Effective Missing Value Imputation with Matrix Factorization

Brian Cho[1], Alexander Sim[2], John Wu[2]

[1]Yale University, [2]Lawrence Berkeley National Laboratory

Abstract

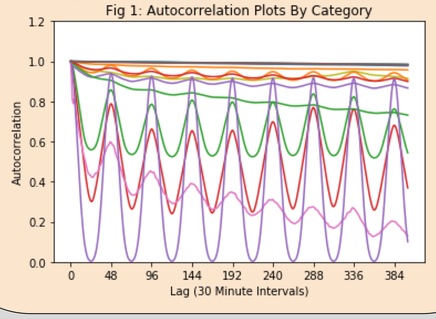
Inconsistencies in sensors cause a significant number of missing time series values in various fields, including the monitoring of building energy use. While methods such as mean imputation and interpolation are common, new methods of imputation based on recommendation systems allow for novel ways to impute time series data through matrix reorganization of time series. In this study, we empirically compare common imputation methods with imputation with multiple matrix imputation techniques, and further optimize the accuracy of imputation through altering methods of preprocessing and hyper-parameters.

Background

- Data collected from a large office building housing a large set of computer systems and offices, with 269 different sensors in 16 distinct sensor categories
- Each sensor recorded in 30 minute intervals over a 2 year time span

- **Initial Characteristics of the Data:**
  - Missing Value Proportion: Over 90% of the sensors are missing between 20-40% of their recordings
  - Gap Sizes: Of the gaps present in the data, 98.2% are within 1 week (336 recordings) in length
  - Periodicity:
    - Most sensor categories demonstrate correlation with values a multiple of 48 (length of a day) recordings apart

Fig 1: Autocorrelation Plots By Category



Research Question

What methods of imputation most effectively fill the missing values of time series data?

Results: Method Comparison

- How similar does the imputed data look compared to the actual data?
- How different are the imputed and actual values (N-RMSE)?

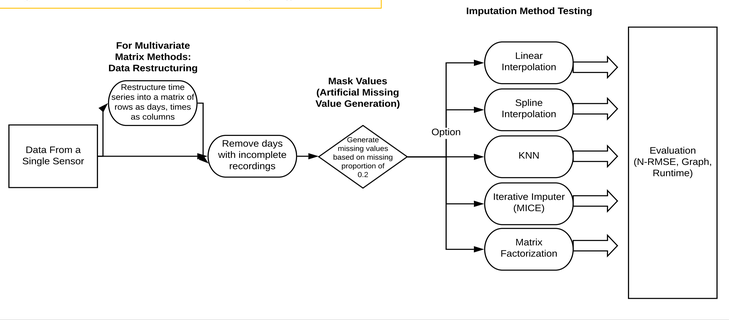
Figure 3: Error Score Metric N-RMSE

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$NRMSE = \frac{RMSE}{\bar{y}}$$

Preprocessing Method: Imputation Comparison

Figure 2: Method of Testing Imputation

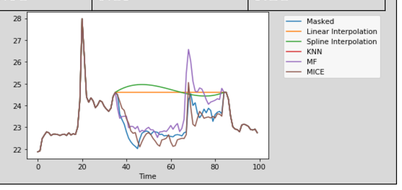


Results: Comparison of Methods

Figure 4: Mean N-RMSE Scores Across Categories

Periodic	Common Methods		Matrix Methods		
	Linear	Spline	KNN	MICE	**MF**
Yes	**2.86**	7480	1.55	1.15	**1.07**
No	**0.17**	2990	0.81	0.15	**0.14**

Figure 5: Plot of Imputed Values On Masked Dataset (Gap = 1 Day)



Optimization: Gap Length

By considering the length of gaps, we preprocess the data with l = 48 (day), 144 (3 day), 336 (week) to determine the best method of reorganization to fill a given gap length via MF.

Figure 6: Averaged N-RMSE Over Sensor Categories

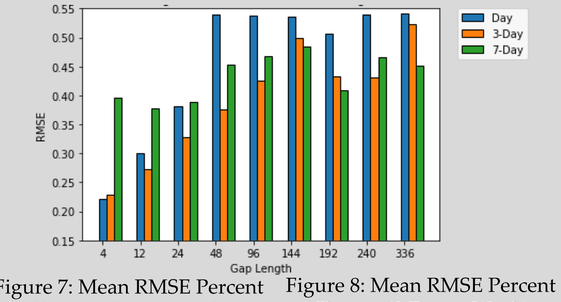


Figure 7: Mean RMSE Percent Change (Days - Week)

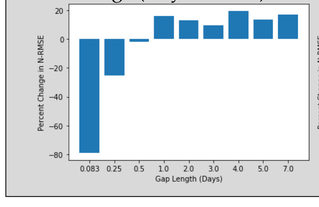
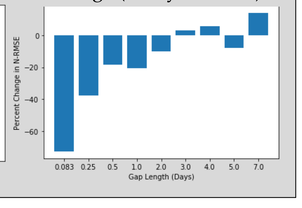


Figure 8: Mean RMSE Percent Change (3 Days - Week)



Further Optimization: Compression and Factorization Methods

Figure 9: Error Values for K-Rank Approximation

Rank	Mean N-RMSE		
	Day	3 Day	Week
Rank 4	0.215	0.231	0.264
¼ of Full Rank	0.205	0.229	0.235
½ of Full Rank	0.208	0.227	0.236
¾ of Full Rank	0.206	0.223	0.236
Full Rank	0.207	0.207	0.230

Figure 10: Plot of Imputed Values of Various Rank Approximation

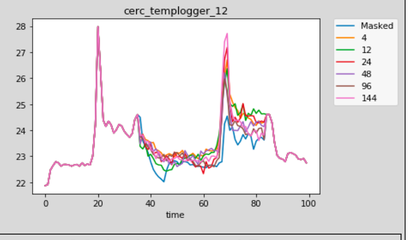


Figure 11: Factorization Method Errors and Runtimes

Method	N-RMSE	Runtime (s) (Median)
SGD: Control	0.428	96.0
Singular Value Decomp.	0.856	1.1
SGD, higher learning rate	0.420	32.8
SGD, low epochs	0.407	69.0

Conclusions

- **Best Method: Matrix Factorization via Gradient Descent**
- Matrix methods are far better with periodic data, slightly better with nonperiodic data
- MF yields the best results across periodic and nonperiodic time series data, both in terms of error score and visually (Figure 4, 5)
- Length of the gap becomes larger than the length of a row: more effective to impute values with longer row organization
- Gaps sized significantly smaller than row lengths filled poorly, leaving future work to be done on finding the optimal length of a row in respect to the gap size
- Low-rank approximations are effective for imputation, but take longer with SGD algorithm
- Optimal runtimes: by increasing learning rate and decreasing the number of epochs, we decrease runtime by 65% with only an 8% error rate increase

Further Reading

For further information and reading, please scan the QR code below.

Acknowledgements

This work was also supported by the Office of Advanced Scientific Computing Research, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, and also used resources of the National Energy Research Scientific Computing Center (NERSC). I would additionally like to thank my mentors Alex Sim and John Wu for their incredible guidance throughout this project, and Teresa Dayrit, Yuan Gao, and Zhe Walter Wang for work done on this dataset.