Intro

Welcome everyone to the final presentation of project imputation. Together with Adrien, Albert, Jesus, Julien, Michael and myself Ramon, we are the project team IMP.

Today we will cover the following topics: An introduction to our project, the methodology, our conclusions and the future work we proposed based on our results and conclusions.

Terms:

Imputation:

is the process of replacing missing data with substituted values. ...

Time series data:

A time series is a series of data points indexed in time order.

BMS:

A System that controls and monitors building appliances such as a thermostat or heatpump

Data measurement scale:

Measurement refers to the assignment of numbers e.g. Interval, Ratio, nominal or ordinal

Scenario:

Is a Combination of data measurement scale and gap sizes

Trends:

Behavior of patterns In data.

What is the project about:

Imputation:

This project was set up by the research group energy in transition.

Why is imputation of missing values in data important?

Missing data is a common occurrence in time series data, causes include faulty sensors or errors in data storage. Having a complete data set is essential for the business decision making and forecasting of Building Management Systems (BMS). Missing data introduces an element of ambiguity while analyzing data and that can affect properties of statistical estimators and results into misleading conclusions. Quality of data is main concern of data scientists and researchers working in the field of data science. One of the main concerns of the data scientists is how to handle missing data. Most statistical and machine learning algorithms are not robust enough to handle missing values and get affected by missing data. All of these reasons make data imputation important and are some of the reasons why this research was conducted.

Test methods:

Imputation methods tested in our project are selected from previous research that has been done into the imputation of time series data. The methods that were selected are : Last Observation Carried Forward (LOCF), K-Nearest Neighbour (KNN), Recurrent Neural Network (RNN) and Hot Deck (HD). More explanation about these methods will be given by Jesus in the coming sheets.

Write guideline:

The goal of this research was to write guidelines that would describe which imputation methods should be used for imputing missing data values in building management time-series data.

Our project creates a guideline for imputing the gaps in BMS datasets by comparing four methods. The guideline contains the best method per gap size and scales of measurement. Which Julien will talk about more later in the presentation.

Datasets used:

We used two datasets with mainly ratio and some interval and nominal data. The first dataset is a building management system time-series dataset from factoryzero. This data contains data collected from 120 greenhouses. The second dataset was meteorological time series data from the Royal Netherlands meteorological institute or KNMI. Some differences are that the BMS data had 5 minute intervals and the KNMI had hourly intervals. For our project the main dataset used was the BMS dataset.

Here we see the columns that we used from the datasets for our research. As you can see here the majority of columns consists of ratio data with 2 interval columns and 1 nominal.

Now we will head over to Jesus who will tell us more about testing of methods.