# Introduction

Missing data is a common occurrence in time-series data, for this specific case causes include faulty sensors or errors in data storage. Missing data can cause downstream applications to malfunction and can thus have serious consequences. Missing data in Building Management Systems (BMS) can cause underperforming building services (e.g.lower comfort or higher power usage) in the worst case a building breakdown as control decisions are based on the collected time series data.

There are limited studies to clarify on how to deal with missing data in BMS datasets. Previous research that has been done has either been limited to only imputing energy usage [] and or is aimed at larger buildings such as school buildings [2]. The intention of this paper is to build on top of previous research into BMS time series data and highlight imputation methods suited for terraced houses and the other data held by the systems.

Methods tested in this paper have been obtained from previous studies that evaluate and compare the performance of imputing time series data. Whilst hot deck has been outperformed by machine learning in past studies by machine learning-based methods [3], it is applicable in this study due to the amount of similar data available. Versions of Recurrent Neural Networks (RNN) show promising results in previous studies when imputing multivariate time series data []. KNN regression has shown impressive results in previous studies [2] despite it being a relatively simple algorithm. Imputation by use of mean, mode or median is also tested to form a baseline of imputation performance.

The aim of this paper is to compare the performance of three methods: K Nearest Neighbor, Recurrent Neural Networks and Hot Deck to impute missing BMS time series data. The imputation performance has been evaluated making use of various criteria to facilitate the choice of the most suitable method for each scenario.

The method section contains a description of the datasets used to evaluate the imputation methods’ performance and a description of the selected imputation methods and imputation performance criteria. The result section presents evaluation results for each imputation method per gap and each of the seven selected columns.

# Method

## Datasets

Building management systems store data from the building such as fluid temperature, flow rate and working mode and outdoor weather information like solar radiation and temperature. In order to validate the performance of the imputation methods two data sets have been used, twenty-five weather stations from the Royal Netherlands Meteorological Institute (KNMI) and BMS data of hundred-twenty residential Net-Zero energy houses. The NZEB BMS time series data set contains data over 2019 and is supposed to have five-minute interval data measurements (105096 rows). The KNMI data set contains over the years 2018 to 2020 from January first to December thirty-first and is measured at hourly intervals (17545 rows). The only change made to the datasets was converting the timestamps to Python Date Time objects.

## Columns selected

For the efficiency of research seven columns are selected across the data sets to evaluate imputation performance. The selected features from the NZEB BMS data set are power usage, heat pump operation mode (op\_mode), heat pump flow temperature and C02 sensor C02 measurements. KNMI columns that were selected for imputation are solar radiation, temperature and humidity. The NZEB BMS columns are selected based on having every single data classification in the research for evaluating the difference in performance on. The KNMI columns were selected for being correlated to each other which is confirmed using the Pearson method.

**Table 1. Imputed columns, columns with Dataset origin, unit of measurement and classification**

|  |  |  |  |
| --- | --- | --- | --- |
| **Column name** | **Device** | **Unit of measurement** | **Classification** |
| Temperature | KNMI | C (in 0.1c) | Interval |
| Global Radiation | KNMI | j per cm2 | Ratio |
| Humidity | KNMI | % | Ratio |
| Flow\_temp (BMS) | Alklima Heat Pump | C | Interval |
| op\_mode (BMS) | Alklima Heat Pump | 0-6 modes | Nominal |
| Power (BMS) | Smartmeter | W | Ratio |
| C02 (BMS) | C02 Sensor | PPM | Ratio |

## Evaluation method

In order to evaluate the performance of imputation methods under the same reproducible conditions, a pipeline has been developed. The pipeline performed the following tasks: loading data, creating gaps, imputing the artificial gaps, calculating imputation performance criteria and storing evaluation results. The pipeline code and trained models can be found in the appendix of this paper. No changes are made to the KNMI and BMS nZEB datasets.

### Gap creation

To evaluate the performance of each imputation method artificial gaps are created in both datasets. The gaps come in different sizes to evaluate the performance of each imputation method over various sizes. Gaps are created along the rules listed in the table below and are generated using a set random seed. The random seed is also made use of for gap locations and the actual gap size.

**Table 2. NZEB BMS gap sizes with minimum size, maximum size and percentage of data missing**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nr.** | **Min\_size** | **Max\_size** | **% Of data** |
| 1 | 5 min | 60 min | 15 |
| 2 | 1 hour | 6 hours | 4 |
| 3 | 6 hours | 24 hours | 1.5 |
| 4 | 24 hours | 72 hours | 0.5 |
| 5 | 72 hours | 168 hours | 0.01 |

**Table 3. KNMI gap sizes, minimum size, maximum size and percentage of data missing**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nr.** | **Min\_size** | **Max\_size** | **% Of data** |
| 1 | 1 hour | 6 hours | 15 |
| 2 | 6 hours | 24 hours | 5 |
| 3 | 24 hours | 72 hours | 1.5 |
| 4 | 72 hours | 168 hours | 0.005 |

### Evaluation

Imputation methods use the now incomplete data for imputation and after completion, it returns the imputed data. The imputed data is then compared with the original data and gets evaluated based on the selected performance metrics <Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) >. After evaluating the performance indicators get stored on the pipeline for comparison.

## Imputation methods

Three imputation methods have been applied during this research: Hot Deck, Recurrent Neural Networks and KNN The methods have been selected through recommendations from previous research and to facilitate the characterizations and advantages as well as disadvantages of each method.

WHICH PAPERS AND WHY DISADVANTAGES NEED TO BE HERE TOO FROM PAPER!

### Hot Deck

#### Introduction to hot-deck

Hot-deck imputation is a method for handling missing data in which each missing value from a recipient is replaced with an observed value from a similar unit (the donor). This method applies perfectly to this study since there are multiple units (different houses or different weather stations’ data).

This method is used extensively in practice, but the theory behind the hot deck is not as well developed as that of other imputation methods, leaving researchers and analysts with limited guidance on how to apply the method, the main challenge being the donor selection.

In some versions, the donor is selected randomly from a set of potential donors, which is called the donor pool. In other, more deterministic, versions a single donor is identified, and values are imputed from that case, usually the “nearest neighbor” based on a dataset-dependent metric (i.e.: the mean when imputing temperature time series).

#### Implementing the donor selection

##### In theory

In the case of this research, the donor selection was based on pattern recognition.

It works by taking an extract containing data before and after a series of missing values (a gap) found in the recipient.  
  
To find the best matching segment of data from a donor, the recipient’s extract would then be compared to similarly-sized extracts from the same time period in a donor.

Using the difference in the mean of the donor’s extracts and recipient’s extract, the values from the donor’s extracts can be shifted towards those of the recipient — except when imputing classification data —.  
  
The sum of the absolute difference between the extracts can now be used to sort the comparisons: the smaller the sum, the better the pattern matches.  
  
The operation can then be repeated throughout each donor of the donor pool, for each gap, to find the best possible match before finally importing data into the recipient.

#### In application

This donor selection method has been applied in two versions for this paper.

Whereas the first iteration had a focus on precision, using interpolation to have the most accurate value between two data measurements, for example. The second iteration focused on improving processing time, by vectorizing the search algorithm.

But the processing time improvements had a negligible cost in precision. Which was even more diminished by the ability to compare the recipient’s extract to superior amounts of data from each donor for every gap (equivalent to a month plus the gap size).

### Recurrent Neural Networks

### KNN

## Imputation evaluation criteria

Given the fact that the aim of the research is to create a selection of the most suitable imputation methods for certain applications, it is necessary to use several criteria to properly characterize the performance of the analyzed imputation methods. The selected criteria are Raw Bias (RB), Percent Bias (PB), Absolute Bias (AB), Variance in Error, Maximum Error, Root Mean Squared Error (RMSE) and Mean Squared Errors (MSE).

# Results

# Conclusion

# Discussion

# References:

1.

2. <https://sci-hub.se/10.1109/SEGE.2019.8859963>

3. <https://ieeexplore.ieee.org/abstract/document/7229721>

4.

5.

6.

7.

8.

9.

10. <https://sci-hub.se/https://doi.org/10.1198/000313005X74016>