Master Thesis carried out to obtain the following degrees:

Master Degree in Smart Cities and Communities (SMACCs)

&

Master Degree in Research in Energy Efficiency and Sustainability in Industry, Transportation, Building and Urban Planning (EESITBUP)

Title: Analysis of imputation methods for data gaps in high resolution smart meters in buildings

Student: Bekbol Ismagulov

Main academic Supervisor: Dr. Aitor ERKOREKA, University of the Basque Country (UPV/EHU)

Academic co-supervisor: Prof. Véronique FELDHEIM, University of Mons

Partner supervisor: Dr. Roberto GARAY, Tecnalia

Academic Course: 2020/2021

Date: June 21, 2021

Acknowledgements

While I was working on my master's thesis I have received a great deal of support and advice which helped me to complete the project.

First and foremost I would like to thank my supervisors, Dr. Aitor Erkoreka and Prof. Véronique Feldheim for their guidance, support, and patience. Their immense knowledge and advice directed me to complete the Master's thesis and explore the subject deeper by enhancing my horizons and opening up new skills with their advice.

I would particularly like to single out my supervisor Dr. Roberto Garay for his kind assistance in providing the real-world data upon which this thesis was possible to be carried out. I appreciate his support and directions.

I would also like to thank Prof. Mikel Lumbreras for his assistance in understanding data and thus, obtaining answers for the several questions I had been doubting about as well as his suggestions on the possible scenarios to carry out.

In addition, I would like to express my gratitude to Master's and Ph.D. students Neigeline Rivollat, Shahin Sultana Eity, and Rim Doukha for their support on the academic matter questions and to overcome the data issues I had during the process of working on my thesis as well as providing me with the useful resources facilitating the improvement and hence the completion of my research.

Abstract

Missing data is one of the most common issues of the raw data in data analysis. Missingness could be ignored if it is considered not to have a significant impact on the analysis. In other cases, imputation methods are applied to handle them as machine learning models performed on the data with missing values may have a drastic decrease of the quality with the existence of the missing points. This thesis aims to determine the accuracy of the predictions of single and multiple imputation methods on the energy data as well as considering the impact the weather variables have on them.

To test the methods, the case study was conducted on four separate smart energy meter data from residential buildings located in Tartu, Estonia and each data set also comprised weather variables collected independently by the University of Tartu. The artificial missing values were entered in the clean data to examine the imputation techniques which allowed to compare the outcome with the original complete data set. The results demonstrated the higher accuracy for multiple imputation methods as opposed to the univariate analysis and the importance of highly correlated variables for the prediction of missing points.

We conclude that the increase of the variables included for the prediction of the analysis of the missing values is likely to increase the accuracy of the method as well. Despite multiple imputations appear to have the best accuracy, the challenges related to the concurrent missing values for all variables coming from the same sensor should be considered.

Keywords: Big data, data analysis, treatment of missing data, energy meters, univariate imputation methods, multivariate imputation methods

Bekbol Ismagulov

June 10, 2021

Table of Contents

List of Figures

List of Tables

Nomenclature

Building I	The first data set of smart energy meter
Building II	The second data set of smart energy meter
Building III	The third data set of smart energy meter
Building IV	The fourth data set of smart energy meter
DH	District Heating
DHW	Domestic Hot Water
EDA	Exploratory Data Analysis
EMB	Expectation-Maximization with Bootstrapping
GS	Global Score
SIF	Solar Irradiation Flux
IOT	Internet of Things
IQR	Interquartile Range
kNN	k-Nearest Neighbour
MAR	Missing at Random
MCAR	Missing Completely at Random
MICE	Multivariate Imputation by Chained Equations
MNAR	Missing Not at Random
MND	Multivariate Normal Distribution
NA	Not Applicable / Not Available
R	Missingness
r	Pearson Correlation Coefficient
RMSE	Root Mean Square Error
SH	Space Heating
WD	Wind Direction
WS	Wind Speed
X	Complete variables

Y_{com}	Complete data
Y_{mis}	Missing values
Y_{obs}	Observed values
Z	External causes on missing values

1 Introduction

1.1 Introduction

The growth of data science and big data intensifies the importance of each of its steps starting from the collection of the data, its aggregation to the machine learning methods applied and communication. Understanding data DNA is vital as it can provide in-depth knowledge about the studied field for further analysis and detect numerous insights for future research. Certainly, humanity appears not to live without the word “data” anymore due to its joint establishment not only in statistics now, but it surrounds us in business, economics, and engineering and generally, in all possible modes. On the other hand, one crucial aspect of working with data is dealing with the difficulties one may encounter during the process which might be in the form of incomplete data sets. As the data is not supplied by solely one source, it will further need the collection of many sets and thus it is common that some gaps may appear in the end.

Depending on how the faulty data is dealt with may have a huge impact on the further analysis helping to avoid any biases which might occur in accordance with the data incorrectness. It had been practicing for a considerable period, specifically at the end of the last century and beginning of the 21st century, that the missing points were deleted from the data without further investigations. Even nowadays, variations of deletion methods are utilised, and sometimes it might be the easiest decision to make, yet every case is required to have an individual approach. Simultaneously, there exist numerous methods to fill the gaps from the simplest ones to the complex ones based on machine learning. Consequently, it is vital to observe which methods perform the best and have the most accurate result when missing points are detected. In this master thesis, data gaps will be also referred to as “missing data”, “missing values”.

This thesis takes a case study to examine the imputation methods to fill the gaps in the data set. For that purpose, four smart meter data were selected from the residential buildings connected to the District Heating (DH) system to test various scenarios.

1.2 Missing Data

Customarily, the rows of the data set are defined as observations, measurements, subjects, or cases depending on the context [1]. For the case of smart energy metering of the residential buildings, each row specifies the hourly measured data for the provided variables such as Heating Consumption, Domestic Hot Water (DHW) Flow Temperature, DH Flow Temperature, DH Return Temperature, and so on. Whereas the columns section provides for us what is measured precisely, and they are denoted as variables. Yet, the data set is not always well-filled and distributed but rather comes with skewness, missing observations due to varied reasons, anomalies as well as different data structure errors required to be checked and treated. After all, fit it into a model we would like to build.

One of the major aspects of these types of misleading errors we may encounter is the missing data in the set of values which can be spread out with different mechanisms.

Missing data may hinder the full understanding of the phenomena we are interested in studying as the models either will not be performed properly or will not work at all. Despite the fact, there are existing techniques that can attempt to identify the core of the issue of missing values, they are unsuccessful, and missing data remains the main challenging task in data science. To point out, just removal of them will not lead to a better decision. Decision support systems such as neural networks, many computational intelligence methods as well as widely applied support vector machines are predictive frameworks and rely on the input data to predict an output. However, the presence of gaps in the data set makes it almost impossible to perform those prediction models [2].

Most research performed related to missing data is carried out in the social sciences in terms of surveys but not from an industrial or engineering perspective. Nevertheless, integration of the Internet of Things (IoT) and other smart technologies where they require the transmission of data over a frequent period of time may face intrinsic problems needed to be overcome.

In terms of missing data patterns, they can be univariate where one variable comprises of missing observations while the rest is complete; monotone missing pattern occurs when the gaps in one variable causes gaps in other variables as well monotonically; arbitrary missingness where gaps appear at random order (Figure 1.1) [3].

The missing data mechanism was first introduced by Rubin [4] and there are three types of missingness. Let's denote the complete data Y_{com} and it has two parts Y_{com} = (Y_{obs},

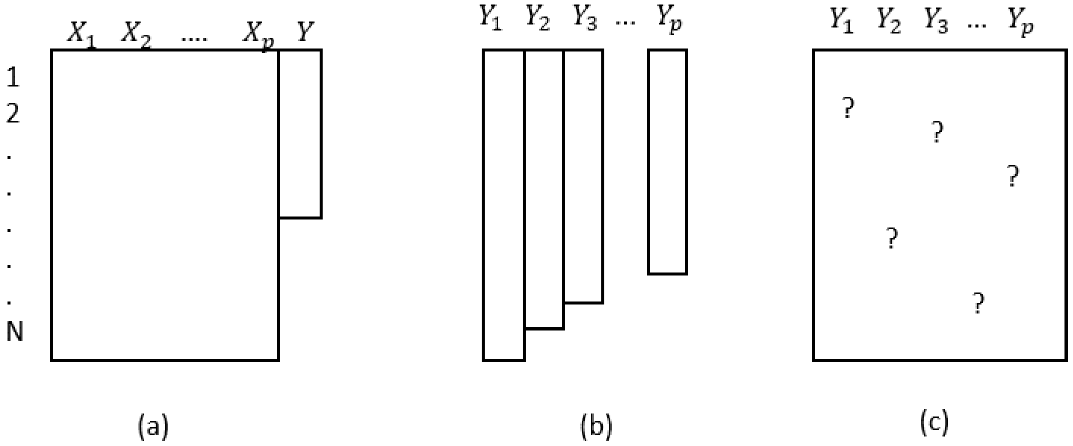
Y_{miss}) where the former is observed and the latter is missing values and Z is external causes. When Missing at Random (MAR) the distribution of the missing values does not depend on Y_{miss} (missing values) but does depend on the observed values and factor Z [3], [5]:

$$P(R|Y_{com}) = P(R|Y_{obs}) \qquad (1.1)$$

A special case of MAR is Missing Completely at Random (MCAR) and happens when the missingness does not depend on Y_{obs} either and the only cause can be external (Z):

$$P(R|Y_{com}) = P(R) \qquad (1.2)$$

But if the missing distribution does depend on Y_{miss}, it is another case named as Missing not at Random (MNAR).



X_1 X_2 .... X_p Y
1
2
.
.
.
.
.
N
Y_1 Y_2 Y_3 ... Y_p
Y_1 Y_2 Y_3 ... Y_p
?
?
?
?
?
(a)
(b)
(c)

Figure 1.1: Missingness patterns: (a) univariate pattern, (b) monotone pattern, (c) arbitrary pattern. Columns and rows represent variables and observations respectively [3]

Figure 1.2 demonstrates the connection between complete variables (X), partially missing values (Y), missingness (R), and external factor (Z).

Missing values are an integral part of most big data because the data is not delivered only from one source but aggregated from various sources. Besides, even within one measurement meter, there might be cases where those gaps appear for the recorded period. Hence, there is not any measurement, and it appears as an empty place or NA in the data set. The missing information is either erroneous (indicated as an error by the smart meter; comprising of NaN values) or missing measurements which is mentioned previously.

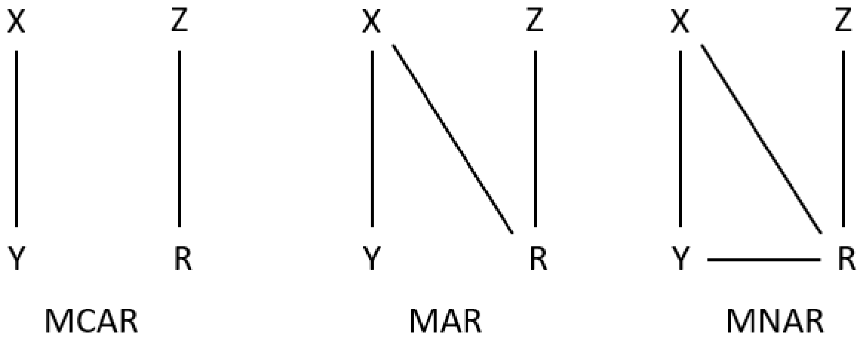


Figure 1.2: Three types of missingness presented by Rubin [3]

There could be several reasons related to the missing data occurrence and possible origins are as followed:

- The failure in the connection when the data is transmitted. Smart meters like any IoT sensors may need constant connection with the Internet through which the corresponding measured data is sent and when it fails to deliver it, the gaps will appear in the data set.
- Another potential reason behind it could be the aggregation process of the data collected [6]. Not all the time the data is supplied by one department but may be collected from several of them. For this reason, during the procedure of merge from one of the departments delivering the data, there may be a failure at certain points leading to gaps for the respective time series.
- When blackouts occur in the system the data will not be gathered from the buildings and thus that time could be stored as missing NA which implies missing observations as well for the corresponding period.

1.3 Aim and objectives

Currently, the most available study regarding the missing data is about social and economic sciences implying that the data is gathered from the surveys and may happen due to the lack of response from the interviewees to certain questions. Hence, most of the tested interpolations of any types are on how to deal with the gaps in their own field.

On the contrary, from the industrial perspective, the parameter is predominantly measured on a time basis meaning if there is malfunctioning occurring on the sensor then it may cause huge data loss if not tackled accordingly. On top of that, there are dedicated sensors which measure only one variable, such as humidity and carbon dioxide concentration, yet there are also meters which record many variables at a time. Therefore, if one variable is failed to be measured then it causes the breakdown of detection of observations of other variables as well. Whereas the data collected from the buildings, as aforementioned, measure the energy-related variables of DH and by that, we may lose significant data about the behavior of the occupants which is a pivotal source for demand-side management. Consequently, it should be considered to surmount the issue by applying the imputation methods if the percentage of missingness are to be discovered high.

What makes the current case study distinct is its aggregation of energy data with weather variables. It enables to carry out our analysis not only with univariate and multiple imputation methods but also inspect how the weather data would facilitate solving the issue with missingness when gaps appear in the data set.

The aim of the thesis is to investigate the accuracy of single and multiple filling the gap techniques in accordance with the energy data and the impact of weather variables on the prediction of missing points. This will allow us to analyse on which basis the missing data should be treated and more importantly, with what type of methods.

In compliance with the aim, the following objectives are structured:

- 4 residential building data sets will be pre-processed and treated separately
- Correlation analysis will be performed in each data set for the energy and weather data to classify the relationship between variables
- Univariate and multiple imputation methods will be performed with the time-gap and correlation-coefficient scenarios
- The accuracy scores will be defined to identify the accuracy of each of the scenarios

Besides, the data set is delivered clean which means there is no missing data and that is instrumental as it allows us to enter artificial missing values and then compare the outcome with the original data sets.

1.4 Literature Review of Relevant Studies

There are numerous textbooks and articles [1], [2], [7]–[9] with the discussion of the potential origin of the data capture failure. Noticeably, they provided the cause from survey perspectives with a mechanical collection of the answers of the respondents.

Handling the missing data is essential to hinder the issues that may appear during the process of analysis and working with models and there exists an enormous number of possibilities to deal with them. According to [10], when the missingness in the data set is very small which can be between 10 % and 15%, then they simply can be removed without a significant impact on the data set. However, it might create a bias if missing data is around one-third of the whole data set [11].

Dealing with the missing data can be performed by deleting them with some changes to imputation which can have many variations. Deletion can vary from "complete deletion", "list-wise deletion" and "complete case analysis" to "specific deletion" where for the first group the observations containing missing data in one or more of their attributes is deleted. On the contrary, for the latter, it can be specified with a certain limit of missingness. While "variable deletion" or "pairwise deletion" will delete the variables having missing data in one of its rows from the case, but includes it for the analysis of other variables in the case without missing observations [12]. Nevertheless, dealing with missingness may not be the best solution especially with big data sets and with a high proportion of missing values. Hence, imputation methods based on the existing values can replace the missing points in a mixed variety of ways. They can be simple such as mean, median, or more complex ones where the prediction model is built to fill the gaps in the data set.

Mean imputation can be carried out by replacing the missing values with mean, median, or mode and its main drawback is if the data set is huge, it replaces all the gaps with one single value. Thus, the data shape and distribution may be altered as well [13]. Another type of single imputation is the k-Nearest Neighbours (kNN) method where the distance function determines the similarity of two points and replaces missing values by copying similar values from the data set [13]. Hot deck imputation takes the observed value with similar characteristics to the point where there is a missing value and substitutes it [14]. It means the data from the current data set is used to analyse and fill the gaps, while the cold deck can use data from other data sets as well.

In terms of univariate imputations, one important aspect of the case study presented is it is a time-series data and hence it could be analyzed on its basis. There are numerous ways of filling the gaps techniques relying on inter-attributes dependencies implying multivariate analysis. However, packages dedicated to univariate time-series imputation can be tedious to discover. In fact, some packages may have certain functions which can work with time-based data univariately such packages as zoo, forecast, spacetime [15], and xts can handle them with certain inbuilt techniques [16]. Zoo package has some functions which can work with the missing data replacing them with either Linear Interpolation or Last Observation Carried Forward (LOCF) methods [17]. timeSeries and missRanger package also contain some very basic tools working with the missing data[18], [19].

There is another package fully dedicated to the imputation of time-series data called imputeTS. It includes varied functions, namely interpolation, LOCF, weighted moving average as well as mean and mode. It has more options for visualizations for the observation of the imputations before and after the methods are applied. All these tools make this package suitable to work with the data when there are time-dependent data with missing values.

Currently, several methods can handle missing data imputations for multivariate analysis such as missForest, Amelia, MICE, VIM, HMISC. Habitually, the data set comprises many variables and their impact on the other variables during the process of filling the gaps may differ. The existence of many algorithms designed for multivariate analysis arises the question about the efficiency of those methods compared to each other. It can be assessed from many perspectives such as time to compute, the size of the data set as well as on the richness of the data set [12].

One of the studies [12] conducted research into multivariate imputation methods, namely VIM, missForest, MICE, and HMISC. In terms of the time consumption of the imputation process, HMISC performed better than others and VIM was concluded to be better for smaller data. Admittedly, for large data sets, HMISC and MICE are more suitable as well as when it comes to the accuracy of the data. Variance analyses show all the methods perform similarly, yet missForest was the worst among the four approaches.

Similarly, another study [11] performs the analysis of imputation methods such as kNN, missForest, MICE, and Phylopars in life-history trait data sets. According to its result, kNN performed less well than the rest where MICE, missForest, and Phylopars showed virtually similar performance. Even though with the addition of some data MICE gave a better result than missForest, the latter does not require a deep knowledge about the data set to apply the method.

Besides, in a study performed by Lia et al. [20], MICE faced some issues with nominal and ordinal data while missForest was among the top, yet faced some difficulties when there were not strong correlations between variables.

Nowadays as the data science importance is ascending, the requirement to handle the recorded data in a proper way to have a clean data frame is increasing aggressively. There are many inbuilt packages for software R such as VIM, AMELIA, MICE, and MCDA [21]. They are specifically devoted to the analysis of multivariate data and there must be more than one variable to be able to run and utilize these techniques. The above methods are applied to impute missing values which can be quite frequent while handling the data set and replacing the gaps with the predicted values.

As it can be seen, there is not enough research done on the analysis of energy data with the most recent one based on the Danish case [22] where the methodologies of data analysis and clustering techniques were presented. The data was based on dwellings connected to the DH System and thus its data for a year was considered. The whole process led to the typical hourly-based daily profiles of the buildings on heat consumption and temperatures. Despite taking into account the basics of handling the missing data, there was not a devoted analysis of the applied methods.

The rest of the thesis is structured as follows: Section 2 is dedicated to the methodology and general principles of the imputation methods. Section 3 introduces the domain in which the aggregated four data sets are described with the provision of preliminary analysis and the correlation of the variables. Subsequently, the result and analysis of the proposed scenarios on the case study with existing missing data filling methods are implemented in Section 4 and we make a conclusion with future works at the end.

2 Methodology

The whole process of data and its analysis was performed using the free software environment and language for statistical computing and graphics "R"[23]. R software has become one of the main tools in data science and used by numerous researchers with different backgrounds and coming from different disciplines. Its simplicity to use and sole dedication for data analysis makes it a perfect tool to perform all data analysis steps along with Exploratory Data Analysis (EDA) and building models. It is open-source and free with varied packages for a variety of purposes. Besides, it is a cross-platform implying it is supported by many operating systems which makes it preferable due to its flexibility. Moreover, the most common methods in statistics such as hypothesis testing, variance analysis, regression methods, and descriptive statistics are inbuilt in the system [24].

First, the data was visualized for the whole data set length to identify the patterns and observe how the data was measured throughout the year for each data set independently. After, the correlation analysis was performed considering the combination of energy and weather data based on the Pearson correlation coefficient to create scenarios as a relationship between variables.

As the data is clean in terms of missing values, artificial missing data with 6 time-gap scenarios were introduced for each energy variable. This administrates to analyse how well the methods are performing with the predictions.

Furthermore, single imputation techniques will be applied using the time series imputation called "Impute TS" (R package) provided with a mixed variety of interpolation methods to replace the missing values with artificially introduced missing data. Besides, it is one of the few methods which will impute the missing values in a univariate form while they are not enough available packages at the moment. Univariate means only one attribute is measured over time, hence, only one variable will be imputed to observe the change and be able to compare the outcome of each parameter[25].

The next step consists of the multiple imputation methods to predict the possible values of the gaps. Imputations for multivariate analysis follow mainly the following steps to get to predict and fill the missing points in the data set (Figure 2.1) [26]:

1. Imputation: Generate a set of m >1 values where each set will impute the missing values in the original data set by the default set value (m) and it will create corresponding copies of the observed values.
2. Analysis: Using complete-case methods, the analysis of the created m dataset is carried out.
3. Combination: Pool – the process of integration of m analyses.

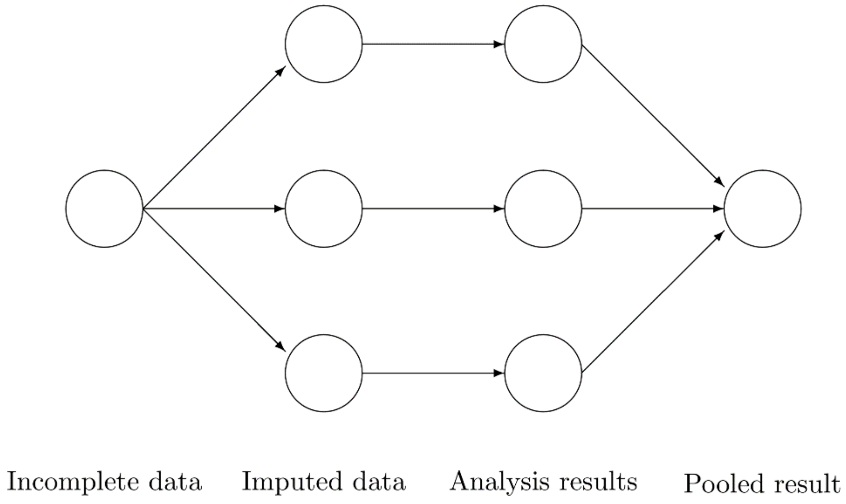


Figure 2.1: Main steps of multiple imputations [27]

For multiple imputations MICE, Amelia, and missForest packages were selected. First and foremost, MICE and missForest are some of the most studied existing imputation methods for different research fields. While Amelia is one of the least investigated methods and hence, it was decided to check its accuracy against the other two methods and identify if it is appropriate to use to fill the gaps.

MICE stands for Multivariate Imputation by Chained Equations. It is one of the widely used and researched multivariate analysis methods and was originally described by Bullen. The basic idea in R software is to create a copy of the original data set with missing values, say as m = 5, and after filling the gaps where the missing values occur, it treats each copy independently. Thus, all those copies are averaged to give a single data set with filled gaps [28]. The principle of the method is based on the following way: if we have X_1, X_2…X_k variables and X_1 has some missing observations, then it will be regressed on the rest of the variables. After the prediction, the gaps in X are replaced by obtained values. If X_2 has missing data in it, X_1, X_3, X_4 to X_k columns will contribute to building the prediction model. Subsequently, the missing points are replaced with the estimated values [29], [30].

missForest (Figure 2.2) was initially proposed by Stekhoven et al. [31] due to the lack of methods which can handle working with both categorical and continuous variables based on a Random Forest. Missing values are treated as the response variables and resampling-based classification with regression trees used to involve the observations from other variables for the prediction of the missing values [20].

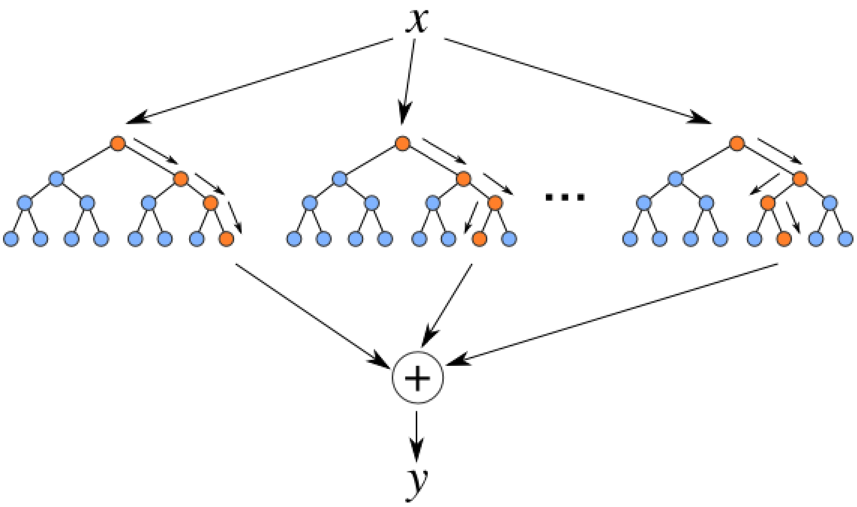


Figure 2.2: Illustration of the working mechanism of missForest [32]

Amelia also performs multiple imputations to work with missing data and those kinds of methods can alleviate the bias while increasing the efficiency of the process. It is presented on bootstrap-based Expectation-Maximization with Bootstrapping (EBM) algorithm and it can work with many variables. There are two assumptions stated as all variables are Multivariate Normal Distribution (MVN) and the observations are MAR [33]

The univariate and multivariate imputation methods were tested for time-gap and correlation-based scenarios. The latter was applicable only for multiple imputation methods as opposed to the single imputation methods. Because for single imputations only the data from a variable where the gaps appear are utilized to compute and fill the missing points.

Following the analysis of the gaps, their accuracy was computed with an Root Mean Square Error (RMSE) value and Global Score (GS) was proposed as a measurement of the accuracy performance of each technique applied.

3 Data

For the case study, four different data sets were selected from the buildings located in the city called Tartu (Estonia). Each data set contains 26 variables with hourly measured observations aggregated from two different sources. Smart meter detected the following parameters coming from the DH, which are DH Flow Temperature, DH Return Temperature,Volume (m^3), Volumetric Flow Rate (l/h), Heating Power, Space Heating (SH) Flow Temperature, SH Return Temperature, and DHW Flow Temperature. For the sake of privacy, no extra information was revealed about the buildings and their types. Consequently, all the assumptions and conclusions were drawn purely based on the analysis and visualizations.

Figure 3.1 provides the measurement setup and how the DH is designed. According to the scheme, we are to see the main variables measured by the smart meter.

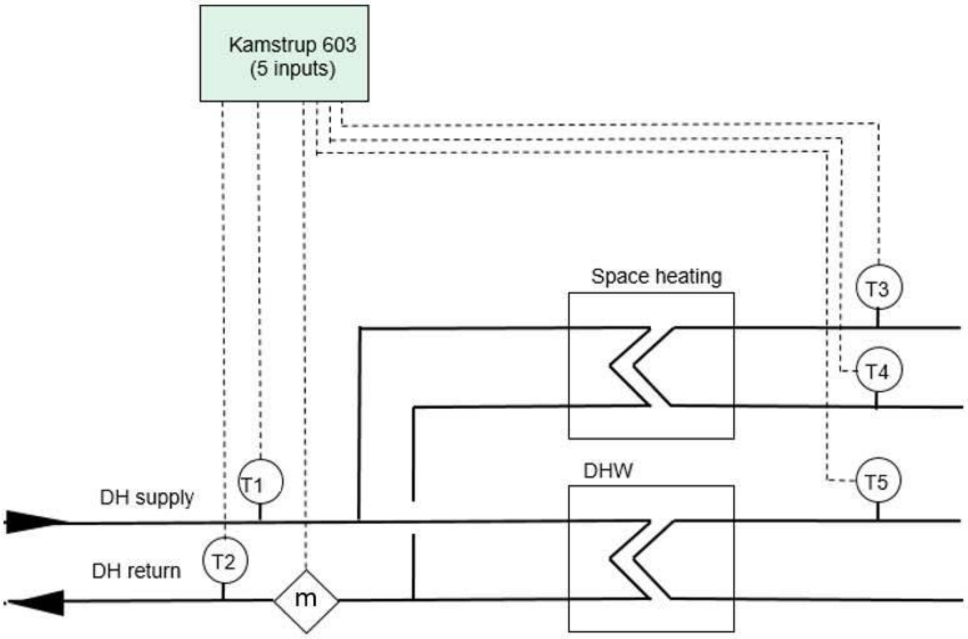


Figure 3.1: The measurement setup of the DH system for the provided data sets

Where T1 – DH Flow Temperature (denoted as Fl_T in visualizations), ºC; T2 – DH Return Temperature (denoted as Ret_T in visualizations), ºC; T3 – SH Flow Temperature (denoted as SH Fl_T in visualizations), ºC; T4 – SH Return Temperature (denoted as SH Ret_T in visualizations), ºC; T5 – DHW Flow Temperature (denoted as DHW Fl_T in visualizations), ºC, m – Volumetric Flow Rate (denoted as Vol Fl Rate in visualizations), l/h. (Heating Power is denoted as Power, Ambient Temperature as Temperature in the visualizations)

The general overview scheme is about how the system is operated. However, not all the data presented in the set are useful or used for the analysis. In accordance with the scope of the study, the focus is on the smart meter data, excluding or neglecting most of the rest of the data with little or no impact on the main variables.

3.1 Annual profile of data set variables

In the meantime, as aforementioned the data set is the aggregation of two different sets. Therefore, the weather data from the weather station managed by the University of Tartu is applied with an hourly measurement of the parameters as well. It starts on January 1st, 2019 throughout the year until 31 December 2019.

Table 3.1: Smart Energy Meter and Weather Variables

Name	Unit
Heating Power	kWh
DH Flow Temperature	°C
DH Return Temperature	°C
Volumetric Flow Rate	l/h
DHW Flow Temperature	°C
SH Flow Temperature	°C
SH Return Temperature	°C
Ambient Temperature	°C
Wind Speed	m/s
Wind Direction	-
Solar Irradiation Flux	W/m^2

The set contains the annual Ambient Temperature, Wind Direction (WD), Wind Speed (WS), and Solar Irradiation Flux (SIF). Thus, the weather data was aggregated with the smart meter data, primarily, to observe the correlations within the main variables and, after, they are applied in the process of prediction of the missing observations. A list of intrinsic smart meter and weather variables is provided in Table 3.1.

Figure 3.2 illustrates the measured smart meter data throughout the year for Building I. One can see that at the beginning and the end of the year the Heating Power increases while the mid of the season witnesses a dramatic decrease of the corresponding usage plateauing at 0. This happens as the period falls approximately between the hours of 4000 and 6000 which are virtually between June and September implying that it is not a heating season. According to the analysis of the annual measurement we can state that the heating season commences nearly at the end of September. Furthermore, it can be proved by the rise of the Heating Power consumption around that time.

The rest of the parameters follow a similar pattern as there is no necessity to heat the buildings at the time. Hence, the DH Flow Temperature goes down during the summer as well. On the other hand, DHW Flow Temperature remains around 60 °C on average, fluctuating between 50 °C and 70 °C and one can say it is because DHW Flow Temperature is utilised for various purposes as showering and other machines in the house where hot water is required. However, for a specific period of time, the measurement outcome abruptly falls to almost 0 at the mid of November. This case (where the records are virtually 0) could be studied further by treating those measurements accordingly or applying some techniques so that it will not hinder or cause difficulties during the future analysis. Besides, a similar radical change is investigated in the DH Return Temperature as well. Customarily, they can be observed with an unaided eye and treated accordingly.

The Building II measurements are substantially correlated and comparable to what is observed for Building I (Figure 3.3). During that same period, the overall consumption of the Heating Power drops to zero again and the rest of the variables illustrate the virtual correspondence, and all the temperatures coming from DH fall considerably due to the known reasons. However, what may seem out of the range or odd is the absence of DHW Flow Temperature for the corresponding building. The data set contains the column with DHW Flow Temperature, yet, with no measurements recorded. These are not gaps but gathered as 0 for the whole year. Consequently, we could assume that residents of the building rely on home-built boilers for DHW needs, and thus, there is no data for the variables. As a result, there is no possibility to even predict them as not even a small portion of raw data is provided to be able to fill the rest.

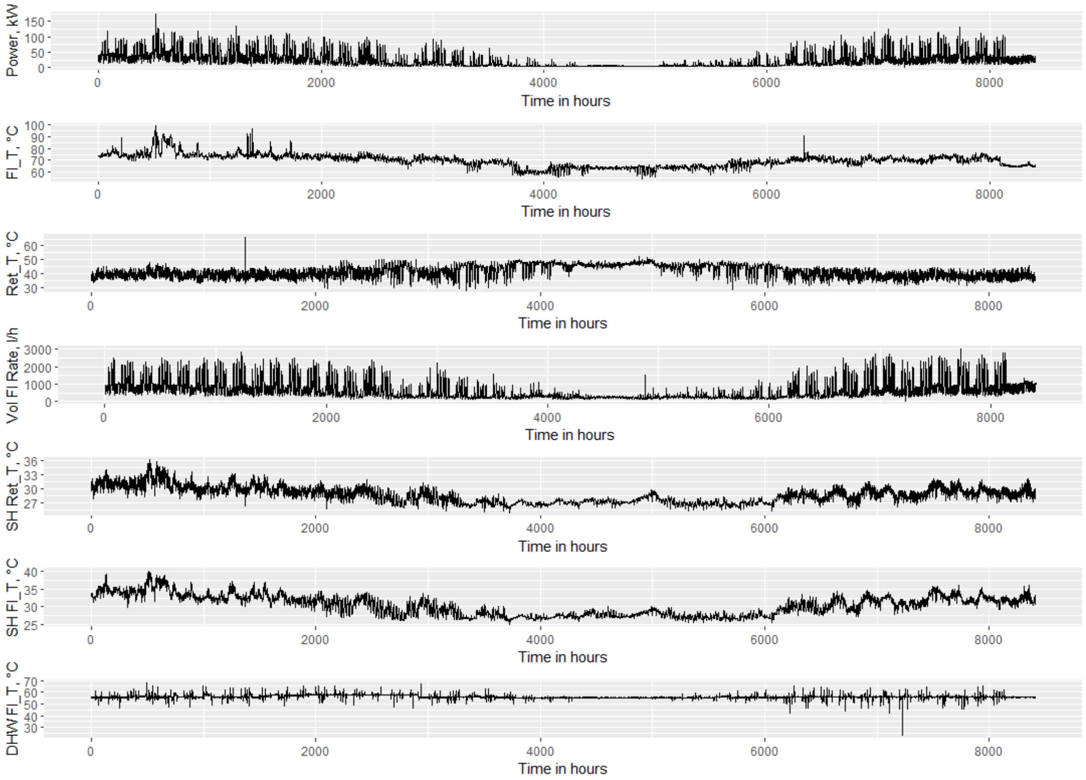


Figure 3.2: Smart meter data measurements of the primary variables at Building I

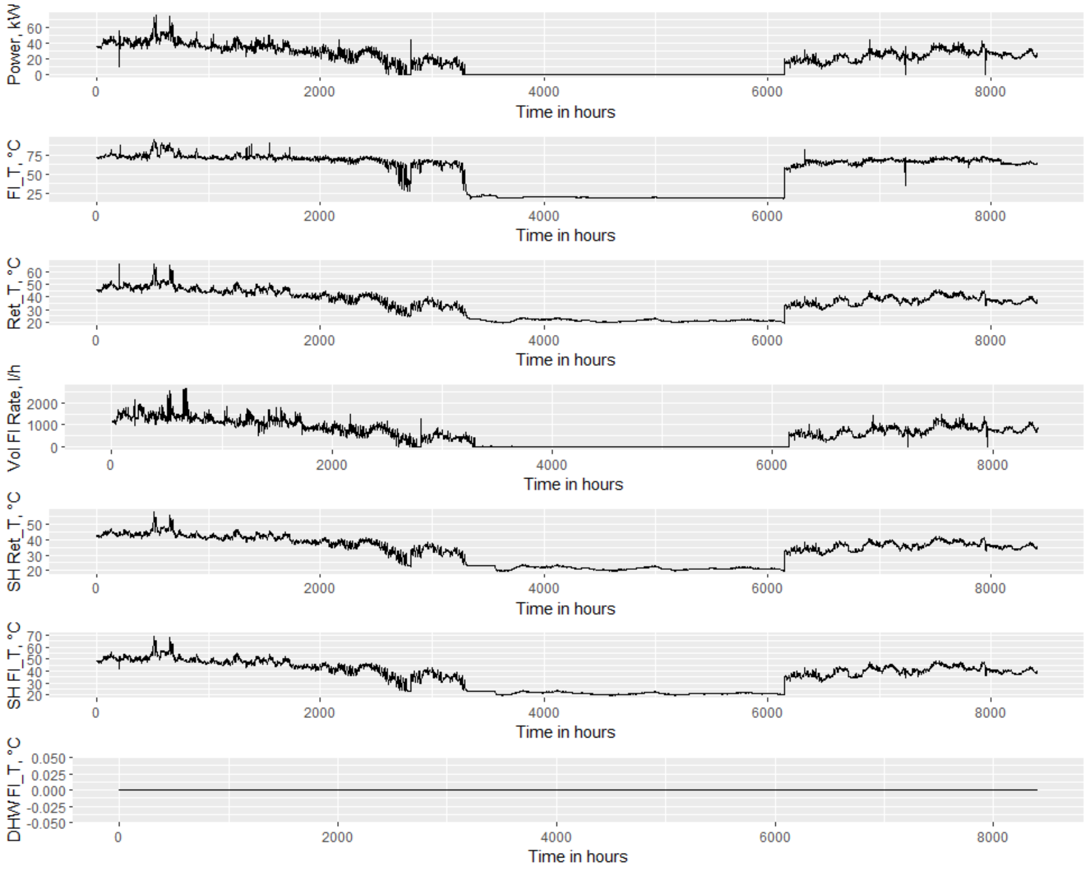


Figure 3.3: Smart meter data measurements of the primary variables at Building II

One thing about which we have no knowledge is the type of the buildings where the smart meters are located. Thus, we can assume and observe their yearly or hourly based profiles and their corresponding usages. The measurement for Building III has also a similar pattern alike the first smart meter data as expected. Nevertheless, analysing it numerically we can state that at the highest the Heating Power consumption of Building I was well above 150 kW in January (around 514 hours which is the 21st day of the month) and the rest of the measurements during the heating season consumption on average was higher than 70 kW. In contrast, the data coming from the third smart meter at the peak displays more than 100 kW in September (seemingly when the heating season starts) for a short period of time (which is substantially higher than other measurements of that variables in Building III), and according to its appearance, it is likely to require further investigations to examine for its veracity.

Apart from that, the observations from Building III follow the same model by decreasing when the heating season terminates and rising by the start of the new heating season (Figure 3.4). Based on the Heating Power consumption, one could presume that it might be a small building with a heating meter built for data analysis. Otherwise, the DHW Flow Temperature records show the drop of the temperature in summer as well which was not the case for the aforementioned residential housings. In addition, there are also some measurements to seem to vary considerably in DH Return Temperature and DH Flow Temperature in winter (between 1500 hours and 2000 hours) than the rest of the data for the specific period. Yet, it might be considered within the normal range of obervations if they are separated and analysed for the month where those measurements (potential outliers) are identified.

The illustration of the last building (Figure 3.5), Building IV, does not provide any novelty regarding the measurements since the form of the heating season is certainly met here as anticipated. In terms of the Heating Power consumption, on average it utilised more than 10 kW compared to the Building I, while the mean temperature DHW is almost 10 °C less than the one in Building I. However, in contrast to the Building III measurements, for both Building I and Building IV the DHW Flow Temperature remains stable across the year.

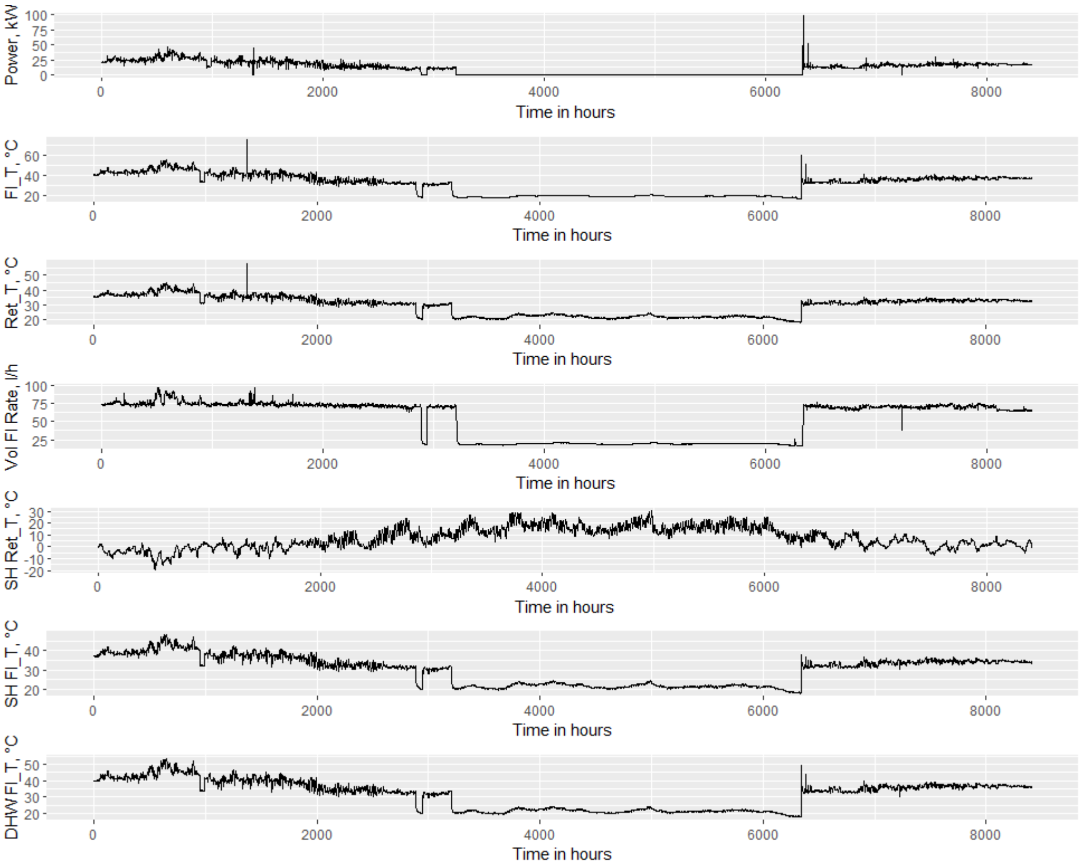


Figure 3.4: Smart meter data measurements of the primary variables at Building III

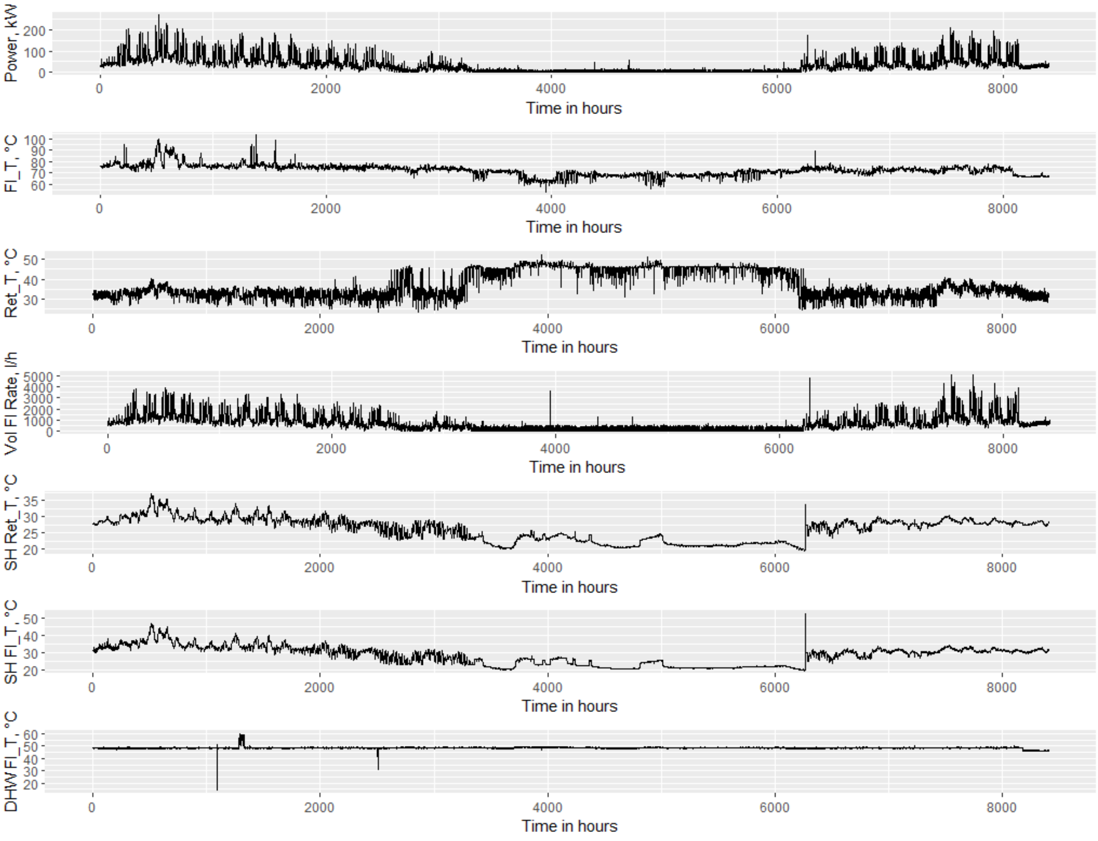


Figure 3.5: Smart meter data measurements of the primary variables at Building IV

Hourly profiles of the Heating Power and other variables facilitate to derive the very initial idea about how the observations are recorded. Besides, it identifies if there are rare observations or patterns in the data since simply looking at the big data does not aim to observe those abnormalities.

On the other hand, visualizations help to institute to “torture” the data and realise what are the following steps to take. For instance, at the first glance, we are to make assumptions that first, second, fourth smart meters might be collecting data from larger residential buildings, at the meantime the third metering is responsible for considerably smaller construction type than others.

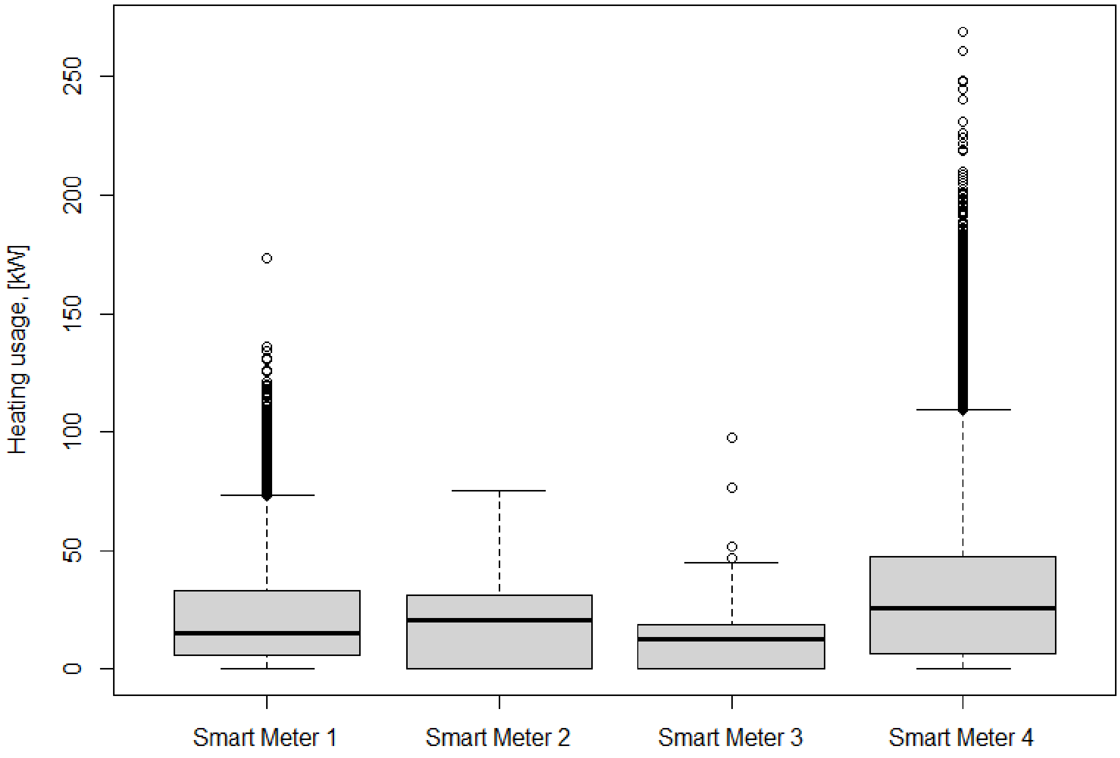


Figure 3.6: Multiple boxplots of Heating Power for four buildings for comparison

Figure 3.6 demonstrates the summary of Heating Powers for all the smart meter data based on a 5 number summary. As it can be seen from both Table 3.2 and the illustration, the Heating Power consumption of the first and last buildings is comparably higher than the ones in the middle. Apart from that, the medium of the data distribution in Building IV is the highest among all, while minimum consumption equals zero at each smart meter variable simply proving that there was no Heating Power in the summertime.

Table 3.2: Summary of the distribution of the Heating Power

Smart Meter	Minimum, kW	Lower Quartile, kW	Median, kW	Upper Quartile, kW	Maximum, kW
Building I	0	5.9	14.95	33	173
Building II	0	0	20.7	31.4	75.2
Building III	0	0	12.6	18.7	97.3
Building IV	0	6.1	25.35	47.4	269.2

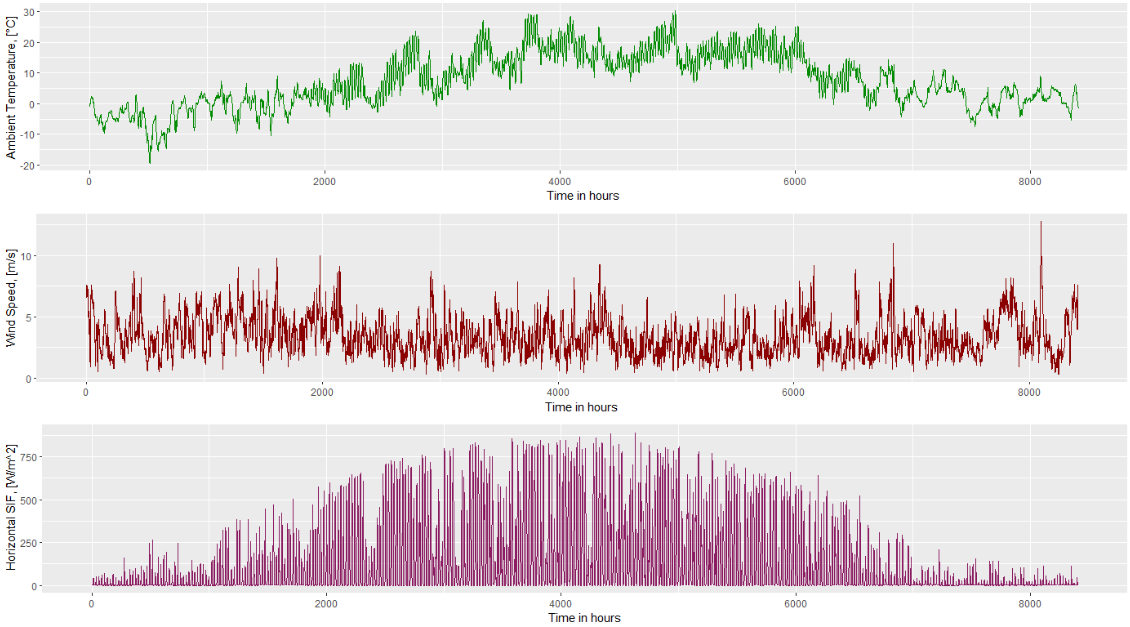


Figure 3.7: Meteorological data throughout the year gathered by the weather station of the University of Tartu

Regarding the points laying out of the range, it is easy to notice that Building II does not show any observations on this matter. It means every measurement is within the Interquartile Range (IQR) and does not exceed the minimum value or drops below the minimum observation. On the contrary, the same variable in other buildings contains possible outliers, or at least the analysis of the boxplot technique detects the presence of those points. Especially, the first and last groups are shown to have a considerable quantity of them. Yet, markedly, this is the image of the whole year data and thus, may need a separate monthly investigation. Within that context, some observations may fit the range and would be excluded from being anomalies.

The weather data for all the data set is the same as the case study buildings are located in the same city. Hence, Figure 3.7 displays the Ambient Temperature, Wind Speed, and SIF measurement annually. As anticipated, due to its geographical location it tends to be colder and go sub-zero regarding the Ambient Temperature during winter months and late autumn. In contrast, spring and summer witness the increase of the temperature reaching up to 30 °C at peak. According to [34], the maximum average temperature is around 23 °C which is mostly set in July. In terms of Wind Speed, it fluctuates considerably going from virtually no wind up to 11 m/s during December. Additionally, SIF tends to be substantially higher starting from around April to October. All these weather variables are helpful for the investigation of the smart meter data as their consumption may directly correlate with some of the weather data. It could be expected that the Heating Power and DH Flow Temperature would be high-negatively correlated with the Ambient Temperature.

Table 3.3: Mean weather variables for each season in Tartu

Season	Temperature, °C	Wind speed, m/s	Solar Irradiation Flux, W/m^2
Winter	-1.4	4	23
Spring	7	3.5	164
Summer	17	3	407
Autumn	7	3	66

Table 3.3 summarizes the mean weather variables divided into four seasons. This way we could detect the actual measurements and changes due to some considerable alterations within the seasons.

Before proceeding on with any mode of analysis, it is essential to check for the variable types and gaps the data set may contain. Completion of these steps will first allow us to see with what variable types we will deal with and then select methods for implementation of various analyses.

All variables and their respective hourly observation numbers are the same for four data sets. Hence, one assessment for one of the sets is sufficient to identify the data type and select the required parameters. At the first glimpse, each set has 26 variables and 8410 observations measured where not all data is useful for analysis. Consequently, solely the data from smart meter and weather data are selected to perform the next steps. Therefore, the whole data set consists of completely quantitative values.

There are statistical methods to identify the gaps based on if they are MAR, MCAR, or MNAR and require some hypothesis before running them. Therefore, it might be cumbersome and necessitate a certain amount of time. Whereas the visualizations make it easier to observe as there are currently numerous techniques to perform them.

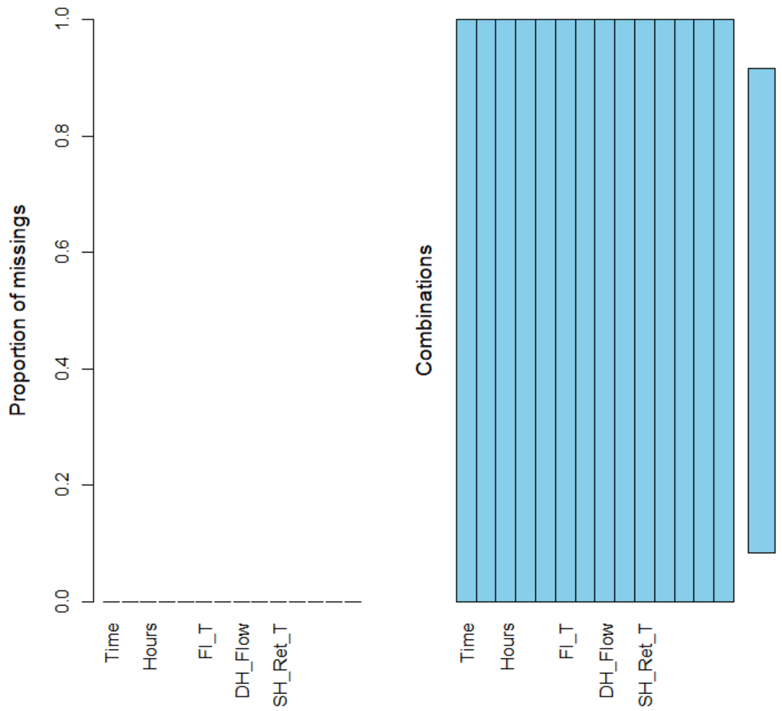


Figure 3.8: The missingness plotting outcome for Building I

Applying one of the existing methods we can obtain the first result related to Building I. Preliminary result display that using the following plotting technique where the outcome is divided into two sections, we can state that no missing value was detected for Building I. Figure 3.8 shows that the left side is empty where the portion of missingness in the variables would be displayed. However, data set as a combination of both smart meter and weather data appear to not have any gaps for the corresponding measurements. Hence, each variable in the smart meter and weather data is complete for the first set of observations.

Using similar plotting and analysis methods we will derive the required summary for the rest of the buildings. All the techniques come to terms with no missingness for all of the data set. The following illustration demonstrates the possibility of the gaps in the second set of data which proves what was stated previously (Figure 3.9). The rest of the illustrations for other data sets are provided in Figure A and Figure B in Appendix.

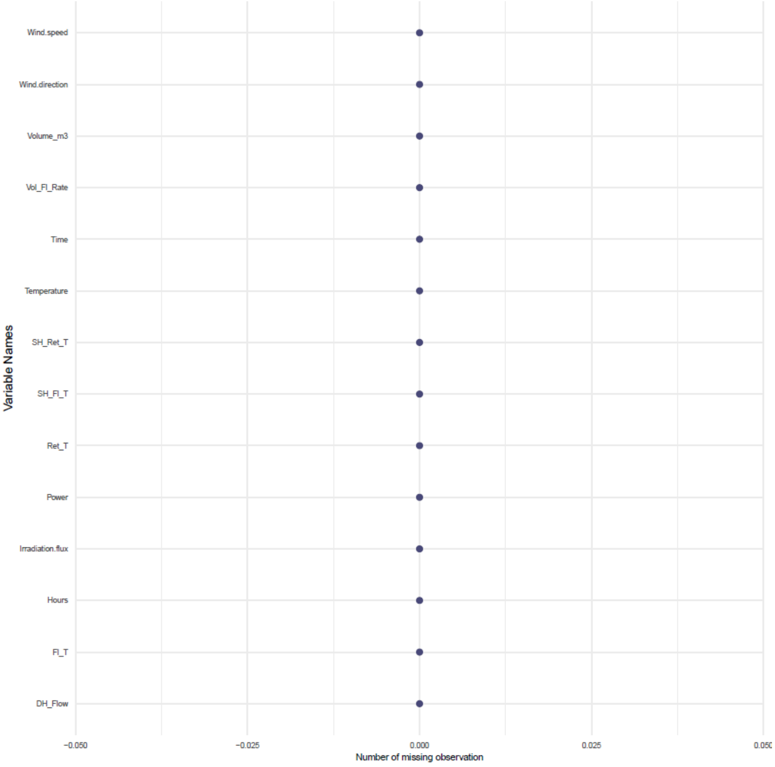


Figure 3.9: Examining of gaps in Building II applying an alternative plotting mode

3.2 Correlation analysis

Initially, having a clear vision about the connection between variables could be a productive start providing essential hints and an overview about how the measured elements are correlated with each other. In data science tools, there are numerous modes of obtaining a correlation matrix for the analysis and understanding of the connections. Thus, it displays the correlation coefficients based on which some conclusions can be drawn and take directions for the following steps.

The chart can show the interconnection based on the coefficients where there can be both positive and negative correlations between the elements of a dataset.

Another method that illustrates the relationship between variables is a scatter plot which indicates the correlation between chosen points by building a plot spreading the observations to the x and y-axis respectively as well as being instrumental to detect outliers. It is also helpful during the process of building polynomial multiple regression models.

3.2.1 Building I. Correlation analysis

Correlation analyses are carried out applying the Pearson correlation formula. According to the formula and its coefficient definitions, we can summarize the output as in Table 3.4

Table 3.4: Association strength and their corresponding correlation coefficients for Pearson correlation formula (Note: “-” sign represents only the coefficient being negative) [35].

Association strength	Correlation coefficient (r)	
	Positive	Negative
Strong	r > 0.5	r < - 0.5
Medium	0.3 < r < 0.5	- 0.3 > r > -0.5
Weak	r < 0.3	r > -0.3

Figure 3.10 illustrates the correlogram plot displaying the variable associations within Building I so that we can identify if those values are correlated; if so, whether it is a negative or positive correlation. The knowledge about variable relationships is to be instrumental for further analysis. A blue color displays a positive correlation, and red color is for a negative correlation. The color intensity and the size of the shape are proportional to the correlation coefficients. The legend on the right side gives an understanding of how the colors are described for the correlation analysis.

As we can see on the correlogram, there is a high positive correlation within the variables from the smart meter data, namely DH Return Temperature, Heating Power, DH Flow Temperature and SH Flow, and SH Return Temperatures. Each of the correlation coefficients is higher than 0.5. All data from the smart meter is positively highly correlated but DH Return Temperature. It was shown it has negative correlations with other measured variables as expected. In contrast, the only variable from the weather station that appears to have a strong relationship with DH data is the Ambient Temperature. Hence, it is certain that when the outdoor temperature increases the DH Flow Temperature and Volumetric Flow Rate drop, leading to the decrease of the Heating Power consumption as well. The rest of the weather data demonstrated weak or no correlation for all of the cases.

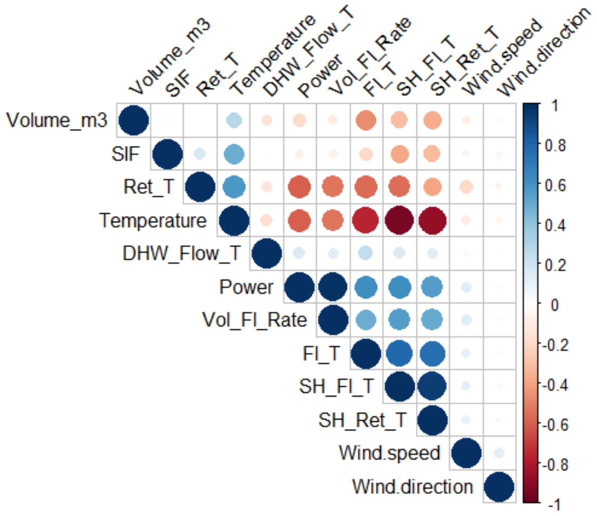


Figure 3.10: Correlogram of Building I data between the smart meter and weather variables

Using the quantitative analysis, we were able to draw the exact correlation coefficients for each pair and it is provided in Table 3.5. It summarizes all the relationship coefficients following a similar pattern as the above-provided correlogram. The greener the coefficient, the stronger the coefficient positively. The redder the colour, the stronger the connection between variables in a negative way. Hence, it can be summarised that all the data for Building I from the smart meter are almost highly correlated with each other along with the Ambient Temperature.

Table 3.5: Correlation matrix for the aggregation of smart meter and weather data for Building I

Volume	1.00										
Power	-0.19	1.00									
Flow T	-0.47	0.62	1.00								
Return T	0.00	-0.60	-0.56	1.00							
DHW Flow_T	-0.15	0.15	0.24	-0.13	1.00						
SH Flow T	-0.30	0.62	0.78	-0.55	0.16	1.00					
SH Return T	-0.36	0.55	0.75	-0.39	0.14	0.95	1.00				
Ambient T	0.28	-0.60	-0.77	0.58	-0.17	-0.94	-0.87	1.00			
WS	-0.09	0.13	0.11	-0.20	0.04	0.12	0.10	-0.11	1.00		
WD	-0.03	0.03	0.03	-0.07	-0.02	0.04	0.03	-0.06	0.12	1.00	
SIF	-0.02	-0.09	-0.19	0.17	0.00	-0.38	-0.31	0.50	0.05	-0.04	1.00
	Volume	Power	Flow T	Return T	DHW Flow_T	SH Flow T	SH Return T	Ambient T	WS	WD	SIF

3.2.2 Building II. Correlation analysis

Furthermore, the same analysis was performed for Building II. Here, we attempted to display the singular connection between two variables selecting them arbitrarily: one from smart meter data and another from weather data. The result is illustrated in Figure 3.11. Here the scatter plot can be used to observe the linear regression between two variables which are the Ambient Temperature and DH Flow Temperature. However, what we seek is the relationship they have, and it can be straightforwardly identified by the number displayed within the plot. It equals -0.82 and the figure illustrates the winter time when there is a high correlation between these variables. Consequently, it is a negative correlation implying that when the outdoor temperature is low the DH Flow Temperature is high and vice versa.

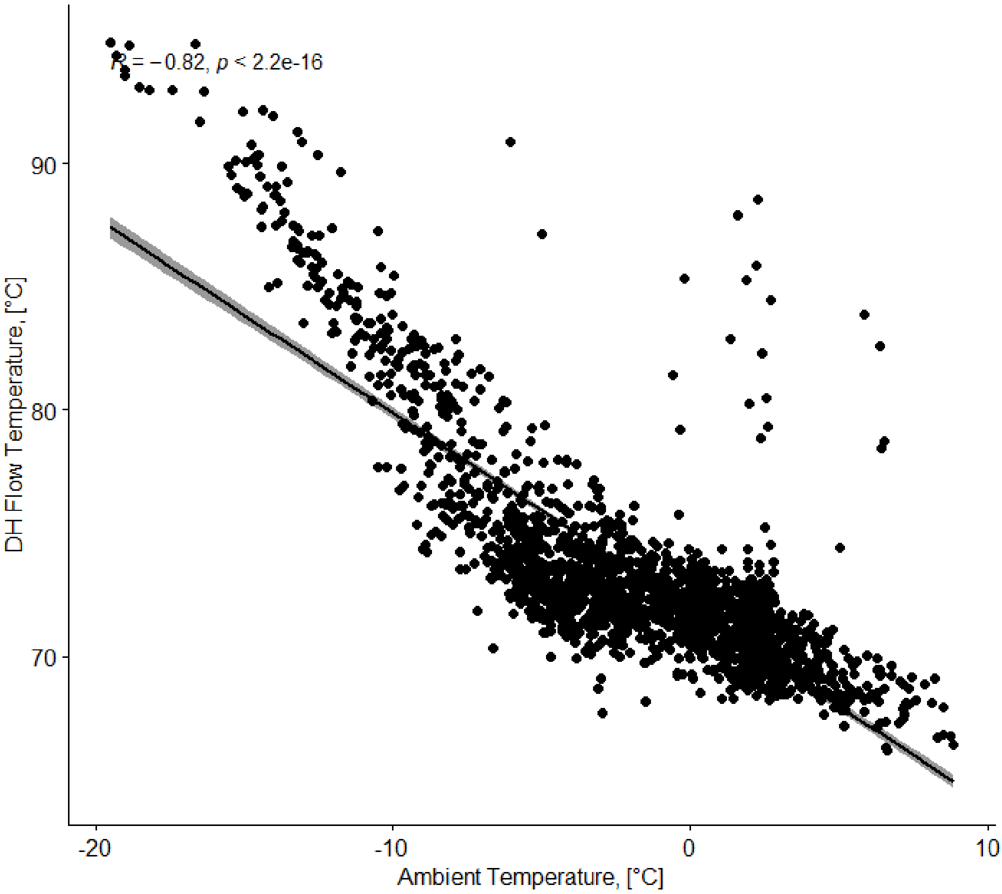


Figure 3.11: Visualisation of the DH Flow Temperature vs Ambient Temperature in Building II. Scatter plot applied for a linear regression check

The benchmarking of the correlation coefficients given in Table 3.6 was obtained statistically, while the solo coefficient provided for the Ambient Temperature and DH Flow Temperature are computed within the plot. What is significant is the precision of both methods and their calculations are the same. However, the plotting technique is cumbersome and time-consuming compared to the statistical summary.

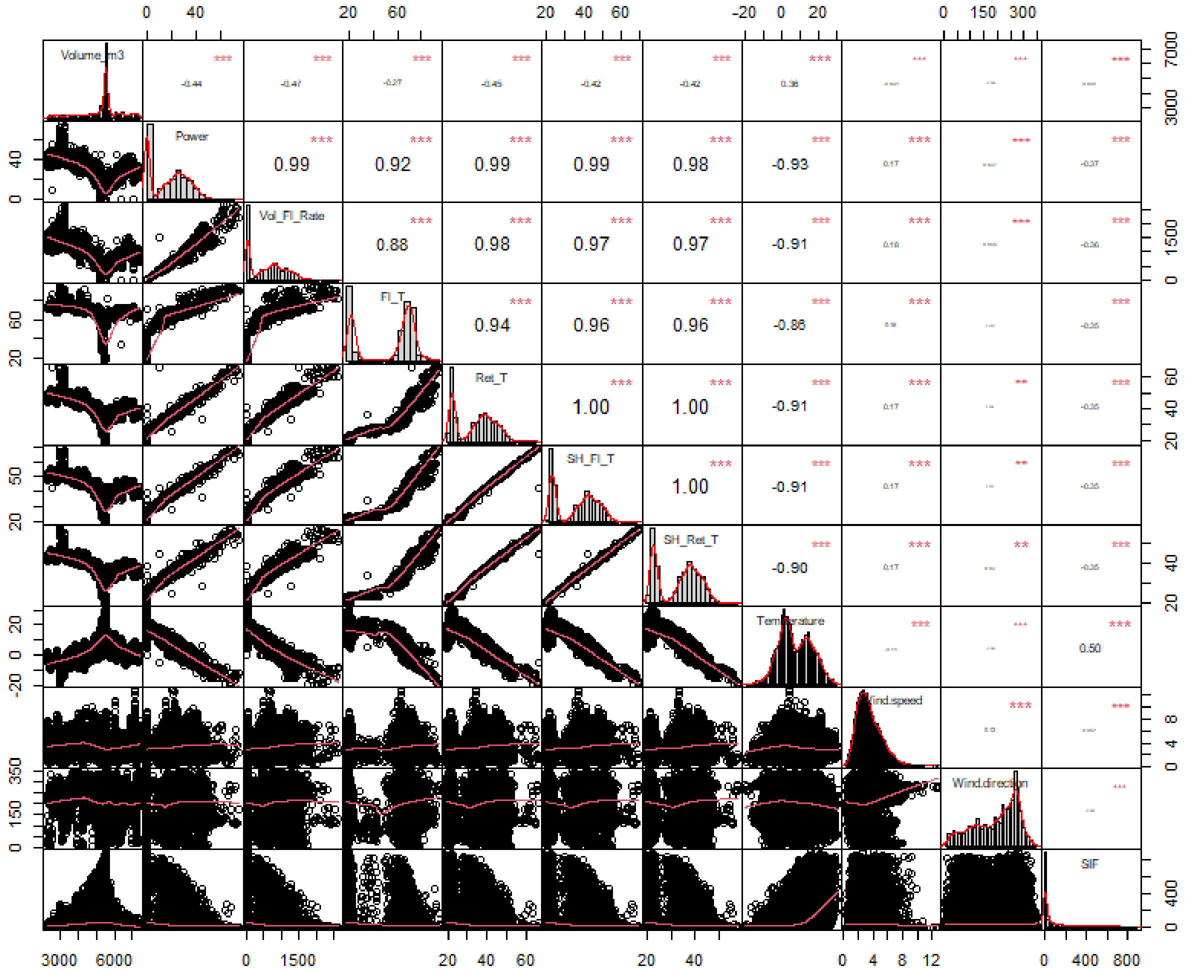


Figure 3.12: The correlation chart of Building II variables. (Variables distributions are in diagonal; Bottom of the diagonal displays bivariate analysis; Top of the diagonal shows the correlation coefficients and significance level as stars)

According to the correlation coefficients in Table 3.6 and Figure 3.12, we can detect a similar pattern as it happened in Building I. Yet, the relationship between the smart meter variables in accordance with each other seemed to be significantly strong where the r value was at least 0.92 for the correlations for Heating Power usage with other variables. Similar relationships are detected for the correlation of other variables as well. In the case of the connection of DH Return Temperature with SH Flow and SH Return Temperatures, it was virtually equal to 1. Regarding the weather data, there is a strong negative correlation between each variable of smart meter data with the Ambient Temperature and more than 0.9 for each case. What makes the current benchmarking different from the previous analysis is the medium correlation of SIF with the smart meter data where it is around -0.35.

One can see in Figure 3.12 that the high correlation in Heating Power with DH Flow and DH Return Temperatures during winter time when the system is operating. Whereas SH Flow and SH Return Temperatures have a correlation coefficient of almost 1 and this might occur since it is the same heat exchanger and hence, they both have the same temperature drop. Besides, the relationship between the Ambient Temperature and the smart energy meter data is pivotal because when the gap appears in the energy meter, the only weather variable which is highly correlated with the energy data is the Ambient Temperature. This relationship provides an opportunity to fill the gaps solely based on the weather data and this high correlation betwen Ambient Temperature and the energy data is valid excluding the not heating season.

Table 3.6: Correlation matrix for the aggregation of smart meter and weather data for Building II

Volume	1.00									
Power	-0.44	1.00								
Flow T	-0.27	0.92	1.00							
Return T	-0.45	0.99	0.94	1.00						
SH Flow T	-0.42	0.99	0.96	1.00	1.00					
SH Return T	-0.42	0.98	0.96	1.00	1.00	1.00				
Ambient T	0.36	-0.93	-0.86	-0.91	-0.91	-0.90	1.00			
WS	-0.09	0.17	0.16	0.17	0.17	0.17	-0.11	1.00		
WD	-0.05	0.04	0.00	0.03	0.03	0.03	0.06	0.12	1.00	
SIF	-0.05	-0.37	-0.35	-0.35	-0.35	-0.35	-0.50	0.05	-0.04	1.00
	Volume	Power	Flow T	Return T	SH Flow T	SH Return T	Ambient T	WS	WD	SIF

3.2.3 Building III. Correlation Analysis

The correlation analysis result of Building III was particularly close to those of Building II (Table 3.7). Correlation values of smart meter data variables for each type of benchmarking were more than 0.9 and positively correlated.

In contrast to the DHW Flow Temperature correlation coefficients in Building I, the same parameters for Building III showed a strong positive correlation with energy meter variables. Even though, there may not be a high relationship in real life as usage of the DHW does not depend on the season or other parameters, but for the current data set it appeared to have a strong correlation and hence linear relationship was identified with smart meter data and the Ambient Temperature.

Table 3.7: Correlation matrix for the aggregation of smart meter and weather data for Building III

Volume	1.00										
Power	-0.43	1.00									
Flow T	-0.27	0.91	1.00								
Return T	-0.41	0.98	0.94	1.00							
DHW Flow_T	-0.41	0.97	0.95	0.99	1.00						
SH Flow T	-0.41	0.98	0.94	1.00	1.00	1.00					
SH Return T	-0.41	0.98	0.95	0.99	1.00	1.00	1.00				
Ambient T	0.37	-0.87	-0.80	-0.86	-0.83	-0.84	-0.84	1.00			
WS	-0.09	0.15	0.14	0.15	0.15	0.15	0.15	-0.11	1.00		
WD	-0.05	0.00	-0.03	0.00	-0.01	-0.01	-0.01	-0.06	0.12	1.00	
SIF	0.06	-0.38	-0.33	-0.37	-0.36	-0.36	-0.36	0.50	0.05	0.04	1.00
	Volume	Power	Flow T	Return T	DHW Flow_T	SH Flow T	SH Return T	Ambient T	WS	WD	SIF

Outdoor temperature with SIF has respectively strong and medium correlations with heating meter variables. However, the rest of the weather variables did not have any high correlation coefficients with the energy data. Figure 3.13 shows the same correlogram as for Building I applying the same rules regarding the colour and the size of the shapes.

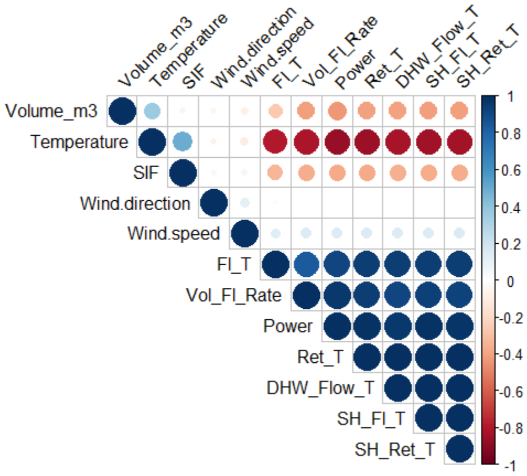


Figure 3.13: Correlogram of Building III data between the smart meter and weather variables

3.2.4 Building IV. Correlation analysis

On the other hand, one can see in the profile of DH Flow Temperature and Heating Power correlation with DH Return Temperature that there is the same strong relationship but negative. A similar pattern can be identified for DH Return Temperature with SH Flow and SH Return Temperatures as well. Repeatedly, no connection of DHW Flow

Temperature with other variables was identified, neither was the weather data except the Ambient Temperature which had high negative correlations with DH variables.

The correlation coefficient between the Ambient Temperature and the smart meter data ranges between 0.66 – 0.86 (Figure 3.14). Despite including the summer season for the analysis, high correlations are still maintained. This occurs due to the strong correlation during the heating season between the Ambient Temperature and the energy variables and hence, if only the cold period was considered, then coefficient would near 1.

Table 3.8 and Figure 3.14 the correlation analysis of Building IV. Unlike in Building II and Building III, there are not the same level of strong correlations around 0.9 to 1 between smart meter data and the Ambient Temperature. Yet, the coefficients are similar to those of Building I where there was a high correlation within smart meter variables, and it is mostly between positive 0.6 to 0.7.

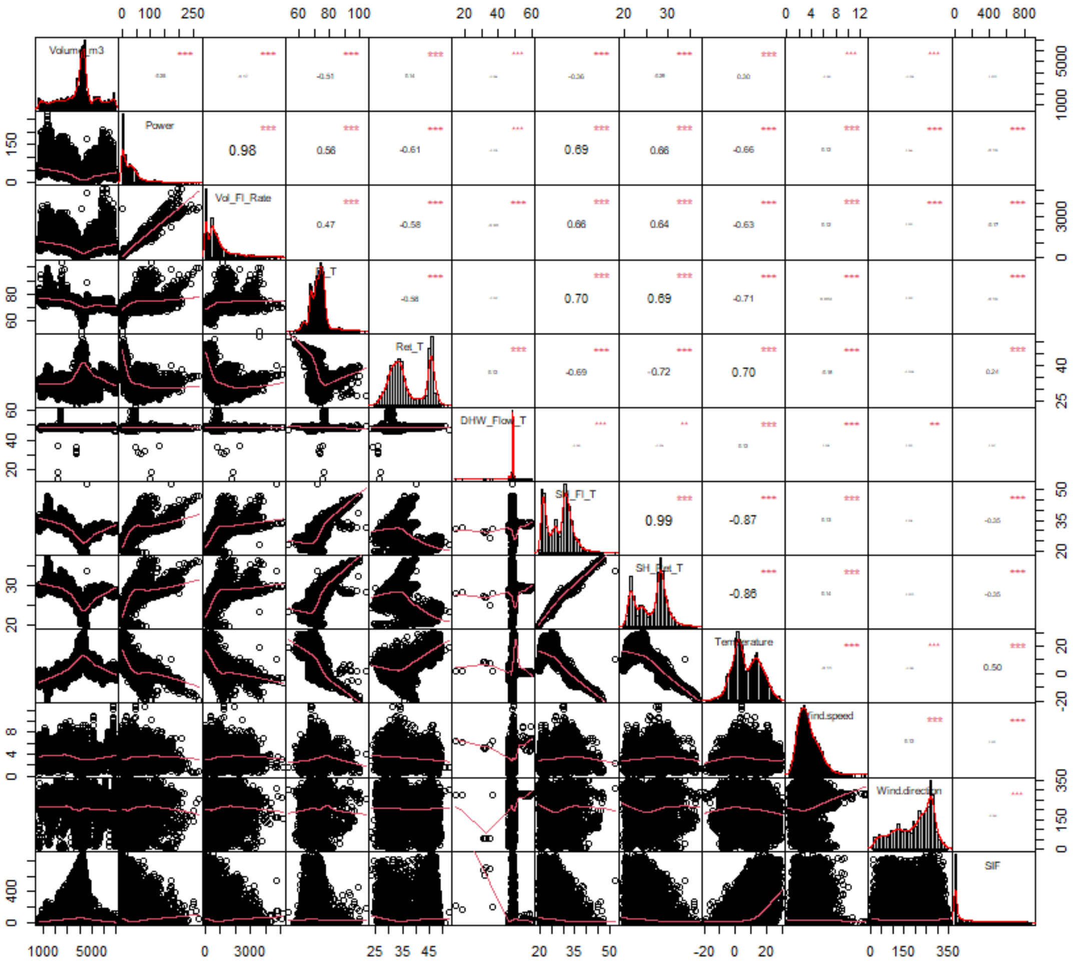
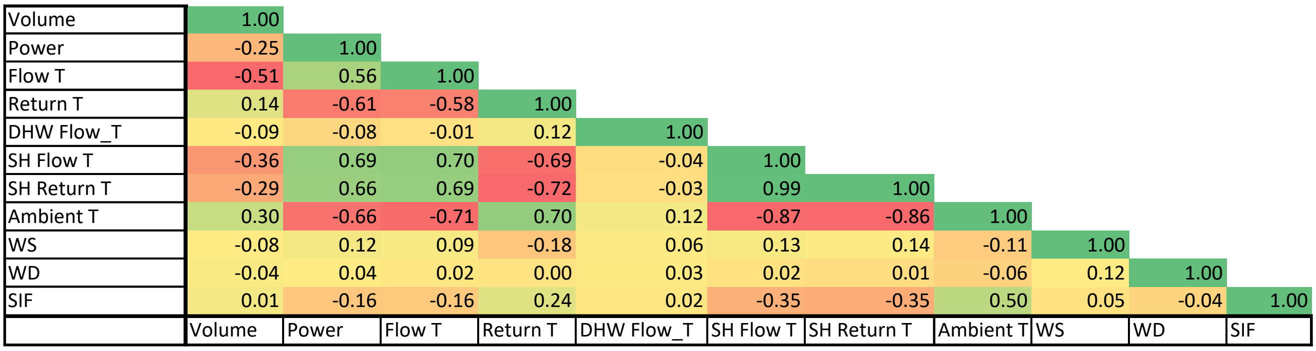


Figure 3.14: The correlation chart of Building IV variables. (Variables distributions are in diagonal; Bottom of the diagonal displays bivariate analysis; Top of the diagonal shows the correlation coefficients and significance level as stars)

On the other hand, one can see in the profile of DH Flow Temperature and Heating Power correlation with DH Return Temperature that there is the same strong relationship but negative. A similar pattern can be identified for DH Return Temperature with SH Flow and SH Return Temperatures as well. Repeatedly, no connection of DHW Flow Temperature with other variables was identified, neither was the weather data except the Ambient Temperature which had high negative correlations with DH variables.

The correlation coefficient between the Ambient Temperature and the smart meter data ranges between 0.66 – 0.86 (Figure 3.14). Despite including the summer season for the analysis, high correlations are still maintained. This occurs due to the strong correlation during the heating season between the Ambient Temperature and the energy variables and hence, if only the cold period was considered, then coefficient would near 1.

Table 3.8: Correlation matrix for the aggregation of smart meter and weather data for Building IV



Volume	1.00										
Power	-0.25	1.00									
Flow T	-0.51	0.56	1.00								
Return T	0.14	-0.61	-0.58	1.00							
DHW Flow_T	-0.09	-0.08	-0.01	0.12	1.00						
SH Flow T	-0.36	0.69	0.70	-0.69	-0.04	1.00					
SH Return T	-0.29	0.66	0.69	-0.72	-0.03	0.99	1.00				
Ambient T	0.30	-0.66	-0.71	0.70	0.12	-0.87	-0.86	1.00			
WS	-0.08	0.12	0.09	-0.18	0.06	0.13	0.14	-0.11	1.00		
WD	-0.04	0.04	0.02	0.00	0.03	0.02	0.01	-0.06	0.12	1.00	
SIF	0.01	-0.16	-0.16	0.24	0.02	-0.35	-0.35	0.50	0.05	-0.04	1.00
	Volume	Power	Flow T	Return T	DHW Flow_T	SH Flow T	SH Return T	Ambient T	WS	WD	SIF

4 Result and Analysis

In the data sets of the case study, as it was examined previously there were no missing values in any of the four data sets. In order to test the various imputation methods, we are required to introduce missing values. This can be implemented artificially and with certain consequences for each variable in each set of data.

4.1 Introduction of artificial NA values to data sets

A period of two weeks was chosen for each variable in the smart energy meter for each data set to enter the gaps. For the sake of simplicity and to avoid any inconveniences that might affect the output result, it was decided to select two weeks without any outliers by applying the boxplot technique to test for anomalies on variables individually. So that all the data seemed to be normally distributed within their range. It means 6 variables from smart energy meter data from 4 total sets were selected summing up to 48 weeks (Heating Power, DH Flow Temperature, DH Return Temperature, SH Flow Temperature, SH Return Temperature, DHW Flow Temperature). However, he selected weeks are not the same for each case.

One can see in Figure 4.1 five boxplots illustration of DH Flow and DH Return Temperatures, SH Flow and SH Return Temperatures, and DHW Flow Temperature in Building I. 2 weeks of January in DH Return and DH Flow Temperatures, 2 weeks of April in SH Flow and SH Return Temperatures, and 2 weeks of July in DHW Flow Temperature were chosen and then tested. No outlier was identified for any of the variables which was the preliminary necessity before commencing to proceed on the next steps. As it can be observed, the selection of weeks was decided depending on the cleanliness of the two weeks and vary from variable to variable and data sets as well. Implying 2 weeks of January picked for DH Return Temperature of Building I may not be the same for the rest of the buildings as they are considered on an individual basis. The rest of the boxplot figures for other buildings are provided in Figure C, Figure D, and Figure E in Appendix.

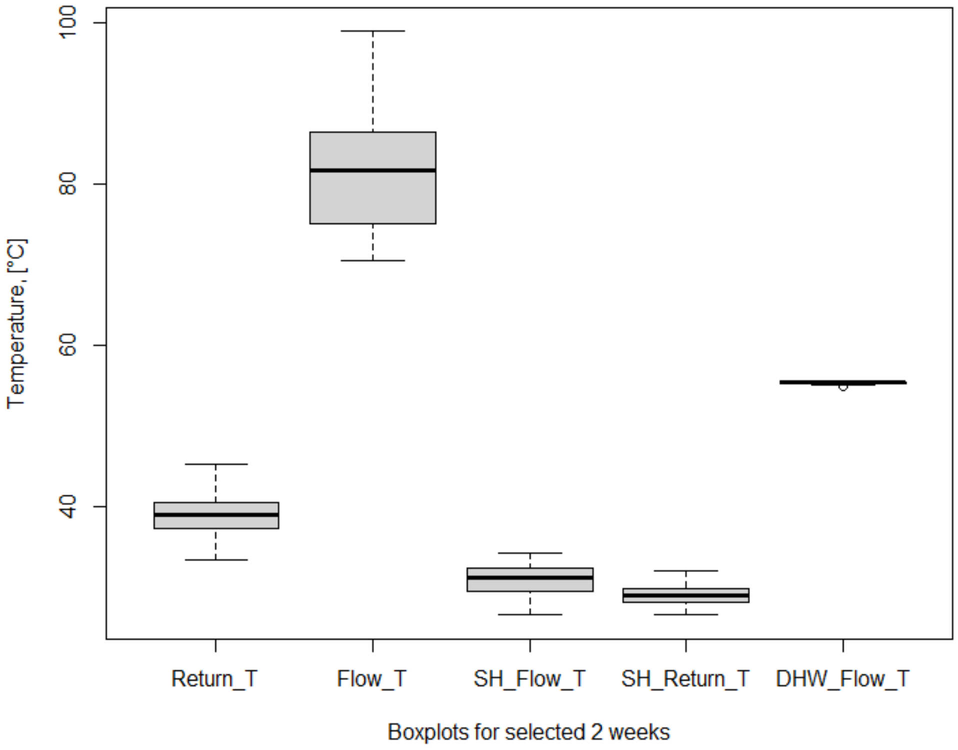


Figure 4.1: Outlier analysis based on a boxplot summary of the smart energy meter data in Building I.

Furthermore, a total of 30 hours of the gap were introduced in the following scenarios: 15 gaps of 2 consecutive (uninterrupted) hours each, 10 gaps of 3 consecutive (uninterrupted) hours each, 6 gaps of 5 consecutive (uninterrupted) hours each, 5 gaps of 6 consecutive (uninterrupted) hours each, 3 gaps of consecutive (uninterrupted) hours each, 2 gaps of consecutive (uninterrupted) hours each. Each of the combinations sums up to total of 30 hours (Table 4.1). This way, we entered 6 different time-gap based scenarios for each variable within their selected individual weeks but with various consecutive number gaps following those rules.

Table 4.1: Introducing the gaps with total 30 hours gaps with various combinations

Total 30 hours missingness scenarios					
15 gaps of 2 hours	10 gaps of 3 hours	6 gaps of 5 hours	5 gaps of 6 hours	3 gaps of 10 hours	2 gaps of 15 hours

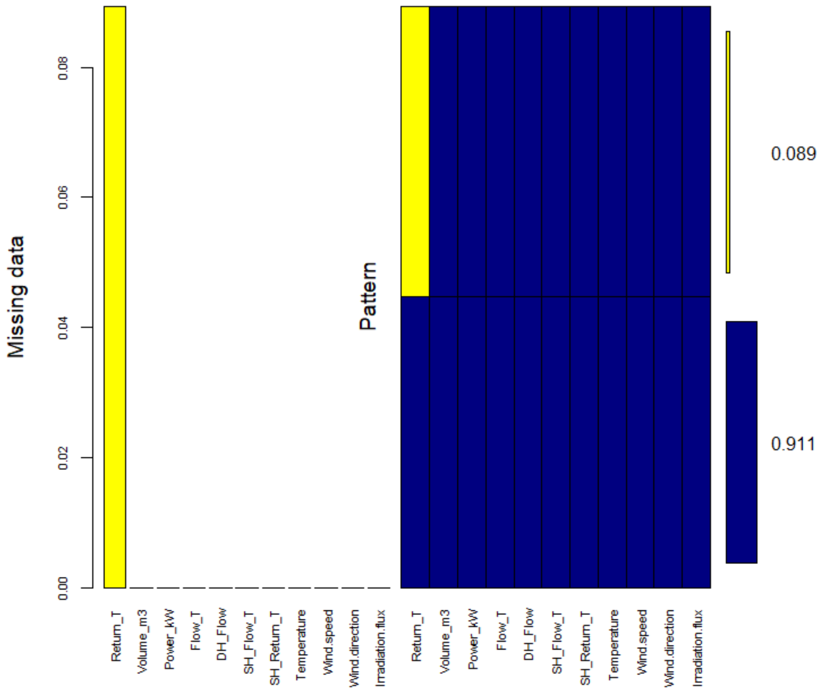


Figure 4.2: Input of missing points in DH Return Temperature of Building I

For demonstration purposes, the DH Return Temperature in Building I was considered to show the missingness when the gaps are introduced and how they are distributed. Initially, 15 gaps of 2 hours were introduced in the DH Return Temperature while all other variables from smart energy meter and weather data did not undergo any alterations. Furthermore, portion of missingness in DH Return Temperture accounted for 0.81 %. As can be expected, the share (percentage) of the gaps remains unchanged for all time-gap scenarios. One can observe in Figure 4.2 that the percentage of missing points is 8.9 % which is solely in DH Return Temperature for two weeks period only, whereas the complete observations of DH Return Temperature without missing rows consist of 91.1 % of the variable. On the contrary, the other variables are fully complete with no gaps, and thus there is no corresponding yellow column on the left “Missing data” plot.

Figure 4.3 and Table 4.2 summarise the 15 gaps of 2 hours scenario of DH Return Temperature in Building I where the time and date the missing points occur, and their visualizations are provided. Following this, we enter 30 hours gaps for DH Flow Temperature, Heating Power, SH Flow, and SH Return Temperatures with 2 hours consecutive missingness for all of the datasets. After, 3 hours, 5 hours, 6 hours, 10 hours, and 15 hours scenarios were implemented the same way.

Table 4.2: Dates and times for which 2 hours gaps are introduced in DH Return Temperature of Building I

Date	Time
1 January 2019	02:00 – 04:00
1 January 2019	18:00 – 20:00
3 January 2019	00:00 – 02:00
4 January 2019	05:00 – 07:00
4 January 2019	17:00 – 19:00
5 January 2019	13:00 – 15:00
6 January 2019	14:00 – 16:00
8 January 2019	20:00 – 22:00
9 January 2019 – 10 January 2019	23:00 – 01:00
11 January 2019	04:00 – 06:00
12 January 2019	01:00 – 03:00
12 January 2019	21:00 – 23:00
13 January 2019	11:00 – 13:00
14 January 2019	00:00 – 02:00
14 January 2019	15:00 – 17:00

4.2 Univariate Imputation Methods

As the gaps were inputted with certain orders, the next step is to apply imputation techniques in order to observe how they perform. For that purpose, univariate imputation techniques were tested at first. Specifically, starting from the simplest ones as Mean and Median with LOCF and NOCB growing into more complex computation methods such as Interpolation, Moving Average. Each method was performed for 6 time-gap scenarios in each variable and after, their accuracy was computed using an RMSE to analyse their performance. RMSE was calculated only for the places where the missing points were introduced. In other words, 30 artificial gaps were considered for the calculation of RMSE and it was computed in the following way:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{30}(x_i - \widehat{x_i})^2}{30}} \tag{4.1}$$

Where i – variable i, x_i – an observed value, $\widehat{x_i}$ – a predicted value.

Table 4.3 summarizes the result of all the scenarios applied to fill the gaps in the DH Return Temperature in Building I based on 2 hours, 3 hours, 5 hours, 6 hours, 10 hours, 15 hours gaps with a total of 30 hours gaps. For demonstration purposes, the analysis of DH Return Temperature in Building I continued to be examined and illustrated for all the scenarios. Other variables of four data sets followed the same scenarios with the same methods.

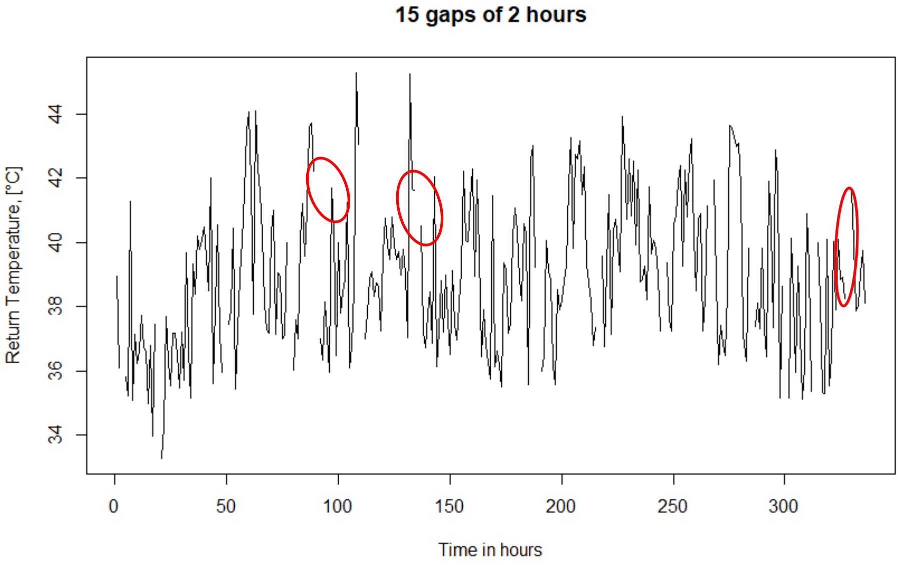
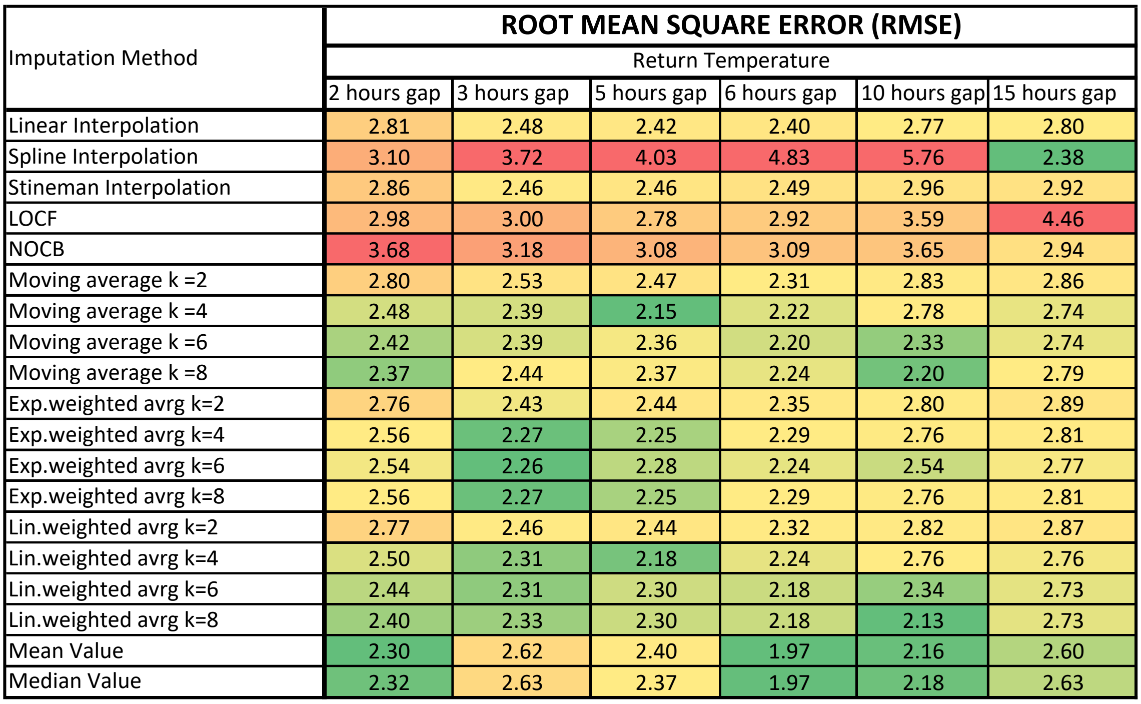


Figure 4.3: 15 gaps of 2 consecutive hours representation of DH Return Temperature

One can see in Table 4.3 that Mean, Median, Linear Weighted Average (k=8 which means in total 8 values: 4 above and 4 below the gap values are contributed to the computation of the missing point) demonstrated the best accuracy for 15 gaps of 2 consecutive hours having RMSE score of 2.32 °C, 2.30 °C, and 2.44 °C respectively, whereas the accuracy of the moving average techniques with the different size of windows was slightly lower (average of 2.51 °C of four variations) compared to the best methods for the given scenario. On the contrary, NOCB, LOCF, and Spline Interpolation showed the worst accuracy when they replaced the gaps with calculated values (respective 3.68 °C, 2.98°C, and 3.10 °C)

At the opposite side of the table, one can observe 2 gaps of 15 consecutive hours. As opposed to 15 gaps of 2 consecutive hours, the accuracy of Spline Interpolation improved considerably reaching 2.38 °C which was more than twice as less as of 10 hours gaps result. Thus, it happened to be the best accuracy for this condition. Whereas Mean and Median remained to have better RMSE along with Moving Average group and its variations. On the other hand, LOCF and NOCB perpetuated to display one of the least accurate techniques throughout the time-gap scenarios. Besides, for most cases, the wider the width of the Moving Average window, the better the result of RMSE. Even though it is not true for all the scenarios, but the trend is met for the majority of the cases.

Table 4.3: RMSE of univariate imputation methods applied for the DH Return Temperature gaps introduced artificially



Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	Return Temperature					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
Linear Interpolation	2.81	2.48	2.42	2.40	2.77	2.80
Spline Interpolation	3.10	3.72	4.03	4.83	5.76	2.38
Stineman Interpolation	2.86	2.46	2.46	2.49	2.96	2.92
LOCF	2.98	3.00	2.78	2.92	3.59	4.46
NOCB	3.68	3.18	3.08	3.09	3.65	2.94
Moving average k =2	2.80	2.53	2.47	2.31	2.83	2.86
Moving average k =4	2.48	2.39	2.15	2.22	2.78	2.74
Moving average k =6	2.42	2.39	2.36	2.20	2.33	2.74
Moving average k =8	2.37	2.44	2.37	2.24	2.20	2.79
Exp.weighted avrg k=2	2.76	2.43	2.44	2.35	2.80	2.89
Exp.weighted avrg k=4	2.56	2.27	2.25	2.29	2.76	2.81
Exp.weighted avrg k=6	2.54	2.26	2.28	2.24	2.54	2.77
Exp.weighted avrg k=8	2.56	2.27	2.25	2.29	2.76	2.81
Lin.weighted avrg k=2	2.77	2.46	2.44	2.32	2.82	2.87
Lin.weighted avrg k=4	2.50	2.31	2.18	2.24	2.76	2.76
Lin.weighted avrg k=6	2.44	2.31	2.30	2.18	2.34	2.73
Lin.weighted avrg k=8	2.40	2.33	2.30	2.18	2.13	2.73
Mean Value	2.30	2.62	2.40	1.97	2.16	2.60
Median Value	2.32	2.63	2.37	1.97	2.18	2.63

At first, it was expected that the accuracy would get gradually better by the increase of the hour gaps meaning that RMSE for 15 hours gap should be considerably smaller than for 2 hours gap (the smaller the RMSE vslue, the better the accuracy). Yet, this anticipation was not met for almost any of the methods but Spline Interpolation, and one can see the consistent decrease or increase was not detected either. Therefore, the RMSE value fluctuates increasing and decreasing without certain order.

Overall, for DH Return Temperature in Building I, Mean, Median, and Linear Weighted Average (k=8) showed the best accuracy for all the scenarios having a total sum of 14 °C each. In contrast, Spline, despite peaking at the top at the end, it decreased substantially having a total of 23.8 °C for the sum of 6 scenarios. Regarding the highest RMSE values, LOCF and NOCB scored second and third after Spline with values of 19.7 °C and 19.6 °C accordingly.

Figure 4.4 shows the best and worst three methods compared to the original data for 15 gaps of 2 hours. For the sake of simplicity and to better observe the alterations, only the first 50 hours were presented, where the dark grey represents the original week without any missing points in January for DH Return Temperature in Building I. Their respective missing points are provided in Table 4.2.

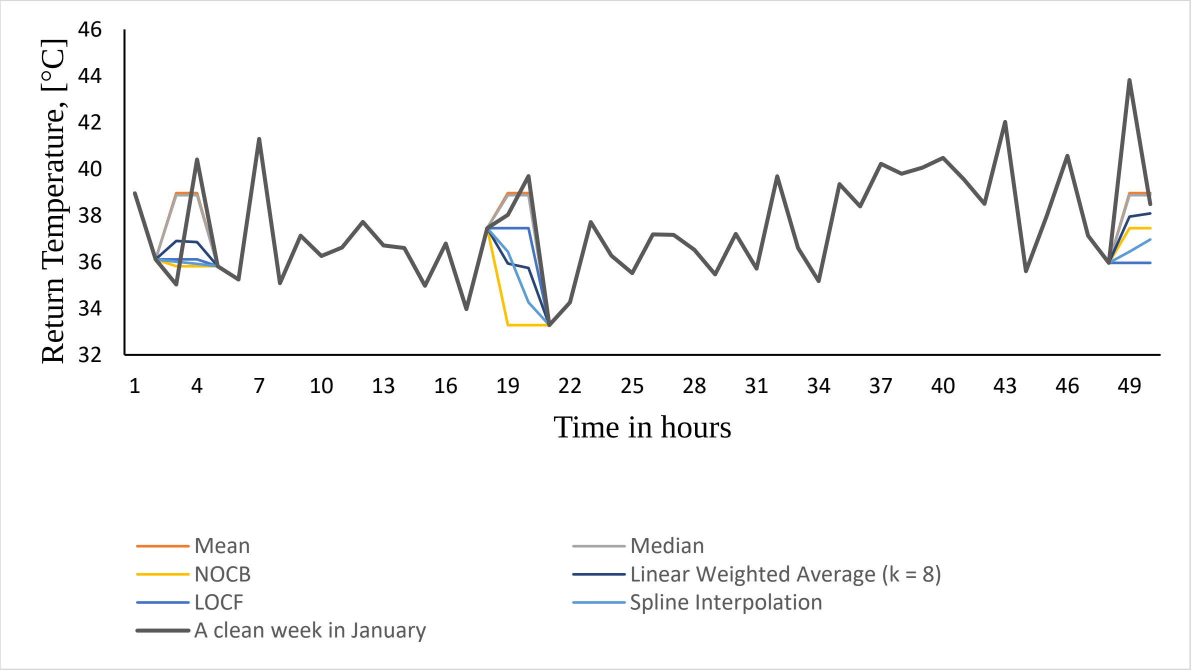


Figure 4.4: Univariate imputation methods filling the gap performance of DH Return Temperature on 15 gaps of 2 hours. The original week, 3 worst, and 3 best methods illustrations

As it is provided in Table 4.2, first, the missing points appear between 00:00 – 02:00 at midnight, which can be seen as the discrepancy of the line plots at around that time in Figure 4.4. The measured DH Return Temperature was 35.03 °C and 40.4 °C for the first 2 hours gap. Thus, Mean, Median and Linear Moving Average (k=8) filled those gaps with 38.95 °C (Mean), 38.87 °C (Median Interpolation) and 36.9 °C, 36.85 °C (Stineman Interpolation), whereas LOCF, NOCB, and Spline Interpolation computed and replaced those gaps with 36.11 °C, 35.81 °C, and 36.01 – 36.65 °C respectively. In the same way, the other two gaps, which were between 18:00 – 20:00 on February 15 and 00:00 – 02:00 on February 17, were filled with all the applied methods carrying out the univariate imputations.

Linear and Stineman Interpolations for the taken precise example did not fill the gaps with high accuracy compared to those of Mean, Median, and Linear Weighted Average (k=8). However, for other variables within the same data set as well as for variables in other data sets, Linear and Stineman Interpolations demonstrated the best accuracy among all the univariate imputation methods tested.

4.3 Physical Computation

The next step consists of the computation based on the physical knowledge we have, and we attempted to calculate DH Return Temperature applying the Heating Power formula:

$$Q = \dot{V}\rho_w c_w (T_{Flow} - T_{Return}) \tag{4.2}$$

Where $\dot{V}$ − Volumetric flow rate [l/h], ρ_w − density of water [kg/m^3], c_w − specific heat capacity of water [J/kg °C], T_{Flow} − DH Flow Temperature [°C], T_{Return} − DH Return Temperature [°C].

One can see that there is a correlation between DH Flow Temperature, DH Return Temperature, and Heating Power. Consequently, this allows us to calculate from this relationship DH Return Temperature applying the physical formula. Yet this scenario is instrumental only for these three variables as there is not enough physical knowledge about the rest of the variables. The core idea of performing this case is to test if the multiple imputation techniques are good enough in their predictions to find the correlation between these variables (e.g., as if they can identify this physical relationship between DH Flow Temperature, DH Flow Temperature and Heating Power) and demonstrate the similar results. Hence, the scenario was based on as if when there is a gap in DH Return Temperature other variables are complete which allowed us to perform the computation.

Knowing that we have gaps in DH Return Temperature which we introduced artificially, using we can fill them by calculating those points mathematically. Hence, our final formula would be:

$$T_{Return} = T_{Flow} - \frac{Q}{\dot{V}\rho_w c_w} \tag{4.3}$$

Initially, the Heating Power was calculated on an hourly basis as the data set is structured. This was carried out to test if the Heating Power in the data set has the same values as Heating Power computed applying the existing formula. Hence, despite having a slight alteration which was negligible, it was virtually the same having an RMSE value of 0.22 °C.

Table 4.4: Comparison of the DH Return Temperature measured and computed using the physical formula for the first 20 hours in January

Time	Measured Return Temperature [°C]	Calculated Return Temperature [°C]
00:00	38.95	38.95
01:00	36.11	36.11
02:00	35.03	35.03
03:00	40.4	40.4
04:00	35.81	35.81
05:00	35.24	35.24
06:00	41.29	41.29
07:00	35.08	35.08
08:00	37.13	37.13
09:00	36.25	36.25
10:00	36.62	36.62
11:00	37.72	37.72
12:00	36.71	36.71
13:00	36.6	36.6
14:00	34.97	34.97
15:00	36.79	36.79
16:00	33.97	33.97
17:00	37.45	37.45
18:00	38.02	38.02
19:00	39.69	39.69

After, as all the needed measurements are provided in the data set along with the constant values, the physical calculation was performed to observe the change within the measured and computed DH Return Temperatures. Table 4.4 summarizes the first 20 hours of the data set to compare the output result of the two methods. One can see that the gaps are shaded in blue which was between 02:00 and 04:00 in the morning in January, second consecutive gap took place in the evening of 18:00 to 20:00. Hence, according to the table, the measured DH Return Temperature at midnight was 35.03 °C and 40.4 °C, which later was substituted with missing points. According to the calculations, those missing points would be filled with the exact same values as they were measured. The same was displayed for the second gap: 38.02 °C and 39.69 °C for both measured and computed observations. This method can be valuable if we have the required variables to compute the missing points and apply them to examine the accuracy of software-based prediction methods. This would give us an opportunity, in case there is missing data, to observe how accurate the predictions as the mathematical solution of the gaps should be closer to the recorded observations.

4.4 Multiple Imputation

Multiple imputations were performed applying several techniques coming from different packages which are MICE, Amelia, and MissForest. As the name suggests, multiple imputation methods operate with a data frame as opposed to the univariate imputation methods where only one single variable was considered at a time.

All multiple filling the gap techniques follow the same missing point consecutiveness on 6 time-gap scenarios applied for univariate analysis. Apart from that, as it is a data frame, it was decided to include weather variables as well. Because depending on the correlation between variables which was performed previously, their accuracy varies considerably. Consequently, focused on the correlation analysis of the variables, the following additional scenarios were created for multiple imputation methods:

- *Scenario I:* Only highly correlated variables are included for the imputation where the correlation coefficient is more than 0.5 for both negative and positive correlations
- *Scenario II:* Combination of high and medium correlated variables are applied to fill the gaps with a correlation coefficient of higher than 0.3 for both negative and positive correlations
- *Scenario III:* Including all variables with high, medium, and small correlations

4.4.1 Scenario I. Multiple imputations of highly correlated variables

The variables with a Pearson correlation coefficient higher than 0.5 (Heating Power, DH Flow Temperature, SH Flow Temperature, and DHW Flow Temperature) were included for the computation of the first scenario based on Table 3.5, in addition to the first conditions where two weeks of 30 hours gap were selected for each variable in each data set. Besides, only the Ambient Temperature from weather variables was included for predictions as had a high coefficient of 0.58 and positively correlated. Table 4.5 summarizes the accuracy rate of each multivariate technique regarding the 6 different consecutive gap scenarios of DH Return Temperature.

Table 4.5: RMSE of multiple imputation methods with solely highly correlated variables

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE 2 weeks: High correlations	3.34	3.00	3.16	2.30	3.03	2.84
Amelia 2 weeks: High correlations	2.82	3.23	2.87	3.56	3.19	3.77
missForest 2 weeks: High correlations	1.42	1.76	1.80	1.96	1.46	1.60

One can see that for 2 hours gap missForest demonstrated the best accuracy having an RMSE of 1.42 °C which was well above two times better than that of MICE imputation. Throughout the gap scenarios, missForest remained to show the best RMSE value, every time being less than 2 °C in total, whereas Amelia and MICE had substantially worse accuracy than missForest. For shorter gaps, Amelia performed better than MICE. In contrast, the bigger the consecutive gaps, the worse the RMSE value of Amelia.

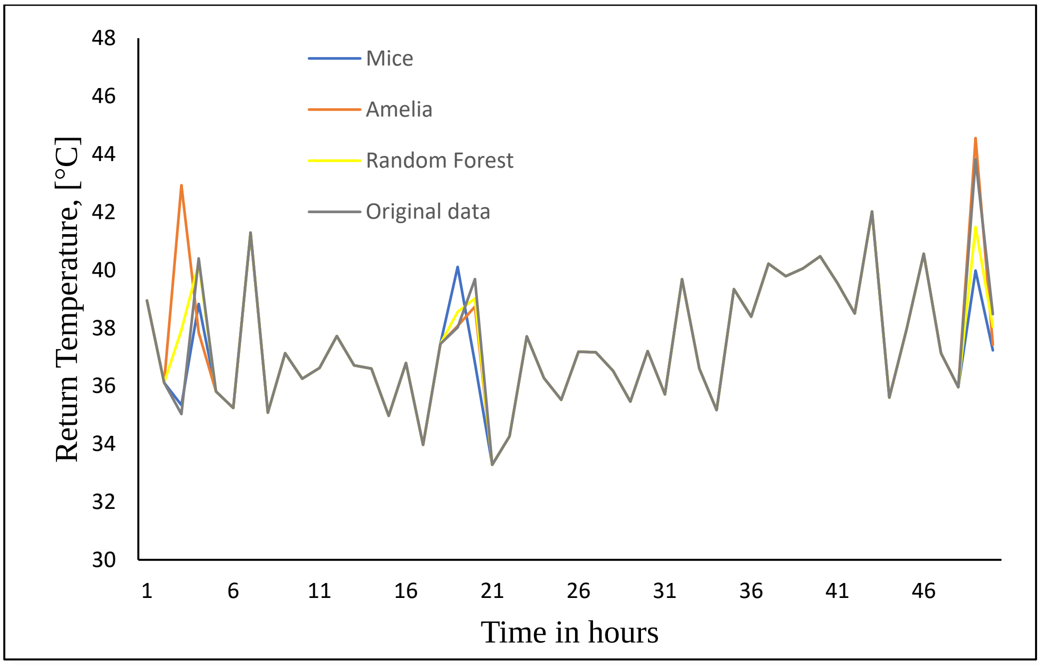


Figure 4.5: Illustration of multiple imputation techniques with highly correlated variables for the first 50 hours of the week of 15 gaps of 2 hours in DH Return Temperature of Building I for two weeks period

Figure 4.5 illustrates how the multiple imputation methods filled the gaps in the first 50 hours of January of DH Return Temperature in Building I. The week was shortened to demonstrate the result on a smaller scale to better observe the alteration. As it can be seen in the plot, for the first gap between 02:00 and 04:00, the closest predictions were from MICE which filled the gap with 35.34 °C and 38.83 °C respectively in accordance to the original data of 35.03 °C and 40.04 °C. Similar results were predicted by missForest as well with 37.92 °C and 40.22 °C. Whereas Amelia showed a slightly different outcome of 42.92 °C and 37.83 °C. For the second gap between 19:00 and 21:00, missForest performed better than others. Even though MICE predicted the values of first gaps (02:00 – 04:00) with higher accuracy, the accuracy worsened for further predictions and this is the reason why it had an RMSE of around 3.34 °C for 15 gaps of 2 hours which was the worst for that case.

4.4.2 Scenario II. Multiple imputations of highly and medium correlated variables

Furthermore, the combination of both strong and medium correlations was selected based on Table 3.5. Hence, the coefficient must be higher than 0.3.

Table 4.6: RMSE of multiple imputation methods with high and medium correlated variables

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE 2 weeks: High and Medium Correlations	2.20	2.57	2.69	2.52	2.91	2.45
Amelia 2 weeks: High and Medium Correlations	2.25	2.52	2.23	2.42	2.65	2.69
missForest 2 weeks: High and Medium Correlations	1.27	1.43	1.59	1.69	1.15	1.44

Based on Table 3.5, we can conclude that there are 5 variables (Heating Power, DH Flow Temperature, SH Flow Temperature, SH Return Temperature, and Ambient Temperature) highly and medium correlated with DH Return Temperature. Compared to the first scenario, one can observe the addition of SH Return Temperature with a coefficient of 0.39 and it was absent for the first case.

One can see in Table 4.6 that missForest performed the best among three multiple imputation methods with a total RMSE of 8.57 °C in comparison with MICE and Amelia 15.3 °C and 14.8 °C respectively. For individual gaps, for almost all the cases missForest had almost two times higher accuracy than others. Whereas MICE and Amelia showed virtually similar RMSE values throughout the time-gap scenarios.

In comparison with how the observations were predicted in the scenario I, one can observe the same process for scenario II (Figure 4.6). This time, all methods performed better than in the case of the first scenario. For instance, the predicted values using missForest for the first gap were 35.88 °C and 40.4 °C as the original data set appeared to be 35.03 °C and 40.4 °C. Despite not having the same sharp accuracy as missForest, MICE and Amelia had better results as well compared to the previous scenario with only high correlated variables.

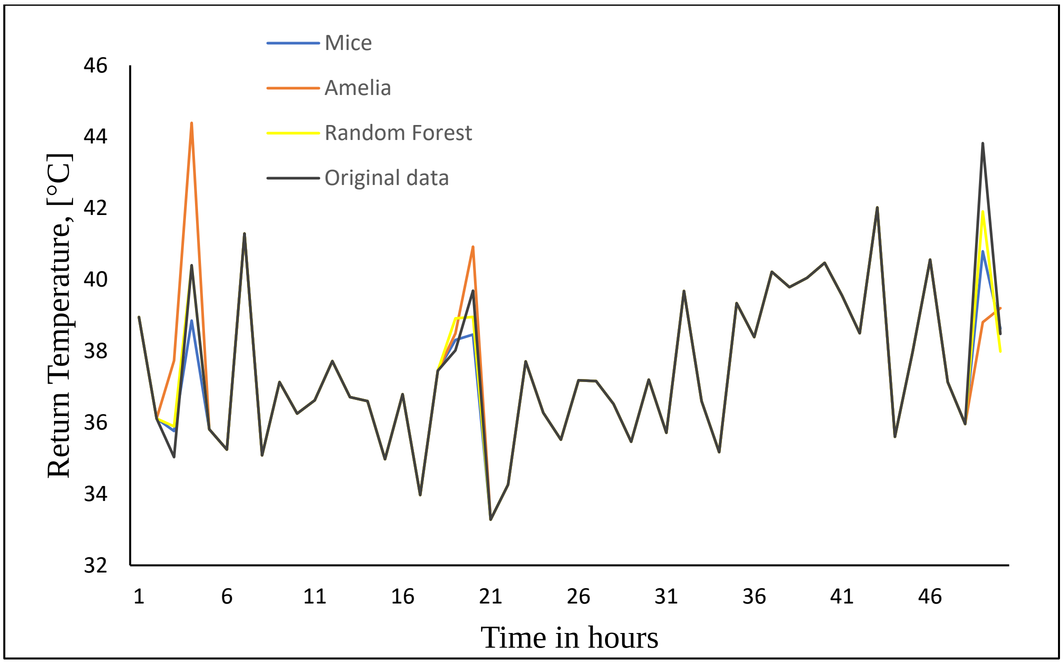


Figure 4.6: Illustration of multiple imputation techniques with high and medium correlated variables for the first 50 hours of the week 15 gaps of 2 hours in DH Return Temperature of Building I for two weeks period

4.4.3 Scenario III. Multiple imputations all correlated variables

Table 4.7 summarizes the RMSE values of multiple imputation methods with all correlations included with a condition of each of them being greater than 0. Hence, all the variables contributed to predict the missing values and replace them with those predictions. As anticipated, missForest performed better than other techniques. On the other hand, MICE and Amelia did not interchange the worst accuracy throughout the scenarios unlike for the previous cases. Total RMSE values for 6 scenarios of time-gaps for MICE, Amelia, and missForest were 13.9 °C, 12.8 °C, and 8.41 °C respectively. Hence, at almost each of the scenarios, missForest demonstrated two times better overall accuracy for DH Return Temperature in Building I. On top of that, multivariate imputation methods appear to be variable sensitive. In other words, it implies that with the increase of the number of variables in the analysis of the missing points, the overall accuracy also improves for the scenario, and it is summarised and can be identified in Table 4.8.

Table 4.7: RMSE of multiple imputation methods with all variables (correlation coefficient greater than 0)

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE 2 weeks: All correlations	2.40	2.13	2.41	2.55	1.78	2.64
Amelia 2 weeks: All correlations	1.92	1.98	2.33	2.37	1.75	2.45
missForest 2 weeks: All correlations	1.28	1.54	1.54	1.53	1.04	1.48

On the other hand, as it happened with univariate analysis as well, there was not any descending or ascending order by the increase of the hour gaps. For instance, one can see in Table 4.7 that for 2 gaps of 15 hours, the accuracy of missForest was 1.48 °C, while for 15 gaps of 2 hours RMSE showed 1.28 °C. The intermediate scenarios fluctuate without any specific order. The same could be observed using Amelia and MICE with no consistency.

Table 4.8: Total RMSE values of each method at each scenario for two weeks

Imputation methods	Total RMSE value, [°C]
Scenario I	
MICE: High correlation	17.67

Table 4.9: Total RMSE values of each method at each scenario for two weeks (continued)

Amelia: High correlation	19.44
missForest: High correlation	9.99
Scenario II	
MICE: High and Medium correlations	15.33
Amelia: High and Medium correlations	14.76
missForest: High and Medium correlations	8.57
Scenario III	
MICE All: All correlations	13.91
Amelia All: All correlations	12.79
missForest: All correlations	8.41

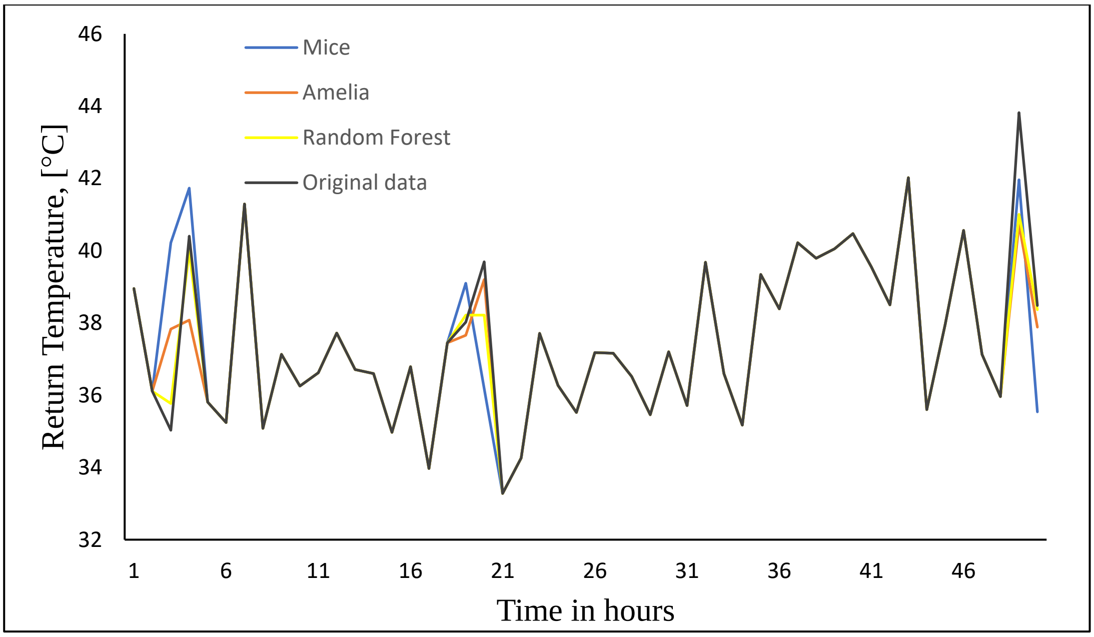


Figure 4.7: Illustration of multiple imputation techniques with all correlated variables for the first 50 hours of the week of 15 gaps of 2 hours in DH Return Temperature of Building I for two weeks period

Last but not least, the illustration of three methods on how they filled those gaps within the first 50 hours of 15 gaps of 2 hours in January for DH Return Temperature in Building I is displayed in Figure 4.7. Similarly, to the previous scenarios, the prediction performed by missForest was better for most of the cases. It predicted the first 2 hours gap with 35.76 °C and 40 °C, the result of Amelia showed the following with 37.83 °C and 38.08 °C, whereas MICE filled those gaps with 40.22 °C and 41.73 °C (the original data was 35.03 °C and 40.4 °C respectively for the gap happening between 02:00 to 04:00 in the morning). One can see that a similar pattern was perpetuated and that is the reason for having those accuracy rates for each of the methods.

4.5 Weather Variable Correlations only

One issue that could be encountered during the process of provided multiple imputation methods might be the inability of having full data with just losses of measurements in one variable only. In other words, as the predictions are made for the smart energy meter variables where missing values were introduced to only one of the variables at a time, while in real life if there is to be a gap for example in Heating Power then there are gaps for that period for all the variables in the smart meter. This happens because some fail in the sensor recordings leads to the loss of all the observations for that precise time. Hence, if it is a real-life scenario, then the only way we could predict those missing values would be relying on the weather data. Because they would be the only remaining data for the case study, assuming the rest of the smart energy meter observations were lost as well.

As a result, for this scenario, only the observations measured from the weather station are considered. Moreover, it was shown multiple times that the only variable having a strong correlation with the energy data was the Ambient Temperature which was predominantly high-negatively correlated for all cases. Despite the weak connection between other weather variables with the energy meter data, it was decided to include all of them for the analysis without the need of dividing them into correlation groups unlike was carried out for the previous three scenarios.

The result of the computation is summarised in Table 4.10 for DH Return Temperature in Building I. Alike all the scenarios performed earlier, missForest had a better accuracy showing the best RMSE value for each time-gap scenario. Whereas the accuracy of MICE was worse than Amelia and missForest for all time scenarios but for 5 gaps of 6 hours. The total RMSE values of MICE, Amelia, and missForest were 18.9 °C, 18 °C, and 12.1 °C. One can detect that there was not a considerable change for the first two methods as their total accuracy remained around the same number. On the other hand, the prediction of missing values in accordance only with weather variables worsened the RMSE value of missForest substantially. It was discovered that missForest was more variable sensitive than other multivariate imputation methods used together for the scenario.

Table 4.10: RMSE of multiple imputation methods with solely weather variables

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE: Weather Variables	3.23	3.49	2.97	2.99	3.02	3.20
Amelia: Weather Variables	3.14	3.17	2.96	3.19	2.92	2.62
missForest: Weather Variables	2.54	2.19	1.98	1.62	1.65	2.12

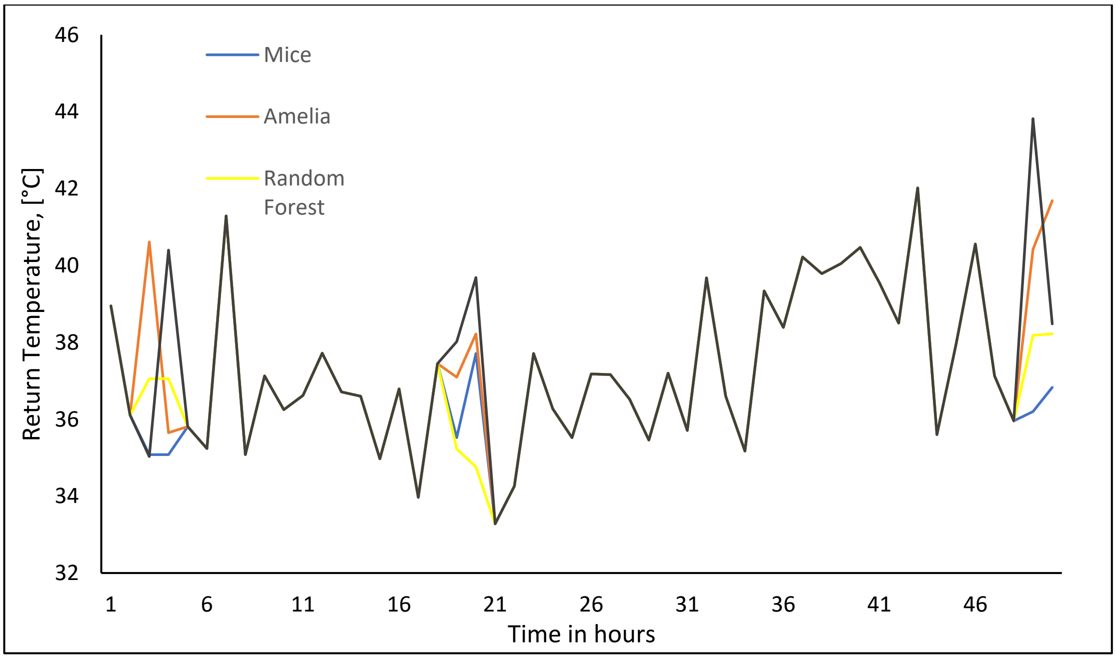


Figure 4.8: Illustration of multiple imputation techniques with only weather variables for the first 50 hours of the week of 15 gaps of 2 hours in DH Return Temperature of Building I for two weeks period

Figure 4.8 shows a similar analysis as was performed previously for the first 50 hours of the week on 15 gaps of 2 hours for DH Return Temperature in Building I. For the given period, none of the methods appears to show a constant better result, yet the total accuracy for all of the time scenarios was predicted better by applying missForest, despite the worsened RMSE value compared to the previous scenarios.

4.6 Multiple Imputation on eight weeks data

In the beginning, for all types of scenarios performed, two weeks without outliers were selected to hinder further obstacles. Next, now for the sensitivity analysis, those weeks are to be extended into eight weeks to observe how the multiple imputations accuracy will be altered with the expansion of the data set from which they can learn. This implies that instead of 336 observations from 11 smart energy meter data weather variables, it will now rely on 1344 observations with the same number of variables.

The same gap positions and their 6 time-gaps scenarios are preserved so that they will have the same conditions and the only distinction is the number of data they can apply to predict the missing points. Besides, three scenarios for multiple imputation analysis suggested considering the correlation analysis will be used as well. Consequently, alike two weeks of multiple imputation analysis, the following scenarios are to be demonstrated:

- *Scenario I with eight weeks:* Only highly correlated variables are included for the imputation where the correlation coefficient is more than 0.5 for both negative and positive correlations
- *Scenario II with eight weeks:* Combination of high and medium correlated variables are applied to fill the gaps with a correlation coefficient of higher than 0.3 for both negative and positive correlations
- *Scenario III with eight weeks:* Including all variables with high, medium, and small correlations

4.6.1 Scenario I. Multiple imputations of highly correlated variables

The first scenario applies the same multivariate methods but with eight weeks keeping the same missing points positions. After applying the techniques, the RMSE value of each of them is computed on the time-gap scenarios. Table 4.11 summarizes the accuracy of each combination of scenarios for DH Return Temperature in Building I and one can identify, regardless of alteration of scenarios and other modifications, missForest is still performing more accurately than the two others with an average of 1.53 °C. On the contrary, the analysis of MICE and Amelia on highly correlated variables displayed three times less accuracy than missForest.

As opposed to the previous scenario, where only the weather variables were included, it is more appropriate to compare the current case with a similar scenario where the conditions were the same, but the weeks were shorter (two weeks). One can observe that the extension of the weeks had a slight positive change on missForest improving the total accuracy by 0.77 °C (respective 9.99 °C for two weeks and 9.20 °C for eight weeks). On the other hand, for the rest of the methods, the total accuracy RMSE value worsened for MICE from 17.7 °C to 21.1 °C and for Amelia from 19.4 °C to 22.1 °C.

Table 4.11: RMSE of multiple imputation methods with solely highly correlated variables for eight weeks period

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE 8 weeks: High correlations	3.74	3.72	3.23	3.31	3.25	3.80
Amelia 8 weeks: High correlations	3.96	3.91	3.19	3.23	3.54	4.28
missForest 8 weeks: High correlations	1.25	1.58	1.70	1.85	1.27	1.54

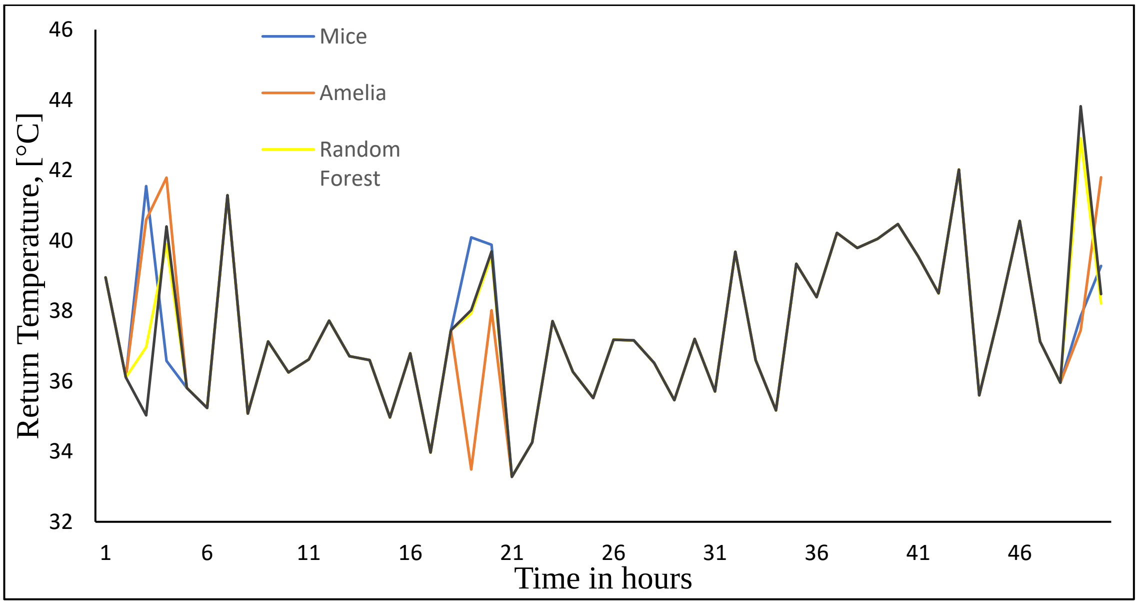


Figure 4.9: Illustration of multiple imputation techniques with highly correlated variables for the first 50 hours of the week of 15 gaps of 2 hours in DH Return Temperature of Building I for eight weeks period

Figure 4.9 illustrates the same gap-filling process for the first 50 hours on 15 gaps of 2 hours for DH Return Temperature in Building I and in accordance with this, Table 4.12 displays those exact points where the values were predicted for the first 50 hours as well.

Table 4.12: The predicted values of multiple imputation methods for the first three gaps on 15 gaps of 2 hours.

DH Return Temperature, Building I				
Hours	Original data [°C]	MICE [°C]	Amelia [°C]	MissForest [°C]
02:00	35.03	41.55	40.59	39.7
03:00	40.4	36.58	41.91	39.89
18:00	38.02	40.09	33.49	37.92
19:00	39.69	39.88	38.01	39.56
00:00	43.82	37.85	37.45	42.9
01:00	38.48	39.28	41.8	38.21

4.6.2 Scenario II with eight weeks. Multiple imputations of highly and medium correlated variables

Furthermore, similarly, for the second scenario with eight weeks, multivariate analysis for high and medium correlation variables was carried out. Table 4.13 contains the RMSE values of three methods and one can observe that with the addition of medium correlation observations, the accuracy of the predictions of the methods improved compared to scenario I for eight weeks.

Table 4.13: RMSE of multiple imputation methods with high and medium correlated variables for eight weeks period

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE 8 weeks: High and Medium Correlations	2.15	2.37	2.38	2.78	2.83	2.96
Amelia 8 weeks: High and Medium Correlations	3.72	3.84	3.59	3.08	3.62	3.48
missForest 8 weeks: High and Medium Correlations	0.87	1.03	1.23	1.35	1.01	1.21

MICE performed better than Amelia for all time-gap scenarios unlike all other scenarios analysed previously and the total accuracy for DH Return Temperature improved from 21.1 °C to 15.5 °C compared to the eight weeks data with only highly correlated analysis. The same pattern was recorded for Amelia and missForest too by 0.8 °C and 2.49 °C accuracy improvement as well, respectively.

In comparison with the scenario of high and medium correlation for two weeks, MICE and Amelia worsened the accuracy, despite having more data observations to predict and increase the RMSE values as opposed to missForest where better accuracy was detected than in the previous scenarios.

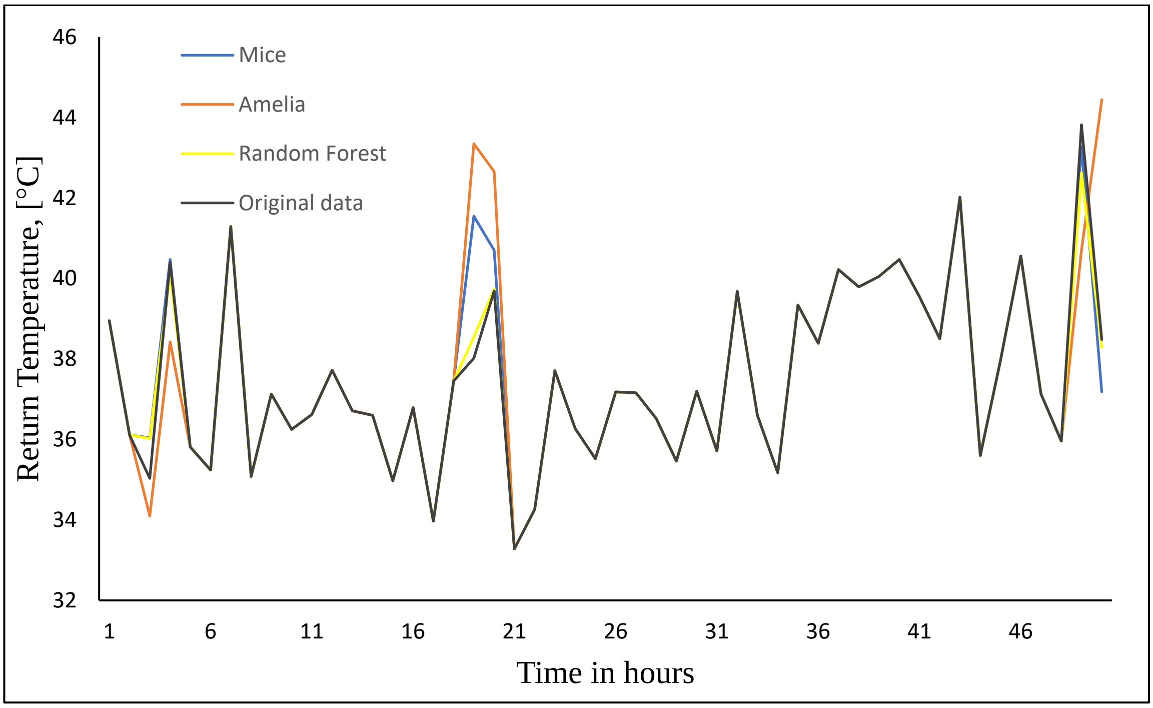


Figure 4.10: Illustration of multiple imputation techniques with high and medium correlated variables for the first 50 hours of the week of 15 gaps of 2 hours in DH Return Temperature of Building I for eight weeks period

Similarly, Figure 4.10 demonstrates the illustration of three methods predictions with high and medium correlations for the first 50 hours on 15 gaps of 2 hours for DH Return Temperature in Building I for eight weeks. It is readily detectable that the orange line representing Amelia has a greater distortion than those of MICE and missForest which is also proved by their accuracy result provided in Table 4.13.

4.6.3 Scenario III with eight weeks. Multiple imputations of all correlated variables

The last scenario of this group is on all correlated variables included for prediction analysis for eight weeks. For the previous scenarios, the result was predominantly improved by the addition of variables to the analysis. The result of this scenario is provided in Table 4.14 and compared to the previous case, where there was a clear leader and outsider, despite missForest showing the best accuracy result, this time the least accurate result for the time-gap scenario was interchanging between MICE and Amelia as it happened for almost all other scenarios.

Table 4.14: RMSE of multiple imputation methods with all correlated variables for eight weeks period

Imputation Method	ROOT MEAN SQUARE ERROR (RMSE)					
	RETURN_T					
	2 hours gap	3 hours gap	5 hours gap	6 hours gap	10 hours gap	15 hours gap
MICE 8 weeks: All correlated variables	2.05	2.27	2.85	4.94	2.92	2.75
Amelia 8 weeks: All correlated variables	2.74	2.69	2.56	2.30	2.59	2.85
missForest 8 weeks: All correlated variables	1.25	1.44	1.35	1.41	1.04	1.46

In general, as opposed to the total RMSE values of two weeks, eight weeks accuracy did not demonstrate the same variable sensitivity. In other words, for two weeks of analysis, each of the methods performed better by the addition of variables where they had the best accuracy rate when all the correlated variables were considered. In contrast, the same consistency was not observed for eight weeks RMSE values as can be seen in Table 4.15. Even though one can see that by expanding the window with the data set and hence having more data, the accuracy bettered for all the respective correlation scenarios in missForest. However, MICE and Amelia showed the opposite outputs each of them decreasing their RMSE values for all of the scenarios of eight weeks.

Hence, one can conclude that the prolongation of the week and thus more observations to learn might be beneficial for missForest, but there was not a substantial difference when the week was expanded. In the meantime, the two other techniques worsened their accuracies.

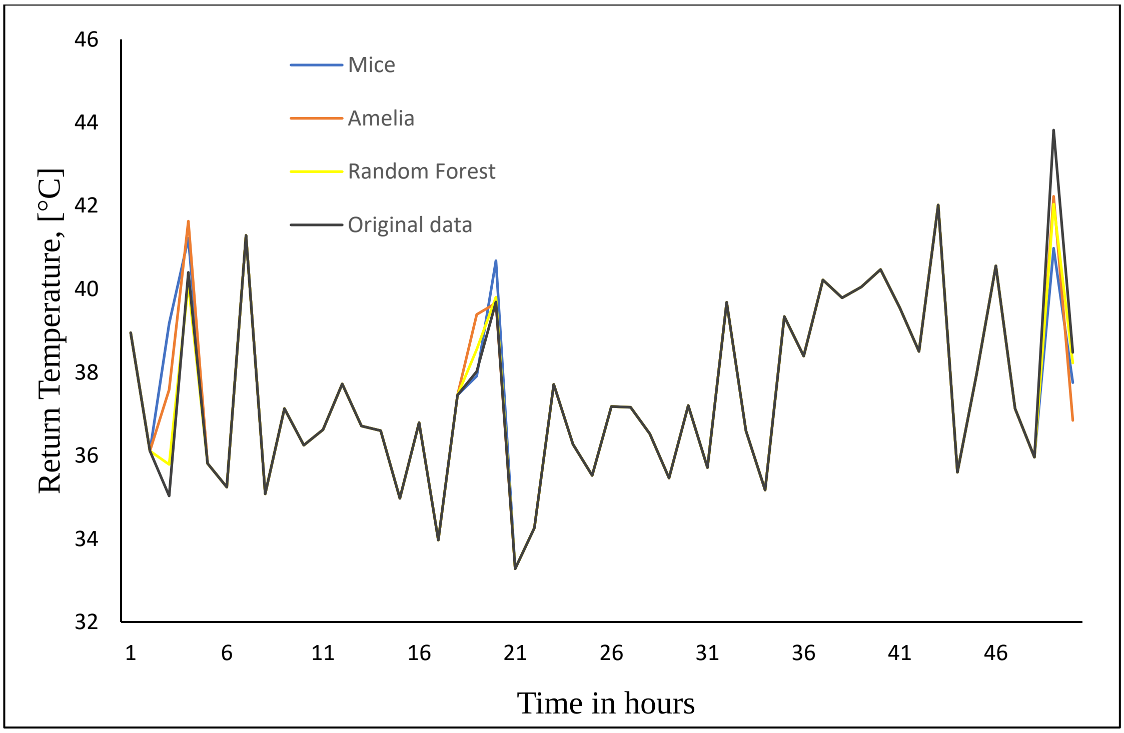


Figure 4.11: Illustration of multiple imputation techniques with all correlated variables for the first 50 hours of the week of 15 gaps of 2 hours in DH Return Temperature of Building I for eight weeks period

The illustration of the methods of the first 50 hours on 15 gaps of 2 hours with all correlated variables for DH Return Temperature in Building I is given in Figure 4.11. Likewise in scenario II for eight weeks, Amelia had the least accurate RMSE value and it can be readily observed in the plot denoted by orange, while the yellow (missForest) appears to be the closest to the original data.

Table 4.15: Total RMSE values of each method at each scenario for two and eight weeks

Imputation methods	Total RMSE value of two weeks [°C]	Total RMSE value of eight weeks [°C]
Scenario I		
MICE: High correlation	17.67	21.06
Amelia: High correlation	19.44	22.11

Table 4.16: Total RMSE values of each method at each scenario for two and eight weeks (continued)

missForest: High correlation	9.99	9.20
Scenario II		
MICE: High and Medium correlations	15.33	15.47
Amelia: High and Medium correlations	14.76	21.33
missForest: High and Medium correlations	8.57	6.71
Scenario III		
MICE All: All correlations	13.91	17.79
Amelia All: All correlations	12.79	15.74
missForest: All correlations	8.41	7.94

4.7 Global Score

After all the scenarios have been implemented, we are required to come up with a formula to assess which method performed and showed the best accuracy across all data sets. As was discussed earlier, overall, mainly two methods were proposed based on univariate and multivariate analysis. For each of them, 6 time-gap scenarios were created and for multiple imputation methods additional seven scenarios were carried out with alterations of correlation coefficient variables and the prolongation of the week. Hence, we shall observe by proposing a formula which would suggest the best method for four data sets with all their variables where those techniques were applied.

As a result, the following GS is suggested which would first obtain the sum of each method for time-gap scenarios in each variable and then, normalise them based on their corresponding specific variable analysis. In other words, each method applied for that variable is divided by the least accurate method in that variable. Consequently, sum of each normalised RMSE value for that specific method for each variable in each of the data set would give us the GS we aim to reach:

$$GS = \frac{\sum_{i=1}^{6} RMSE_{i(Return\_T)}}{max\ (\sum_{i=1}^{6} RMSE_i(N))} + \frac{\sum_{i=1}^{6} RMSE_{i(Flow\_T)}}{max\ (\sum_{i=1}^{6} RMSE_i(N))} + \frac{\sum_{i=1}^{6} RMSE_{i(Power)}}{max\ (\sum_{i=1}^{6} RMSE_i(N))} + \frac{\sum_{i=1}^{6} RMSE_{i(DHW\_Flow\_T)}}{max\ (\sum_{i=1}^{6} RMSE_i(N))} + \frac{\sum_{i=1}^{6} RMSE_{i(SH\_Flow\_T)}}{max\ (\sum_{i=1}^{6} RMSE_i(N))} + \frac{\sum_{i=1}^{6} RMSE_{i(SH\_Return\_T)}}{max\ (\sum_{i=1}^{6} RMSE_i(N))} \tag{4.4}$$

Where $i -$ the number of time-gap of scenarios, which is used for all the cases such as 15 gaps of 2 hours, 10 gaps of 3 hours, etc., $N -$ number of the imputation methods applied for the variable analysis.

It should be noted that GS does not have a unit as it is a sum of the accuracies of Temperatures and Heating Power. Hence, GS is computed for all data sets and after they are summed to obtain the final outcome. GS of four data sets for each of the methods and scenario performed are summarised in Table 4.17. One can see that the table is divided into four sections according to the methods and scenarios applied.

The first is the GS of univariate imputation methods starting from Linear Interpolation to Mean and Median which was based on the time series package and the functions within that package. The total sum of all RMSE values of four data sets and their total 23 variables upon which filling the gaps techniques were applied was provided in the GS section. The colour pattern for most of the univariate analyses is around orange and yellow implying they are one of the least accurate when compared to multiple imputation methods. As expected, univariate techniques analyse only the variable itself and thus filling the gaps based on the observation of those variables excluding the impact of other data in the set. According to the result, one can observe Mean and Median finished with the least accurate result not only for the section but for all types of methods with 19.6 and 20 GS values, respectively. This happened due to the replacement of the gaps with the same value for all the missing points and subsequently, decreasing the accuracy. LOCF and NOCB were not far from the previous methods, despite showing high accuracy for certain variables at times. On the other hand, the Moving Average and its variations showed similar outcomes predominantly ranging between the GS score of 9 and 10. Whereas, Linear and Stineman Interpolations displayed the best accuracy for univariate imputation methods with the score of respectively 8.54 °C and 8.67 °C.

Table 4.17: Total GS of all imputation methods with every scenario applied

Imputation Method	Global Score
Univariate Imputation Methods	
Linear Interpolation	8.54
Spline Interpolation	13.81
Stineman Interpolation	8.67
LOCF	12.11
NOCB	12.99
Moving average k =2	9.49
Moving average k =4	9.48
Moving average k =6	9.98
Moving average k =8	10.83
Exp.weighted avrg k=2	9.31
Exp.weighted avrg k=4	9.14
Exp.weighted avrg k=6	9.08
Exp.weighted avrg k=8	9.14
Lin.weighted avrg k=2	9.37
Lin.weighted avrg k=4	9.26
Lin.weighted avrg k=6	9.46
Lin.weighted avrg k=8	9.89
Mean Value	19.62
Median Value	19.96
Multivariate Imputation methods for two weeks data	
MICE 2 weeks: High correlations	7.18
Amelia 2 weeks: High correlations	6.66
missForest 2 weeks: High correlations	4.69
MICE 2 weeks: High and medium correlations	6.68
Amelia 2 weeks: High and medium correlations	6.48
missForest 2 weeks: High and medium correlations	4.61
MICE 2 weeks: All correlated variables	6.15
Amelia 2 weeks: All correlated variables	5.78
missForest 2 weeks: All correlated variables	4.29
Multivariate Imputation methods for two weeks data based only on weather data	
MICE Weather	13.93
Amelia Weather	14.84
missForest Weather	8.39
Multivariate Imputation methods for eight weeks data	
MICE High 8 weeks: High correlations	8.58
Amelia High 8 weeks: High correlations	8.94

Table 4.18: Total GS of all imputation methods with every scenario applied (continued)

missForest 8 weeks: High correlations	4.57
MICE 8 weeks: High and medium correlations	8.11
Amelia 8 weeks: High and medium correlations	8.95
missForest 8 weeks: High and medium correlations	4.37
MICE 8 weeks: All correlated variables	7.98
Amelia 8 weeks: All correlated variables	8.04
missForest 8 weeks: All correlated variables	4.03

The second part is dedicated to multiple imputation methods for two weeks period, the same as for single imputations. Furthermore, three scenarios depending on the correlation coefficients were tested. One can identify that the GS values were considerably improved compared to the previous section, especially missForest had less GS value than the score of Linear Interpolation which performed the best for its corresponding section. Another consistency was noticed for the addition of more variables where the result bettered each time. This means with all the correlated variables multiple imputation methods showed the best score than for the scenario with only strong or string and medium correlations. For instance, MICE, Amelia, and MissForest improved by 1.03 °C, 0.51 °C, and 0.4 °C at the last scenario in accordance with the first case. Hence, each method appeared to be variable sensitive for the proposed combinations.

On top of that, one of the tested experimental cases was the physical computation on DH Return Temperature. It was performed considering the physical formula and was useful only for DH Return Temperature, DH Flow Temperature, and Heating Power. The idea was to examine if the multivariate analysis would be able to identify that correlation yet one can see that the results were not as accurate as the physical computation. This is because the physical computation result was virtually the same as the measurements, while multiple imputation methods had considerably worse accuracy than that of formula-based calculation.

Whereas the third section demonstrated the analysis on the scenarios where only weather variables were used for the prediction of the gaps. Therefore, according to their GS score, they were not as successful as in the case of the contribution of smart energy meter variables. It is readily observed that their score dropped by two times compared to the scenario where the energy data correlation was included. In addition to that, despite examining the case with many variables, weather-based multiple imputations, in particular MICE and Amelia showed one of the worst GS values along with Mean, Median, and Spline Interpolation.

At last, the GS score of the same multiple imputation methods for eight weeks period is computed. It is more relatable to compare the result with the second section where there were the same conditions but fewer data of two weeks. It is shown that the only method which had an improvement was missForest, yet there was not a considerable increase. On the other hand, the GS value of MICE and Amelia witnessed decreases compared to the GS value of two weeks period.

However, taking into account the particular features of the smart energy meters which record all the variables at once, when the missing point appears in one of them then there will be missing values for all the variables of the smart meter for that time. Hence, the best-performing methods of univariate imputations can be compared solely to the analysis when only the weather data was contributing to the prediction of the missing points in multiple imputations. Consequently, as one can see Linear and Stineman Interpolations (the best GS values among single imputation methods) have GS values of 8.54 and 8.67. Whereas the best GS value for weather data correlation-based multiple imputations was missForest with a GS value of 8.39. As a result, there was not a substantial increase in the accuracy compared to the univariate imputation methods when the multiple imputations applied using only the correlation between the variable with missing points and the weather data.

5 Conclusion and future work

The thesis aimed to assess univariate and multivariate imputation methods on energy data along with the impact of the weather variables on an accuracy basis with the proposed Global Score (GS). The study suggested that, generally, multivariate analysis is more accurate in the prediction and can be more flexible as it composes the examination of the many variables and their effect on the missing points. Whereas the essence of the working mechanism of some sensors may cause some barriers to the way of applying the methods.

The effectiveness of the gap-filling techniques was assessed, first, by Root Mean Square Error (RMSE) value, initially, and then summarised by the proposed GS value which is the sum and normalised score of each variable in each data set. For univariate analysis, the methods using the observations around the gaps appear to have better accuracy than the ones which replace the points with the same value across the data set. On the contrary, LOCF and NOCB which consider only the nearest two points around the gap displayed one of the least accurate RMSE values. Whereas Linear and Stineman Interpolations had the best GS value among univariate imputation techniques and none of them demonstrated consistent increase or decrease when the gap size widened remaining with a total 30 hours gap for all time-gap scenarios.

For multivariate imputations, three different scenarios were examined with high correlations, high and medium correlations, and all correlated variables included. The output proposes that all of the methods had better accuracy than any of the univariate imputations, particularly with more than two times better GS scores compared to the best methods in univariate analysis.

For sensitivity analysis, MICE, Amelia, and missForest were examined as previously. However, this time the range of the extracted two weeks was expanded into eight weeks providing more data to be used for the prediction of the gaps. Furthermore, significant improvements were not identified, while for some of the methods, the accuracy even worsened.

However, when one gap appears in the smart energy meter data this creates the gap for all other variables measured by the sensor. Consequently, in real life, the implementation of the first three multiple imputation scenarios might be difficult or even impossible to achieve. For this reason, the next method on weather variables correlations were implemented as if the data from the smart meter was lost they could be the only variables to rely on and predict the missing observations. Because the weather data was measured in the weather station and thus it is not related to the issue the energy sensor may face. *The outcome suggests that, despite the Ambient Temperature having strong correlations with the energy data, standalone with other weather variables with negligible correlation coefficients, the scenario demonstrates virtually the same accuracy as the univariate techniques. Specifically, the best performed for single imputation case, Linear Interpolation, is comparable to the best multivariate weather variables-based method (which is missForest).*

Apart from that, the physical computation can be a solution, when there are known parameters and existing mathematical formulas, and hence, the gaps can be replaced by the computations. On one hand, it may provide a basis upon which other machine learning prediction methods can be assessed for accuracy as the physical calculation shall fill the gaps with more accurate values. As the results demonstrate, the multiple imputation methods are not ideally intelligent meaning that they are not capable of predicting the exact values of the missing points. On the other hand, for the case study, where all energy variables are gathered from the same sensor, the gaps appear for all variables in those missing points. Subsequently, this method faces the same barrier as the multiple imputation methods for correlation coefficient-based scenarios.

In conclusion, the imputation of the missing data can be promising. Despite multiple imputation methods with the correlated variables demonstrating the highest accuracy, for the real-time case when the data is lost, the only method that manages to fill those gaps could be with the help of other correlated variables. If the major goal is to fill the gaps with proper accuracy and the missing data is significant, those correlated variables must come from other sensors or data sources which are not physically connected to the set where missing points occur.

In a further development of the proposed methods, other existing packages can be examined with certain scenarios, and if possible more data about the building characteristics and occupants could be integrated into the data analysis. This might facilitate the new findings.

References

[1] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, Third edition. Hoboken, NJ: Wiley, 2020.

[2] C. A. Leke and T. Marwala, *Deep Learning and Missing Data in Engineering Systems*, vol. 48. Cham: Springer International Publishing, 2019. doi: 10.1007/978-3-030-01180-2.

[3] J. L. Schafer and J. W. Graham, “Missing data: Our view of the state of the art.,” *Psychol. Methods*, vol. 7, no. 2, pp. 147–177, 2002, doi: 10.1037/1082-989X.7.2.147.

[4] D. B. Rubin, “Inference and missing data,” *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976, doi: 10.1093/biomet/63.3.581.

[5] L. Ehrlinger, T. Grubinger, B. Varga, M. Pichler, T. Natschlager, and J. Zeindl, “Treating Missing Data in Industrial Data Analytics,” in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, Berlin, Germany, Sep. 2018, pp. 148–155. doi: 10.1109/ICDIM.2018.8846984.

[6] I. Izonin, N. Kryvinska, R. Tkachenko, and K. Zub, “An Approach towards Missing Data Recovery within IoT Smart System,” *Procedia Comput. Sci.*, vol. 155, pp. 11–18, 2019, doi: 10.1016/j.procs.2019.08.006.

[7] P. E. McKnight, Ed., *Missing data: a gentle introduction*. New York: Guilford Press, 2007.

[8] J. W. Graham, *Missing data: analysis and design*. New York, NY: Springer, 2012.

[9] D. A. Newman, “Missing Data: Five Practical Guidelines,” *Organ. Res. Methods*, vol. 17, no. 4, pp. 372–411, Oct. 2014, doi: 10.1177/1094428114548590.

[10] W.-C. Lin and C.-F. Tsai, “Missing value imputation: a review and analysis of the literature (2006–2017),” *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/s10462-019-09709-4.

[11] C. Penone *et al.*, “Imputation of missing data in life-history trait datasets: which approach performs the best?,” *Methods Ecol. Evol.*, vol. 5, no. 9, pp. 961–970, Sep. 2014, doi: 10.1111/2041-210X.12232.

[12] M. L. Yadav and B. Roychoudhury, “Handling missing values: A study of popular imputation packages in R,” *Knowl.-Based Syst.*, vol. 160, pp. 104–118, Nov. 2018, doi: 10.1016/j.knosys.2018.06.012.

[13] A. Jadhav, D. Pramod, and K. Ramanathan, “Comparison of Performance of Data Imputation Methods for Numeric Dataset,” *Appl. Artif. Intell.*, vol. 33, no. 10, pp. 913–933, Aug. 2019, doi: 10.1080/08839514.2019.1637138.

[14] R. R. Andridge and R. J. A. Little, “A Review of Hot Deck Imputation for Survey Non-response,” *Int. Stat. Rev.*, vol. 78, no. 1, pp. 40–64, Apr. 2010, doi: 10.1111/j.1751-5823.2010.00103.x.

[15] E. Pebesma, “**spacetime** : Spatio-Temporal Data in *R*,” *J. Stat. Softw.*, vol. 51, no. 7, 2012, doi: 10.18637/jss.v051.i07.

[16] S. Moritz and T. Bartz-Beielstein, “imputeTS: Time Series Missing Value Imputation in R,” *R J.*, vol. 9, no. 1, p. 207, 2017, doi: 10.32614/RJ-2017-009.

[17] A. Zeileis and G. Grothendieck, “**zoo** : *S3* Infrastructure for Regular and Irregular Time Series,” *J. Stat. Softw.*, vol. 14, no. 6, 2005, doi: 10.18637/jss.v014.i06.

[18] M. Mayer, “Package ‘missRanger.’” Mar. 27, 2021. Accessed: May 05, 2021. [Online]. Available: https://cran.r-project.org/web/packages/missRanger/missRanger.pdf

[19] Diethelm Wuertz, Tobias Setz, Yohan Chalabi, and Martin Maechler, “Package ‘timeSeries.’” CRAN, Jan. 24, 2020. Accessed: May 05, 2021. [Online]. Available: https://cran.r-project.org/web/packages/timeSeries/timeSeries.pdf

[20] S. G. Liao *et al.*, “Missing value imputation in high-dimensional phenomic data: imputable or not, and how?,” *BMC Bioinformatics*, vol. 15, no. 1, p. 346, Dec. 2014, doi: 10.1186/s12859-014-0346-6.

[21] M. L. Yadav and B. Roychoudhury, “Handling missing values: A study of popular imputation packages in R,” *Knowl.-Based Syst.*, vol. 160, pp. 104–118, Nov. 2018, doi: 10.1016/j.knosys.2018.06.012.

[22] H. Johra, D. Leiria, P. Heiselberg, A. Marszal-Pomianowska, and T. Tvedebrink, “Treatment and analysis of smart energy meter data from a cluster of buildings connected to district heating: A Danish case,” *E3S Web Conf.*, vol. 172, p. 12004, 2020, doi: 10.1051/e3sconf/202017212004.

[23] “R: What is R?” https://www.r-project.org/about.html (accessed Jun. 09, 2021).

[24] P. Lafaye de Micheaux, R. Drouilhet, and B. Liquet, *The R Software: Fundamentals of Programming and Statistical Analysis*, vol. 40. New York, NY: Springer New York, 2013. doi: 10.1007/978-1-4614-9020-3.

[25] S. Moritz and T. Bartz-Beielstein, “imputeTS: Time Series Missing Value Imputation in R,” *R J.*, vol. 9, no. 1, p. 207, 2017, doi: 10.32614/RJ-2017-009.

[26] N. J. Horton and S. R. Lipsitz, “Multiple Imputation in Practice: Comparison of Software Packages for Regression Models With Missing Variables,” *Am. Stat.*, vol. 55, no. 3, pp. 244–254, Aug. 2001, doi: 10.1198/000313001317098266.

[27] S. van Buuren, *Flexible imputation of missing data*. Boca Raton, FL: CRC Press, 2012.

[28] P. Royston, “Multiple Imputation of Missing Values,” *Stata J. Promot. Commun. Stat. Stata*, vol. 4, no. 3, pp. 227–241, Aug. 2004, doi: 10.1177/1536867X0400400301.

[29] H. Mekala, “Dealing with Missing Data using R,” *Medium*, Sep. 24, 2019. https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17 (accessed May 04, 2021).

[30] L. Li, C. G. Prato, and Y. Wang, “Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forest classifier,” *Accid. Anal. Prev.*, vol. 146, p. 105744, Oct. 2020, doi: 10.1016/j.aap.2020.105744.

[31] D. J. Stekhoven and P. Buhlmann, “MissForest--non-parametric missing value imputation for mixed-type data,” *Bioinformatics*, vol. 28, no. 1, pp. 112–118, Jan. 2012, doi: 10.1093/bioinformatics/btr597.

[32] “harp,” *Harp Random Forests*. https://dsc-spidal.github.io/harp/ (accessed Jun. 16, 2021).

[33] H. Mekala, “Dealing with Missing Data using R,” *Medium*, Sep. 24, 2019. https://medium.com/coinmonks/dealing-with-missing-data-using-r-3ae428da2d17 (accessed May 05, 2021).

[34] “Tartu Monthly Climate Averages,” *WorldWeatherOnline.com*. https://www.worldweatheronline.com/tartu-weather/tartumaa/ee.aspx (accessed Jun. 01, 2021).

[35] D. Mindrila, P. Balentyne, and M. Ed, “Scatterplots and Correlation,” p. 14.

Appendix

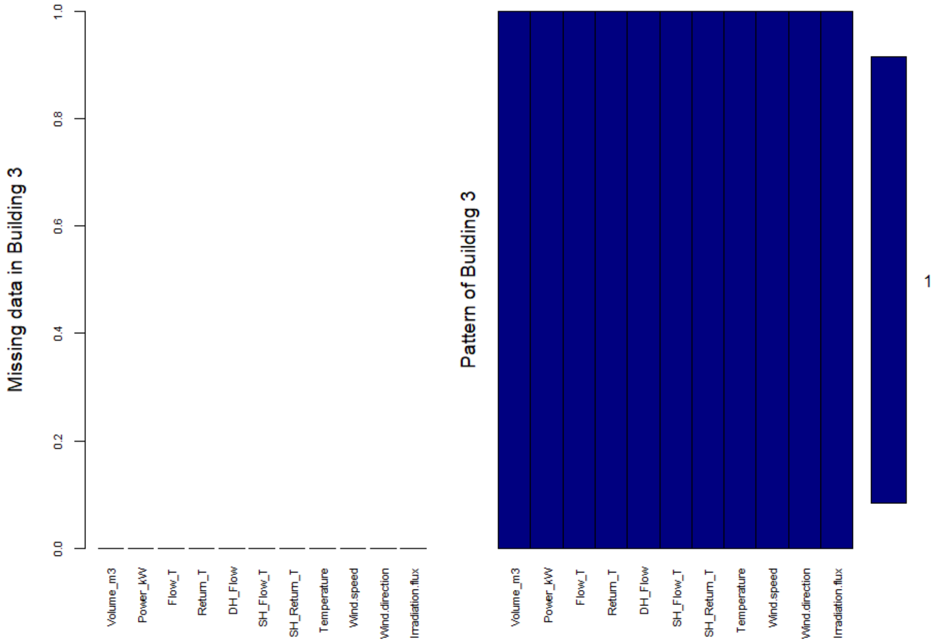


Figure A: Missing points distribution of Building III

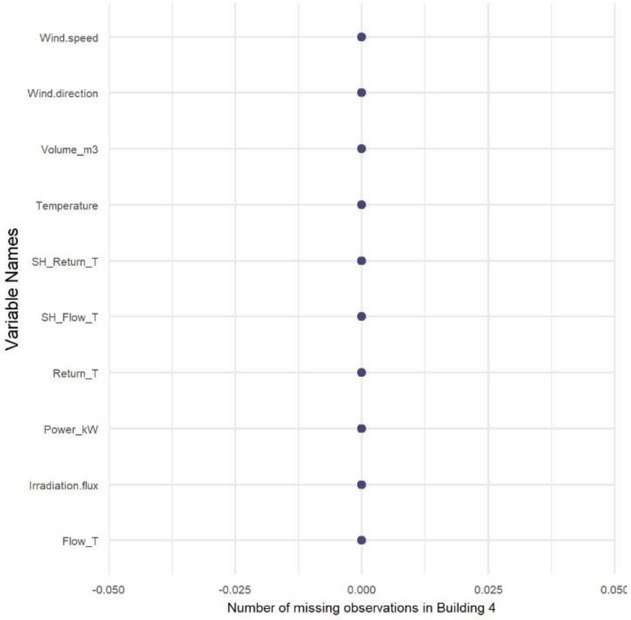


Figure B: Missing points distribution of Building IV

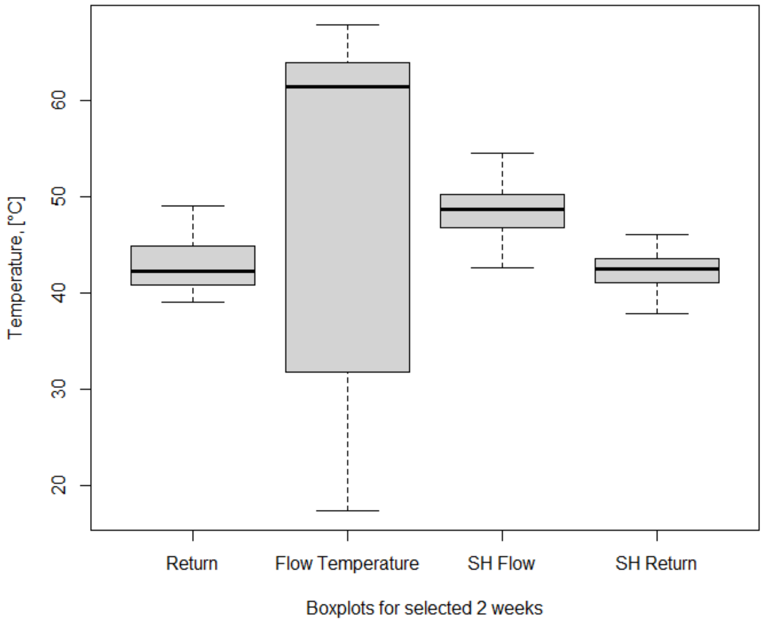


Figure C: Outlier analysis based on boxplot summary of the smart energy meter data in Building II. (2 weeks: March in DH Return Temperature, May in DH Flow Temperature, March in SH Flow Temperature, April in SH Return Temperature)

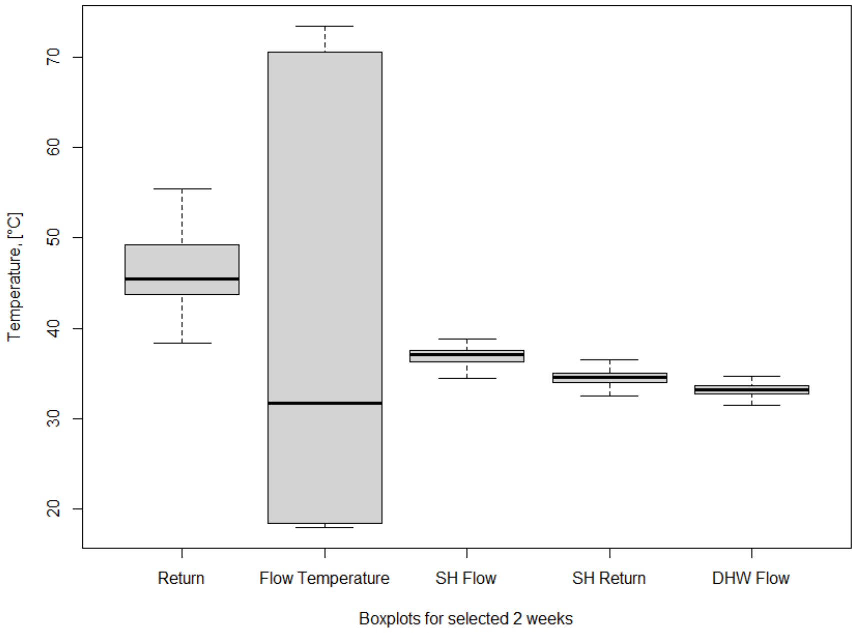


Figure D: Outlier analysis based on boxplot summary of the smart energy meter data in Building III. (2 weeks: January in DH Return Temperature, May in DH Flow Temperature, December in SH Flow Temperature and SH Return Temperatures, and DHW Flow Temperature)

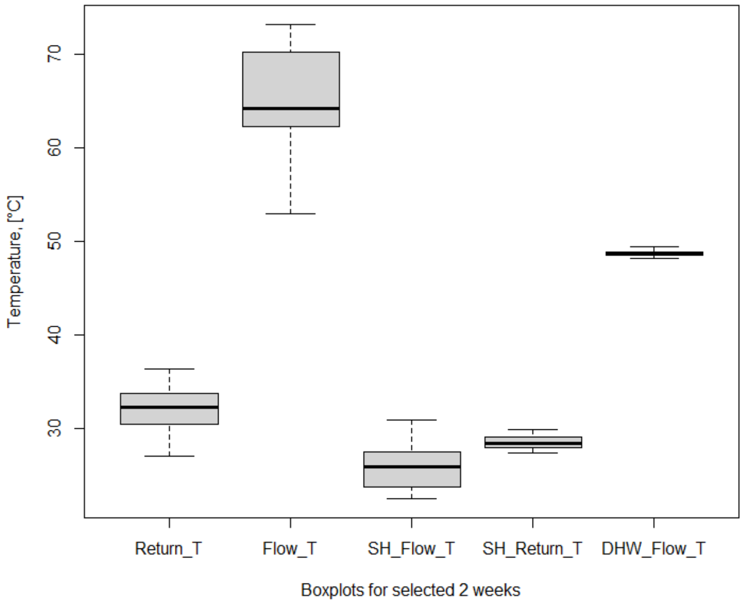


Figure E: Outlier analysis based on boxplot summary of the smart energy meter data in Building IV. (2 weeks: January in DH Return and DH Flow Temperatures, April in SH Flow and SH Return Temperatures, July in DHW Flow Temperature)