None Intrusive load monitoring

Rune A. Heick 11061

Aarhus University, Department of Engineering

1-12-2015

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

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 let 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.

2.1 Quality Criteria

It is not uncommon that different areas of 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 whit several ways of describing quality in spatial data [4]. Method for defining quality in time series data have also been developed [5].

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. Here we asses if there is gaps

in the data, that is caused of unknown interference, and look at how the samples is distributed in the measurement period.

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 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.

The SmartHG data is intended for appliance recognition, and the quality must be assessed with this in mind. First a completeness analysis is done on the data, by analysing the sample availability quality of the data.

This the analysis is done in small segments on one hour. The quality of a given meter in a given hour is defined as the ratio of observed samples over the expected samples.

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

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

As shown on equation 2.1 is the maximum number of samples for a meter m in the period T calculated by taking the period time T_P, corrected with the sample phase (ϕ_{start} 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 to the maximum samples, shown in equation 2.2.

To find the quality of a house in a given period T, that have a set of meters M, we take the mean value of all the meter quality's, as shown in equitation 2.3.

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

Each house has a quality vector Q, with the house quality found whit a period T_P on one hour. This have been done from March t_{start} to October t_{end}.

$$Q^{(M,t_{start},t_{end})} = \{\mu_{q^{(M,T)}} | T \in \{t_{start}, t_{start} + T_p, t_{start} + 2 \times T_p, ..., t_{end}\}\} \tag{2.4}$$

This is shown in equation 2.4 where M is a set of the meters in a given house. This can be graphically shown on the figure 2.1 where the color is a gradient running from light green for the best quality to red for bad quality.

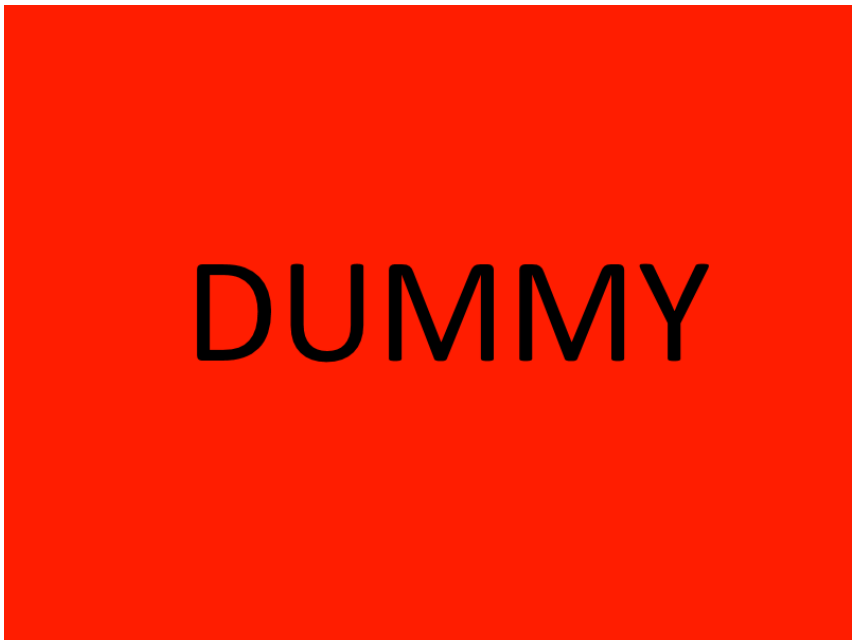

Figure 2.1: Quality of houses in SmartHG project show the figure of quality

Since the data is intended for appliance recognition it is also of interest where in the data there is activity, and where there is not happening anything. We define activity as area in the data where $f(x) \neq f(x+1)$. The standard deviation of a area is a good metric to indicate this behaviour. This is shown on figure 2.2 where green is high activity and red is non activity.

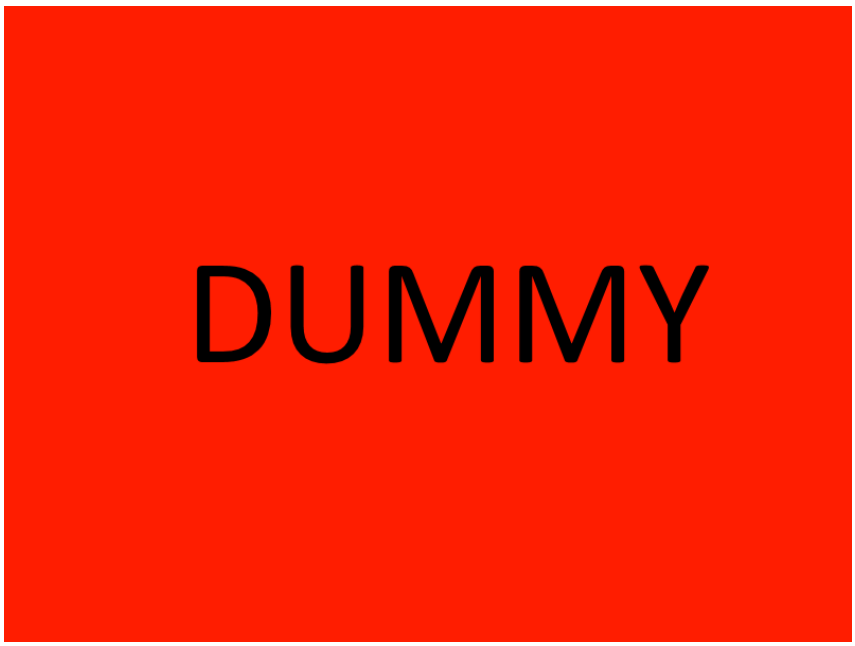

Figure 2.2: Activity Map show the figure of quality

2.3 Related Work

Data Quality is a 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* 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[3].

Common for all the methods above is that they deal whit completeness of the signal.

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. 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. Is the signal so stochastic that this is not the case gap filling is not recommended[**?**].

In the case of the SmartHG project the data can be seen to have a part that is dependency 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 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.

3.1.1 Gap Size

Looking at the different gaps in the dataset we see that the normal gap is relatively small. This can be seen on figure 3.1

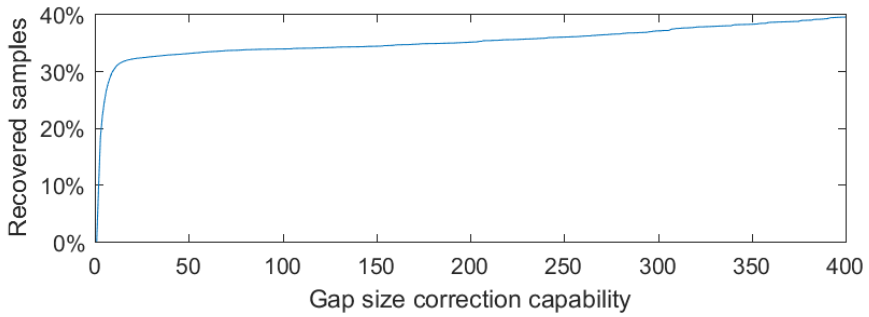

Figure 3.1: Gap size

This is good since the signal part 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.

3.1.2 Past And Future Availability

It is also important how many samples are available on the left and right side of 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.

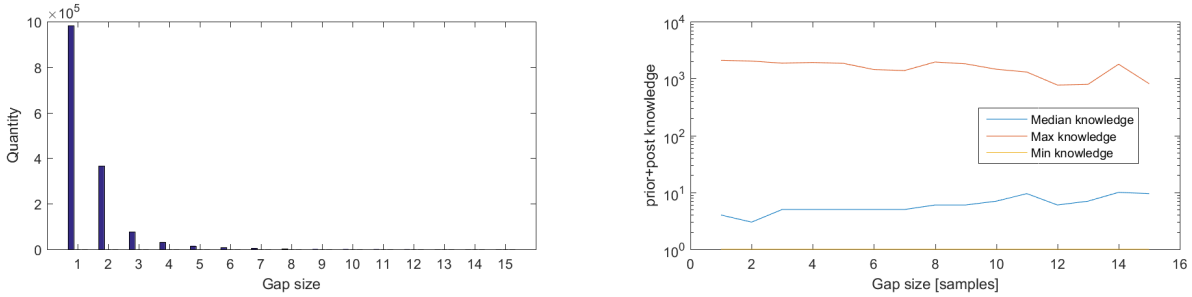

Figure 3.2: Past and future availability

As seen on figure 3.2 Write something more when you know the result

3.2 Gap Filling Methods

Various methods exists for gap filling. Five popular algorithms have been implemented and validated on the SmartHG project data.

3.2.1 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 the gaps, since it can collect information about the signal from multiple fragments signal [**?**]. The Papoulis-Gerchberg algorithm works under the assumption that the signal is a periodic stationary signal whit 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 [**?**]. This approach has fairly good at estimating the true value of M, but it is time-consuming.

3.2.2 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 [**?**]. 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.

3.2.3 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 [**?**].

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 3.2.1.

3.2.4 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 [**?**].

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.

3.2.5 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 ans 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 3.2.4 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 [**?**].

3.3 SmartHG Dataset Reconstruction

3.4 Related Work

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.