Non Intrusive load monitoring

Rune A. Heick 11061

Aarhus University, Department of Engineering

1-12-2015

Abstract

Write an abstract Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Contents

1 Introduction

Write an Introduction Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

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.

In non intrusive load monitoring is a topic that have been focused on a lot in the past years. This have been possible due to the rise of the smart meters, that makes it possible to measure at faster intervals, and collect the data on online services. 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 in the meter data inspiration from related work was 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 whit 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 is 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. Such as 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.

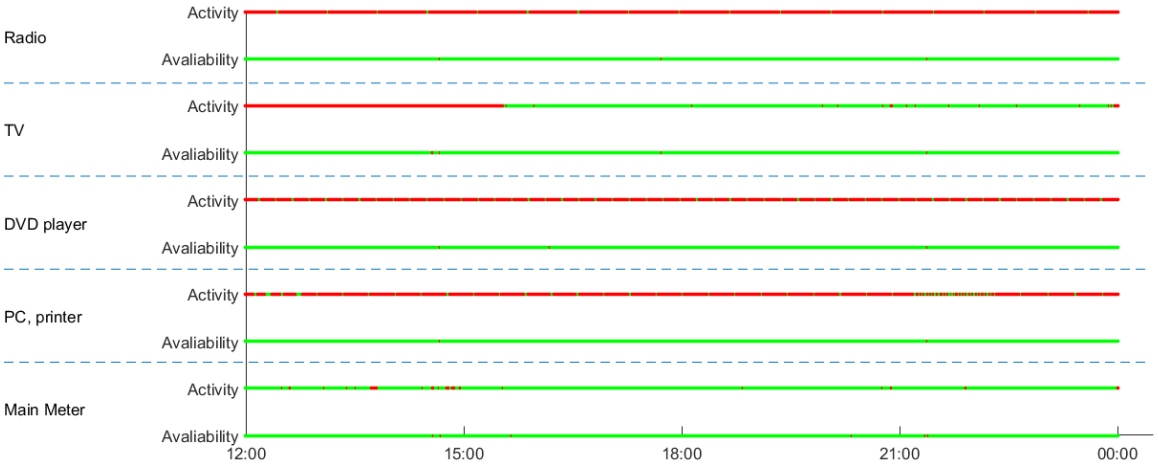

Figure 2.1: 12 hour overview of house 10

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.

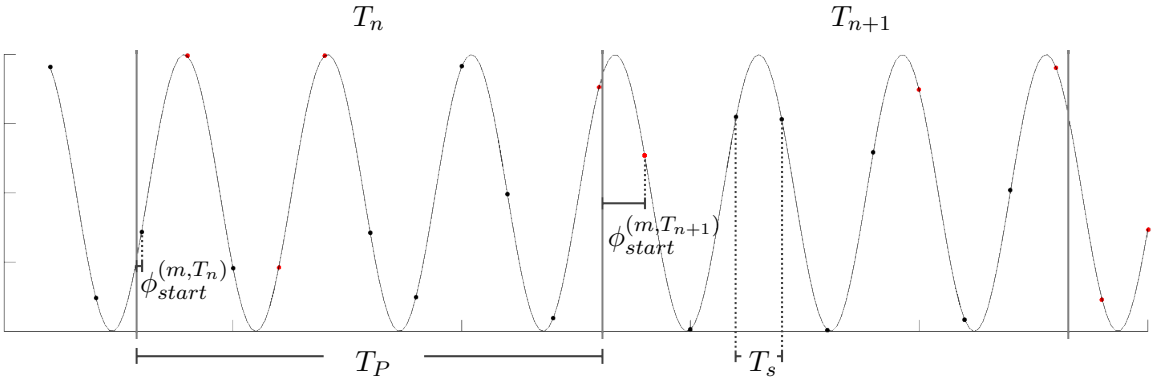

Figure 2.2: Illustration of availability analysis

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.

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

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

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 whit 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 the meters in a given house. This can be graphically shown on the 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.

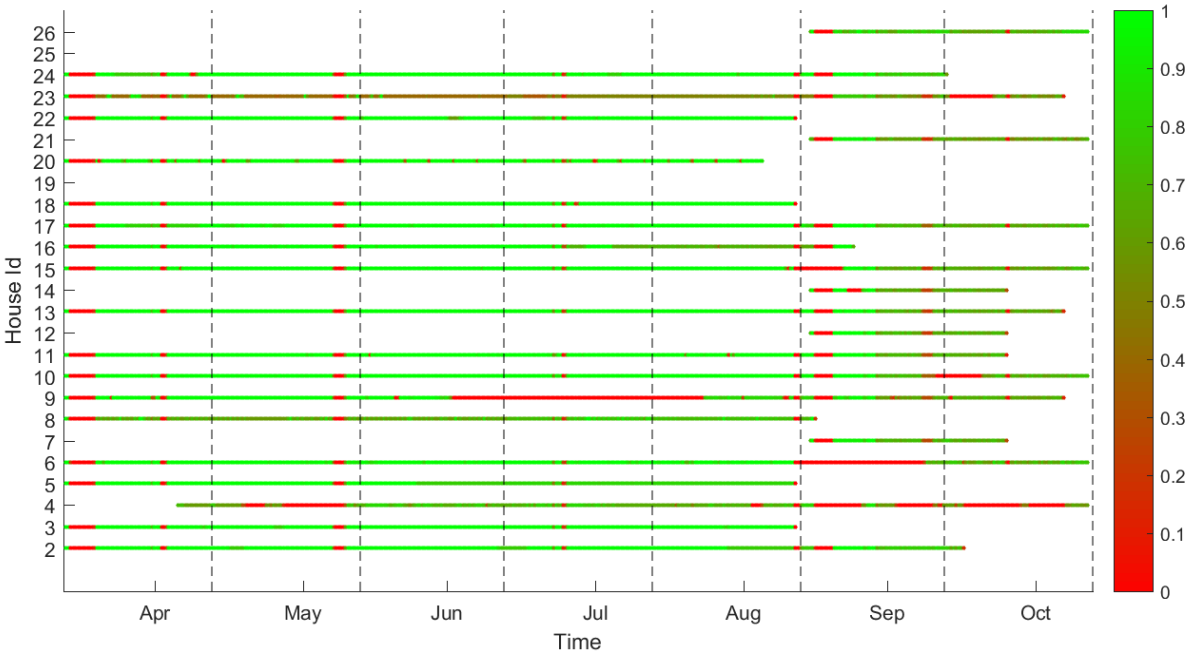

Figure 2.3: Quality of houses in SmartHG project

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, 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 reviving 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 down.

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 not happening anything. 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.

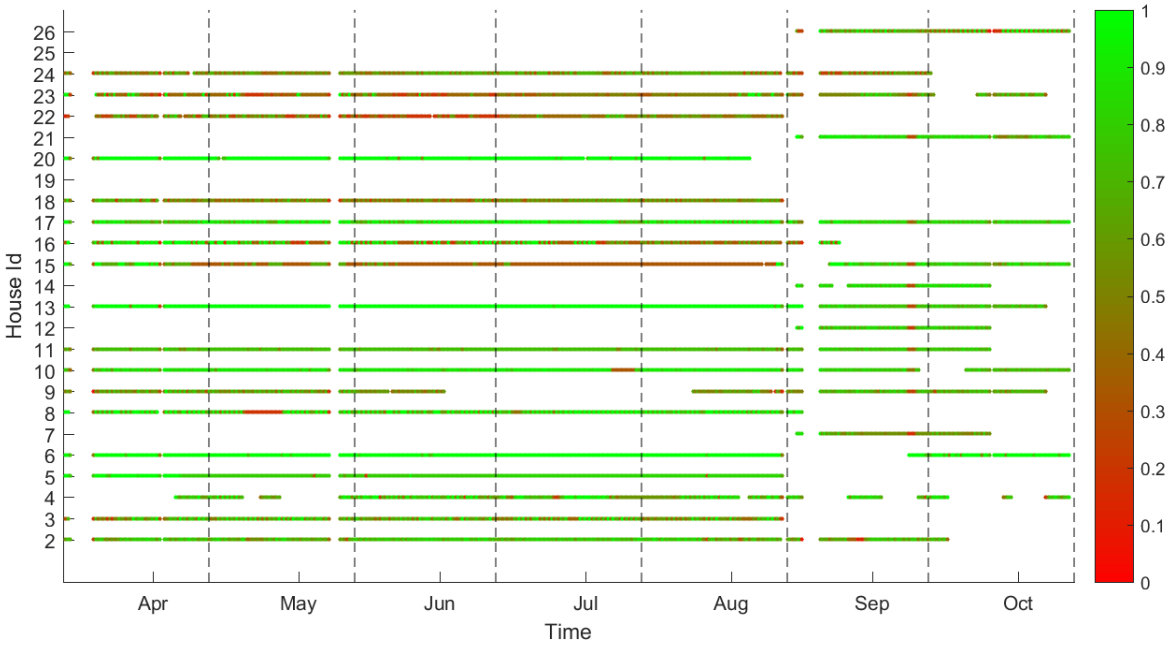

Figure 2.4: Activity of houses in SmartHG project

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 do 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 have been implemented and validated on the SmartHG project data.

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 a extension of a Wiener predictor. Like the Wiener predictor it 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 are the original signal, and the sub-signals is 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 a other artificial gap. While the sub-signals are being accumulated the mean square error of the artificial gap is observed, when this hits its 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 a artificial gap must be introduced. Since it is based on singular spectrum analysis it assumes that the signal consist of stationary processes, like 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. Bye 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 is found the points is connected by using spline [13].

The methods does 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 in to 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].

write a subsection about why these methods have been chosen for gap fixing in the SmartHG data.

3.2 Gaps In SmartHG Dataset

The gaps in the SmartHG project dataset is caused by a lot of different sources. 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. give examples of what can cause the gaps.

3.2.1 Gap Size

Looking at the different gaps in the dataset we see that the normal gap is relatively small. Most of the gaps is between 1-5 samples as seen figure 3.1.

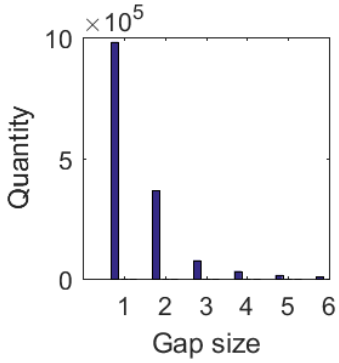

Figure 3.1: Gap quantity

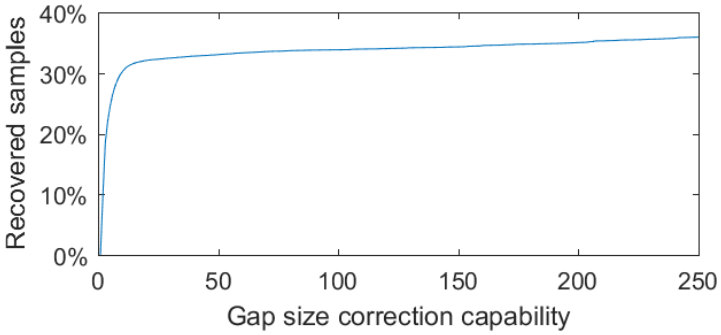

Figure 3.2: Error recovery capability

Units is missing the the graphs, and how is this plots created. Give a more detailed explanation and discussio

This is good since the signal is partly stochastic. The grater the gap, the greater influence does the stochastic part have on the signal. Smaller gaps can therefore be fixed with greater success. Further investigating of the gap quantity shows that approximately 30% of the missing data can be recovered with a gap size correction capability on 20 samples. This is illustrated on figure 3.2.

3.2.2 Past And Future Availability

It is also important how many samples there are available post and prior to the the gap. If there is a lot of gaps in the signal it can create a scenario where there only is a small amount of good data between the gaps. This makes it hard to reconstruct the data, since there is very few points to extract information about the region.

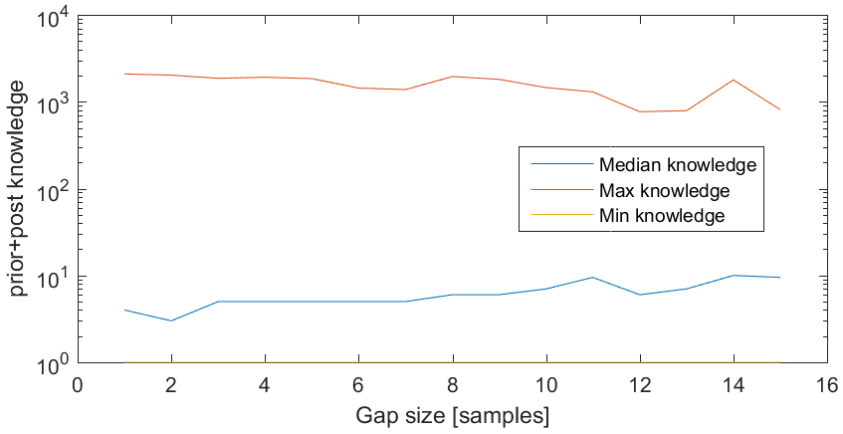

Figure 3.3: Available samples for recovery

The units is not understandable and needs explained.

A bit more discussion about what the purpose of this information is.

In the case of the SmartHG project data 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.

3.3 SmartHG Dataset Reconstruction

Gap fixing på SmartHG dataen.

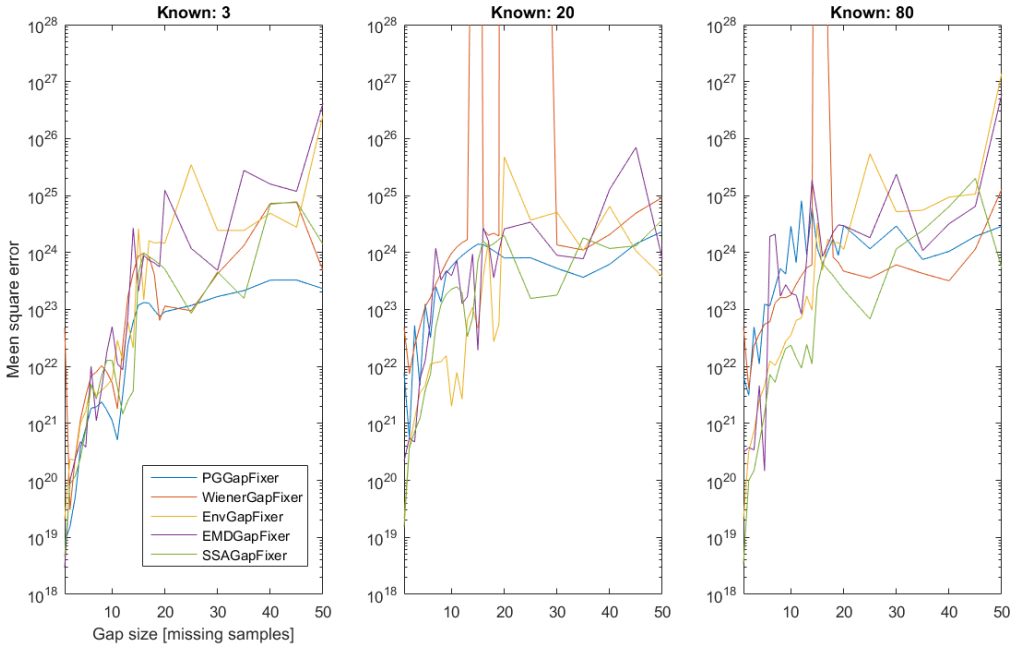

Figure 3.4

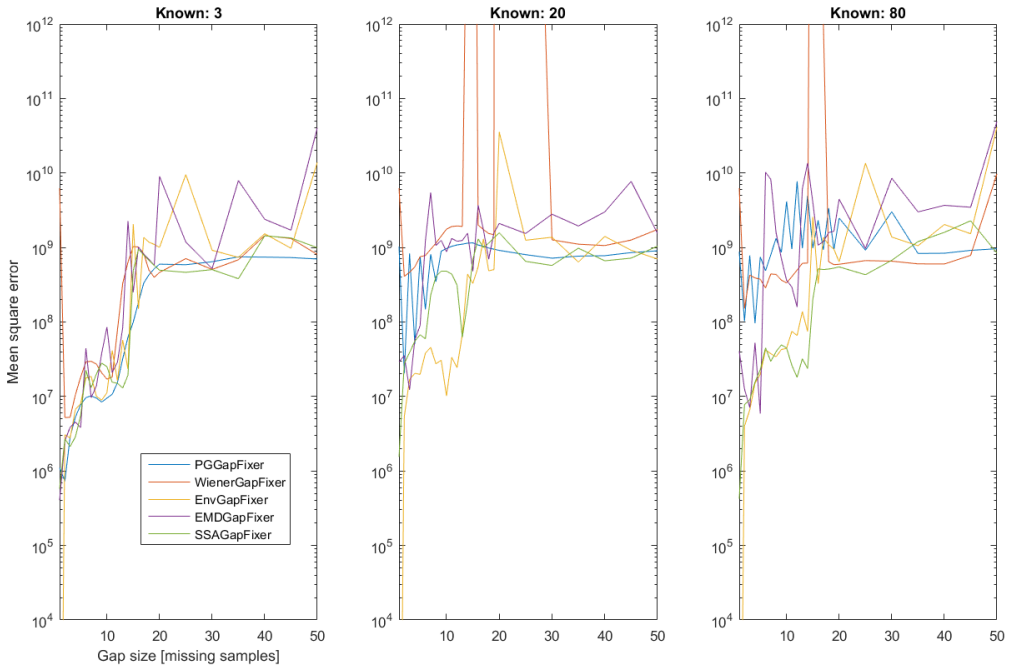

Figure 3.5

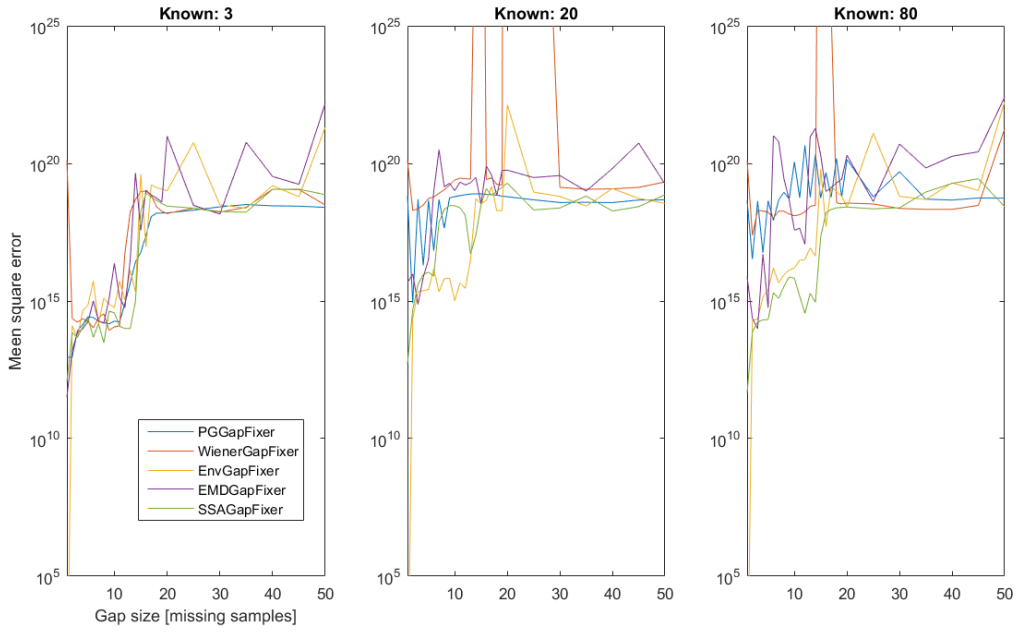

Figure 3.6

3.4 Related Work

Appliance Recognition 4

4.1 Related Work

4.2 Recognition Methods

4.3 Validation of Methods

SmartHG Data Recognition 5

5.1 Non Reconstructed Recognition

5.2 Reconstructed Recognition

5.3 Results

6 Discussion

7 Conclusion

Bibliography

[1] N. Regnauld, "Generalisation and data quality," 2015-08-01T00:00:00Z. Type: article; CC BY.

[2] I. 8402:1994, "Quality management and quality assurance – vocabulary," 1994-03-24.

[3] S. Kelling, D. Fink, F. A. L. Sorte, A. Johnston, N. E. Bruns, and W. M. Hochachka, "Taking a ?big data? approach to data quality in a citizen science project," 2015-10-27; 2015-11. Type: Text; © The Author(s) 2015; Open AccessThis article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

[4] I. 19157:2013, "Geographic information – data quality," 2013-12-15.

[5] G. Pastorello, D. Agarwal, T. Samak, C. Poindexter, B. Faybishenko, D. Gunter, R. Hollowgrass, D. Papale, C. Trotta, A. Ribeca, and E. Canfora, "Observational data patterns for time series data quality assessment," in *e-Science (e-Science), 2014 IEEE 10th International Conference on*, vol. 1, pp. 271–278, 2014. ID: 1.

[6] M. Meijer, L. A. E. Vullings, J. D. Bulens, F. I. Rip, M. Boss, G. Hazeu, and M. Storm, "Spatial data quality and a workflow tool," 2015-08-01T00:00:00Z. Type: article; CC BY.

[7] A. Simader, B. Kluger, N. Neumann, C. Bueschl, M. Lemmens, G. Lirk, R. Krska, and R. Schuhmacher, "Qcscreen: a software tool for data quality control in lc-hrms based metabolomics," 2015-10-24. Type: Software; Copyright 2015 Simader et al.

[8] D. G. Manolakis and V. K. Ingle, *Applied digital signal processing : theory and practice.* New York: Cambridge University Press, 2011.

[9] P. J. S. G. Ferreira, "Interpolation and the discrete papoulis-gerchberg algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2596–2606, 1994. Analyze the performance of an iterative algorithm, similar to the discrete Papoulis-Gerchberg algorithm, and which can be used to recover missing samples in... An analysis of the performance of an iterative algorithm that can be used to recover missing samples in finite-length records of band-limited data.No...

[10] M. Marques, A. Neves, J. Marques, and J. Sanches, "The papoulis-gerchberg algorithm with unknown signal bandwidth," vol. 4141, pp. 436–445, 2006.

[11] D. J. Thomson, L. J. Lanzerotti, and C. G. Maclennan, "Interplanetary magnetic field: Statistical properties and discrete modes," *Journal of Geophysical Research*, vol. 106, no. A8, pp. 15941–15962, 2001.

[12] D. Kondrashov and M. Ghil, "Spatio-temporal filling of missing points in geophysical data sets," *Nonlinear Processes in Geophysics*, vol. 13, no. 2, pp. 151–159, 2006. The majority of data sets in the geosciences are obtained from observations and measurements of natural systems, rather than in the laboratory. These data... The majority of data sets in the geosciences are obtained from observations and measurements of natural systems, rather than in the laboratory. These data sets...

[13] A. Moghtaderi, P. Borgnat, and P. Flandrin, "Gap-filling by the empirical mode decomposition," 2012. We propose a novel gap-filling technique, based on the empirical mode decomposition (EMD). The idea is that a signal with missing data can be decomposed into a...