ASHRAE and IBPSA-USA SimBuild 2016
Building Performance Modeling Conference
Salt Lake City, UT
August 8-12, 2016

IMPUTATION OF MISSING VALUES IN BUILDING SENSOR DATA

Adrian Chong[1], Khee Poh Lam[1], Weili Xu[1], Omer T. Karaguzel[1] and Yunjeong Mo[1]
[1]Center for Building Performance and Diagnostics, Carnegie Mellon University, Pittsburgh, PA, USA

ABSTRACT

In this paper, we present a comparative study of five methods for the estimation of missing values in building sensor data. The methods that were implemented and evaluated include linear regression, weighted K-nearest neighbors (*k*NN), support vector machines (SVM), mean imputation and replacing missing entries with zero. Using data collected from an actual office building, the methods were evaluated using varying parameter settings. Correlation based feature selection is used to evaluate how using different subsets of attributes may affect each method's performance. We also evaluate the effect of including lagged variables as predictors. To test the robustness of each method, the amount of missing values were varied between 5% and 20%.

INTRODUCTION

Installation of sensors in buildings are becoming increasingly common, collecting large amounts of information that can be used for controls and retrofit analysis. Significant discrepancies between simulated and actual measured consumption of buildings has also brought about a movement to uncover these discrepancies by analyzing data collected through extensive sensor networks. Building data has been used for a variety of building energy model calibration techniques such as Bayesian calibration (Chong and Lam 2015; Heo, Choudhary, and Augenbroe 2012) and systematic evidence-based approaches (Raftery, Keane, and ODonnell 2011). Since the calibration of building energy models usually require time-series data, the estimation of missing values forms an important preprocessing step. Building sensor data is also commonly used in various statistical and data mining techniques to predict a building's energy consumption (Dong, Cao, and Lee 2005; Zhao and Magoulès 2012). However, most statistical and data mining methods cannot be applied to data with missing values, further emphasizing the importance of missing value estimation. The data collected from buildings are usually in the form of matrices consisting of successive measurements (rows) made over a time interval for different variables/sensors (columns). Unfortunately, the data collected often contains missing values which can occur for diverse reasons, including power outages at the sensors, malfunctioning of sensors, or network issues. Three approaches are commonly used to deal with missing values in building sensor data. The first approach is complete case analysis, deleting all instances for which the outcome or any of the inputs are missing. Two issues may arise with this approach (Gelman and Hill 2006):

- If the missing values differ systematically from the completely observed cases, complete-case analysis that ignores missing values would be biased
- If there are many variables, there may be very few complete cases, such that most of the data would be deleted resulting in a very small dataset.

The second approach is mean imputation, where missing values are replaced with the mean value of the variable. However, this approach distorts the distribution of the variable, leading to complications in summary statistics including, the underestimation of the variable's standard deviation (Gelman and Hill 2006). Mean imputation also distort relationships between variables by reducing estimates of correlation towards zero. The third approach is replacing missing entries with zero. This approach also distorts the variable's distribution and can result in large errors when actual values are far from zero.

Although much work has been devoted to the imputation of missing values in other fields, there is no published literature concerning the treatment of missing values in building related data. An overview of different techniques for filling up missing values were investigated in the context of non-response in sample surveys, attrition in longitudinal studies and missing data in designed experiments (Little and Rubin 2014). Common imputation-based approaches in survey practice include mean imputation and linear regression (Little and Rubin 2014). Linear regression and weighted *k*-nearest neighbors has also been evaluated in the estimation of missing values in DNA microarrays (Troyanskaya et al. 2001; Kim, Golub, and Park

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

2005). In this paper, we compare and evaluate five algorithms (linear regression, weighted *k*NN, SVM, mean imputation and replacing with zero) for estimating missing values in building data.

METHOD

The data used for this study is time-series data collected from the 9th floor of Toshiba smart community center (office space) in Kawasaki, Japan (Figure 1). Using the data collected, the performance of each algorithm was evaluated at various percentages (5%, 10%, 15% and 20%) of missing values. Data collection took place between July 24, 2015 and September 24, 2015 at 1 minute intervals. The data collected were averaged to obtain hourly values.

Figure 1: Toshiba smart community center

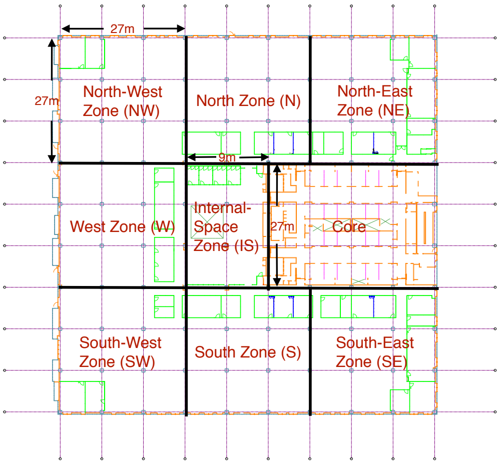

Figure 2: Floor plan of 9th floor of building

Sensor network

The sensor network deployed on the 9th floor of the building can be divided into 9 zones (Figure 2). Each zone (N, NE, S, SW, W, NW, N and IS) aside from the core is installed with sensors that measures zone temperature, lighting, plug and process loads and separate measurements for network related equipment such as routers.
Ventilation is supplied to the N, NE, S, SW, W, NW and N zones by individual variable air volume (VAV) air handling units (AHU). Ventilation to the IS zone is jointly supplied by the AHUs that supplies outdoor air to the N, S and W zones. Table 1 lists the sensors installed in each AHU. Altogether, data from 92 sensors were collected and used for this study.

Table 1: Sensors installed in air handling unit

Description	Unit
Chilled water supply temperature	°C
Chilled water return temperature	°C
Hot water supply temperature	°C
Hot water return temperature	°C
AHU leaving cooling coil air temperature	°C
AHU supply air temperature	°C
Supply fan flow rate	m^3/h
Exhaust fan flow rate	m^3/h
AHU fan power	W

Table 2: Estimating missing values using non-missing values as predictors

Timestamp	Var 1	Var 2	Var 3	Var 4
1/1/2014 0:00	?	0.42	440	455
1/1/2014 1:00	?	?	440	521
1/1/2014 2:00	1.64	0.57	?	?
1/1/2014 3:00	0.97	?	436	451
1/1/2014 4:00	?	?	438	?

Estimation process

Instances (rows) containing missing values were removed to yield 'complete' data sets. The 'complete' data set was separated into two sets: one containing 60 percent of the data for model training, and the other 40 percent of the data for model testing. 5% to 20% of data were deleted at random from the test data sets. Each of the five algorithms were used to estimate missing values by using non-missing values as predictors (Table 2). For example, in Table 2, for the first instance (row), variables 2, 3 and 4 will be used to estimate variable 1. Similarly, for the second instance, variables 3 and 4 will be used to estimate variables 1 and 2 respectively. Since the goal here is not causal inference but simply accurate prediction, it is ac-

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

ceptable to use any inputs as predictors to the algorithm to achieve this goal (Gelman and Hill 2006).

Lagged variables and feature selection

Since observations at time t are likely to be correlated with observations at times $t-1$, $t-2$ and so forth, lagged variables were created and included as predictors to capture the relationship between past and current values. Lags up to 3 timestep into the past were evaluated. For example, in Table 2, each time step would have 8 variables instead of the current 4 if lagged variables (Var 1_{t-1}, Var 2_{t-1}, Var 3_{t-1} and Var 4_{t-1}) at time $t-1$ hour were included. To test the effect of including lagged variables, each algorithm's performance on the following datasets were compared: data with no lagged variables, data including lagged variables of 1 time step ($t-1$), data including consecutive lagged variables up to 2 time steps ($t-1$ and $t-2$) and data including consecutive lagged variables up to 3 time steps ($t-1$, $t-2$ and $t-3$).

Feature selection also known as variable selection is the process of selecting a subset of relevant features for use as predictors. As part of this study, feature selection was carried out to determine how different subsets of predictors may affect each algorithm's accuracy. This was carried out by selecting k attributes that has the highest correlation with the target variable. We test the performance of each algorithm using different subsets (2^1, 2^2, 2^3, 2^4, 2^5, 2^6 and 2^7) of attributes for the prediction. For example, when working with a subset containing 2^3 attributes, we select 8 predictors that has the highest correlation with the output variable. These 8 predictors are then used as predictors for model construction.

The metric used to evaluate accuracy of estimation is the normalized Root Mean Squared difference (NRMSD) between the predicted and actual values (Equation 1) and average NRMSD which is the mean NRMSD across all variables (Equation 2). n denotes the number of missing entries in a variable; y and $\hat{y}$ denotes the true and predicted value respectively; m denotes the total number of variables in the data set.

$$NRMSD = \frac{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y)^2/n}}{y_{max} - y_{min}} \tag{1}$$

$$Average\ NRMSD = \frac{\sum_{i=1}^{m} NRMSD_i}{m} \tag{2}$$

Linear regression

Linear regression is a commonly used method where the scalar output (dependent variable) is a linear function of the predictors/independent variables (Equation 3). Using ordinary least squares, this method estimates the unknown parameter β by minimizing the sum of squared differences between predicted and actual values.

$$y_i = x_i^T \beta + \varepsilon_i \tag{3}$$

y_i denotes the actual values for observation i; x_i denotes the value of the predictors for observation i; and ε_i denotes the prediction error for observation i.

Weighted k-Nearest Neighbors (kNN)

The nearest neighbor algorithm (Cover and Hart 1967) is a nonparametric method that assigns a value to the observation based on the values of the set of points nearest to it. In the kNN algorithm, k nearest points are selected and used for the prediction, and the influence is the same for each of these neighbors (Hechenbichler and Schliep 2004). The nearest neighbors are determined using the Euclidean distance. Since the nearest neighbors are selected using a distance function, the scale of each variable can influence the estimation. As a result, variables on a larger scale will dominate the distance calculation. This is especially in building data where different variables can have significantly different scale. To overcome this, the data is normalize to the same interval $[0,1]$ by subtracting the minimum value and dividing by its range ($z_i = \frac{x_i - min(x)}{max(x) - min(x)}$). In the weighted kNN algorithm, the distances used to select the nearest neighbors are transformed into similarity measures, which are used as weights (Hechenbichler and Schliep 2004). Hence, closer neighbors have higher weights and greater influence on the prediction. We examine the effect of different values of the parameter k on estimation accuracy.

Support vector machines (SVM) for regression

The algorithm used is ε-support vector regression (ε-SVR) with radial basis function (RBF) kernel (Chang and Lin 2011). Consider a set of data points, $\{(x_1, z_1), ..., (x_l, z_l)\}$, where $x_i \in R^n$ is a feature vector and $z_i \in R^1$ is the target output. Under given parameters $C > 0$ and $\varepsilon > 0$, ε-SVR can be defined by

$$\min_{\omega, b, \xi, \xi^*} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^*$$

subject to

$$\begin{aligned} \omega^T\phi(x_i) + b - z_i &\leq \varepsilon + \xi_i \\ z_i - \omega^T\phi(x_i) - b &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\leq 0, i = 1, ..., l \end{aligned} \tag{4}$$

We examine the effect of different values of SVM hyperparameters epsilon ε and cost C on estimation accuracy. The value of ε in the ε-insensitive loss function affects the number of support vectors used to construct the regression function. An increase in ε decreases the number of support vectors that are selected. C is the regularization parameter that determines the tradeoff between maximization of margin and minimization of errors of the SVM on the training data. Hence both ε and C affects model complexity.

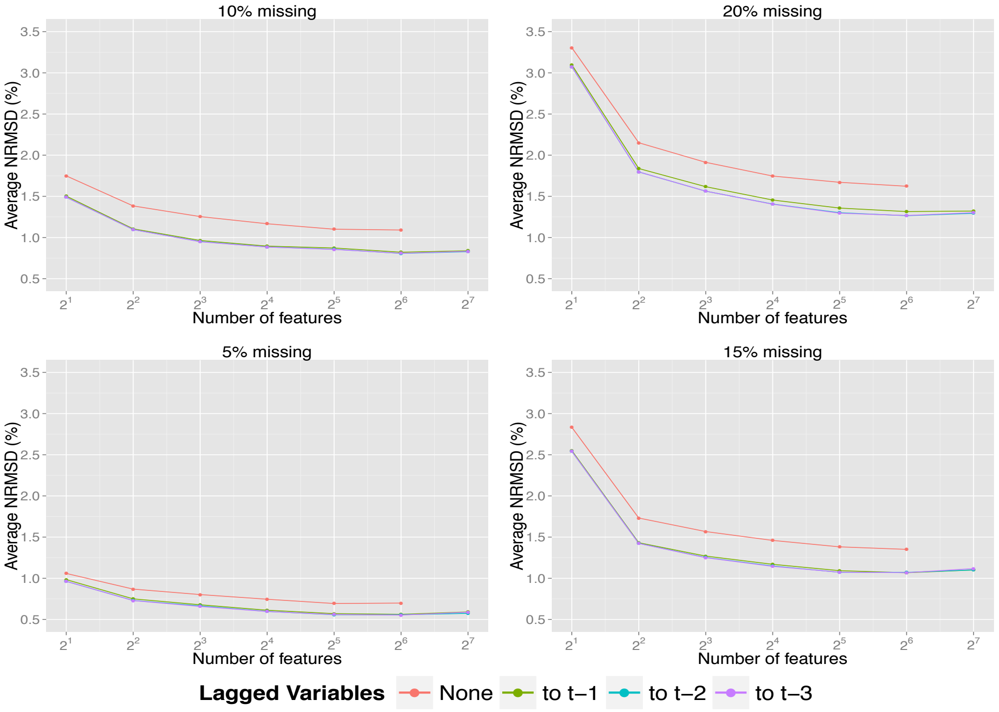

Figure 3: Effect of number of features and including lagged variables with linear regression. Different plots correspond to data sets with different percent of missing entries. Different curves within each plot corresponds to data sets with different number of lagged variables.

DISCUSSION AND RESULT ANALYSIS

Performance of using each algorithm to estimate missing values was assessed over different data sets (both percentage of missing entries and number of periods of lagged variables) and over different number of features/attributes. Note that when no lagged variables are included, there are only 91 attributes available for predicting the target variable (original dataset contains 92 variables/columns). Hence, we do not test the algorithms performance with 2^7 attributes when lagged variables equal none.

Linear regression imputation

Using linear regression to estimate missing values is improved by including lagged variables from time $t-1$ as predictors (Figure 3). This trend can be observed across different percentages of missing values and number of features. This means that regardless of the number of predictors that were used, including lagged variables from time $t-1$ improves the algorithm's performance. However, including more lagged variables ($t-2$ and $t-3$) shows minimal improvements in accuracy. Figure 3 also shows that across different percentages of missing values, including more than 2^6 attributes might result in overfitting and reduce overall accuracy. This method of imputation is accurate with estimated values having on average $0-3.5\%$ deviation from the true values, depending on the number of periods of lagged variables and the percentage of missing entries (Figure 3).

*k*NN imputation

*k*NN estimation was evaluated using different number of features and number of nearest neighbors k. The most accurate estimation is achieved when $k \approx 6$ and approximately 2^3-2^4 features are used for the estimation. Figure 4 shows the performance of using *k*NN estimation with 2^4 features over different values of k and for data sets with different percentages of missing entries. This method of imputation is relatively insensitive to the exact value of k within the range of $2-10$. Within this range, the difference in average NRMSD is within approximately 0.1% (Figure 4). Notably, the performance of *k*NN declines when low number of nearest neighbors ($k<3$) are used for the estimation, probably due to overfitting. Performance of the algorithm also declines when large number of nearest neighbors ($k>10$) are used for the prediction, indicating that some important details are being smoothed out. Performance of *k*NN is also relatively insensitive to the exact number of attributes (used for the prediction) within the range of 2^3-2^5 (Figure 5). Within this range, change in NRMSD is below 0.1%. This is consistent across data

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

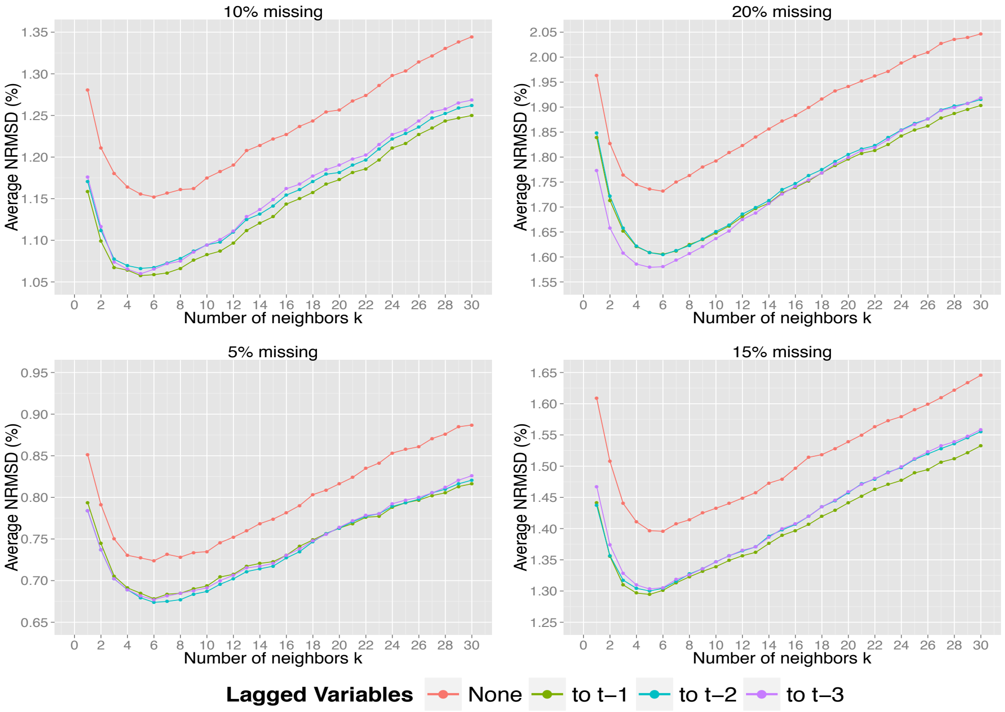

Figure 4: Effect of number of nearest neighbor k used for kNN estimation with 2^4 features. Different plots correspond to data sets with different percent of missing entries. Different curves within each plot corresponds to data sets with different number of lagged variables.

sets containing less missing entries (5%, 10% and 15%). Similar to using linear regression, performance of using *k*NN to estimate missing values is reduced by including lagged variables from time $t-1$ hour as predictors (Figures 4 and 5). However, including more lagged variables ($t-2$ and $t-3$) shows minimal improvements in accuracy. This trend is consistent across different number of features used for the imputation (Figure 5).

SVM imputation

SVM imputation was evaluated using different number of features and hyperparameter (ε and C) values (Equation 4). One way to choose appropriate values for ε and C is to use k-fold cross-validation (Hsu et al. 2003). The most accurate estimation is achieved when $\varepsilon \approx 2^{-7}$, $C \approx 2^5$ and approximately $2^4 - 2^5$ features were used for the imputation. Using SVM for imputation is very accurate after tuning hyperparameters ε and C, showing approximately 1.25% deviation from the true values for the data set with 20% missing entries (Figure 6). Average NRMSD is the lowest when with $\varepsilon = 2^{-7}$ and cost $C = 2^5$. Performance of SVM imputation declines when a larger value of ε and C is used for the estimation. Large values of C places more weight on the minimization of errors on the training data, resulting in overfitting and poorer generalization.

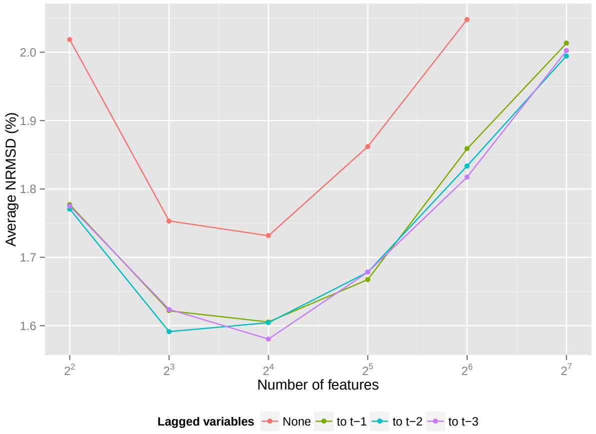

Figure 5: Effect of number of features used for kNN estimation on data set with 20% missing entries. Estimation was carried out using kNN with $k = 2^4$. Different curves corresponds to data sets with different number of lagged variables.

Smaller values of ε result in better performance. However, performance starts to decline as ε is decreased beyond 2^{-7}. This trend observed in Figure 6 is representative of that observed across data sets with different percentages (5%, 10%, 15% and 20%) of missing entries.

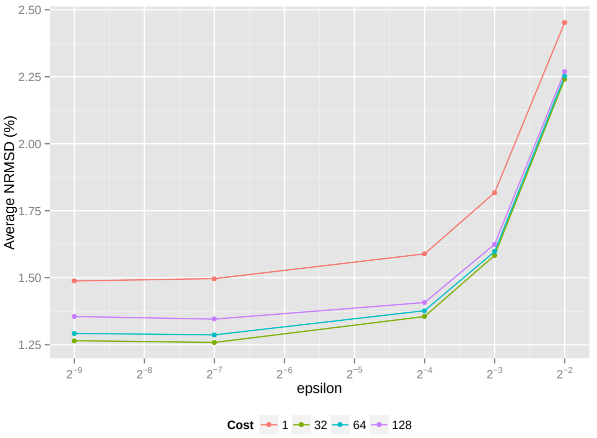

Figure 6: Effect of hyperparameters epsilon ε and Cost C used for SVM Imputation on data set with 20% missing entries. Estimation was carried out using SVM imputation with 2^5 features. Different curves corresponds to data sets with different number of lagged variables.

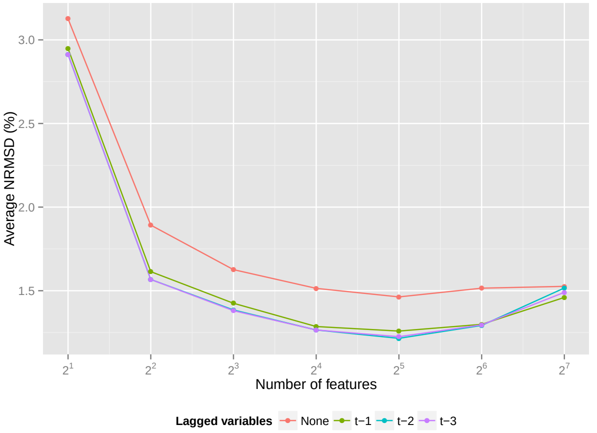

Figure 7: Effect of number of features used for SVM Imputation on data set with 20% missing entries. Estimation was carried out using SVM Imputation with $\varepsilon = 2^{-7}$ and $C = 2^5$.Different curves corresponds to data sets with different number of lagged variables.

Mean imputation and replacing with zero

Replacing missing values with the attribute's average or with zeros are common preprocessing methods before the data is used in various statistical techniques or as inputs to building simulation models. Mean imputation, although a drastic improvement from replacing with zeros has significantly lower accuracy than either linear regression, *k*NN or SVM imputation (Table 3).

Table 3: Performance of mean imputation and replacing missing values with zero.

	Average NRMSD (%)	
Percentage Missing	Mean Imputation	Replacing with zero
5%	6.07	28.25
10%	8.25	39.35
15%	10.62	49.93
20%	11.96	56.19

Comparing algorithms

Replacing missing values with zeros shows the largest spread of NRMSD with a minimum NRMSD of 11.3% and a maximum NRMSD of 321.7% (Figure 8). Mean imputation shows drastically better performance with 91 (of the 92) variables estimated with NRMSD under 20% as compared to 22 (of the 92) variables if missing entries were replaced by zero. Estimation with linear regression, *k*NN or SVM yielded significantly better accuracy than mean imputation, with all variables being estimated with NRMSD under 6%. Linear regression was used with lag variables from period $t-1$ as predictors and with 2^5 attributes selected using a correlation-based feature selection that selects the attributes with the highest correlation with the target variable. *k*NN imputation was carried out with $k = 6$, lag variables from period $t-1$ and 2^4 attributes (selected via feature selection). SVM imputation was carried out using the following parameter setting: $\varepsilon = 2^{-7}$ and $C = 2^5$; lag variables from period $t-1$; and 2^5 attributes (selected via feature selection). When individual algorithms are considered at their optimal parameter settings, using SVM for estimating missing entries is more accurate than other methods such as linear regression and *k*NN. This can be observed from Figure 8 where the distribution of errors for SVM is more right-skewed as compared to *k*NN and linear regression. When errors for individual variables are considered, 25% of the variables were estimated within 0.5% of its true value with SVM at the above mentioned parameter setting. With linear regression and *k*NN, only 11% and 1% of the variables were estimated with NRMSD under 0.5% respectively. Notably, maximum NRMSD is also lower with SVM imputation (4.2%) as compared to linear regression (4.8%) and *k*NN imputation (5.7%). Overall, SVM imputation shows consistent performance across different types of data. For this data set, accuracy using linear regression is comparable and sometimes better because this data set is made up by measurements from sensors measuring the same property in different zones, thus the linear relationship between many variables. For example, lighting power measure-

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.

ments are expected to have the same trend across different office spaces in the building.

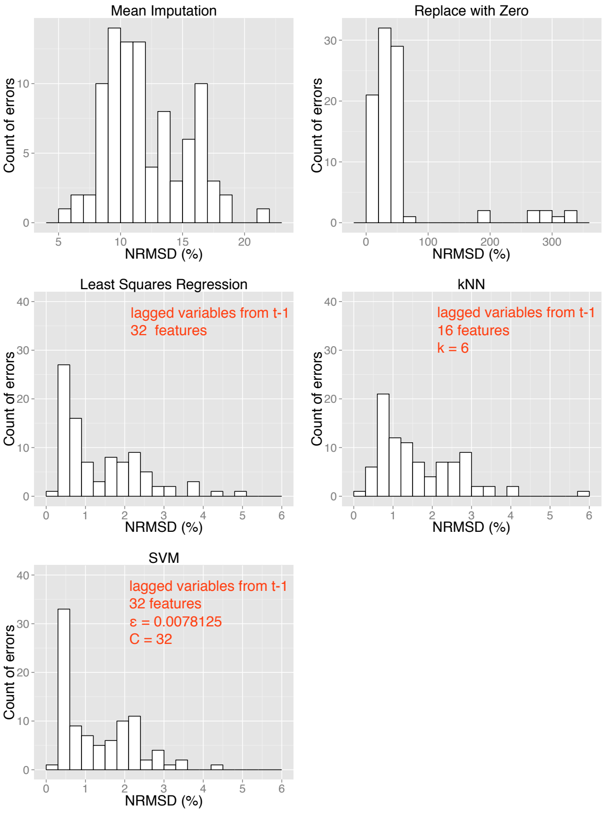

Figure 8: Distribution of errors for different algorithms at their optimal parameter setting for data set with 20% missing entries.

Figure 9 shows the NRMSD of each of the 92 attributes for linear regression, *k*NN and SVM imputation. Parameter settings for each algorithm were as mentioned in the previous paragraph. Linear regression and SVM have comparative performance for attributes 1 to 70 (mostly lighting, equipment and network power measurements) with *k*NN having lower accuracy in general. However, linear squares regression shows significantly higher NRMSD for some attributes between attributes 70 and 92. These attributes contain measurements from AHU sensors. In particular, air and water temperature measurements within the AHU seems to show poorer performance with linear regression. SVM with RBF kernel shows better accuracy for these attributes probably because the relationship between the target and dependent variables are non-linear.

Simulation

We evaluate the impact of missing value imputation in actual application by comparing EnergyPlus hourly simulation output with actual measurements. The objective is to illustrate how different methods of imputation may affect model output, giving a false sense of accuracy. This is especially true in building energy models since these tools usually requires inputs at an hourly resolution. Data from July 24, 2015 to August 31, 2015 was used for model training and data from September 1, 2015 to September 24, 2015 was used as the test data set. 20% of data was randomly removed from the test data set. Missing values were imputed using different methods before they are used as inputs to the EnergyPlus building energy simulation model. The simulation output that was evaluated are the lighting, equipment and network energy consumption (Table 4). Accuracy was evaluated using coefficient of variation of the root mean square error (CVRMSE) and normalized mean bias error (NMBE), metrics that are commonly used to evaluate calibrated building energy simulation models. According to (ASHRAE 2002), if hourly calibration data are used, CVRMSE and NMBE shall be below 10% and 30% respectively. Both mean imputation and replacing with zero are common preprocessing methods before data is used for simulation. Replacing with zero tends to underestimate energy consumption as observed from its negative NMBE (Table 4). This is expected since power measurements are typically positive. Mean imputation yielded results that are significantly better accuracy (both CVRMSE and NMBE). However, imputation with either linear regression, *k*NN or SVM yielded simulation results that is significantly closer to actual values as compared to both methods (Table 4).

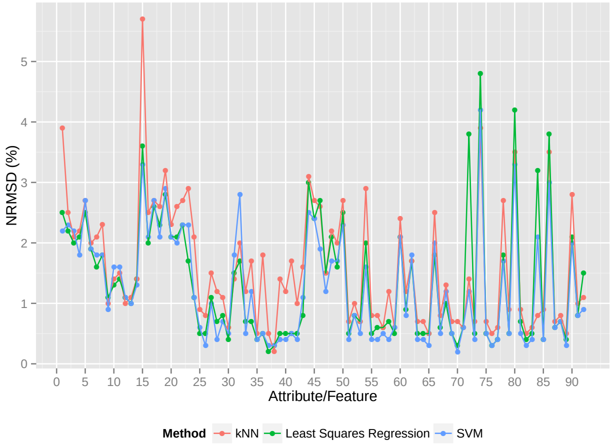

Figure 9: NRMSD of each of the 92 attributes with different imputation method at their optimal parameter settings.

CONCLUSION

The objective of this paper is to illustrate the importance of predicting missing values and how it may affect the

Table 4: CVRMSE and NMBE of simulation with data imputation by different methods.

	Lighting		Equipment		Power Network	
Method	CVRMSE	NMBE	CVRMSE	NMBE	CVRMSE	NMBE
Replacing with zero	30.4%	-18.7%	29.0%	-19.7%	31.1%	-20.4%
Mean imputation	21.3%	1.46%	8.08%	-0.35%	15.1%	-0.45%
Linear regression	2.46%	0.08%	2.18%	-0.04%	0.62%	-0.04%
*k*NN	2.99%	-0.27%	2.26%	0.10%	0.75%	-0.06%
SVM	2.49%	-0.24%	2.12%	-0.08%	0.68%	-0.08%

accuracy of building energy simulation. This paper has shown that linear regression, *k*NN and SVM are more accurate for estimating missing values in building sensor data as compared to replacing with zero or mean imputation. Linear regression shows better accuracy when there is a linear relationship between the target and dependent variables. On the contrary, SVM shows better accuracy when this relationship is non-linear. All three methods show significant improvements over replacing with zero or mean imputation by taking advantage of the patterns and relationships with other variables. Based on the results of this study, we recommend SVM imputation because of its ability to model non-linear relationships. However, where there are many sensors measuring the same property in different zones, linear regression shows comparable and sometimes better performance. We also recommend the inclusion of lagged variables from time $t-1$ as predictors for better performance.

ACKNOWLEDGMENT

The authors would like to thank Toshiba Corp and NREG Toshiba Building Co., Ltd. for kindly supplying the data used in this paper.

REFERENCES

ASHRAE. 2002. "Guideline 14-2002, Measurement of Energy and Demand Savings." *American Society of Heating, Ventilating, and Air Conditioning Engineers, Atlanta, Georgia*.

Chang, Chih-Chung, and Chih-Jen Lin. 2011. "LIBSVM: A library for support vector machines." *ACM Transactions on Intelligent Systems and Technology* 2:27:1–27:27. Software available at `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Chong, Adrian, and Khee Poh Lam. 2015. "Uncertainty analysis and parameter estimation of HVAC systems in building energy models." *Proceedings of the 14th IBPSA Building Simulation Conference*.

Cover, Thomas M, and Peter E Hart. 1967. "Nearest neighbor pattern classification." *Information Theory, IEEE Transactions on* 13 (1): 21–27.

Dong, Bing, Cheng Cao, and Siew Eang Lee. 2005. "Applying support vector machines to predict building energy consumption in tropical region." *Energy and Buildings* 37 (5): 545–553.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Hechenbichler, Klaus, and Klaus Schliep. 2004. "Weighted k-nearest-neighbor techniques and ordinal classification."

Heo, Yeonsook, Ruchi Choudhary, and Godfried Augenbroe. 2012. "Calibration of building energy models for retrofit analysis under uncertainty." *Energy and Buildings* 47:550–560.

Hsu, Chih-Wei, Chih-Chung Chang, Chih-Jen Lin, et al. 2003. "A practical guide to support vector classification."

Kim, Hyunsoo, Gene H Golub, and Haesun Park. 2005. "Missing value estimation for DNA microarray gene expression data: local least squares imputation." *Bioinformatics* 21 (2): 187–198.

Li, Qiong, Qinglin Meng, Jiejin Cai, Hiroshi Yoshino, and Akashi Mochida. 2009. "Applying support vector machine to predict hourly cooling load in the building." *Applied Energy* 86 (10): 2249–2256.

Little, Roderick JA, and Donald B Rubin. 2014. *Statistical analysis with missing data*. John Wiley & Sons.

Raftery, Paul, Marcus Keane, and James ODonnell. 2011. "Calibrating whole building energy models: An evidence-based methodology." *Energy and Buildings* 43 (9): 2356–2364.

Troyanskaya, Olga, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. "Missing value estimation methods for DNA microarrays." *Bioinformatics* 17 (6): 520–525.

Zhao, Hai-xiang, and Frédéric Magoulès. 2012. "A review on the prediction of building energy consumption." *Renewable and Sustainable Energy Reviews* 16 (6): 3586–3592.

This is a work of the U.S. Government and is not subject to copyright protection in the United States. Foreign copyrights may apply.