Data Quality 2

Various projects today is focused on gathering data and analysing it. The gathered data is used for obtaining behaviours, habits and properties of the observed objects. This is done by using powerful statistical leaning algorithms, that are able to deduce these properties from the data. This approach is called data driven development, since the success is mainly determined by the data and not the algorithm.

When data is the central role of the system, the quality of the data are very important. Poor data can lead to wrong assumptions, and have a negative effect on the application. Choosing the correct dataset is therefore a key factor [1]. Looking at the quality of the data can help you chose what dataset to use. Data quality can be described many ways, one of the more formal is from the ISO 8402 standard that describes quality as:

> *"The totality of characteristics of an entity that bear upon its ability of satisfy stated and implied needs"* [2].

This indicate that data quality is something that is very depended on the intended application, and is therefore hard to generalize.

Quality of data is a subject that is gaining more and more attention due to the fact that the quantity of data available is larger than ever before. This forces researchers to choose between datasets, and a notion of quality in the dataset can help them choose. The heavy growth in available data is a result of projects that are moving data gathering from controlled labs, to the public. Many of these projects is citizen science projects, where it is the citizens who collect the data, and not the researchers [3]. This enables researchers to gather enormous amount of data, but they are no longer in control of the conditions the data is collected in, which introduces errors and other quality decreasing factors.

Non intrusive load monitoring is a topic that have been in focus in the past years. This is due to the rise of the smart meters, that makes it possible to measure at faster intervals, and collect the data on online services form real environments. But the smart meters architecture is designed after billing and regulation purposes, and not load monitoring. The network architecture is therefore often based on the unreliable UDP protocol, since it is more important to get the current information fast, than get all information. This courses a lot of packets to be lost in transition, which can degrade the completeness quality of the signal. The missing data can be a problem for load monitoring, since many methods of load disaggregation is based on learning techniques. The quality of the signal can also help identify if the collected data is suitable as a training set.

2.1 Quality Criteria

It is not uncommon that different areas of research has its own quality criteria. This is due to the fact that quality is a very domain specific subject. One of the areas that have been dealing with citizen data for many years is the Geographic information area, that are used for maps, weather prediction and climate research. They have come up with several ways of describing quality in spatial data [4]. Method for defining quality in time series data have also been developed [5]. To better define quality criteria in the smart meter data inspiration from related work is used.

2.1.1 Related Work

Data Quality is an area that recently have become a hot topic, due to the wast quantity of data. Many researchers strive to make tools that better can analyse data quality in different areas. In the area of spatial data is a *"Quality and Workflow tool"* being developed by the University of Wageningen [6]. The objective is to help researchers select the best suited data for a given data driven project. It does this by looking at different quality criteria, given by the user or found in standards for spatial data.

In bioinformatics is a tool named QCScreen developed to help create better dataset to metabolomics studies. In metabolomics studies is dataset often created by joining information from several different experiments of various quality. By using tools that can check the data quality and consistency to determine if a dataset is suitable for further processing, they are able to greatly improve the test results [7].

In the article *Taking a big Data approach to data quality in a citizen science project*[3] they talk about how quality assessment can be used to rate the believe on your data, and how to improve data collected in citizen science. The project focuses on bird observations, done by users on their smartphones. They improve the quality by disallowing the user to send incomplete datasets to the database, and in this way forcing the user to only deliver high quality information. They then cross check the information with information from people in the same area, to see if it varies greatly.

One of the things all the methods have in common is trying to look at the completeness of the data. Some of the most low level criteria is the sample availability. The sample availability describes how many samples there is collected in relation to the expected collection amount, and look at how the samples are distributed in the measurement period. It is also seen that the activity is a good quality metric since a good data set must contain both areas with activity and areas without.

2.2 Quality In SmartHG Citizen Data

As a part of the SmartHG project 25 households have been equipped with meters on selected appliances and the main meter. The data collected from this experiment are prone with errors due to malfunctioning test equipment or unexpected interference from the resident which have resulted in offline measurement equipment for periods of time. Examples on unexpected interference could be if the resident is unplugging the measurable equipment, or turning off the power socket that supply's it. Unstable network does further degrade the signal, since the measurement equipment uses a lossy network.

The SmartHG data is intended for appliance recognition, and the quality must be assessed with this in mind. The completeness and the activity in the data is therefore important.

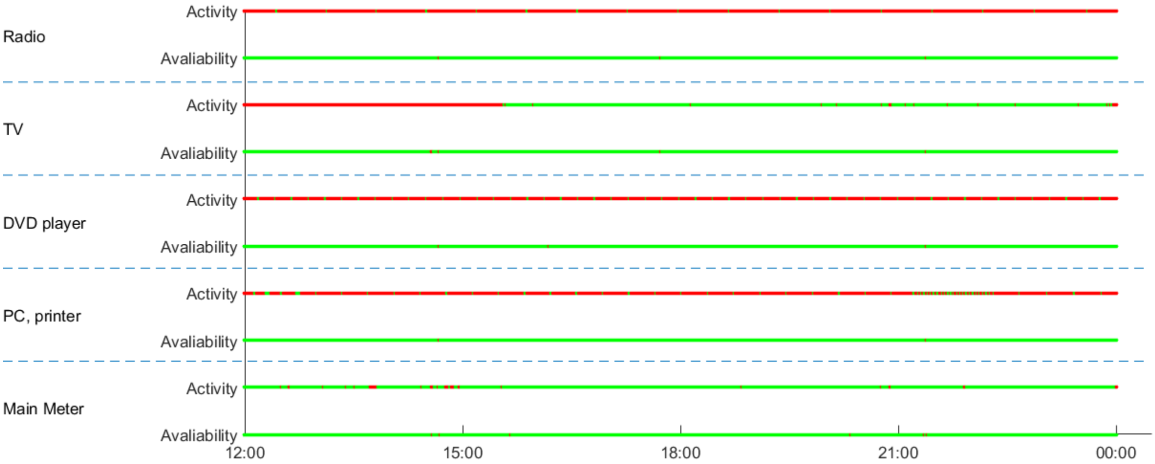

Figure 2.1: 12 hour overview of house 10

On figure 2.1 is a 12 hour overview of the data in house 10 from the 16/8/2015. Here we see the sample availability and Activity of the five different meters in the house. The availability is shown as a line where green is indicating that a sample is received as expected, and red shows a missing sample. On the figure it is shown that there is a few samples missing, which is to be expected due to the lossy network architecture. The Activity is also shown as a line, where green indicates activity and red indicates no activity. Activity is defined as a change in the signal, from prior values. From this we can see that the resident have a lot on activity on the TV from around 16:00 to 00:00, which we can presume means that the television is turned on in this period.

First the availability quality of the data is assessed. The availability quality for a specific period of time T_n for a specific meter m, is defined as the amount of samples observed in that timeslot over the expected sample amount. The resolution period T_P for each time period in T is chosen to be one hour.

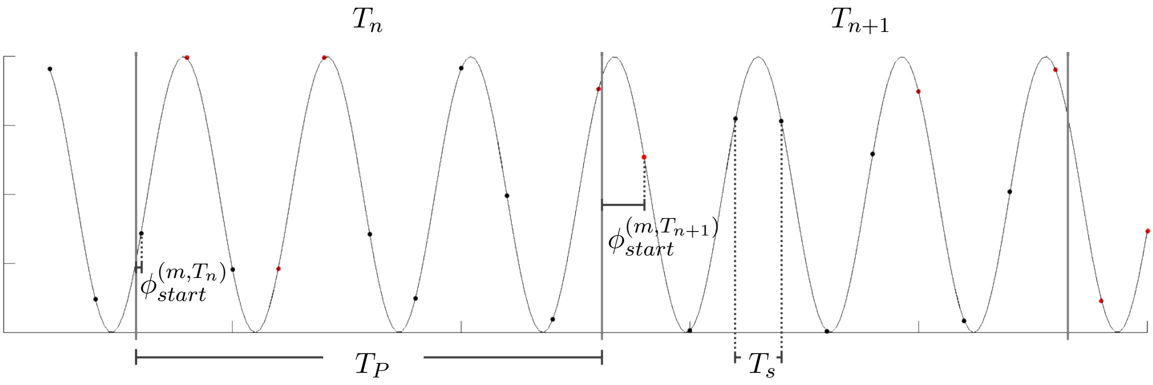

Figure 2.2: Illustration of availability analysis

To calculate the amount of samples expected to be in a specific period of time $N_{max}^{(m,T_n)}$ does the sample phase $\phi_{start}^{(m,T_n)}$ for the given period need to be known. On figure 2.2 a illustration of a signal, the black dots are the samples and the red dots are the ones that are missing. The sample phase is the time from the beginning of the period T_n and to the first expected sample. This is needed since for a timeslot T_n of a length of T_P the maximum expected sample amount can vary with 1. This is shown in figure 2.2 where the period T_n has a potential of having 11 samples, where as period T_{n+1} only can have 10.

$$N_{max}^{(m,T_n)} = \lfloor \frac{(T_P - \phi_{start}^{(m,T_n)})}{T_s^{(m)}} \rfloor + 1 \tag{2.1}$$

$$q^{(m,T_n)} = \frac{N_{observed}^{(m,T_n)}}{N_{max}^{(m,T_n)}} \tag{2.2}$$

As shown on equation 2.1 is the maximum number of samples for a meter m in the period T_n calculated by taking the period time T_P, corrected with the sample phase $\phi_{start}^{(m,T_n)}$ for the given period, and dividing it with the sample time T_s. The quality of the meter is calculated as the ratio of observed samples in the timeslot T_n to the maximum samples, shown in equation 2.2.

To find the quality of a house in a given period T_n, that have a set of meters $\mathbf{M}$ with a cardinality of M, we take the mean value of all the meter quality's, as shown in equitation 2.3.

$$\mu_{q^{(\mathbf{M},T_n)}} = \frac{1}{M} \sum_{m \in \mathbf{M}} q^{(m,T_n)} \tag{2.3}$$

A quality vector Q is constructed for each house. The quality vector contains the house quality found with a period T_P on one hour. This have been done from March T_1 to October T_N.

$$Q^{(\mathbf{M})} = \{\mu_{q^{(M,T)}} | T \in \{T_1, T_2, ..., T_n, ..., T_N\}\} \tag{2.4}$$

This is shown in equation 2.4 where $\mathbf{M}$ is a set of meters in a given house. This can be graphically shown in figure 2.3 where all the houses Q vectors is shown. The color is a gradient running from light green for the best quality to red for bad quality.

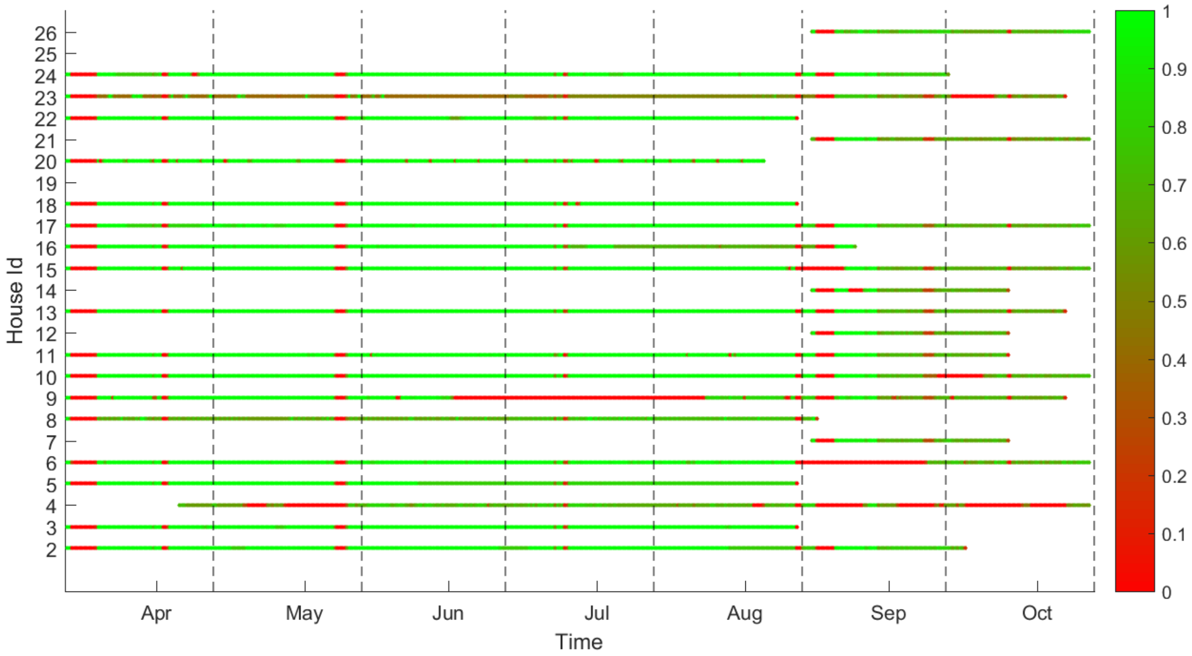

Figure 2.3: Quality of houses in SmartHG project

On figure 2.3 is the availability quality shown for the 25 houses in the SmartHG project. When seen on this scale with a analysis resolution on one hour it is hard to see degradation coursed by single sample missing here and there, but meters that have malfunctioned over a longer times shows itself. Since the quality of a house is the mean of several meter quality's, will this most likely appear as darker green spots, since not all meters are malfunctioning at the same time. But there still are some red areas indicating that all the meters in a house is not working.

There are the red spots that goes through all houses on the exact same time. This indicates that the server receiving the data for all the houses have been down, since it is unlikely that all meters in every house is down at the same time. The conclusion being that red dots most commonly are coursed by the network being unavailable so the client can not sent to the server, or the server is unable to receive.

It is assumed that the first sample received from a meter happen at the time of meter installation, and the last sample received is the time of meter removal. The meter is assumed to be operating in between these two points in time. On the figure does the coloured Q vector starts at installation time, and ends at removal time. This illustrates how some houses have been operational longer than others.

Since the data is intended for appliance recognition it is of interest where in the data there is activity, and where there is nothing happening. Both areas are impotent for the NILM application in training scenarios. We define activity as area in the data where there is change as described in equation 2.5.

$$f(x) + \epsilon < f(x+1) \vee f(x) - \epsilon > f(x+1) \tag{2.5}$$

Where ϵ describes a threshold to filter out changes caused by noise. This can also be described as the standard divination over a area is grater than the threshold. The activity is analysed in the available data, and is shown on figure 2.4 where green is high activity and red is non activity.

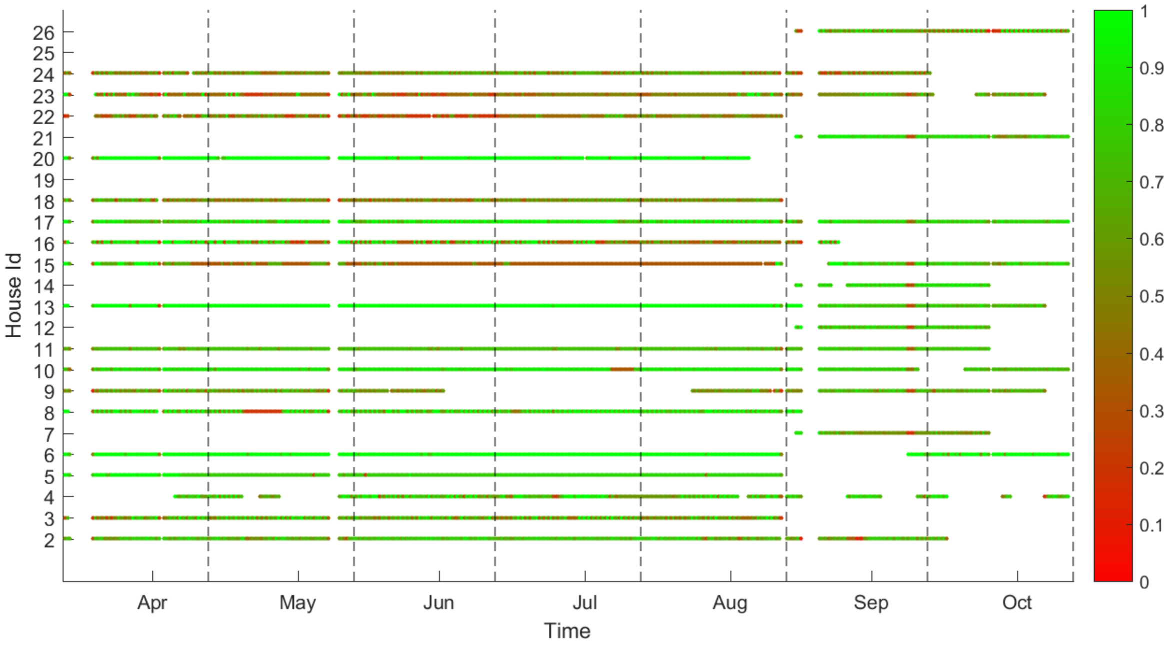

Figure 2.4: Activity of houses in SmartHG project

The activity shown on figure 2.4 is the average of the activity of the meters in the house. There is almost never a meter that does not have at least a little activity in a house, so a complete read area is fairly rare. The most interesting areas in the activity map, is often the places where there is a lot of change in the amount of activity like for house 16.

Gap Reconstruction 3

One of the more common problems in citizen science projects is gaps in data. This can happen either if the network connection is unstable or the test equipment gets prematurely turn off, as discussed in chapter 2. This can greatly degrade the data quality, and lead to errors in the application. One way to deal with this problem is to use mathematical gap filling techniques to come with a qualified guess on how the data would look like in the gap.

In order to use this methods we must assume that the missing data in the gap follows the same behaviour as the data on each side off the gap. If the signal is so stochastic that this is not the case then gap filling is not recommended[8].

In the case of the SmartHG project the data can be seen to have a part that is depended on the previous and future data plus a stochastic part that are determined by the user and the appliance. Due to the stochastic part a perfect reconstruction is not possible, but it is the hypotheses that the non stochastic part is still so dominant that a decent reconstruction is possible.

3.1 Gap Filling Methods

Various methods exists for gap filling. Five popular algorithms are selected for this project, and is validated on the SmartHG project data. These methods have been chosen since they all have a different approach on the gap filling process. For some of the algorithms it is important to keep the frequency spectra as intact as possible, for other it is the jitter power or the exact sample value they are trying to estimate. Some algorithms have also been designed for small gaps, while others have been designed for larger.

In the following sections it will be briefly described what makes the five algorithms unique, and how they work.

3.1.1 Papoulis-Gerchberg Algorithm

The Papoulis-Gerchberg algorithm is a multi gap filling algorithm, meaning it is capable of correcting more than one gap at the time. This makes the algorithm preform good in conditions with many gaps and few available data points between gaps. This is due to its ability to collect information about the signal across multiple gaps[9]. The Papoulis-Gerchberg algorithm works under the assumption that the signal is a periodic stationary signal with a known bandwidth. The signal will therefore consist of M frequency components, and everything outside the band is assumed to be noise. The signals in the SmartHG is not stationary, but for small snippets can approximately stationariness be assumed.

The true bandwidth is also unknown in the signal. The Papoulis-Gerchberg algorithm is very depended on the bandwidth for a correct reconstruction. A modified version of the algorithm that estimates the bandwidth, by varying the frequency components M and analysing the mean square error on the known signal is therefore used [10]. This approach is fairly good at estimating the true value of M, but it is time-consuming.

3.1.2 Wiener Filling Algorithm

The Wiener filling algorithm is an extension of a Wiener predictor, the Wiener predictor assumes that there exist a linear relationship between the next sample and the previous samples. By trying to predict the missing samples from both sides of the gap, and combining the knowledge, it estimates the missing samples [11]. For larger gaps does this methods rely on earlier predictions to close the gaps. This result in errors being accumulated over the gaps. The method is fast, and is therefore suited for large data with small gaps.

3.1.3 Spatio-Temporal Filling Algorithm

The Spatio-Temporal filling algorithm uses singular spectrum analysis to split the signal into a series of sub-signals. The sum of the sub-signals is the original signal, and the sub-signals are ordered so the most dominant is first, and the least dominant is last.

The reconstruction philosophy is that the gap has introduced noise in the signal, but a sum of only the most dominant sub-signals must be close to the original signal without noise. But in order to know how many sub-signals to include in this sum, we introduce an other artificial gap. While the sub-signals are being accumulated the mean square error of the artificial gap is observed. When this mean square error hits its minimum peek, it is assumed that the reconstruction is as good as possible [12].

This method is very popular for gap filling. It has shown to be very noise resistant since it finds the overall trends in the data. It does require quite a lot of data to be known post and prior to the gap since an artificial gap must be introduced. It is based on singular spectrum analysis which assumes that the signal consist of stationary processes. This is a similar constraint to the Papoulis-Gerchberg Algorithm in section 3.1.1.

3.1.4 Envelope Filling Algorithm

Unlike the previous described methods does the Envelope filling algorithm not depend on frequency analyses, but rather on the expected power of the signal. Looking at the envelope of the signal it assumes that all local maxima and minima must lie on the upper and lower envelope. It then looks at the data prior and post the gap and try to estimate the number of local maxima and minima in the gap, and their locations. It does this by looking for patterns in the time series data [5]. When the new maxima and minima are found the points is connected by using spline [13].

The methods do not make any assumptions about the signals stationariness or bandwidth. The method can also be used on none equally spaced time series.

3.1.5 Empirical Mode Decomposition Filling Algorithm

The empirical mode decomposition filling algorithm uses empirical mode decomposition, to break the signal into intrinsic mode functions (IMF). The sum of all IMF's is the original signal. The IMF's is all more low frequent and simpler in structure than the original signal. The hypothesis is that it is easier fixing a gap in a simple signal than a complex one.

The envelope filling algorithm in section 3.1.4 is used to fix the gaps in the IMF's. The IMF's can now be accumulated to get the original fixed signal. Like the envelope filling algorithm does it not make any assumptions about the signals stationariness, bandwidth and can be used on none equally spaced time series. But making a empirical mode decomposition on a signal with a gap in is a non trivial process and can introduce errors [13].

3.2 Gaps In SmartHG Dataset

The gaps in the SmartHG project dataset is caused by a lot of different sources such as bad network connection, unplugged measurement equipment or server breakdown. This makes the type of gaps different from case to case. Three aspects of a gap is important for the gap filling: The size of the gap, the data known before the gap, and the data known after the gap.

3.2.1 Gap Size

On figure 3.1 is the quantity of different gaps shown. The different gap size is measured as samples missing, e.g. a gap size on 3 means that there is 3 samples missing in a row, between two received samples. Looking at the different gaps in the dataset we see that the normal gap is relatively small. Most of the gaps are between 1-5 samples as seen figure 3.1.

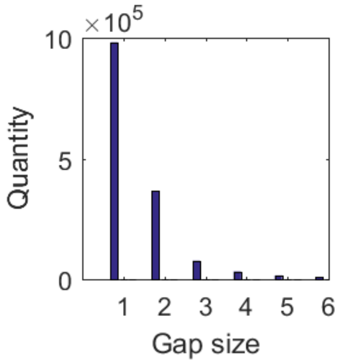

Figure 3.1: Gap quantity

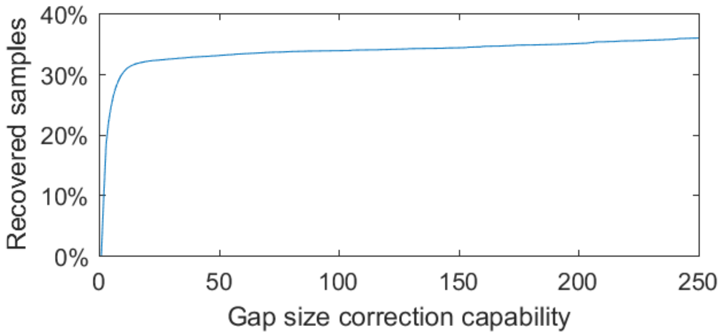

Figure 3.2: Error recovery capability

The aim of data reconstruction is to recover the missing samples. The gap size correction capability is a metric telling how big gaps it is possible to correct. If a application has a gap size correction capability on 5 it means that is is capable of correcting gaps that have the gap size of 5 samples or smaller.

On figure 3.2 is shown how much of the missing signal that can be reconstructed with different gap size correction capability. It is shown that a if a gap size correction capability of approximately 20 samples can be achieved, it is possible to recover 30% of the missing data in the SmartHG dataset. Since the signal is partly stochastic, and it is not possible to recover the stochastic part, the complete signal can newer be reconstructed. The greater the gap, the greater influence does the stochastic part have on the signal. Smaller gaps can therefore be fixed with greater success. It is therefore unlikely that recovery of more than 30% will be possible.

3.2.2 Post And Prior Knowledge

The reconstruction process works by looking at the samples available prior and post of the gap. Common for all reconstruction methods is that they assume that the signal in the gap must have behaved in relation to the samples prior and post for the gap. Therefore is the samples prior and post for the gap called knowledges, since it grants the knowledges used for reconstructing the gap.

But in a signal with lots of gaps it can be interesting to see how much knowledge is available for reconstructing a gap. This is done by seeing how many samples that are available prior and post to the gap. This is important since much data allows for detailed models, that can improve gap reconstruction and little knowledge gives the stochastic part dominance which will lead to error prone reconstruction.

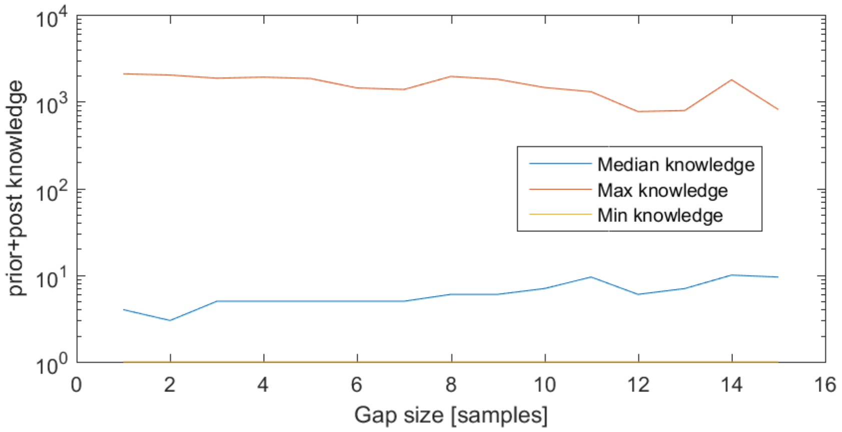

Figure 3.3: Available samples for recovery

In the case of the SmartHG project dataset the data available prior and post to a gap varies greatly but is around the same max and min values for every gap size. The median samples available to fix a gap is around 6 samples as shown in figure 3.3. On the figure is shown that no matter the gap size is the expected knowledge about the same. This can be problematic since lager gaps offend needs more knowledge than smaller gaps. This further indicates that complete recovery of all gaps is impossible.

3.3 SmartHG Dataset Reconstruction

The data reconstruction algorithms mentioned on section 3.1 have been tied on the SmartHG dataset. First have 6 unique error free areas in the data been found, and a artificial gap have been introduced in the 6 scenarios. The gaps have now been reconstructed, and a comparison to the true value is made.

The 6 scenarios have been randomly chosen under the constraints that there where no missing data in them, and they all are different in activity level from the other chosen scenarios. Both accumulative scenarios and non accumulative scenarios have been chosen.

There are several ways of comparing the different algorithms, in this report 3 have been selected based on the likelihood of the importance in a learning algorithm, which is the target application for the data. The first is a simple sample by sample comparison, where it is seen how much each sample differs from the true value. The other is the frequency, where it is analysed how much the frequency response is different. The last is a method called jitter compression, where we see that the power of the jitter is the same, more on this in section 3.3.3.

3.3.1 Sample Comparison

The sample by sample comparison is created by finding the mean square error of the samples compared to there true values.

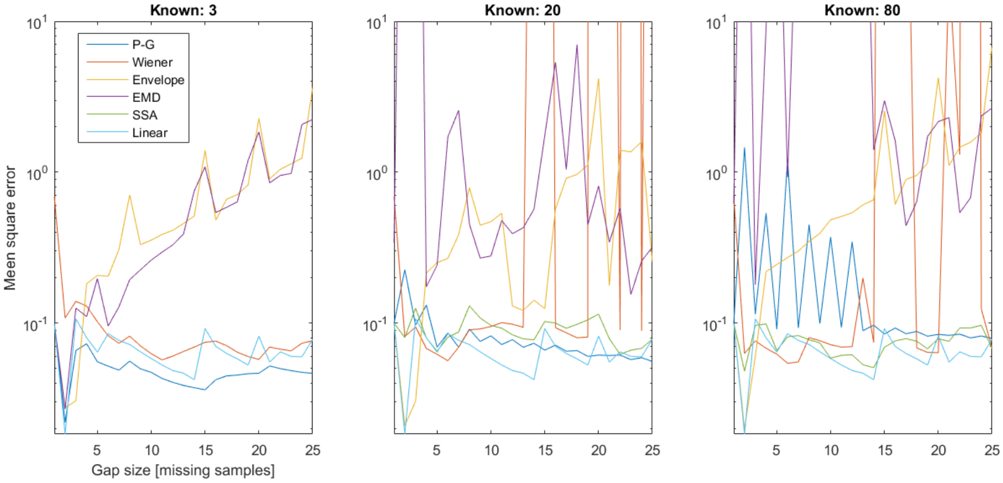

Figure 3.4: Sample comparison of the reconstruction methods

On figure 3.4 is the result shown for 3 different cases. The Known attribute indicates how much prior and post knowledge in samples is used to reconstruct the gap. On the x axis is shown the gap size, as expected does the prediction grow worse the greater the gap size. As a baseline is a linear predictor also added. The linear predictor makes a simple line between the two samples at each edge of the gap, and places all missing samples on the line.

As seen on the figure the best reconstruction is made by the Papoulis-Gerchberg algorithm, with relatively few samples as knowledge. The wiener algorithm is also close to the linear base line. The logical statement is the more information or knowledge we have the better should the prediction become. This does not seems to be the case, since a lot of the algorithms amperes to preform better at lower knowledge rate. This can be explained by the stochastic part of a signal, when the knowledge is small it is more likely that the information around the gap is relevant, the more knowledge grows the more likely is it the a stochastic event will bring false information to the model. This can be illustrated as on figure 3.5.

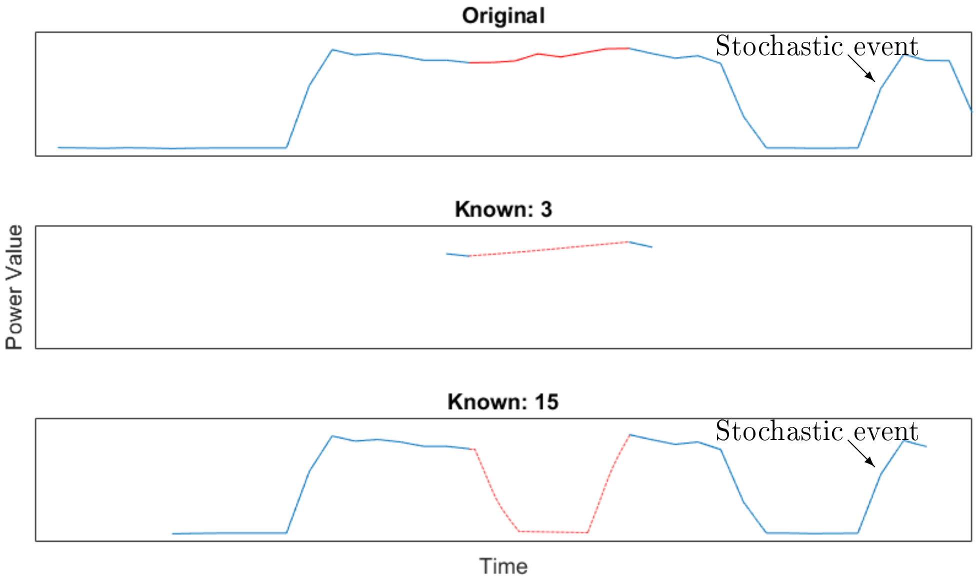

Figure 3.5: Stochastic effect on prediction.

On the figure 3.5 is shown a original signal with a gap, indicated by the red area. The true value is illustrated by the full red line in the original signal. The two subsequent graphs is the reconstruction, the blue area is the known area and the dotted red line is the predicted values. When the known area is small there is not enough frequency information to make any grant changes, so the prediction will look more or less like a linear interpolation. This is compared to the true value not a bad guess. When there is more samples known we are able to make more complex estimates of the signal. On the figure it is shown how when we have a knowledge of 15 we also see the stochastic event. Taking this in to the model, changes the prediction to a less correct value, due to the frequency's added by the stochastic event.

Since the results shown in figure 3.4 at higher knowledge shows that the reconstruction methods for the most part preforms worse than the linear basecase, we can conclude that the signal is quite stochastic and not as periodic as one would hope.

3.3.2 Frequency Comparison

An other metric to compare is how well the frequency response is preserved in the reconstruction, since this is often more important that the true value itself for smaller devices whit many states like computers. Such devices is often recognised by there usage pattern and not the true power draw. By looking at the Fourier transform of the original signal and the reconstructed power of the different frequencies was compared.

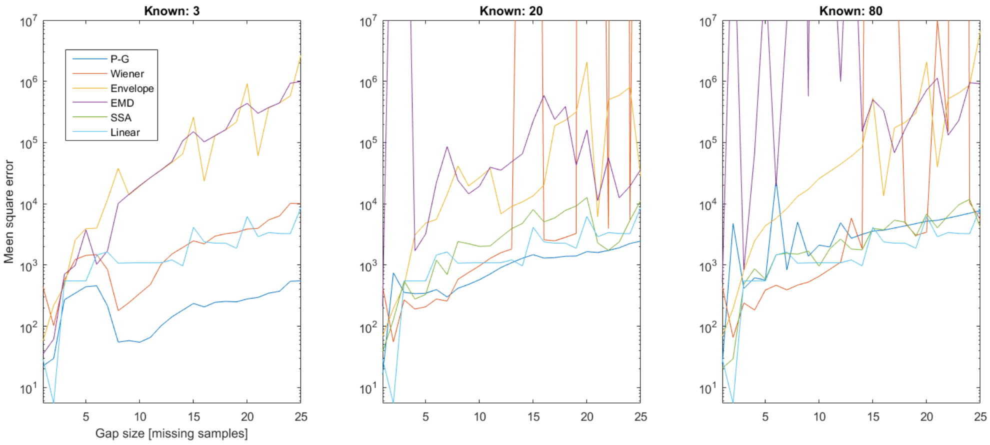

Figure 3.6: Frequency comparison of the reconstruction methods

As shown on figure 3.6 both the Wiener and the Papoulis-Gerchberg Algorithm seems to preform well. Like for the sample comparison the most successful reconstruction seems to be at low knowledge.

3.3.3 Jitter Comparison

The jitter is a different metric as it tells how well we have modelled the "noise" of the signal. In a jitter analysis the signal is thought of as a slowly changing signal with a faster noise signal embedded in it.

The slow signal is assumed to be found by taking a low pass filter over the signal to filter out the noise. The jitter analysis measures the power in the jitter from this signal as illustrated on figure 3.7.

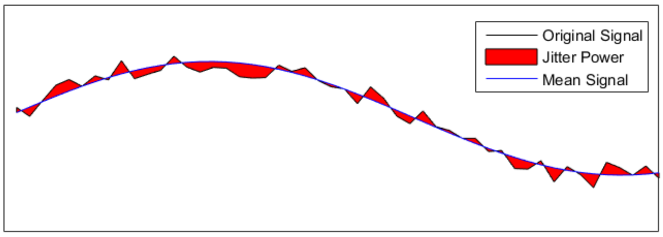

Figure 3.7: Jitter Power Illustration

It the estimated jitter power have been compared with the real jitter power for all the reconstruction methods.

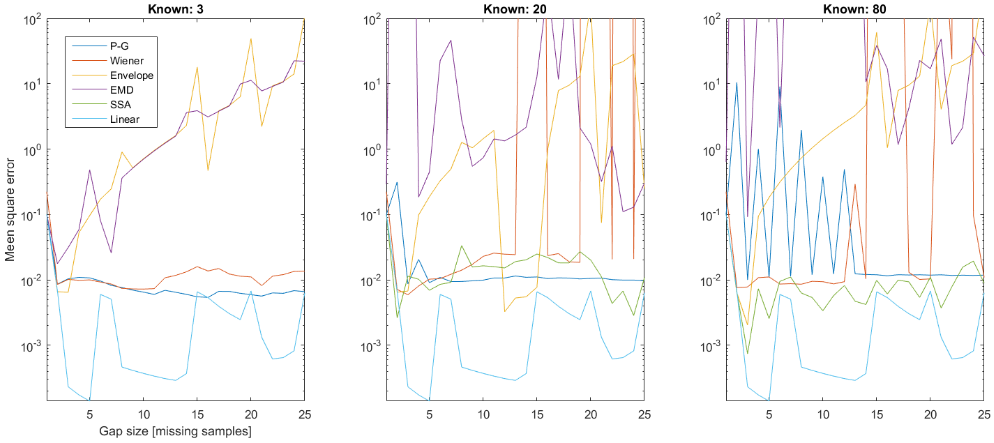

Figure 3.8: Jitter Power comparison of the reconstruction methods

As seen on figure 3.8 does non of the methods model the noise very well, as is to be expected due to the nature of the selected methods. This is will probably not be a problem for the SmartHG data, since it is sampled at a very low sample rate, and noise therefore is not likely to be used as a parameter. For data there have been sampled with high frequency it have shown that the noise created by the different appliances can be used to detect them [14].