to 40% missing ratios were created by removing the datapoints at the randomly generated indices. Similarly, 'continuous missing' validation sets were created by generating a random index, then removing the next instances of datapoints until the desired missing ratio wass met. Figure 1 presents the two categories of validation datasets created.

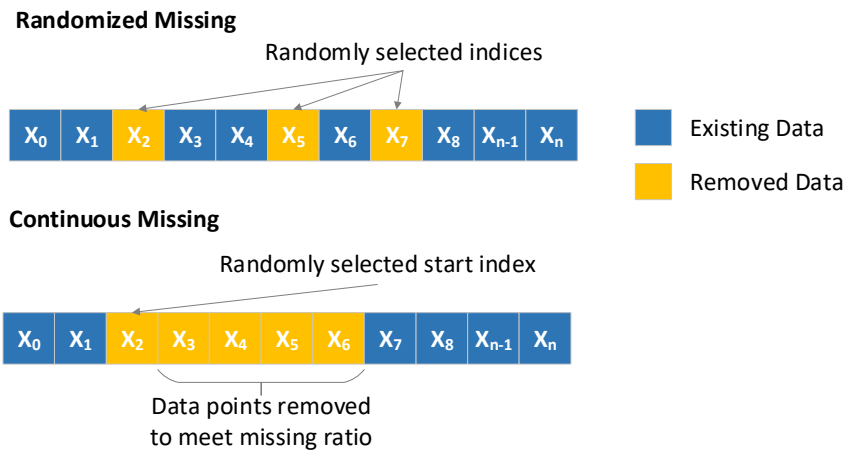

Figure 1: Categories of validation dataset for evaluating data imputation methods

4.3 Energy Forecasting Model

Following the data imputation, a building energy forecasting model was used to evaluate the impact of the imputation method on the energy forecasts. The energy modeling procedure utilizes a multi-input-single-output regression model called Multivariate Adaptive Regression Splines (MARS). MARS is a nonparametric regression which constructs underlying relationship from a set of coefficients and basis function that are determined by training data. Since MARS operates as multiple piecewise linear regression, it can handle the nonlinearities in the datasets. Mathematically, MARS is defined as:

$$f(x) = \gamma_0 + \sum_{i=1}^{m} \gamma_i h_i(x) \tag{6}$$

where, x represents the features, m is the number of features, γ is the constant coefficient of the combination whose value is jointly adjusted to give the best fit to the data and the basis function $h_i(x)$ (Gints, 2011). The basis function h_i can be represented as:

$$h_i(x) = \prod_{k=1}^{K_m} \left[S_{k,m} \cdot \left(x_{v(k,m)} - t_k, m \right) \right]_+^q \tag{7}$$

where K_m is the number of splits given to the m^{th} basis function, $S_{k,m} = \pm 1$ indicates the right/left sense of the associated step function, $v(k,m)$ is the label of the features, and $t_{k,m}$ represents values (often called knot locations) of the corresponding variables. The superscript q and subscript $+$ indicate the truncated power functions with polynomials of lower order than q. More details of fundamentals of MARS can be found in Gints (2011).

The forecasting model was trained to predict the building electricity and heating energy using data collected from the case study building from March 1st to 7th, 2021, and tested using data collected from March 8th to 11th, 2021

4.4 Performance Metric

Normalized Root Mean Squared Error (NRMSE) was used as the performance indicator to measure both the imputation accuracy and the energy forecast accuracy. The NRMSE is defined as:

$$NRMSE(y,x) = \frac{1}{mean(x)} \sqrt{\frac{\sum_{i=1}^{n} (y_i - x_i)^2}{n}} \tag{8}$$

where, y is the imputed or predicted value, x is the actual value and n is the number of imputed data points. The overall performance metrics of the imputation method was calculated by averaging the metric values for each sensor.

5. RESULTS AND DISCUSSION

The performance of the data imputation method and the impact on the short-term energy forecast was evaluated using the NRMSE. The results are summarized in Table 2 and Table 3, respectively. The best performing imputation method under each category is highlighted in bold.

Table 2: Averaged imputation error for each validation dataset

Validation Dataset	Imputation Method				
	Linear	kNN Imputer	Lagged kNN	Fourier	Fourier-Lagged kNN
10p-Continous	0.772	0.291	**0.207**	1.229	0.664
20p-Continous	0.630	0.486	**0.273**	0.935	0.525
30p-Continous	0.874	0.436	**0.338**	1.092	0.594
40p-Continous	1.089	0.455	**0.210**	1.113	0.545
10p-Scatter	0.087	0.161	0.098	0.111	**0.081**
20p-Scatter	0.105	0.193	**0.063**	0.166	0.099
30p-Scatter	0.127	0.292	**0.086**	0.257	0.153
40p-Scatter	0.118	0.430	**0.093**	0.269	0.160

Table 3: Averaged energy forecasting error for each validation dataset

Validation Dataset	Imputation Method				
	Linear	kNN Imputer	Lagged kNN	Fourier	Fourier-Lagged kNN
10p-Continous	0.025	0.025	**0.024**	0.025	0.025
20p-Continous	0.042	0.038	**0.036**	0.038	0.039
30p-Continous	0.066	0.076	**0.063**	0.074	0.072
40p-Continous	0.082	0.091	**0.072**	0.076	0.079
10p-Scatter	**0.021**	**0.021**	**0.021**	0.022	**0.021**
20p-Scatter	**0.020**	**0.020**	**0.020**	**0.020**	**0.020**
30p-Scatter	0.024	0.024	0.024	0.024	**0.023**
40p-Scatter	0.025	0.025	**0.024**	0.025	0.025

From the results presented above, several conclusions can be drawn in terms of imputation and energy forecasting accuracies:

- The lagged-kNN method has the best performance in 7 out of the 8 validation sets evaluated for data imputation. Identifying the strongly correlated building system measurements using the kNN algorithm helps the method to utilize the shared mutual information among the different measurements, hence, resulting in more accurate estimation of missing values.
- Incorporating the time-lagged cross-correlation in the lagged-kNN method helps to significantly improve the imputation accuracy as well. Since building system measurements usually have a delayed response to changes in external weather conditions or supervisory control signals, using a set of time-lagged nearest neighbors to perform the data imputation result in better performance.

- The combined Fourier-lagged kNN method has comparable performance with the lagged-kNN method for lower missing ratios in the randomized scattered set, however, the error increases when applying the combined method for larger missing ratios. Since the Fourier imputation uses past values of each variable to impute each missing value, this method requires a large amount of historical data which is not satisfied by the 8-days of data collected from the BAS. The inaccuracies from the Fourier method, thus, leads to larger overall error for the combined method.
- While statistical methods (such as linear interpolation, backward fill, forward fill etc.) are simple and relatively easy to implement, these methods do not perform well for datasets with high missing ratios and/or when data is continuous missing. These methods also lack the capability to utilize the strong correlations that exist between various building system measurements.
- Results from the building energy forecasting model again show that the dataset imputed using the lagged-kNN has the best performance. However, the forecasting accuracy is comparable among all imputation methods. Since the forecasting model was tested using limited data under similar building operating and weather conditions, the choice of imputation method does not have a significant impact on the short-term building energy forecast.

6. CONCLUSIONS

This paper evaluates different data imputation methods using data collected from a multi-stream building dataset. Since building system sensors are often correlated with each other, there is potential in utilizing information from other variables to impute the missing values in a dataset. Recent development in machine learning models such as kNN methods allow to identify such existing correlations that improve the estimates of the missing values by leveraging the shared information.

To evaluate the effectiveness of such methods, five different imputation methods of varying sophistication were tested by creating artificial validation sets that consists of different ratios and patterns of missing data. Results demonstrate that incorporating time-lagged cross correlations within the kNN framework help to significantly improve the imputation accuracy. In terms of the impact of the chosen imputation method on building energy forecasting, results from limited testing showed comparable accuracy between the different imputation methods. Although the choice of the imputation method showed minimal impact on the performance of the forecasting model reported in the paper, additional validation is required for forecasting models that use larger datasets under diverse system operating conditions. Furthermore, the impact on other data-driven applications such as fault detection or fault diagnosis, where the strategies require complete dataset and are more sensitive to the data quality needs to be further investigated in the future.

REFERENCES

Allison, P. D. (2003). Missing Data Techniques for Structural Equation Modeling. *Journal of Abnormal Psychology*. https://doi.org/10.1037/0021-843X.112.4.545

Garnier, A., Eynard, J., Caussanel, M., & Grieu, S. (2012). Missing data estimation for energy resources management in tertiary buildings. *2nd International Conference on Communications Computing and Control Applications, CCCA 2012*. https://doi.org/10.1109/CCCA.2012.6417902

Gints, J. (2011). *ARESLab: adaptive regression splines toolbox for Matlab/Octave*. 1–19. http://www.cs.rtu.lv/jekabsons/

Halleberg, D., & Martinac, I. (2020). Indoor Climate as a Service: A digitalized approach to building performance management. *IOP Conference Series: Earth and Environmental Science*, *588*(3), 0–8. https://doi.org/10.1088/1755-1315/588/3/032013

Hammar, K., Wallin, E. O., Karlberg, P., & Hälleberg, D. (2019). The RealEstateCore Ontology. In C. Ghidini, O. Hartig, M. Maleshkova, V. Svátek, I. Cruz, A. Hogan, J. Song, M. Lefrançois, & F. Gandon (Eds.), *The Semantic Web -- ISWC 2019* (pp. 130–145). Springer International Publishing.

Inman, D., Elmore, R., & Bush, B. (2015). A case study to examine the imputation of missing data to improve clustering analysis of building electrical demand. *Building Services Engineering Research and Technology*, *36*(5), 628–637. https://doi.org/10.1177/0143624415573215

Jin, Z., Yezheng, W., & Gang, Y. (2006). A stochastic method to generate bin weather data in Nanjing, China. *Energy Conversion and Management*, *47*(13–14), 1843–1850. https://doi.org/10.1016/j.enconman.2005.10.006

Kasam, A. A., Lee, B. D., & Paredis, C. J. J. (2014). Statistical methods for interpolating missing meteorological data for use in building simulation. *Building Simulation*, *7*(5), 455–465. https://doi.org/10.1007/s12273-014-0174-7

Ma, J., Cheng, J. C. P., Jiang, F., Chen, W., Wang, M., & Zhai, C. (2020). A bi-directional missing data imputation scheme based on LSTM and transfer learning for building energy data. *Energy and Buildings*, *216*. https://doi.org/10.1016/j.enbuild.2020.109941

Ouyang, T., Zha, X., & Qin, L. (2017). A combined multivariate model for wind power prediction. *Energy Conversion and Management*. https://doi.org/10.1016/j.enconman.2017.04.077

Qu, L., Li, L., Zhang, Y., & Hu, J. (2009). PPCA-based missing data imputation for traffic flow volume: A systematical approach. *IEEE Transactions on Intelligent Transportation Systems*, *10*(3), 512–522. https://doi.org/10.1109/TITS.2009.2026312

Rahman, A., Srikumar, V., & Smith, A. D. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *Applied Energy*, *212*(December 2017), 372–385. https://doi.org/10.1016/j.apenergy.2017.12.051

Rahman, S. A., Huang, Y., Claassen, J., Heintzman, N., & Kleinberg, S. (2015). Combining Fourier and lagged k-nearest neighbor imputation for biomedical time series data. *Journal of Biomedical Informatics*, *58*, 198–207. https://doi.org/10.1016/j.jbi.2015.10.004

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Yang, J., Tan, K. K., Santamouris, M., & Lee, S. E. (2019). Building energy consumption raw data forecasting using data cleaning and deep recurrent neural networks. *Buildings*, *9*(9). https://doi.org/10.3390/buildings9090204

Zhang, L. (2020). A pattern-recognition-based ensemble data imputation framework for sensors from building energy systems. *Sensors*, *20*(5974), 1–16. https://doi.org/10.3390/s20205947

Zhang, L., & Wen, J. (2019). A systematic feature selection procedure for short-term data-driven building energy forecasting model development. *Energy and Buildings*, *183*, 428–442. https://doi.org/10.1016/j.enbuild.2018.11.010

ACKNOWLEDGEMENT

The work presented in the paper has been carried out with the support of National Science Foundation (NSF) under the Partnerships for Innovation – Research Project (PFI-RP): Data – Driven Services for High Performance of Sustainable Buildings (Award no. 2050509).