Energy Technologies Area
Lawrence Berkeley National Laboratory

Effective Missing Value Imputation Methods for Building Monitoring Data

Brian Cho[1], Teresa Dayrit[2], Yuan Gao[2], Zhe Wang[3], Tianzhen Hong[3], Alex Sim[3], and Kesheng Wu[3]

[1]Yale University, New Haven, CT, [2] Stanford University, Stanford, CA, and
[3]Lawrence Berkeley National Laboratory, Berkeley, CA

December 2020

Effective Missing Value Imputation Methods for Building Monitoring Data

Brian Cho
Yale University
New Haven, CT, USA
brian.cho@yale.edu

Teresa Dayrit, Yuan Gao
Stanford University
Stanford, CA, USA
{*tedayrit,gaoy*}*@stanford.edu*

Zhe Wang, Tianzhen Hong, Alex Sim, Kesheng Wu
Lawrence Berkeley National Laboratory
Berkeley, CA, USA
{*zwang5, thong, asim, kwu*}*@lbl.gov*

***Abstract*—To understand behaviors of natural and man-made events, such as energy consumption of buildings, which accounts for 40% of energy uses in the US, we deploy automated monitoring devices to record periodic observations. However, such experimental and observation data often contains problems and irregularities that have to be cleaned up before analyses. Due to various conditions affecting sensor operations, the communication channels, recording steps, or the recording media, the recorded data might have missing values, errors, or anomalous values. An effective way to clean up these problems is to replace these missing values, errors and anomalous values with expected values, a process generally known as imputation. In this work, we survey commonly used missing value imputation techniques and compare their performance on a set of building monitoring data. To compare the different types of sensor measurements with widely varying characteristics, we use normalized root mean squared error (NRMSE) as the key metric for the effectiveness of the imputation methods. We additionally consider periodicity and run time when considering comparing methods. Through extensive testing, we find that for small gap sizes, up to 8 consecutive missing values, linear interpolation performs the best; for larger gaps stretching up to 48 consecutive missing values, K-nearest neighbors provides the most accurate imputations; for even larger gaps, more computational intensive methods, such as matrix factorization, achieve the smallest NRMSE. Additionally, we observe that these computationally intensive algorithms not only provide accurate imputations for large gaps, but are also more robust across all types of sensors.**

***Index Terms*—Matrix factorization, interpolation, imputation, building monitoring**

1. Introduction

Time series data is a common form of data recording, present across nearly all domains of research, from economics to meteorology. This type of data proves to be incredibly useful, with researchers performing analyses for tasks such as pattern recognition, forecasting, and system optimization; however, the results are highly dependent on the reliability of the data and completeness of the information.

A crucial application of time series analysis lies in building performance analysis, with buildings accounting for approximately 40% of the primary energy consumption in the United States. With the advancement of Internet of Things (IoT) and smart sensing and metering technologies, the building operation, control, and maintenance process is getting digitized; building environment (e.g., space air temperature, humidity, CO_2 concentration; outdoor weather parameters including air temperature, humidity, solar irradiance, wind speed and direction), energy (e.g., electricity and/or natural gas consumed by lighting, HVAC (heating, ventilation, and air conditioning), plug-in equipment, service water heating), and operational data (e.g., from building automation and control systems), in the form of time series data, have been collected and saved. Building performance diagnostics and improvements through data analytics demonstrate a huge potential to improve occupant satisfaction while reducing energy consumption and carbon emissions in buildings. However, due to the sensor failures and network disconnection, the time-series data collected by the sensor networks in many buildings is inconsistent in quality, resulting in noisy data with significant proportions of missing values that hinder the building performance analysis and improvements. Furthermore, due to the time required to calibrate, reinstall and restore a sensor, the data gap size, i.e., stretches of consecutive missing values, could be up to several weeks in some circumstances.

Many applications need to deal with similar missing values and related data quality issues, and an effective approach is to replace these missing or anomalous values with the expected ones, a process generally known as *imputation* [1], [2], [3], [4]. Based on the characteristics of the data and the underlying application, what we expect the missing values to be, i.e., exactly how to impute the values, will be different. For example, a simple approach would be to assume the missing values connect the nearest good measurements through a linear function, which leads to linear interpolation as an imputation method [3], [5]. In many cases, where the measurements change slowly, this classic approach is very effective. A question we need to investigate is whether building energy data measurements are changing slow enough that this approach might actually work. Similarly, there are many different imputation methods in the published literature [1], [2], [3], [4]. The plethora of

imputation approaches poses the pressing question of which methods are effective for building energy applications, and the conditions in which these methods perform well.

In this work, we plan to examine most commonly used imputation approaches for time series data including the linear interpolation method mentioned above, and common interpolation methods such as the cubic splines. In addition to these univariate methods, we will also examine multivariate imputation methods, such as multiple imputations via chained equations, matrix factorization, and K-nearest neighbors, which have proven to be effective in a wide variety of applications, e.g., electricity [6], genetics [7], and vehicle safety [8]. Within the category of time-series data, researchers have also found high accuracy imputation methods through expectation maximizing algorithms, as in the case with multivariate meteorology data [9]. There are also many techniques under various names such as recommender systems and matrix completion techniques, which could be considered imputation methods, that have been successfully used in different applications [10], [11].

Despite various studies on missing value imputation effectiveness across a wide variety of domains, there has been little work on imputing missing values specific to buildings' time-series data, which contain unpredictable data gaps at varying sizes. Building data exhibits some unique characteristics that could be leveraged for missing data imputation. For instance, building data is highly periodical due to occupancy schedule and weather variation. Identifying those patterns and behaviors through data-driven approach could significantly improve data imputation accuracy. Common univariate time series imputation methods do not leverage such characteristics and often produce inconsistent results that deviate significantly from the actual measured values. Furthermore, the use of multivariate methods across different sensor recordings assumes specific relationships between sensors, which may cause misleading results for future work on the imputed dataset. Given these issues, it becomes essential to find effective imputation methods that only consider the data of a single sensor.

In this study, we empirically evaluate the effectiveness of both univariate and multivariate imputation methods for a single sensor's recordings across different contexts of sensor category, missing rates, and gap sizes. Univariate imputation methods include linear interpolation (LIN) and spline interpolation (SPL), and multivariate imputation methods include K-nearest neighbors (KNN), multiple imputations via chained equations (MICE), iterative singular value decomposition (SVD-EM), and matrix factorization (MF) via stochastic gradient descent (SGD).

We select these methods due to their distinct approaches to generating missing value imputations, from established, simple interpolations generally used for time series to machine learning approaches developed for recommendation systems. The two univariate methods use polynomials up to degree 1 and 3 for linear and spline interpolation respectively. KNN uses a weighted mean that considers similarity among samples, and MICE uses multivariate linear regression to generate its imputations. SVD-EM combines a low-rank, orthogonal approximation with an expectation-maximising procedure, and MF uses SGD, a simple yet powerful machine learning algorithm. While no means a comprehensive testing of all missing value imputation methods, the chosen methods provide a good indication of which general imputation approach works best with building sensor data, which naturally leads to further optimization by testing more advanced methods with similar approaches.

For multivariate methods, we reorganize the structure of a single sensor's time series data into a matrix, using the naturally occurring periodicity (weekly in this case) to determine its dimensions to leverage the historical periodic pattern for data imputation. In preliminary tests, which tested row lengths in multiples of days, we saw that the most accurate imputations are produced through reorganizing the time series into weeks, the dimension that corresponds to the periodicity present in the data.

To evaluate imputation performance, we consider imputation accuracy with Normalized Root Mean Square Error (NRMSE) [12] between imputed and masked values, visual similarity of the imputations, and runtime. This version of NRMSE measures the imputation error against the natural variations in the original data and has a number of theoretical advantages over other normalization approaches [13].

Key contributions of the work include:

- We systematically examine the options to reorganize the data to best take advantage of the inherent structure in the building monitoring data to best utilize the strengths of the well-known imputation methods.
- For a variety of imputation methods, we tested each of them to determine the best parameters for optimal performance.
- Through extensive testing, we determine that the best imputation method is significantly affected by the gaps. For gap sizes up to 8, the linear interpolation produced the smallest NRMSE; for large gaps up to 48, KNN was the most accurate; while for even large gaps, MF is the most effective.
- While computationally expensive algorithms require significantly longer runtimes, these methods not only provide the most accurate imputations for large gaps, but also demonstrate robustness across sensor category, missing rate, and gap size.
- The findings in this study can be further applied to other time series, as the methods in this study are tested across both periodic and nonperiodic data. Unlike common univariate methods, multivariate methods demonstrate the ability to accurately impute missing values for both periodic and nonperiodic data across large stretches of missing values.

2. Background

2.1. Dataset

The building energy data used in this study was collected from Lawrence Berkeley National Laboratory (LBNL)'s

TABLE 1. SENSOR DATA INFORMATION BY CATEGORY

Category	Sensor Type	Description	Number of Sensors	Missing Rate
Electricity Consumption Data	mels_S	Miscellaneous electric load for the South Wing	1	0.35
	lig_S	Lighting load for the South Wing	1	0.21
	mels_N	Miscellaneous electric load for the North Wing	1	0.21
	lig_N	Lighting load for the North Wing		0.98
	hvac_N	Heating Ventilation and Air Conditioning load for the North Wing	1	0.11
	hvac_S	Heating Ventilation and Air Conditioning load for the South Wing	1	0.11
HVAC Operation Data	hp_hws_temp	Heat pump heating water supply temperature	1	0.19
	rtu_fltrd_sa_fr	Roof Top Unit filtered supply air flow rate	4	0.20
	rtu_sa_temp	Roof Top Unit supply air temperature	4	0.20
	zone_fan_spd	Supply air fan speed of specified zone	44	0.15-0.45
	zone_hw_valve	Heating water valve position of specified zone	51	0.14-0.24
Temperature Data	zone_temp	Zone temperature of exterior zone	51	0.14-0.24
	cerc_templogger	Zone temperature of interior zone	16	0.07-0.13

Building 59, a four-floor office building that houses two floors' office spaces, NERSC computing facility, and mechanical equipment. The indoor conditions of the two-floor office spaces are maintained through the four rooftop units as part of the HVAC (Heating, Ventilation, and Air-Conditioning) system. The collected dataset contains measurements from 269 sensors over the timespan of two years, from January 1st, 2018 to January 1st, 2020. At the sampling rate of half-hour, we have 35,040 expected measurements per sensor. Table 1 describes the specific categories, sensor counts, and data missing rate.

Plots of sensors in Figure 1 in each of the categories explicitly demonstrate key characteristics of this dataset that determine how we evaluate the imputation errors. Sensors measure different types of data and have significantly different standard deviations across the two-year period: temperature data tends to remain in a narrow range, while other sensors vary significantly. Therefore, it becomes imperative to normalize results when comparing or averaging across sensors. Within a sensor's recordings, differences in behavior between the first and second year, as shown by the energy and HVAC operation data, are present.

A crucial consideration of missing value imputation is the mechanism in which missing values occur within a dataset [14], which determines whether missing value recovery is possible. The three mechanisms of missingness are as follows: Missing Completely at Random (MCAR), where the likelihood of a data point being missing is unrelated to the values of any variables, whether missing or observed, Missing at Random (MAR), where the likelihood of a data point being missing is unrelated to the missing values but may be related to the observed values of other variables, and Missing Not at Random (MNAR), which indicates that the likelihood of a missing data point depends on the actual value of this datapoint. In the case of MNAR, missing value recovery is near impossible with the use of previous data, and would provide inaccurate estimates with our method choices.

The studied dataset demonstrates a MAR pattern. For this dataset, missing values are not completely random, but can be fully accounted by the time variable. The missing rates in year 1 are significantly higher than year 2 (Figure 2) due to frequent commissioning work of the building in the first year. However, these missing values do not seem to be related to the actual value of the missing data.

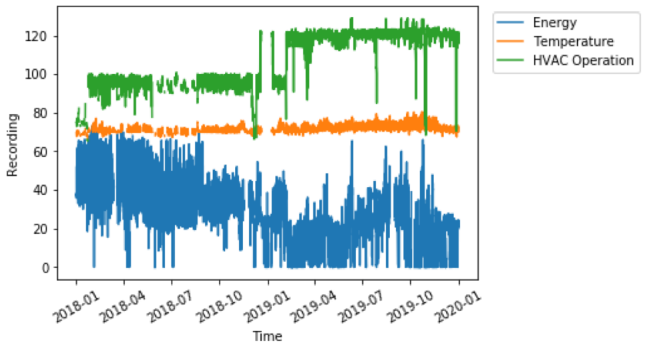

Figure 1. Plots of Sensor Recordings By Category: the unit of energy is kW, unit of temperature is degree Fahrenheit, unit of HVAC operation is percentage

Naturally, a significant number of sensors demonstrate periodicity. Some periodic sensors have high degrees of autocorrelation with values that are apart by a multiple of 48 (number of recordings in a day), which indicate the presence of cycles in multiples of days. Other periodic sensors demonstrate significant correlations with values 336 (number of recordings in a week) apart, demonstrating weekly cycles. The nonperiodic data demonstrates autocorrelation properties that suggest its readings are near constant: autocorrelation values up to 384 recordings apart remain close to 1, as shown in Figure 3.

Given that the longest cycles in the data are weekly, we plan to take advantage of the natural periodicity present through reorganizing one-dimensional time series into a data matrix. Please refer to Section 3.1 for further details.

2.2. Imputation Methods

The first step of the imputation process is to reformat the data from a vector to a matrix, so that the periodical patterns could be better leveraged. Then six different approaches have been compared and applied to impute the data. The following section introduces the general framework of each imputation method. We note that these methods were chosen

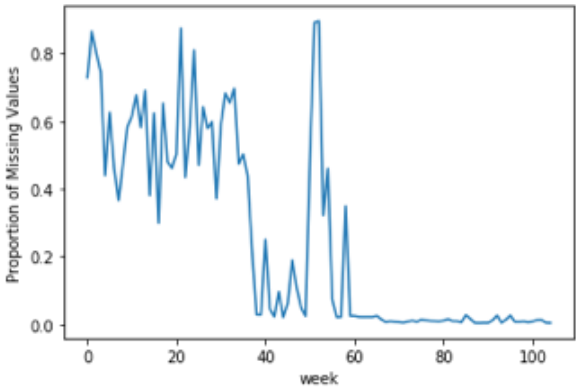

Figure 2. Average Missing Rate Across All Sensors In Respect to Weeks

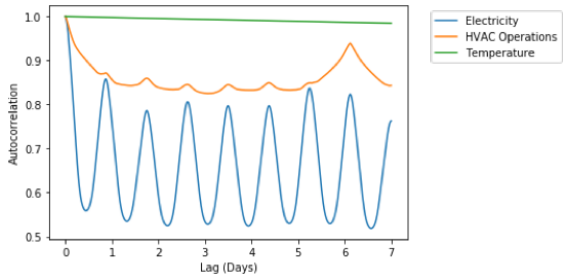

Figure 3. Autocorrelation Plots for a Single Sensor in Each Category

to stretch across various categories of imputation methods, from simple linear fits to machine learning algorithms. We will only describe basic implementation principles behind those imputation methods in this section. For implementation details for all methods other than interpolation methods, please refer to the fancyimpute python package documentation.

2.2.1. Linear Interpolation. Linear interpolation is a commonly used method to impute missing values. After flattening the reformatted data back to a vector, this method takes the last value before the gap and first value after the gap and linearly interpolates the missing values between them.

2.2.2. Spline Interpolation. Spline interpolation expands upon interpolation approaches by using a polynomial of up to degree 3 for an interval of missing values. It then chooses polynomial pieces such that the results fit smoothly together, resulting in a piecewise function for the data called a spline [15].

2.2.3. K-Nearest Neighbor (KNN). KNN imputes values using the weighted mean of the k most similar rows, weighted by their similarity [16]. KNN is a generalization of the classic linear interpolation and is widely used in cases where relations among the dimensions of the data are complex. Since the relation among the sensors are complex and unknown as this time, we believe KNN is potentially a good approach for imputation of building energy data. In this study, we set $k = 5$ and use the Pearson correlation coefficient r as the metric for similarity. Given the worst case scenario where a pre-processed sensor only has 10 complete rows (refer to section 3.1), $k = 5$ maximally allows for half of the most similar sensors to contribute to the weighted mean. Furthermore, initial tests indicate that increasing k results in marginal differences in both accuracy and robustness of KNN.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \quad (1)$$

2.2.4. Multiple Imputations through Chained Equations (MICE). MICE is a multivariate imputation method for missing values. MICE was applied by Ruggles et al. to fill in the missing values of grid-level electricity demand data data [17]. Because the hourly electricity usage exhibit similar daily patterns as many of the measurements in our data set, we believe MICE is potentially a good technique for imputing missing values in our use case. MICE generates estimates from modeling each feature with missing values as a function of other features [5]. Missing values are first imputed using the mean of the column. In the following iterations, each column is set as the response variable sequentially, with other columns forming the observed explanatory variables. Missing values present in the set response variable column are imputed through a multivariate linear regression model. One sweep through each column is an iteration, this method runs until a maximum number of iterations is reached. We set the maximum number of iterations to 10; preliminary results indicate greater numbers of iteration result in longer runtimes and minimal changes in NRMSE scores.

2.2.5. Iterative Singular Value Decomposition (SVD-EM). Expectation Maximization (EM) procedures stand as a direct alternative to MICE due to making minimal assumptions about the distribution of the underlying data [18]. Rather than using a multivariate regression model, with columns as distinct variables, SVD-EM extracts crucial weekly trends through SVD, and further refines these trends through the an EM procedure. With initial data analysis, we know that periodic, weekly trends are present in the sensor data, making SVD-EM a good approach for testing. This imputation method first initializes the missing values as the column means, similar to MICE. Then, rank-k SVD approximation of the matrix re-imputes the missing values; this procedure is terminated until a maximum number of iterations is reached, or there is minimal change in the matrix with respect to its Frobenius norm. [19]. In this study, we set $k = 1$, and maximum number of iterations to 1,000. For this choice of rank, we choose rank-1 approximations due to preliminary NRMSE results from testing across different rank approximations across gap lengths, with missing rate fixed at 0.4 and across all sensor categories (Figure 4). Note that these results with rank approximation may be specific to building energy data with prevalent periodic trends.

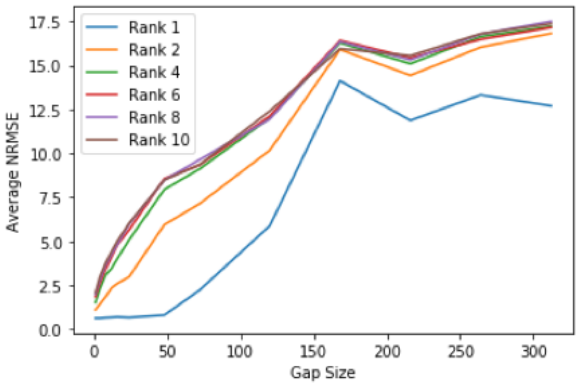

Figure 4. Average NRMSE Across Gap Size with Different SVD-EM Ranks

2.2.6. Matrix Factorization via Stochastic Gradient Descent (MF). Matrix Factorization was widely used to impute missing data. For instance, Zhou et al. used Matrix Factorization to recover missing traffic data [11]. MF works based on the assumption that different days of measurements (different rows of matrix) are generated from a shared subspace, therefore the data matrix of different days can be decomposed using a common factor [20]. MF is widely used to develop a recommendation system based on the assumption that the same customer would prefer products with similar attributes [10]. MF decomposes the incomplete matrix of sensor data into rank-k matrices W and H. W and H are found by minimizing the difference between approximated values from rank-k matrix $W * H$ and observed values present in the sensor data matrix through stochastic gradient descent [21]. The $W * H$ matrix is then a complete matrix, with all missing values filled. Figure 5 provides more specifics on the stochastic gradient descent algorithm.

There are multiple MF methods, such as LU Matrix Decomposition, QR Matrix Decomposition, Non-negative Matrix Decomposition and etc. LU Matrix Decomposition only works for square matrices, which is not the case in this study. Other matrix decomposition methods impose other limitations, such as orthogonality in QR decomposition, non-negativity in non-negative decomposition. We have tested all common versions of matrix decomposition and found the matrix factorization through SGD produced the best answer for our data set.

In later tests, we will always use this matrix factorization with SGD. We set k to be a full-rank approximation of the sensor data matrix, with $k = 336$. Each step (learning rate) is set to 0.001 with 10,000 epochs, and the sparsity penalties for W and H are $L1$ and $L2$ respectively. We choose rank-336 approximations with this method due to two distinct reasons: low-rank approximations cause increases in runtime (Table 2) and inconsistent results across sensor categories (Figure 6). By testing for average NRMSE scores across larger gap sizes, with a set missing rate of 0.4 for all sensor categories, we see that full-rank approximations are more robust, with acceptable runtimes.

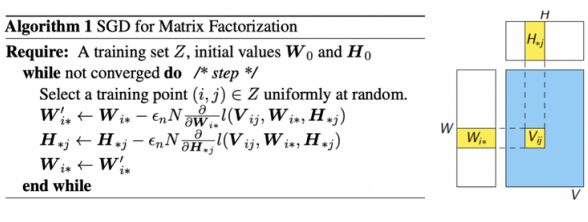

Figure 5. Gradient Descent for Matrix Factorization

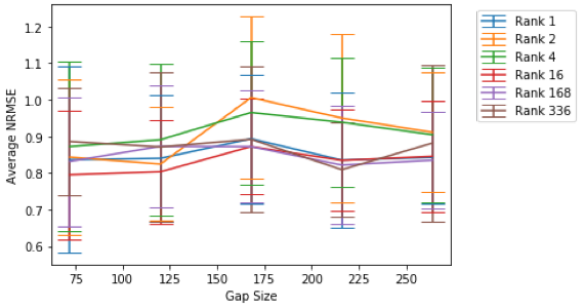

Figure 6. Average NRMSE Across Gap Size with Different SVD-EM Ranks

3. Experiment and Analysis

We examined imputation effectiveness both with variable missing rates and gap sizes. While many common studies focus on testing against only missing rates, building energy data differs in that many missing values occur consecutively due to the time needed to re-calibrate, reinstall and restore the sensors. As a result, gaps in the building data are often much larger than 1, and therefore testing effective imputation for large gaps that span up to a nearly a week in length is necessary for this particular type of data. These experiments were run on CORI, a Cray XC40 with a peak performance of about 30 petaflops, at the National Energy Research Scientific Computing Center.

3.1. Data Preprocessing

Given that many of the sensor recordings exhibit significantly different behavior between the two years, we split the recordings for each sensor into two distinct recordings, each representing a separate year.

For imputation methods other than linear and spline interpolation, they require the data to be in a matrix form. Therefore, we pre-process the year-long time series data with the algorithm in Figure 7.

A single sensor, with recorded values y, is reshaped into a Lxl matrix, with the dimensions of the matrix being determined by the time period in which we consider a single cycle

TABLE 2. AVERAGE RUNTIMES(S) FOR SINGLE SENSOR IMPUTATION, MF RANK TESTINGG

1	2	4	16	168	336
10.3	8.81	8.92	8.09	7.96	6.18

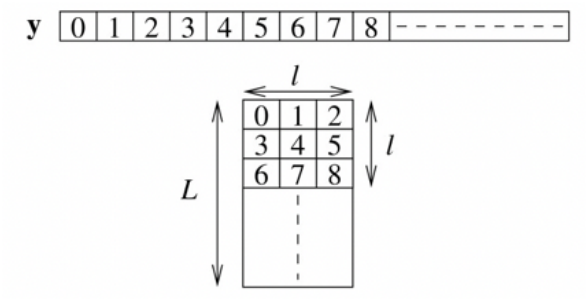

Figure 7. Matrix Reorganization Method (figure obtained from [22])

(Figure 7). We set $l = 336$, indicating weeks as a single observation/row as the default method for comparison in order to capture weekly trends, such as differences between weekdays and weekends. To test error with NRMSE, any rows containing missing values are then removed from the reorganized sensor data, resulting in a matrix containing complete weeks for each sensor.

For testing across all methods, we only consider year-long recordings that contain 10 complete rows after matrix reorganization.

We have tested different time windows for the matrix reorganization, ranging from a few days to many weeks. Since we exclude rows with missing values from the current tests, we have fewer rows when each row covers a longer time window. While with a longer time window, we anticipate to see more variety of patterns and therefore potentially more accurate imputations. Tests show when the time window is shorter than a week, we miss important weekly patterns and the imputations suffer low accuracy. While increasing time window to multiple weeks does not actually increase imputation accuracy likely due to reduced number of examples in the training data. In short, we use one week as the default time window for data reorganization throughout the remaining of this work.

3.2. Masking Algorithm

A masking algorithm was used to generate artificial missing values from the pre-processed data of complete daily recordings. This algorithm generates uniformly sized gaps of a set length up to a certain proportion. For this study, we test across gap lengths of 1, 4, 8, 12, 16, 24, 72, 120, 168, 216, 264, 312, and missing proportions 0.1, 0.2, 0.3, and 0.4.

3.3. Comparison of Imputation Methods

The aforementioned imputation methods are evaluated on the basis of Normalized Root Mean Square Error (NRMSE), runtime, and graphical observation. Of these evaluation criteria, we primarily focus on NRMSE scores and graphical observation of the imputed values to determine effectiveness. Runtimes, representative of the computational resources required for the data imputation, are used to distinguish optimal methods when given comparable error scores and reasonable plots for the missing values.

3.3.1. Normalized Root Mean Square Error (NRMSE). While Root Mean Square Error (RMSE) is commonly used to compare missing value, normalization of RMSE is crucial in this study. We have many measurements of in-door temperatures. Since these measurements are for an office building with well-regulated temperature, these temperature values are very close to each other. In contrast, the variables related to HVAC operation and electricity usage vary in a much wider range. Additionally, the different types of sensors are also measured in different units, which lead to large variations among the measurement values. Furthermore, some measurements have a natural variation over time as described before, while others remains nearly constant through the duration of the study. Therefore, it is important to normalize the error measures so as to not give some variables more importance in the judging the effectiveness of an imputation methods. There are variety of normalization procedures in the literature. We have experimented with normalizing the sensor measurements as well as normalizing the errors. Based on our observations as well as the theoretical analyses from published literature [12], [13], we have selected to normalize RMSE by the standard deviation of the original data:

$$NRMSE = \frac{1}{\sigma}\sqrt{(\frac{1}{n})\sum_{i=1}^{n}(y_i - x_i)^2}$$

Given that we test multiple sensors per category, we consider average NRMSE of each category, and look at the standard deviation of NRMSE to record performance consistency. If the standard deviation of NRMSE is small, then the imputation method performs consistently well on different sensors of the same sensor type.

3.3.2. Visual Similarity. While error rates often give a good basis to determining imputation effectiveness, it fails to give a complete picture of how visually similar the imputed data is. By plotting the imputed data, we can further distinguish imputation methods with similar NRMSE scores by examining how similar the imputed data is to the actual recordings.

3.3.3. Runtime. Considered after two aforementioned criteria, runtime roughly indicates the computational expense of each method. With similar NRMSE and visual similarity, methods with shorter runtimes are preferred. Generally, common univariate methods are expected to run very quickly, while multivariate methods vary in computational expense, making runtime a valid method to determine effective imputation approaches.

4. Results

We compared the six different imputation methods under the context of different missing rates and gap sizes. For

missing rate testing, we fix the gap size to 48; for gap size testing, we fix the missing rate to 0.4. By providing results for each sensor category, we further distinguish category specific results for imputation effectiveness.

4.1. Missing Rate Testing

Table 3 shows the average error of different sensor types for each imputation method, which indicates accuracy. Across all sensor categories and missing rates, we observe that multivariate methods provide the lowest imputation errors: KNN, MICE, SVD-EM, and MF. We not only care about the NRMSE (in Table 3), but also the standard deviation of NRMSE (in Table 4), as we hope the imputation method has consistent performance across different sensors of the same sensor type. As shown in Table 4, not only do these four methods (KNN, MICE, SVD-EM, and MF)) produce the lowest average NRMSE scores, but do so consistently by having lower standard deviations for NRMSE in all categories.

4.2. Gap Size Testing

Compared to missing rate, variable gap sizes tell a more interesting and complicated story of the imputation effectiveness of different methods. Table 5 demonstrate the average NRMSE scores obtained by testing imputation methods against variable gap sizes, with a fixed missing rate of 0.4. Again, despite stratifying by sensor type, imputation effectiveness between the methods remain relatively similar. Across all categories, we see that linear interpolation generally provides the smallest average NRMSE values if the gap lengths are less than 8 (i.e., 4 hours). Between the gap size of 8 and 48, KNN and SVD-EM generally produce the lowest average NRMSE scores. For gaps larger than 48, equivalent to one day missing, MF and MICE provide the most accurate results, with MF delivering a slightly lower NRMSE scores on average. Figure 8 shows a summary of LIN, KNN, and MF effectiveness across different gap sizes. By considering the standard deviation of NRMSE scores obtained across all categories, we see that low standard deviations occur with lower NRMSE scores, demonstrating consistently accurate imputation across all sensors.

4.3. Imputation Plots

While NRMSE values demonstrate the accuracy of the imputation, plots of the imputed values also provide valuable information through visually comparing the imputation results with actual data.

For gaps with less than 8 in length, as shown in figure 9, we notice that both linear and spline interpolation provide unrealistic imputations compared to the actual data; however, because the missing gap is small, these imputations do not result in significant changes to the overall shape of the recording across the year.

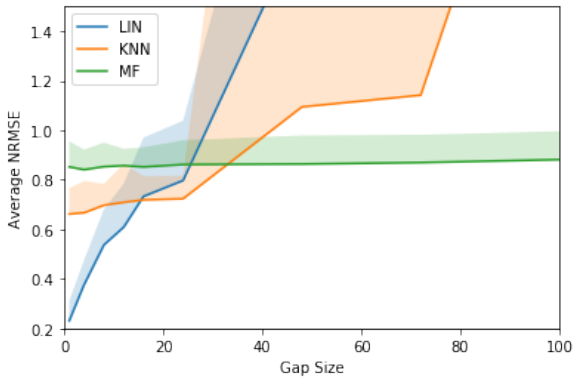

Figure 8. Average NRMSE with Respect to Gap Size

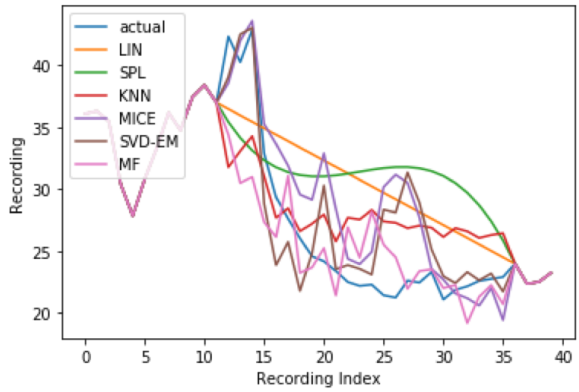

Figure 9. Imputed Values for 4-hour Gap

Figures 10 and 11 show that with larger gaps of 48 and 216 consecutive recordings, the two interpolation methods provide unrealistic imputations to the dataset.

For all sizes of gaps, matrix methods KNN, MICE, SVD-EM, and MF provide visually similar imputed values to the actual dataset.

4.4. Consistency of Imputation Effectiveness

4.4.1. Robustness. Among the imputation methods, we find MF and MICE to be the most robust methods under different missing rates and gap sizes. Across the three sensor categories, MF and MICE report average NRMSE increases by only 0.09 and 0.07 respectively when the missing rate increases from 0.1 to 0.4. This trend reemerges in gap size testing: compared to other methods, the average increase of NRMSE for MF and MICE are 0.04 and 0.25 respectively when gap sizes increase from 1 (0.5 hour) to 312 (6.5 days). While all imputation methods perform similarly across different sensor categories, MF and MICE demonstrate robustness with respect to increasing gap sizes and missing rates by averaging the smallest increases in NRMSE among the other methods tested.

4.4.2. Runtime. Table 7 demonstrates the average runtime of each imputation method in seconds. Notably, we observe

TABLE 3. AVERAGE NRMSE ACROSS MISSING PROPORTIONS

Method	Electricity				HVAC Operations				Temperature			
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
LIN	1.33	1.09	1.22	3.22	0.978	0.97	1.11	1.08	0.997	1.053	1.159	1.20
SPL	485	73.0	1040	6990	43.2	40.4	500	2160	155	1840	282	7660
KNN	0.698	0.803	0.837	0.846	0.655	0.736	0.801	0.834	0.605	0.645	0.678	1.60
MICE	0.903	0.930	0.927	0.961	0.975	0.993	0.985	0.971	0.688	0.747	0.780	0.828
SVD-EM	0.698	0.860	0.722	0.776	0.777	0.751	0.774	0.927	0.593	0.619	0.748	2.14
MF	0.747	0.815	0.830	0.847	0.824	0.829	0.842	0.883	0.745	0.803	0.832	0.861

TABLE 4. STANDARD DEVIATION OF NRMSE ACROSS MISSING PROPORTIOS

Method	Electricity				HVAC Operations				Temperature			
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
LIN	0.704	0.227	0.256	2.64	0.599	0.430	0.795	0.586	0.467	0.535	0.629	0.961
SPL	1070	147	12400	10900	232	240	3430	15800	434	14800	908	46000
KNN	0.163	0.102	0.136	0.108	0.324	0.269	0.500	0.413	0.166	0.164	0.113	3.39
MICE	0.129	0.111	0.165	0.162	0.446	0.349	0.291	0.262	0.171	0.139	0.096	0.088
SVD-EM	0.081	0.147	0.143	0.131	0.278	0.220	0.189	0.494	0.101	0.112	0.207	3.187
MF	0.140	0.138	0.204	0.173	0.328	0.232	0.218	0.232	0.145	0.138	0.123	0.115

TABLE 5. AVERAGE NRMSE ACROSS GAP SIZES

		1	4	8	12	16	24	48	72	120	168	216	264	312
Electricity	LIN	0.367	0.491	0.668	0.701	0.852	0.861	3.22	3.51	1.13	10.2	23.0	1.09	16.2
	SPL	0.474	1.165	221	8.913	26.5	234	6990	66000	153	99700	3.90e+06	172000	80700
	KNN	0.781	0.770	0.809	0.792	0.795	0.820	0.846	0.872	1.07	1.73	1.635	1.66	1.24
	MICE	0.800	0.857	0.910	0.932	0.940	0.964	0.961	0.969	0.920	0.961	0.956	0.884	0.964
	SVD-EM	0.690	0.697	0.712	0.714	0.728	0.758	0.723	0.773	1.02	1.04	0.981	1.42	1.49
	MF	0.980	0.925	0.920	0.920	0.896	0.914	0.847	0.899	0.924	0.891	0.859	0.880	0.887
HVAC Operations	LIN	0.208	0.391	0.534	0.640	0.720	0.818	1.08	0.992	1.26	1.76	2.64	2.60	2.36
	SPL	0.225	0.624	5.46	10.9	93.4	488	2150	5630	361	830000	148000	4220	1.01e+06
	KNN	0.656	0.667	0.710	0.730	0.743	0.744	0.834	0.973	1.65	2.84	2.42	2.77	2.73
	MICE	0.945	0.858	0.844	0.886	0.911	0.957	0.970	0.979	0.948	0.945	0.962	0.930	0.882
	SVD-EM	0.776	0.778	0.778	0.766	0.852	0.810	0.813	0.910	1.52	2.34	2.10	2.16	2.17
	MF	0.814	0.826	0.833	0.835	0.850	0.848	0.883	0.866	0.861	0.873	0.836	0.868	0.850
Temperature	LIN	0.118	0.248	0.409	0.486	0.627	0.713	1.20	1.34	3.76	1.17	2.38	5.33	6.56
	SPL	0.118	0.642	3.32	11.6	74.9	6.22	7660	4960	4430	137000	12100	3.90e+06	2.48e+06
	KNN	0.549	0.564	0.573	0.605	0.619	0.607	1.603	1.58	9.21	25.0	20.0	23.6	24.8
	MICE	0.297	0.413	0.503	0.597	0.626	0.697	0.828	0.799	0.855	0.887	0.834	0.917	0.834
	SVD-EM	0.563	0.574	0.594	0.597	0.606	0.652	0.855	5.14	11.4	23.2	20.3	23.6	24.1
	MF	0.762	0.772	0.806	0.817	0.810	0.823	0.861	0.844	0.885	0.893	0.824	0.847	0.899

TABLE 6. STANDARD DEVIATION OF NRMSE ACROSS MISSING PROPORTIONS

		1	4	8	12	16	24	48	72	120	168	216	264	312
Electricity	LIN	0.143	0.114	0.080	0.099	0.178	0.117	2.64	3.75	0.145	7.60	24.2	0.140	18.8
	SPL	0.213	0.510	485	10.8	17.7	278	10900	145000	181000	1.01e+06	7.16e+06	352000	1.80+06
	KNN	0.127	0.113	0.110	0.081	0.095	0.092	0.108	0.145	0.255	0.653	0.655	0.722	0.457
	MICE	0.126	0.126	0.156	0.157	0.173	0.126	0.162	0.184	0.172	0.267	0.239	0.151	0.206
	SVD-EM	0.136	0.147	0.137	0.122	0.139	0.150	0.127	0.119	0.428	0.227	0.415	0.272	0.646
	MF	0.140	0.177	0.160	0.175	0.174	0.138	0.173	0.131	0.146	0.111	0.148	0.190	0.154
HVAC Operations	LIN	0.118	0.154	0.194	0.222	0.257	0.254	0.586	0.391	1.77	3.16	8.39	10.2	10.8
	SPL	0.154	0.673	32.1	80.270	562	3790	15800	53900	1500	6.50e+06	859000	257000	9.83e+06
	KNN	0.224	0.230	0.223	0.235	0.226	0.241	0.413	1.04	2.57	4.63	3.41	4.66	3.97
	MICE	0.874	0.403	0.261	0.271	0.248	0.261	0.262	0.286	0.244	0.252	0.292	0.251	0.249
	SVD-EM	0.329	0.330	0.352	0.213	0.521	0.316	0.237	0.435	1.35	2.45	2.01	2.48	2.68
	MF	0.195	0.187	0.198	0.190	0.189	0.195	0.232	0.210	0.209	0.230	0.216	0.217	0.232
Temperature	LIN	0.080	0.099	0.143	0.175	0.237	0.241	0.961	1.27	11.3	0.565	2.38	11.1	10.5
	SPL	0.056	0.709	8.74	28.0	17.6	13.6	46000	9990	11300	356000	31000	1.58e+07	1.02e+07
	KNN	0.104	0.128	0.086	0.155	0.096	0.093	3.390	2.74	7.75	10.0	7.02	7.75	8.37
	MICE	0.091	0.114	0.102	0.109	0.082	0.094	0.088	0.080	0.112	0.144	0.067	0.158	0.111
	SVD-EM	0.117	0.119	0.157	0.136	0.119	0.164	0.110	0.114	0.104	0.286	0.504	1.13	0.701
	MF	0.103	0.0805	0.0968	0.0673	0.0792	0.098	0.115	0.112	0.116	0.083	0.100	0.133	0.162

TABLE 7. AVERAGE RUNTIMES(S) FOR SINGLE SENSOR IMPUTATION

LIN	SPL	KNN	MICE	SVD-EM	MF
0.03	0.05	0.02	11.83	0.09	6.18

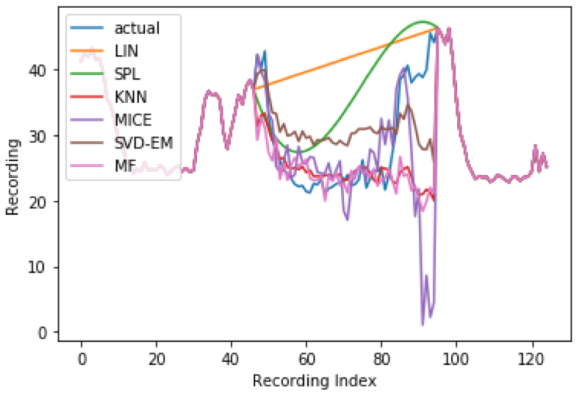

Figure 10. Imputed Values for 1-Day Gap

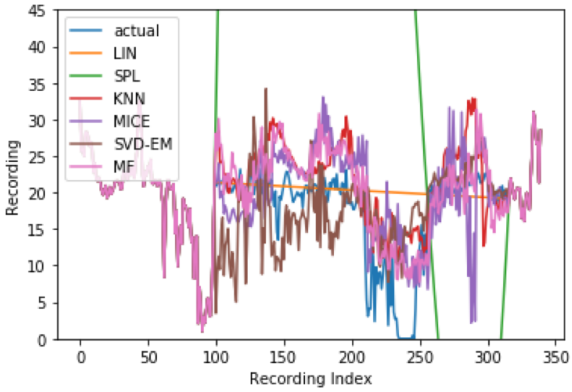

Figure 11. Imputed Values for 4.5-Day Gap

that the two most robust and best performing methods for large missing rates and gap sizes, MICE and MF, require significantly longer computation time. Given similar average NRMSE values and NRMSE standard deviations, MF outperforms MICE with a 48% decrease in average runtime. Furthermore, we are able to further differentiate KNN and SVD-EM through runtime, as they provide similar average NRMSE scores and realistic imputations. As compared to SVD-EM, KNN reduces runtimes by 78%, making it the optimal choice for imputing gaps within a day (gaps of size 48).

5. Conclusion

Data cleaning and gap filling is the prerequisite of applying data analyses to enhance building performance. In this study, we applied and compared six data imputation techniques from the perspective of imputation errors, imputation shape, and computation time under the context of building environment, energy and operational data. We find LIN, KNN, and MF to be the most effective imputation approaches for gap sizes within 4 hours, a day, and 6.5 days (gap sizes 8, 48, 312 respectively). For gap sizes of 4 hours, 1 day, and 6.5 days, LIN, KNN, and MF stand out as the fastest methods that achieve realistic imputations with minimal error, as indicated by the criterion of runtime, imputation plots, and NRMSE.

By testing across different sensor categories, missing rates, and gap sizes, we observe that the effectiveness of the six imputation methods is most sensitive to the gap sizes (the consecutive length of data that is missing). Common univariate time series imputation methods work well when the gap sizes are within 8 consecutive recordings; however, they struggle with larger gaps due to their inability to accurately fit periodic data. By reorganizing time series into matrices, where each row represents a weekly cycle, multivariate imputation methods can far more effectively impute large gaps in the data. While computational expensive algorithms such as MICE and MF require significantly higher computational resources, these methods prove to be the most robust: across different sensor categories, missing rates, and gap sizes, MICE and MF consistently generate imputations that have an average NRMSE of 1 standard deviation for each sensor.

Future work lies in further optimizing the computational expensive methods such as MF with the tuning of hyperparameters, constraints on layer weights, or rank optimization to see reductions in both NRMSE and computational time. Future work also includes further evaluation of these imputing techniques using other building datasets.

Acknowledgements

This work was supported by Laboratory Directed Research and Development (LDRD) funding from Lawrence Berkeley National Laboratory and by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy, under Contract No. DE-AC02-05CH11231. This work also used resources of the National Energy Research Scientific Computing Center (NERSC).

References

[1] J Michael Brick and Graham Kalton. Handling missing data in survey research. *Statistical methods in medical research*, 5(3):215–238, 1996.

[2] Graham Kalton and Daniel Kasprzyk. Imputing for missing survey responses. In *Proceedings of the section on survey research methods, American Statistical Association*, volume 22, page 31. American Statistical Association Cincinnati, 1982.

[3] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[4] William Young, Gary Weckman, and W Holland. A survey of methodologies for the treatment of missing values within datasets: Limitations and benefits. *Theoretical Issues in Ergonomics Science*, 12(1):15–43, 2011.

[5] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002.

[6] Tyler H Ruggles, David J Farnham, Dan Tong, and Ken Caldeira. Developing reliable hourly electricity demand data through screening and imputation. *Scientific Data*, 7(1):1–14, 2020.

[7] Marcilio Cp De Souto, Pablo A Jaskowiak, and Ivan G Costa. Impact of missing data imputation methods on gene expression clustering and classification. *BMC Bioinformatics*, 16(1), 2015.

[8] Siam Rafsunjani, Rifat Sultana Safa, Abdullah Al Imran, Shamsur Rahim, and Dip Nandi. An empirical comparison of missing value imputation techniques on aps failure prediction. *International Journal of Information Technology and Computer Science*, 11(2):21–29, 2019.

[9] Ceylan Yozgatligil, Sipan Aslan, Cem Iyigun, and Inci Batmaz. Comparison of missing value imputation methods in time series: the case of turkish meteorological data. *Theoretical and Applied Climatology*, 112(1-2):143–167, 2012.

[10] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[11] Weiming Zhou, Haifeng Zheng, Xinxin Feng, and Dong Lin. A multi-source based coupled tensors completion algorithm for incomplete traffic data imputation. In *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*, pages 1–6. IEEE, 2019.

[12] Maxim Vladimirovich Shcherbakov, Adriaan Brebels, Nataliya Lvovna Shcherbakova, Anton Pavlovich Tyukov, Timur Alexandrovich Janovsky, and Valeriy Anatol'evich Kamaev. A survey of forecast error measures. *World applied sciences journal*, 24(24):171–176, 2013.

[13] Hoshin Vijai Gupta and Harald Kling. On typical range, sensitivity, and normalization of mean squared error and nash-sutcliffe efficiency type metrics. *Water Resources Research*, 47(10), 2011.

[14] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[15] Garrett Birkhoff and Henry L. Garabedian. Smooth surface interpolation. *Journal of Mathematics and Physics*, 39(1-4):258–268, 1960.

[16] Jason Van Hulse and Taghi M. Khoshgoftaa. Incomplete-case nearest neighbor imputation in software measurement data. *2007 IEEE International Conference on Information Reuse and Integration*, 2007.

[17] Tyler H Ruggles, David J Farnham, Dan Tong, and Ken Caldeira. Developing reliable hourly electricity demand data through screening and imputation. *Scientific Data*, 7(1):1–14, 2020.

[18] G Molenberghs and G Verbeke. Multiple imputation and the expectation-maximization algorithm. *Springer Series in Statistics Models for Discrete Longitudinal Data*, page 511–529, 2005.

[19] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[20] Sun Jing-Tao and Zhang Qiu-Yu. Completion of multiview missing data based on multi-manifold regularised non-negative matrix factorisation. *Artificial Intelligence Review*, pages 1–18, 2020.

[21] Rainer Gemulla, Erik Nijkamp, Peter J. Haas, and Yannis Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 11*, 2011.

[22] Melanie Marazin, Roland Gautier, and Gilles Burel. Blind recovery of k/n rate convolutional encoders in a noisy environment. *EURASIP Journal on Wireless Communications and Networking*, 2011(1), 2011.