Research proposal Applied Data Science project

Imputation methods for green building management data

Names & student codes:

Ramon van der Elst	- 16077466
Albert Corson	- 21084858
Adrien Lucbert	- 21132356
Jesús Martinez De Juan	- 20150261
Michael Weij	- 18095593
Juliën van der Niet	- 18069681

Group: Imp

Date: 9/9/2021

Abstract

Building Management Systems (BMS) energy time series data has gaps which are problematic for further use of the data e.g.: models and visualization. To fill in the gaps imputation techniques can be used, but what techniques are relevant for this specific use case and suited for the quirks of building management system data?

In this study four imputation methods will be studied, classified, tested and described to create guidelines for imputing missing data based on existing methods. Research has been done to consider optimal ways of handling missing data to minimize the bias potential. The consideration of how to optimize handling missing data will be critical to recommend several approaches to prevent bias for Factory Zero.

The final research paper will not create new imputation techniques but it will specify them for the desired use case. First existing techniques will be explored in the first sub question. The second and third sub questions will then specify the techniques found in the first question for gap size and data classification. The findings from the sub questions will be validated using the data from Factory Zero and KNMI time series data.

The paper will contain a guideline describing how and when to use and which imputation methods for missing data in time series data.

Introduction

Building management systems create data on which models are trained to further optimize our future homes. The data these solutions produce has gaps of various sizes, from 5 minutes to entire days that are missing. These gaps make visualizing and training models based on the buildings less accurate than they could be.

This is where imputation comes in, there are various methods for imputing data e.g.: taking the average, nulling the missing data or interpolation. The research done for this paper will look at imputation techniques and which ones should be used in which scenarios. The result of this will be a guideline of what imputation techniques to use on data from green building management solutions.

Factory Zero a Dutch company creating green building management solutions, has kindly lent us their data on which we will base and verify our findings. Factory Zero's data contains 120 houses, all of these houses contain gaps in the data. The building data contains polling data on: thermostat readings, heat pump usage, co2 sensor polling's and other devices used in their solutions. Every poll of the sensor has a corresponding timestamp which makes the gaps in time difference noticeable.

The research group Energy and Transition has done multiple research projects on BMS time series data before this project started. During the previous research projects researchers encountered problems with missing data. Factory Zero a dutch company creating green building management system solutions, has partnered with the research group and has provided their data for the research group to study.

Imputation of data is a well-studied field in data science and has many existing papers already published on the topic. The goal of this research will not be to add on too but to specify the existing knowledge for the purpose of imputing green building management time series data. The guidelines as mentioned before will contain existing imputation techniques optimized specifically for our and similar use cases . Preliminary research findings suggest that there will not be a one size fits all imputation technique for the gaps that are present in the test data.

Using these guidelines will in turn further optimize the visualizations, models and in turn the buildings itself as there is more accurate data present.

Background & context

Before this research project took place there were other research teams that used data provided by Factory Zero. All of the previous research teams ran into the same problem; that missing data impaired their ability to get accurate results from research.

Which is why they created another research assignment to try to impute the data from their data sets. This research team however is no expert in the domain of data science and imputation of data, which is why orientation with the subject matter is required.

To fulfill the wish of Factory Zero the result of the paper will take the form of a comprehensive guideline. This guideline contains a list of scenarios found in the given data and what imputation techniques to use to solve the problem.

Whilst this research makes use of Factory Zero data, the research team will try to make the guideline as applicable to similar building management data sets as possible.

The research group (NAME HERE!) is researching the energy usage and efficiency in Building Management Systems (BMS). In previous research projects gaps of data were encountered in the data. These gaps will impair research into BMS data and getting accurate results from models trained on the data sets will be less accurate.

Which is why the research group created a research project dedicated to creating a comprehensive guideline to impute missing data in BMS time series data. The guideline will contain a list of techniques that should be applied in scenarios found in the BMS datasets.

Whilst this research project will make use of the Factory Zero BMS data sets as a reference the guideline will be applicable to all BMS time series data.

Research questions

To focus the research into the imputation of building management data the following research question was formed: "*Which imputation techniques should be applied for data imputation in building energy time series data?".* The answer of this question will take the form of a guideline in which methods and techniques to use in what scenarios are recommended.

To support and divide the work on the main research question three sub questions were formed. Each sub question answers to what techniques are available in the domain of data science, and then answers as to what techniques are suited in the sub domain of imputation building management time series data.

Which leads to the first sub question: "*What imputation methods are known for imputing time series data?"*. This sub question allows the research team to orientate and familiarize themselves with data science and the imputation of time series data. The answer of this question will contain a comprehensive list of imputation methods with an explanation to why they were selected.

After exploring the domain of data science the research will specify more into the sub domain in imputing building management time series data. This will be done by research the following two catogories:

- Gap sizes.
- Types/classifications of data

The question to answer the gap size problem is as follows: *"Which imputation techniques are best suited for what gap sizes?".* This question will take the answer from sub question one and narrow down the list to what technique to use for what gap size. The preliminary research into the factory zero data resulted in a list of five realistic gap sizes from five minutes to a full day missing. This question needs to be answered because from literature reviews it was found that there most likely will not be a technique that can impute all data accurately.

The third and final question is as follows: "*What imputation techniques are best suited for which types of data?"*. The previous question stated that there are many sizes of gaps in the data from Factory Zero, the same however could be said about the different kinds of types/classifications of data. Whilst the data set contains only numerical data in terms of data types in programming it does hold many classifications of data e.g.: Ratio, Interval or Nominal. To answer this question results will be taken from both question one and two.

Literature review

Missing data imputation of high-resolution temporal climate time series data:

As far as the review is concerned, E Afrifa-Yamoah (2020) states the importance of missing data and imputation in climate and time series data. In his study, multiple approaches to the imputation of missing values were evaluated, including a structural time series model with Kalman smoothing, an autoregressive integrated moving average (ARIMA) model with Kalman smoothing and multiple linear regression. In the study data from a 12 month weather time series of the coast of Western Australia were used. The assumption that data was missing at random was made and led to the creation of artificial gaps to be studied using a five-fold cross validation methodology.

The conclusions of this study reveal that the methods studied have demonstrated their suitability in imputing missing data in high-resolution temperature, humidity and wind speed data. However, the study only used sub-samples of relatively short time series and this could have contributed to the general performance of the univariate time series modelling approaches. It is recommended that longer climate time series datasets with varying patterns of complexity are studied to assess the techniques further under several varying scenarios of missing data.

Overall this study states useful methods and techniques for our study and in ways comes close to what we want to achieve. The problem that this study doesn’t talk about is which imputation methods would be best fitting for different types of gaps in time series data. There are methods compared but not ranked or classified.

Missing Data: Five Practical Guidelines:

Daniel A.Newman (2014) study talks about missing data in three missing data mechanisms being: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). The importance and focus of the study is about the missing data treatments that lead to bias and less accurate hypothesis. Companies still routinely choose the more biased and error-prone techniques (listwise and pairwise deletion), likely due to poor familiarity with and misconceptions about the less biased/less error-prone techniques (maximum likelihood and multiple imputation).
The five guidelines provided by this study talk about: Using all available data, do not use single imputation, construct-level missingness, item-level missingness and person-level missingness. The five practical guidelines offered in the study are built upon statistical theory (see reviews by Allison, 2002; Enders, 2001b, 2010; Dempster, Laird, & Rubin, 1977; Little & Rubin, 2002; Newman, 2003; Rubin, 1976, 1987; Schafer, 1997; Schafer & Graham, 2002). This offers a set of compromised standards that are midway between current research practice and statistical best practice.

What this study does well is describe the guidelines for handling missing data and therefore gives a good understanding on the general way to take on missing data in our project. What this study can be improved on by our study is focussing the guidelines on a more specific topic like time series data where the guidelines are used to show the best imputation methods for different types of missing data.

A deep learning approach for missing data imputation of rating scales assessing attention deficit hyperactivity disorder:

This research article talks about the importance of imputing missing data in ADHD behavioral studies. A deep learning method has been used to impute missing data in ADHD rating scales and evaluate the ability of the imputed dataset. Even though most of this research is interesting it doesn’t always connect to the research we want to conduct ourselves. What does connect to our research is the deep neural network for missing data imputation.

The interior architecture that was used here is deep neural networks (DNN), which are stacked modules that have multiple hidden layers. It is also known as a multi-layer perceptron (MLP). This deep learning approach can impute missing data with both the case and control groups together in a dataset. The findings in the research provide evidence that deep learning approaches can impute missing data with high accuracy in an aggregate dataset from multiple samples.

What this article provides for this research is the ways we could use a deep learning approach like theirs. This will be a good approach to impute missing data in a big dataset to get high accuracy. What this article lacks is an explanation why they went for this approach and not another approach. There is no clear indication how they got to the conclusion that this neural network approach really is the best way to work on missing data.

Methodology and design

Method

The primary research method for this study is literature review and testing four different imputation techniques for specified gap sizes. This is specifically done for building time series data. Imputation technique identification is the very first step toward this study. We will first review various types of data imputation methods. The second stage we research what 'average' gap sizes are in the given Factory Zero dataset and pick 20 of the houses with the least amount of gaps. Then the gaps will be classified to create four classes. In this stage we also create our own gap program that will be used to randomly generate gaps in existing data.. Based on this understanding a classification method will be developed to categorize the gaps and imputation methods. In the third stage of the study imputation methods will be tested on the categorized gap sizes to identify the most fitting methods for each category of gaps. Finally, once the imputation methods per category have been identified and tested , a conceptual framework/guideline will be written. This study will be conducted between September 2021 and February 2022.

Study design

This research is a mix of experimental and descriptive design. This can be explained by the fact that there will be work based on experimenting with imputation methods and different gap sizes as well as describing outcomes of the best imputation methods to use for different types of gaps in time series data. This study will be qualitative research since the focus is to gain a rich, detailed and specific understanding of which imputation method is best used for different categories of gap sizes in time series data.

Reference

Afrifa-Yamoah, E. (2020, 1 januari). *Missing data imputation of highâresolution temporal climate time series data*. Geraadpleegd op 27 september 2021, van https://rmets.onlinelibrary.wiley.com/doi/full/10.1002/met.1873

Jakobsen, J. C. (2017, 6 december). *When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts*. Geraadpleegd op 27 september 2021, van https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1

Newman, D. A. (2014). Missing Data. *Organizational Research Methods*, *17*(4), 372–411. https://study.sagepub.com/sites/default/files/newman.pdf

Cheng, C. (2020). *A Deep Learning Approach for Missing Data Imputation of Rating Scales Assessing Attention-Deficit Hyperactivity Disorder*. Frontiers. https://www.frontiersin.org/articles/10.3389/fpsyt.2020.00673/full

Research Schedule

During this project the team will be working with a SCRUM based incremental development method. To visualise and clearly communicate what each member is working on, a Jira board was created. Jira has important integrations for software development and is just as well suited for research furthermore all members were familiar with this tool.

SCRUM sprints will be used during the research and will have the length of 1-3 weeks. During each sprint week a meeting will be organized with the contact from Factory Zero. The purpose of this meeting is to keep the project aligned with Factory Zero's wishes and get immediate feedback on the progress made that week.