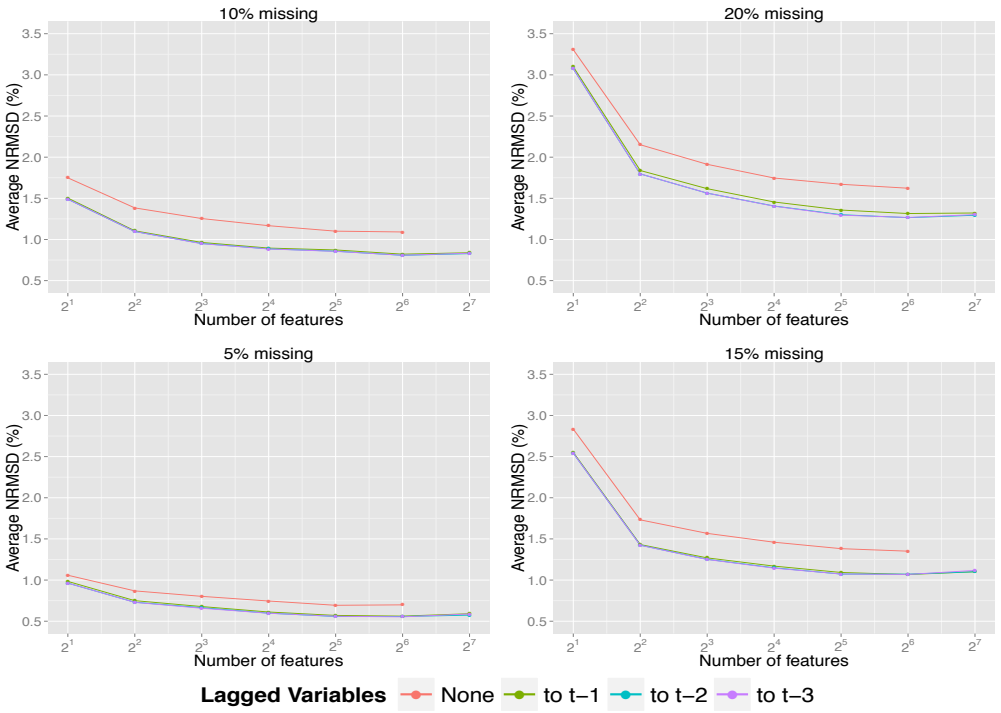

Figure 3: Effect of number of features and including lagged variables with linear regression. Different plots correspond to data sets with different percent of missing entries. Different curves within each plot corresponds to data sets with different number of lagged variables.

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

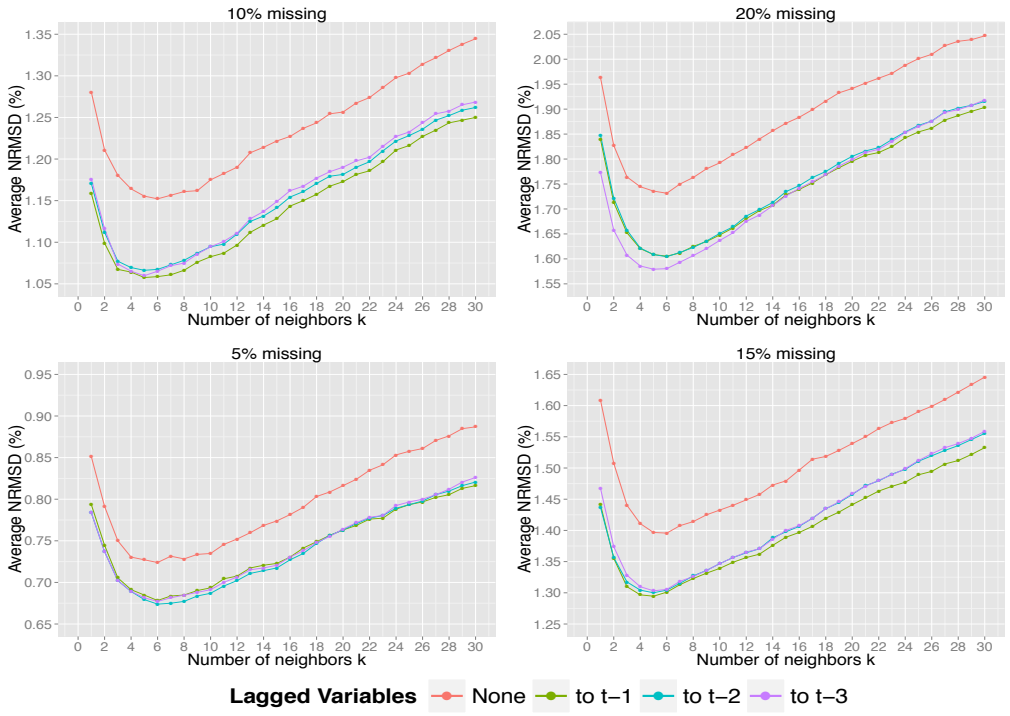

Figure 4: Effect of number of nearest neighbor k used for kNN estimation with 2^4 features. Different plots correspond to data sets with different percent of missing entries. Different curves within each plot corresponds to data sets with different number of lagged variables.

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.

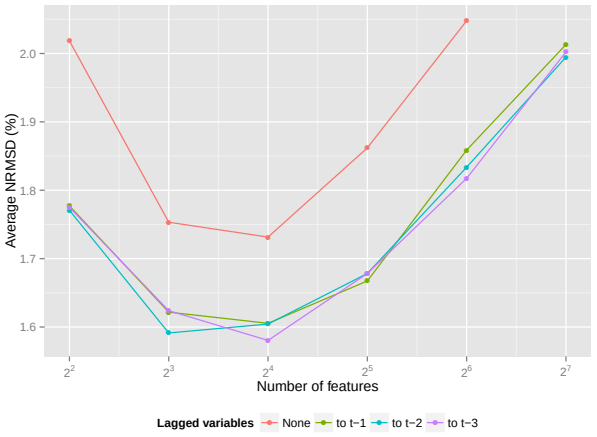

Figure 5: Effect of number of features used for kNN estimation on data set with 20% missing entries. Estimation was carried out using kNN with $k = 2^4$. Different curves corresponds to data sets with different number of lagged variables.

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.

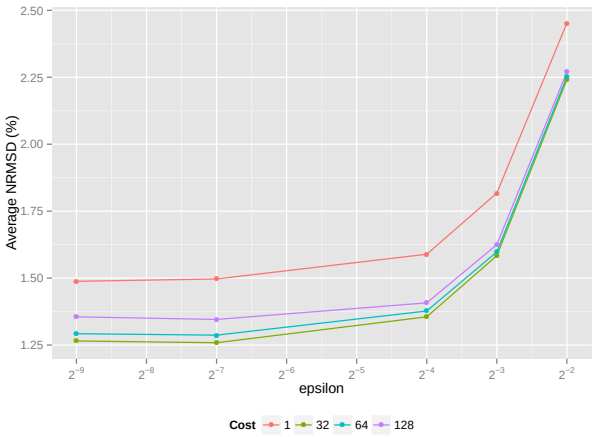

Figure 6: Effect of hyperparameters epsilon ε and Cost C used for SVM Imputation on data set with 20% missing entries. Estimation was carried out using SVM imputation with 2^5 features. Different curves corresponds to data sets with different number of lagged variables.

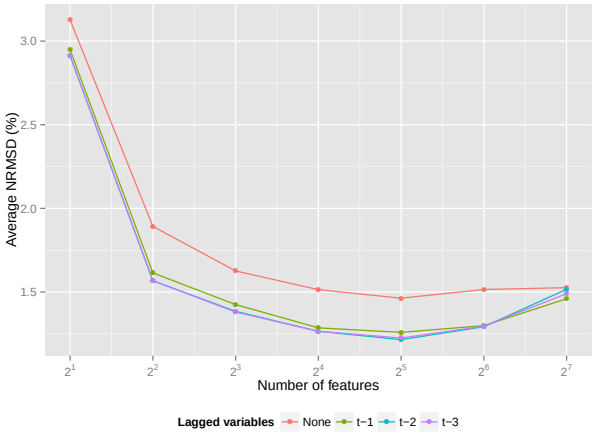

Figure 7: Effect of number of features used for SVM Imputation on data set with 20% missing entries. Estimation was carried out using SVM Imputation with $\varepsilon = 2^{-7}$ and $C = 2^5$.Different curves corresponds to data sets with different number of lagged variables.

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-

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

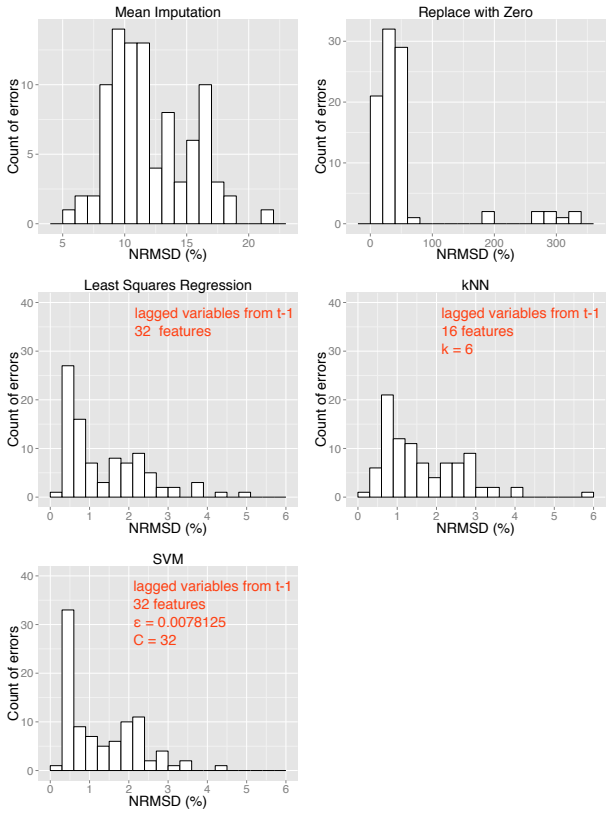

Figure 8: Distribution of errors for different algorithms at their optimal parameter setting for data set with 20% missing entries.

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

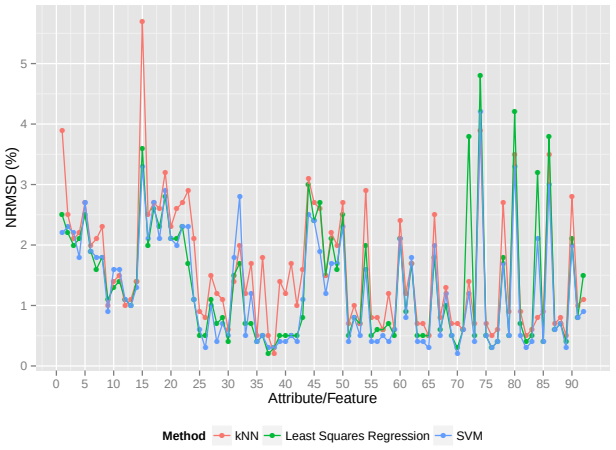

Figure 9: NRMSD of each of the 92 attributes with different imputation method at their optimal parameter settings.

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).

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.

